Estimating Casualty Reductions from Road Safety Measures

Neil Thorpe and Lee Fawcett

Newcastle University, Newcastle-upon-Tyne, U.K.

lee.fawcett@ncl.ac.uk

Transport Operations Research Group, March 5th 2008

Structure of this part of the talk

1. **Sources of error in road safety scheme evaluation**
   - Scheme effects
   - Non–scheme effects
2. **Regression–to–mean (RTM)**
   - What is it?
   - Consequences of ignoring it
   - Possible solutions
3. **Modelling casualty frequencies**
   - 'Empirical' Bayes and how it caters for RTM
   - 'Full' Bayes
4. **Current/further research ideas**

Sources of error in road safety scheme evaluation

Mountain et al. (2004) identify two main sources of error in road safety scheme evaluation:

- Effects due to the safety scheme being implemented
- Non–scheme effects

1. **Scheme effects:**
   - Effects of the safety scheme
   - changes in exposure to risk
2. **Non–scheme effects:**
   - Trend in accidents
   - Regression–To–Mean

Regression–To–Mean (RTM)

Hypothesis: Placing a piece of paper under a dice causes it to decrease the number of sixes it rolls.

The experiment:

- Take ten dice.
- Roll each of them ten times noting the number of sixes for each.
- Place a piece of paper under the three highest scoring dice.
- Roll each of these dice another ten times and count the number of sixes observed again.
- The second total is almost always lower than the first, *proving* that the piece of paper decreased the number of sixes rolled by the dice.

Regression–To–Mean (RTM)

Dice 1	2	3	5	1	2	1	4	3	4	2	**0**
Dice 2	1	6	5	4	4	3	4	5	2	1	**1**
Dice 3	3	2	6	5	4	3	1	5	4	2	**1**
Dice 4	1	3	2	5	5	2	1	6	5	3	**1**
Dice 5	6	6	2	1	3	5	6	4	3	3	**3**
Dice 6	1	6	5	4	4	3	1	5	2	1	**1**
Dice 7	2	4	5	3	6	4	5	3	6	5	**2**
Dice 8	6	3	2	1	6	6	5	2	4	6	**4**
Dice 9	1	1	2	1	3	5	4	3	2	2	**0**
Dice 10	1	6	5	4	4	3	6	5	2	1	**1**

Reroll **5**, **7**, **8** (after placing a piece of paper under them).

Rolls after paper:

- Dice 5: $0 = 100\%$ decrease
- Dice 7: $2 = 0\%$ decrease
- Dice 8: $1 = 75\%$ decrease
- Average $= 66\%$ decrease!

Conclusion: putting paper under dice decreases the roll by an average of 66%!

Regression–To–Mean (RTM)

Traditional **Before/After studies** in road safety scheme evaluation might replace the "piece of paper" with safety cameras and the "outcomes on the dice" with number of casualties to 'prove' that safety cameras work.

People generally don't trust the conclusion of the dice experiment, so why should we trust the results of simple Before/After studies on safety cameras?

Regression–To–Mean (RTM)

- It's a pile of rubbish – the analysis uses an *artificial selection* of the data, not either *all* the data or a *randomly chosen sample* of the data.
- Dice/Safety camera locations are chosen with an unusually high number of sixes/casualties.
- Thus, the number of sixes/casualties is bound to reduce in any subsequent period, regardless of any intervention.
- This phenomenon is known as **Regression–To–(the)–Mean (RTM)**.

Regression–To–Mean (RTM)

"Speed cameras for blackspots", Hartlepool Mail, 18/12/05

"Speed cameras are set to be deployed at two new accident blackspots..."

"Five people have been killed or seriously injured after collisions on Elwick Road and King Oswy Drive, in Hartlepool, in the past three years..."

"...forty–four people have been injured..."

