There's a fine line between a numerator and a denominator

Paul Hewson
paul.hewson@plymouth.ac.uk

1st March 2016

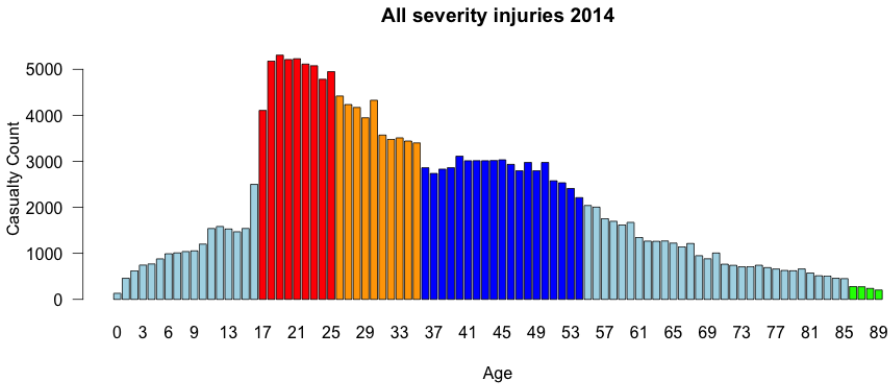
Table of contents

Numerators and Denominators

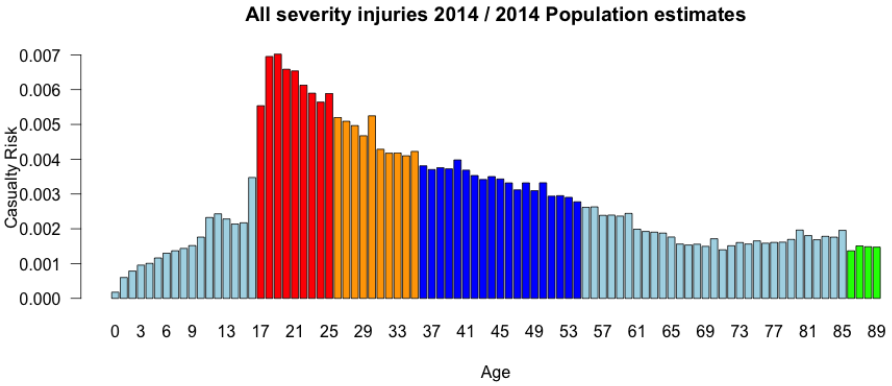
$$\text{Risk} = \frac{\text{Injury Count}}{\text{Exposed Population}}$$

$$\text{Rate} = \frac{\text{Injury Count}}{\text{Amount of exposure}}$$

Counts

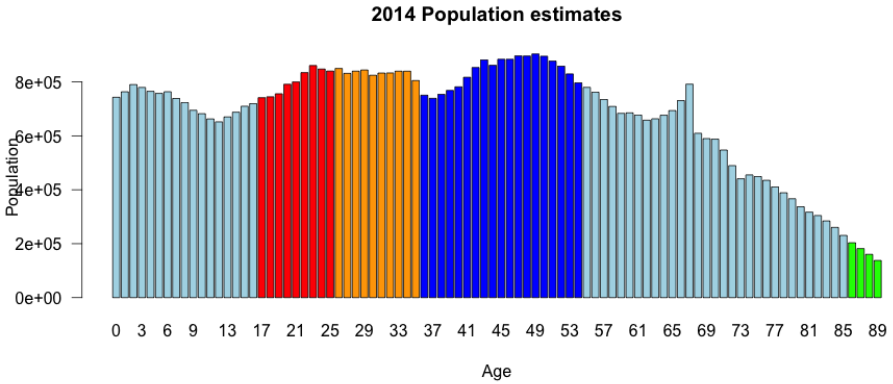

All severity injuries 2014
Casualty Count
5000
4000
3000
2000
1000
0
0 3 6 9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73 77 81 85 89
Age

Something we can all agree on

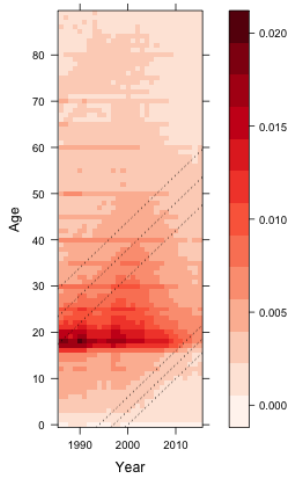


Why the middle aged bulge?

Why the middle aged bulge?



