Utilising Data and Knowledge for Policy Impact

A Case Study Using Bayesian Approach

Mohamed Ismail

Analytical Research Ltd
Surrey, UK

August, 2016

Analytical
Research

Rationale for New Estimates

- Potential for inaccurate results
- More new data generated
- Potential for better sample representation
- Importance of choosing a suitable method

Purpose & Context

- Purpose
  - * To produce valid & credible estimate.

- Context
  - * Previous attempts
  - * Political sensitivity
  - * Money/policies
  - * Legal issues
  - * Law enforcement
  - * There are also data issues
  - * Will be carefully scrutinised

Analytical Research

What is available?

DATA

- National Administrative Data
    - * Large sample size
    - * Intensive information
    - * Lack of data entry validation
    - * Considerable data-minning is required
    - * Need to establish communication with the data vendor

What is available?

DATA

- National Administrative Data
    * Large sample size
    * Intensive information
    * Lack of data entry validation
    * Considerable data-minning is required
    * Need to establish communication with the data vendor

- National Survey:
    * Fill in the gaps
    * Provide useful information on unpaid travel time
    * Can't be linked directly to the administrative data

Analytical
Research

KNOWLEDGE

- Previous Estimates

  * Authoritative government organisations

  * Others from market research and small surveys

  * Wide ranging estimates based on different sample sizes

Knowledge
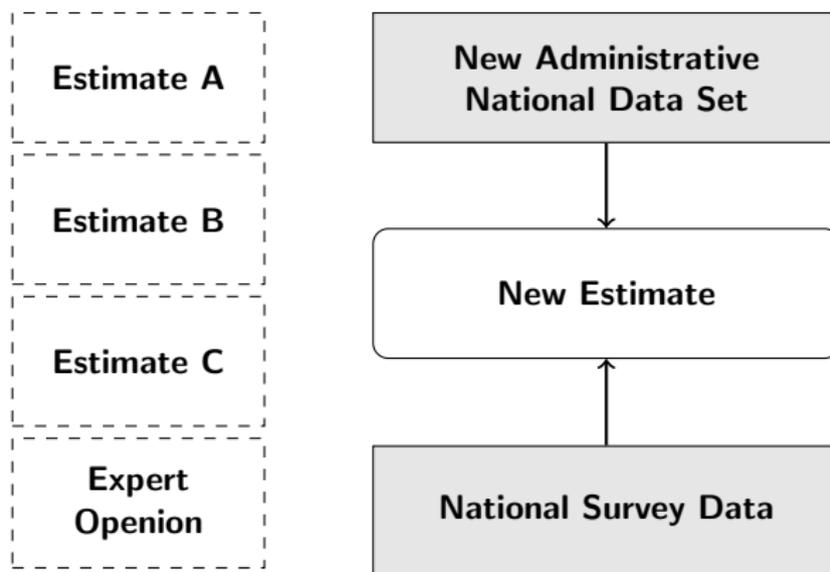
- Previous Estimates
  - * Authoritative government organisations
  - * Others from market research and small surveys
  - * Wide ranging estimates based on different sample sizes

- Expert Opinion:
  - * Different from previous estimates
  - * Based on related research

Analytical
Research

Information summary

Estimate A		New Administrative National Data Set
Estimate B		
Estimate C		New Estimate
Expert Openion		National Survey Data

Accounting for prior knowledge

- Previous point estimates that varies
- All previous samples are drawn from same population
- Researcher's opinion based on observations
- We need full distribution estimate that account for all the above
- A Bayesian approach is most suitable for this job

Analytical
Research

How can we make use of all the above?

Bayes' Theorem

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

$$P(Parameter|Data) \propto P(Parameter) \times P(Data|Parameter)$$

PROS

- A formal approach for including prior knowledge in the analysis
- Flexibility in constructing appropriate model for the data
- Transparent, all modelling decisions are clear
- Full distributions instead of point estimates
- CI have more intuitive meaning

CONS

- Tailored estimation could be challenging to implement
- Computationally intensive
- Need defense of decisions

Analytical
Research

Hierarchical Bayesian Model

- A framework for capturing dependencies between parameters
- Treating all estimates as arising from a random process governed by hyperparameters
- More accurate than treating previous estimates as fixed prior
- Full posterior distributions for all parameters including previous estimates

Analytical
Research

Model specifications

- All estimates are based on samples drawn
  from same population

$$P(\theta_i | y_i) \propto P(y_i | \theta_i) \times P(\theta_i)$$

- θ_i are drawn from a distribution with unkown parameter
  vector ϕ

- ϕ is an unknown hyperparameter with its own posterior
  distribution

$$P(\theta_i, \phi | y_i) \propto P(y_i | \theta_i, \phi) \times P(\theta_i, \phi)$$

$$\underbrace{P(\theta, \alpha, \beta | y)}_{\text{Posterior}} \propto \underbrace{P(y | \theta)}_{\text{Likelihood}} \times \underbrace{P(\theta | \alpha, \beta)}_{\text{Prior}} \times \underbrace{P(\alpha, \beta)}_{\text{hyperprior}}$$