Regression–To–Mean (RTM)

"The first camera will be located in Elwick Road, from York Road to Elwick Rise. It was chosen after research found that:

- **60** per cent of motorists drive above the 40mph limit...
- **3** people have been killed or seriously injured in the last three years...
- **35** people have been injured..."

"The second camera will go up in King Oswy Drive, from Easington Road to West View Road...

- ...**35 per cent** of people using the road exceeded the speed limit
- **9** people had been injured in the past three years..."

Regression–To–Mean (RTM)

"Safety measures save lives", Hartlepool Mail, 13/12/07

Casualties

	Before	**After**	**% change**
Elwick Road	35	13	**–63%**
King Oswy Drive	9	5	**–44%**

Deaths

	Before	**After**	**% change**
Elwick Road	3	0	**–100%**
King Oswy Drive	0	0	**–0%**

Regression–To–Mean (RTM)

So how can we remedy this problem?

- **Ideal experiment**
  - Randomly select some speed camera sites
  - Run two parallel universes, one with cameras and one without cameras, and compare the results
  - For obvious reasons this might not be practical!
- **Randomly select sites and put cameras in half of them**
  - We could, by chance, choose to put cameras in sites with similar conditions
  - There is also the chance that sites with cameras might affect sites without cameras
- **Think about things statistically**

Statistical modelling

Let's assume, as is often the case, that Y – the number of casualties at any site – follows a **negative binomial** distribution with mean and over–dispersion **parameters** μ, and κ (respectively), i.e.

$$Y \quad \sim \quad \text{NB}(\mu, \kappa)$$

- We could somehow estimate μ and κ...
- ... and then calculate the probability of observing $r = \mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots$ casualties using the formula for this negative binomial model:

$$P(Y = r) \quad = \quad \binom{r + \mu - 1}{\mu - 1} \left(\frac{\kappa}{\kappa + 1}\right)^{\mu} \left(\frac{1}{\kappa + r}\right)^{r}.$$

Statistical modelling

But how do we estimate μ? How do we estimate κ?

- μ is the **mean** parameter
- κ is the **over–dispersion** parameter

There are mathematical/statistical techniques available for estimating these **parameters** for all of our data together, across all sites

However, not all sites will have the same value of μ (mean number of casualties); this will be site–specific depending on things like:

- Road type/classification
- Speed limit/*observed* speed
- Traffic flow
- Other geographical/topological features

Statistical modelling

Recent attempts to use this model have been two–pronged:

1. Estimate μ separately for each site

This is often done using standard regression techniques, e.g.:

$$\mu = \exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\}$$

where $x_1, x_2, \ldots$ are **predictor variables** (such as average speed, traffic flow, road type etc.)

2. Estimate κ

- ‘Plug–in’ the estimated value of μ and find κ using a mathematical numerical method
- Then use the NB formula to estimate the probabilities

Statistical modelling

Problems/Issues:

- Estimating μ is a fiddly, time–consuming, manual job
  - We need to choose predictor variables using significance tests (yawn!)
  - Certain modelling assumptions need to be verified for the model to work, and often we need to **transform** our data
- We have shown that estimating κ can be sensitive to:
  - the type of mathematical estimation procedure used
  - the statistical package used to estimate it!

The Bayesian framework

The main objective of any **Bayesian** analysis is to supplement our probability calculations with other sources of information (e.g. the data!)

For example, we might use the procedure previously outlined to calculate the probability of observing **3** casualties at a particular site using the negative binomial model, that is

$$\Pr(Y = 3)$$

However, would it not be better to use:

$$\Pr(Y = 3|\mathbf{y} = (0, 1, 1, 0, 0, 0, 1, 2, ...)) \textbf{ ?}$$

The Bayesian framework

The Bayesian approach to modelling caters naturally for such **conditional probabilities**.