Exploring Population Risk (all casualties): Lexis Diagram



Fitting models

Fitting Age-Period-Cohort models

- Casualty Count $= y$
- Age of casualty $= a$
- Cohort (deduced birth year of casualty) $= c$
- Period (year this happened) $= p$

Assume that the casualty count can be modelled as a realisation of a Poisson random variable

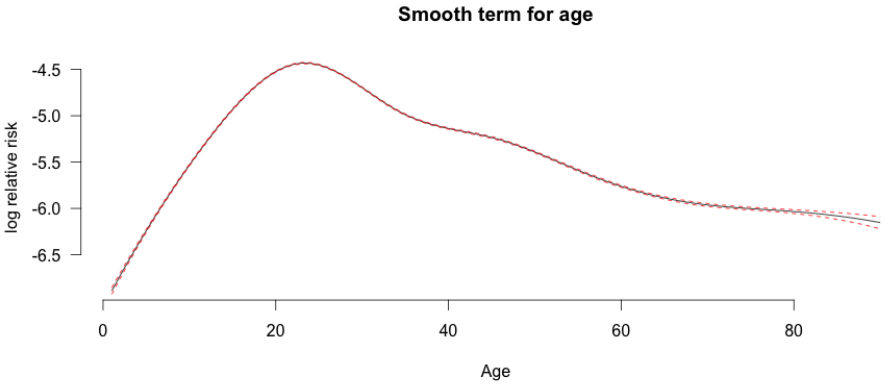
$$Y_{apc} \sim Poisson(\text{Denominator} \times \lambda_{apc})$$

Now assume that we can model the logarithm of this model parameter:

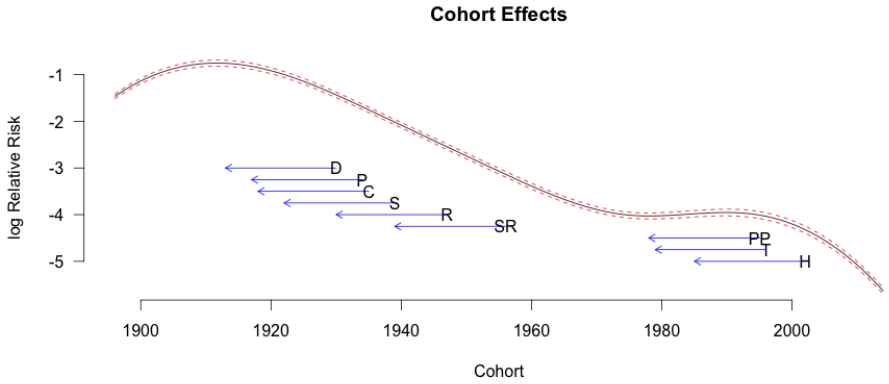
$$\log(\lambda_{apc}) = \text{Slope}_a + \text{Slope}_p + s(a) + s(p) + s(c) + \epsilon$$

(Check, if I fit model 1985-2013 do I make a good job predicting 2014?)
What do $s(a)$ and $s(c)$ look like?

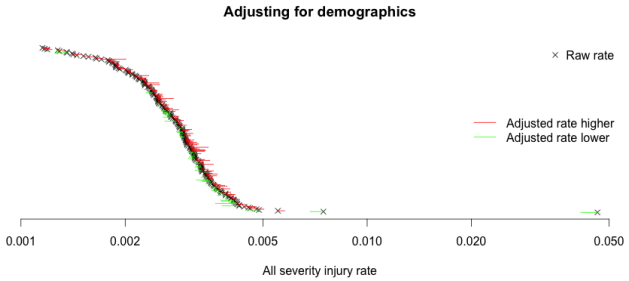
Age: as expected



Cohort

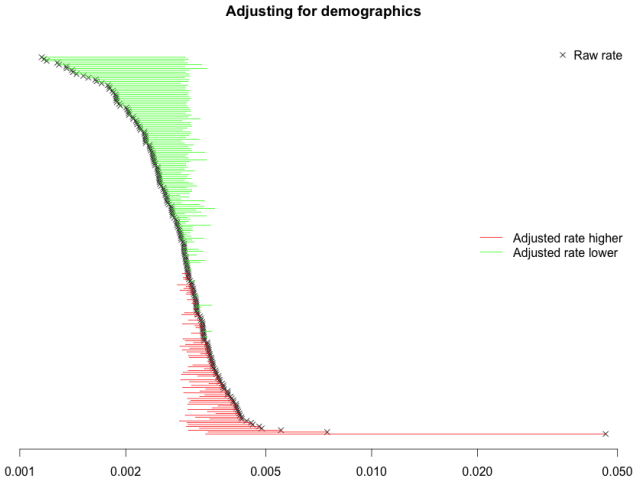

Cohort Effects
log Relative Risk
-1
-2
-3
-4
-5
D
P
C
S
R
SR
PP
T
H
1900
1920
1940
1960
1980
2000
Cohort

Adjusting for demographics



Casualties per 100,000		
HA Code	Adjusted	Raw
E10000009	347.1	307.7
W06000023	478.0	441.2
E06000011	394.3	361.7
E10000011	400.8	369.5
S12000013	201.0	170.7
W06000009	350.8	320.6
...	...	...
E09000019	364.6	420.5
E09000033	684.2	747.5

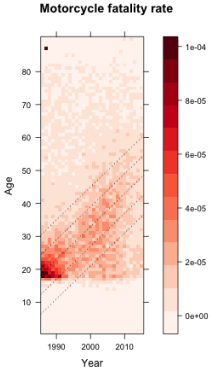
Adjusting to demographics

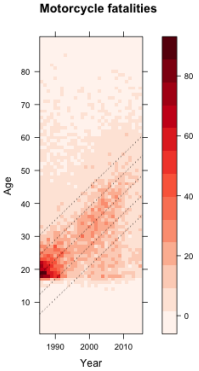


Don't have time today but . . .