Analytical
Research

Model implementation
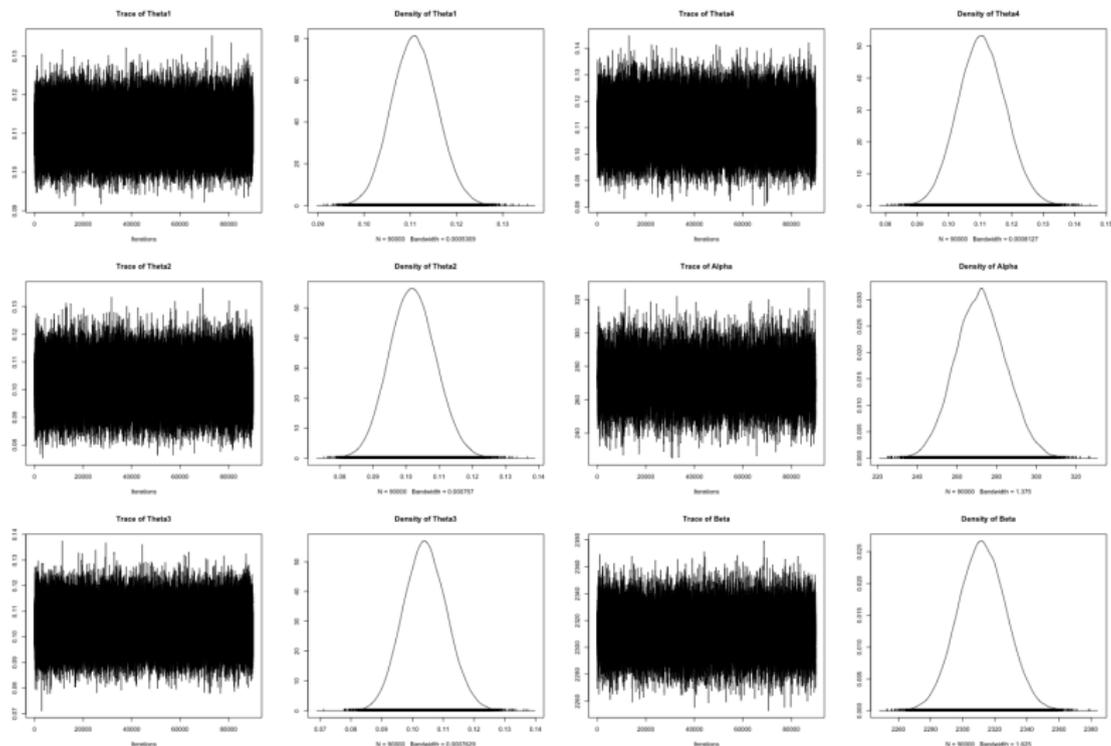
- All posterior densities could be derived, for example:

$$P(\theta_i|\alpha, \beta, y_i) \propto$$

$$(1 - \theta_i)^{\beta + n_i - y_i - 1} \prod_{i=1}^{I} \theta_i^{y_i}(1 - \theta_i)^{n_i - y_i} \times$$
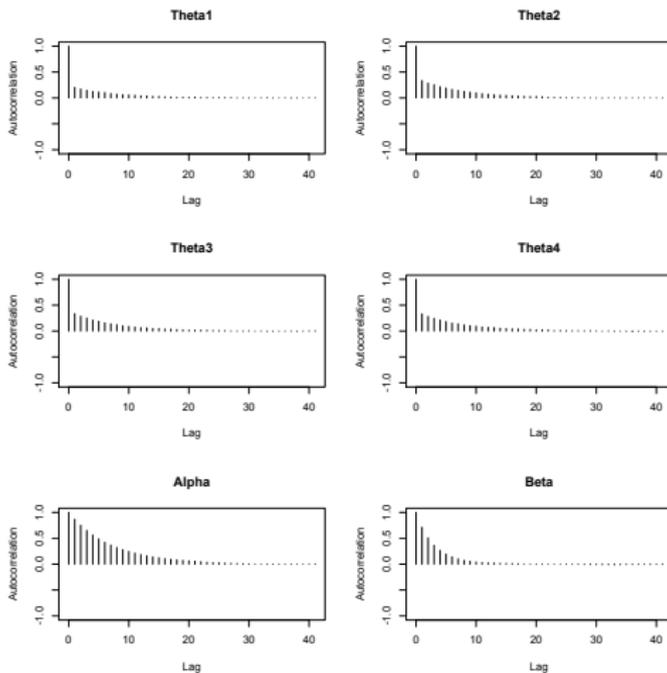
$$\prod_{i=1}^{I} \frac{\Gamma(\alpha + \beta) + n_i}{\Gamma(\alpha + y_i)\Gamma(\beta + n_i - y_i)} \theta_i^{\alpha + y_i - 1}$$

- Stochastic integration via Markov Chain Monte Carlo
- A hyper Metropolis-Hasting/Gibbs sampling algorithm
- Estimated densities of θ_i, $\alpha/(\alpha + \beta)$
- R-Unix environment

Output 1

Output 2

Communication & Interepretation

- Visualisation
- Communication process
- Providing tools for researchers to make decision

Reflection

- This is just an example
- Focusing on the suitability rather than convenience could be rewarding
- Recent developments offer new opportunities
    - * New algorithms for stochastic integration
        - * Sampling instead of numerical approximation
    - * Choice of programming languages
        - * C/C++, R, Stan
    - * Significant hardware advancement
    - * Availability of big data
- Skills Matrix
    - * Subject matter expert
    - * Mathematical statistics
    - * Quantitative programming

Further Reading

Adrian E. Raftery (2000)
Statistics in Sociology, 1950-2000
Journal of the American Statistical Association 95(450): 654 – 661.

Bruce Western (1999)
Bayesian Analysis for Sociologists
Sociological Methods & Research 28(1): 7 – 34

Bruce Western (2001)
Bayesian Thinking about Macrosociology
American Journal of Sociology 107(2): 353 – 378

Bob Carpenter, et all (2016)
Stan: A Probabilistic Programming Language
Journal of Statistical Software (in press)

Thank You for Listening

Contact details:
mohamed@analyticalresearch.co.uk
+44(0)7952774365
www.analyticalresearch.co.uk