Practical issues in model design

CAS Special Interest Seminar on Predictive Modeling

Boston, October 2006

James Tanser

WWW.WATSONWYATT.COM

Watson Wyatt Worldwide

Agenda

- Variates, Polynomials and Splines
- Continuous interactions
- Practical problems with continuous variables
- Over-dispersed Poisson

Agenda

- Variates, Polynomials and Splines
- Continuous interactions
- Practical problems with continuous variables
- Over-dispersed Poisson

Variates, Polynomials and Splines

- Common to model using treating all variables categorically, with discrete levels
  - Allows actual shape to show through
  - Produces step changes in genuinely continuous variables (eg AOI)
  - No extrapolation
- Some variables (age) are both continuous and discrete
  - Step changes are acceptable
  - Some smoothness is desirable

Variates, Polynomials and Splines

- Variates allow each unique data value to have a different effect on the linear predictor, but force some smoothness
- In practice implemented via:
  - Polynomials

  or

  - Splines

Agenda

- Variates, Polynomials and Splines
- Continuous interactions
- Practical problems with continuous variables
- Over-dispersed Poisson

Continuous interactions

- Interaction = multiplication
- Continuous * Discrete
- Continuous * Continuous

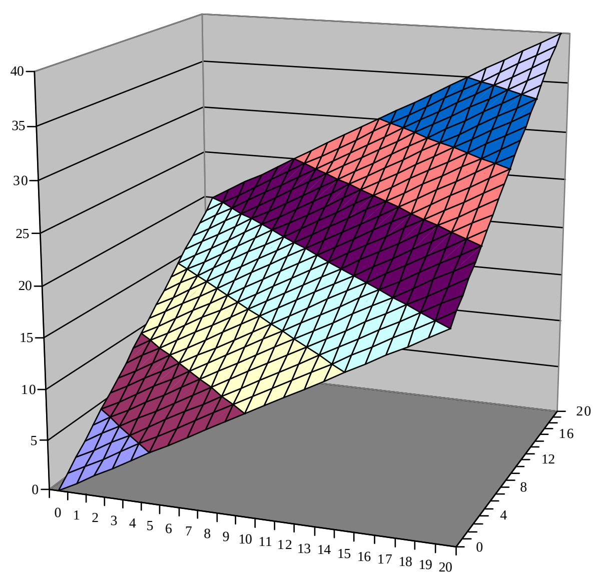
Interaction = Multiplication

- The notation is the clue:
  - A.B or A*B
- For example:
  - Two variates X and Y
  - Model should be linear in X and linear in Y
  - Interaction between X and Y to be included

Interaction = multiplication

- Model 1: Z = aX + bY
  - Linear in X
  - Linear in Y
  - No interaction – "a" does not depend on Y

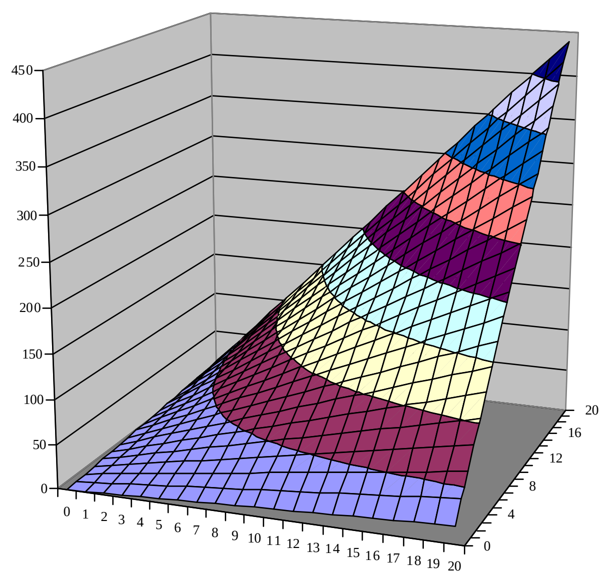
Interaction = Multiplication
Z = aX + bY



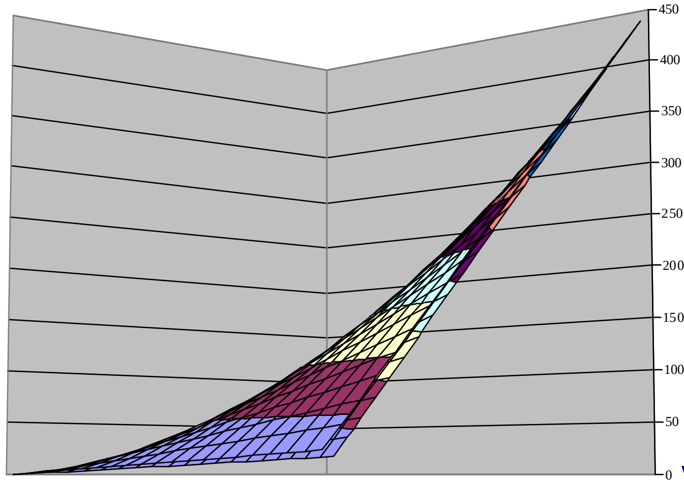
Interaction = multiplication

- Model 1: Z = aX + bY
  - Linear in X
  - Linear in Y
  - No interaction – "a" does not depend on Y
- Model 2: Z = aX + bY + cXY
  - Linear in X (for any given Y)
  - Linear in Y (for any given X)
  - Interaction present – the gradient for X depends on the value of Y (and vice versa)
  - Quadratic in X=Y direction