Following from **Bayes' Theorem**, we have that:

Posterior distribution $\propto$ **prior disribution** $\times$ **model for our data**

- We have a model for our data – the negative binomial distribution
- What we also need to specify are **prior distributions** for each of the parameters in our model

The Bayesian framework

To date, an **Empirical Bayes** approach has been used:

1. Use classical regression techniques to estimate μ at each site
2. Specify a **gamma** distribution for the other parameter, κ
3. Then the **posterior distribution** for casualty frequencies, *given that we have already observed y_b casualties in the before period*, is now also a **gamma** distribution with mean:

$$E(Y|y_b) = \frac{\kappa + y_b}{\kappa/\mu + 1}$$

$$= \alpha\mu + (1-\alpha)y_b$$

The Bayesian framework

So how does this cater for **Regression–To–Mean**?

Recall from the last slide:

$$E(Y|y_b) \;=\; \alpha\mu + (1-\alpha)y_b.$$

This is the **Empirical Bayes (EB)** estimate of casualty frequency.

Instead of comparing **Before** $\Leftrightarrow$ **After**, we now compare **EB** $\Leftrightarrow$ **After**.

- The ‘before’ figure – y_b – will be unusually high
- We don’t discard this value altogether – it is a *real* observation!
- We ‘tone it down’ by calculating a ‘weighted sum’ of this figure and μ

The Bayesian framework

So what's the problem?

We are still faced with the problem of how best to estimate μ.

Remember:

- Estimating μ is a fiddly, time–consuming, manual job
- We need to choose predictor variables using significance tests (yawn!)
- Certain modelling assumptions need to be verified for goodness–of–fit, and often we need to **transform** our data

The Bayesian framework

The prior distribution used to form the posterior is unrealistically simple – authors have chosen this to "keep the maths nice" and to make sure the posterior distribution is of known form (**conjugacy**).

There are problems with assessing the **precision** of our estimates.

A Fully Bayesian approach

'Recent' advances in fitting Bayesian models have made it fairly simple to obtain approximate samples from the **full posterior** distribution – as opposed to using just the mean in the EB approach.

These advances (mainly **Markov chain Monte Carlo**) actually allow us to specify more realistic prior distributions for our model parameters without having to worry about the more complex (often un–do–able!) maths.

Instead of the two–pronged approach of the EB method, we now have a single model–fitting procedure, and we don't have to worry about the clumsy regression modelling issues surrounding the estimation of μ.

A Fully Bayesian approach

Recall:

Posterior distribution $\propto$ **prior disribution** $\times$ **model for our data**

Under the **Empirical Bayes** approach we have: We have:

Posterior distribution $\propto$ $f(\mu)f(\kappa) \times$ Neg. binomial model

Under a **Fully Bayesian approach** we have:

Posterior distribution $\propto$ $f(\mu|\beta_0, \beta_1, ...)f(\kappa) \times$ Neg. binomial model

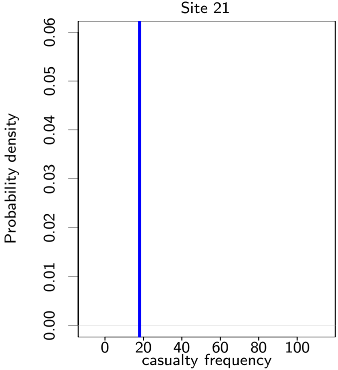
Some results

Empirical Bayes analysis

	Before	**EB**	**After**	**Observed difference**	**Cameras**
Slight	430	431	430	0	–1
Serious	95	90	41	–54	–49
fatal	13	7	5	–8	–2
Totals	**538**	**528**	**476**	**–62**	**–52**