- Have made more subtle adjustments (age / gender / urban)
- Have used spatial smoothing
- This smooths a lot of trends over time (especially if you use age / gender / mode)
- Very powerful forecasting method

Motorbike Lexis Diagrams





Turning risk into rate

- Either asume everyone at every age has equal exposure
- Or
- Find a better denominator. For roads we use AADT without thinking about it.

Information on useage by demography

- Census/DWP/Other (e.g. GP registrations): Age/Gender
- Census: travel to work, car ownership and other obviously useful things
- Census: occupational status, limiting long term illness, marital status and other less obviously useful things
- Other: driving licences, vehicles, economic activity, rainfall, twitter activity, flatness, aggregated GPS logs
- NTS: for a carefully managed random sample:number of trips taken by motorbike in a week, the average annual mileage
- Other surveys e.g. APS cycling activity (not necessarily on road)

How to combine this information to be

- Useful
- Not overstate the accuracy of the combining process

Zero Inflated Poisson Model

- Assume there are two kinds of people $Z = 1$ if they ride a motorcycle and $Z = 0$ if they don't.
- Model $Z \sim Bernoulli(p)$
- Model $Y|Z = 1$ as $Poisson(\lambda)$

For example, according to NTS

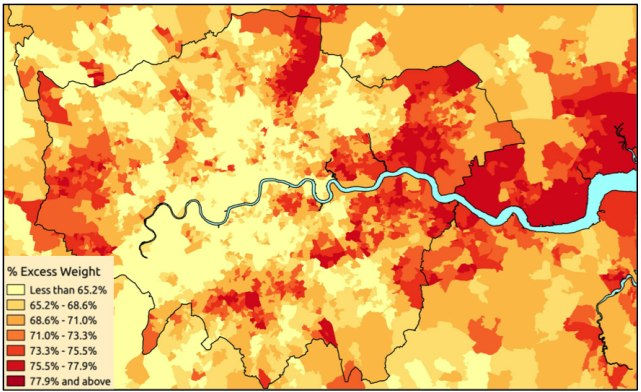
- Baseline $P(Z = 1) = 0.0001011964$
- Adjust this:

16-20	Up a lot
21-39	Up a bit
40-64	Up a bit
Over 65	Up
Ride bike to work	Up massively
Female	Down a lot

Combining with administrative data

- Either: Standard raking buys you a lot of detail e.g. Travel to work by Motorbike only released at LA level, small area micro-data suggests how to achieve age/gender breakdown. Can look at lower level geography (motorbikes grouped with "other") and combine to match local known totals and local authority totals.
- Or: Combine model parameters with census data to create simulations of Britain; multilevel models that account for urban/flat/rainy nature of higher level geography
- Or: Both
- Methods tested by predicting known quantities and checking.

Widely accepted methodology: Obesity

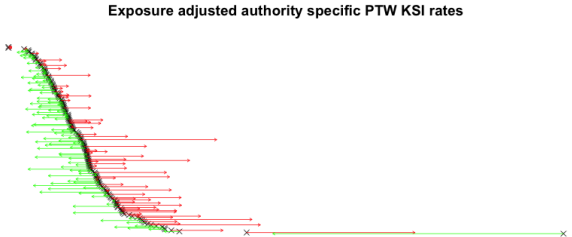


(Also, our Skills for Life 2011 small area estimated highlighted by ONS: Beyond 2011: Producing Socio-demographic Statistics)

This is not perfect knowledge

- We can't "crack" the truth out of disclosure controlled administrative data.
- But given various released sources the range of possibilities can be numerically limited, and this can provide enough information to be useful.
- Run several simulations in order to account for lack of certain knowledge

Adjusting for estimated PTW usage



After adjustment rates got:

Bigger	Smaller
W06000023	W06000004
E06000017	E06000044
E06000046	E09000020
W06000002	E06000008
E10000013	E09000033
E10000027	E09000001

Summary

- Even simple denominators such as population counts can be powerfully informative (if you think about the counterfactuals)
- What can you assume about exposure by age/gender/over time?
- Lots of publicly available datasets provide at worst interesting proxy variables and at best a route into estimating relative exposure
- I'm very very very keen to develop this work further in the real world if anyone is good at project managing from a distance

Paul Hewson
www.plymouth.ac.uk/staff/phewson
paul.hewson@plymouth.ac.uk