Interaction = Multiplication
Z = aX + bY + cXY



Interaction = Multiplication
$Z = aX + bY + cXY$



Continuous interactions

- Interaction = multiplication
- Continuous * Discrete
- Continuous * Continuous

Continuous * Discrete

- Define a new set of variates, one for each factor level, so that variate n is:
  - Variate value if factor at level n
  - Zero otherwise
- Treat each of these variates as usual:
  - Polynomial (same order?)
  - Spline (same knots?)
- Useful to include factor in model for neatness

Design matrix
Spline

18	M	⇒	1	1	1	0	0	…	1
20	F		1	0.52	0.98	0.02	0	…	0
22	F		1	0.17	0.83	0.17	0	…	0
24	M		1	0.02	0.5	0.48	0.02	…	1
26	M		1	0	0.17	0.67	0.17	…	1
28	M		1	0	0.02	0.48	0.48	…	1
30	F		1	0	0	0.17	0.67	…	0
32	F		1	0	0	0.02	0.48	…	0
34	M		1	0	0	0	0.17	…	1
36	F		1	0	0	0	0.02	…	0

Design matrix
Discrete * Spline

1	1	1	0	0	…	0	0	0	0	…
1	0	0	0	0	…	0.52	0.98	0.02	0	…
1	0	0	0	0	…	0.17	0.83	0.17	0	…
1	0.02	0.5	0.48	0.02	…	0	0	0	0	…
1	0	0.17	0.67	0.17	…	0	0	0	0	…
1	0	0.02	0.48	0.48	…	0	0	0	0	…
1	0	0	0	0	…	0	0	0.17	0.67	…
1	0	0	0	0	…	0	0	0.02	0.48	…
1	0	0	0	0.17	…	0	0	0	0	…
1	0	0	0	0	…	0	0	0	0.02	…

Design matrix
Discrete * Polynomial

1	18	324	5832	104976	…	0	0	0	0	…
1	0	0	0	0	…	20	400	8000	160000	…
1	0	0	0	0	…	22	484	10648	234256	…
1	24	576	13824	331776	…	0	0	0	0	…
1	26	676	17576	456976	…	0	0	0	0	…
1	28	784	21952	614656	…	0	0	0	0	…
1	0	0	0	0	…	30	900	27000	810000	…
1	0	0	0	0	…	32	1024	32768	1048576	…
1	34	1156	39304	1336336	…	0	0	0	0	…
1	0	0	0	0	…	36	1296	46656	1679616	…

Continuous interactions

- Interaction = multiplication
- Continuous * Discrete
- Continuous * Continuous

Continuous * Continuous

- Simply create X*Y terms!
- Eg Polynomial order 2 for X and Y:
  - X, X^2, Y, Y^2
  - XY
  - XY^2, X^2Y
  - X^2Y^2
- For splines combine together all the basis functions
  - $f_1(x)$, $f_2(x)$, $f_3(x)$, …, $g_1(y)$, $g_2(y)$, $g_3(y)$, …
  - $f_1(x).g_1(y)$, $f_2(x).g_1(y)$, $f_3(x).g_1(y)$
  - …

Design matrix
Spline * Spline

- Not enough room on slide to show!

Agenda

- Variates, Polynomials and Splines
- Continuous interactions
- Practical problems with continuous variables
- Over-dispersed Poisson

Practical problems

- Number of variates
- Missing values
- Edge effects

Number of variates

- Various tricks used to make modeling with categorical factors quick
  - Only one calculation per factor per row
  - Calculation is addition
- Tricks don't work variates and calculation is multiplication
  - Polynomial with 5 terms is slower than adding 5 new factors

Number of variates

- Splines make it easy to include many variates, slowing down the model
  - Use only at final modelling stages
  - Be parsimonious!
- Interactions with variates creates many (tens or hundreds) of variates very quickly

Practical problems

- Number of variates
- Missing values
- Edge effects

Missing values

- Missing values in a variate often cause entire record to be ignored
  - Replace missing values with zeros
- Care is needed to differentiate "real" zeros and "missing" zeros
  - Create a missing flag and include in all models involving variate
  - Remember spline basis functions transform zero to some other (non-zero) value (extrapolation)