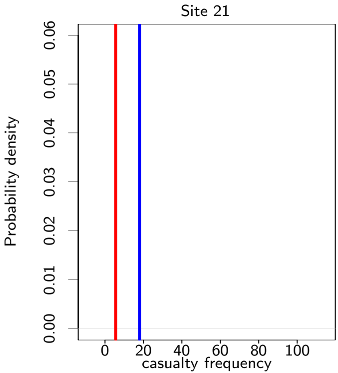
Some results

Fully Bayesian analysis



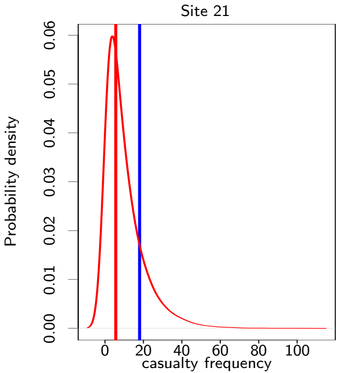
Some results

Fully Bayesian analysis



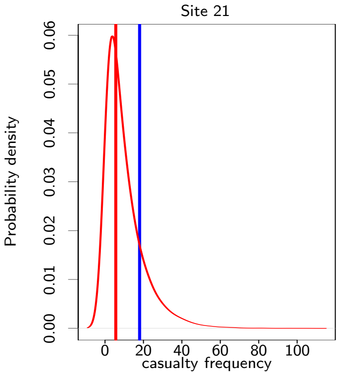
Some results

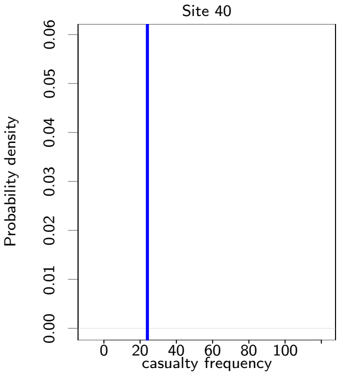
Fully Bayesian analysis



Some results

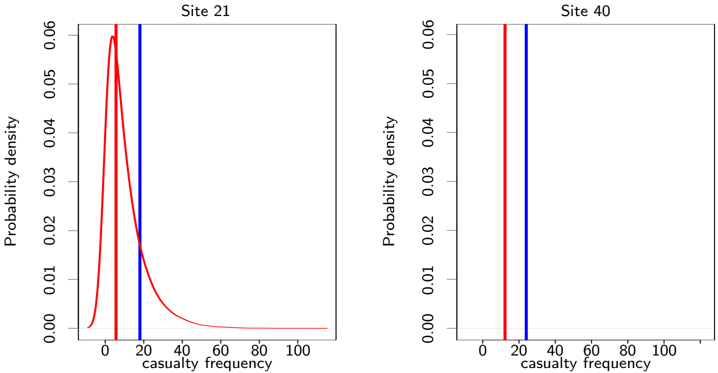
Fully Bayesian analysis





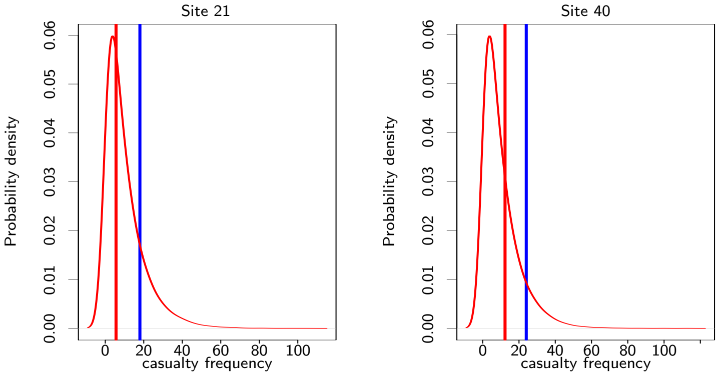
Some results

Fully Bayesian analysis



Some results

Fully Bayesian analysis



Further work

- Look at how best to utilise the output from the **Fully Bayesian approach**
- Investigate models which allow for **trend** as well as RTM
- Investigate the use of more **informative prior information**
- Re–visit the costing procedure to extend these result to estimate financial savings to the NHS
- This work is likely to form the basis of a research proposal for a PhD project