Practical problems

- Number of variates
- Missing values
- Edge effects

Edge effects

- One or two records with extreme variate values can have a disproportionate effect on the model
  - Look at leverage or Cook's distance
  - Understand your data
  - Consider limiting range of variate
  - Be careful when extrapolating

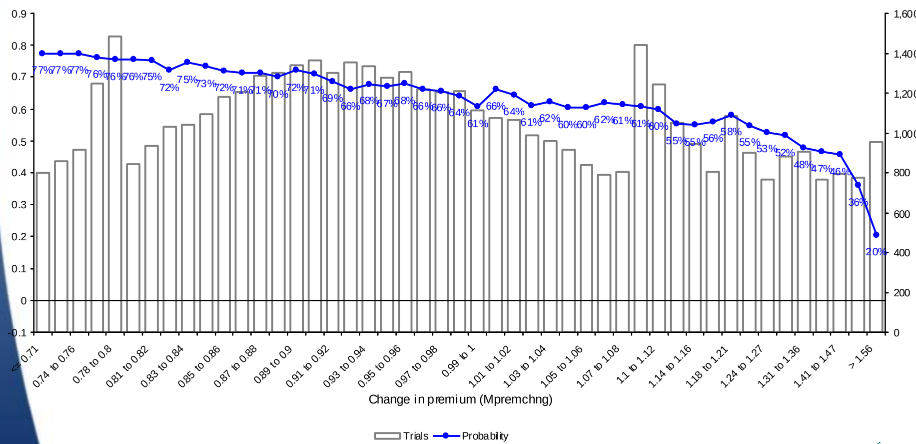
Cautionary example

- Artificial data, loosely based on actual naive analysis
- Retention analysis containing three records with incorrect premium change, all of which renewed
- Problems:
  - Overfitting to edges
  - Knot placement

Simple grouped oneway

Retention job

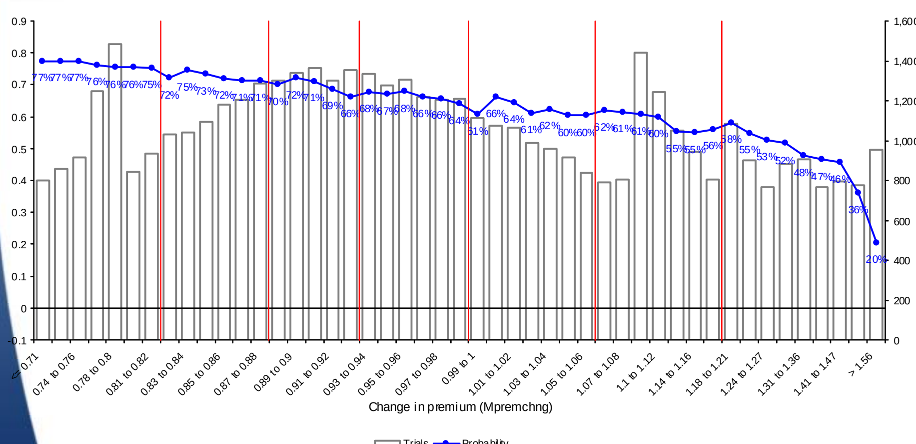
Example of problem factor



Simple grouped oneway

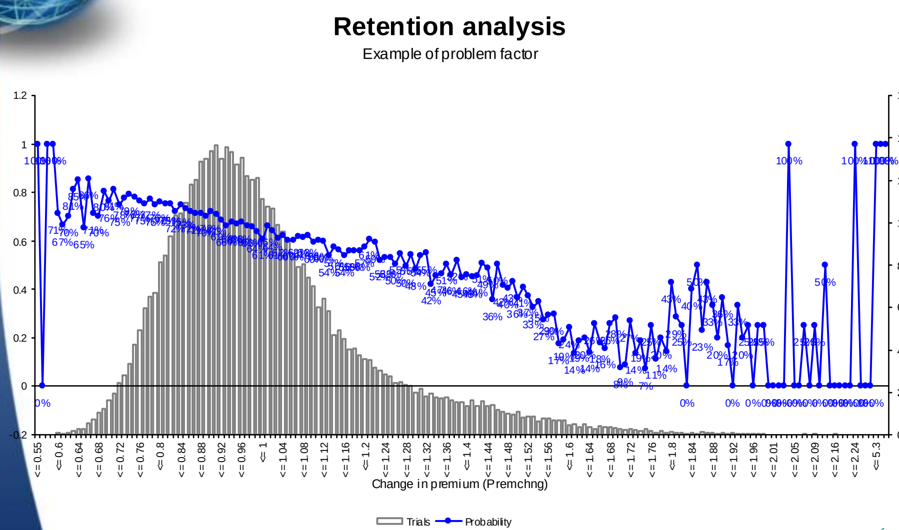
Retention job

Example of problem factor

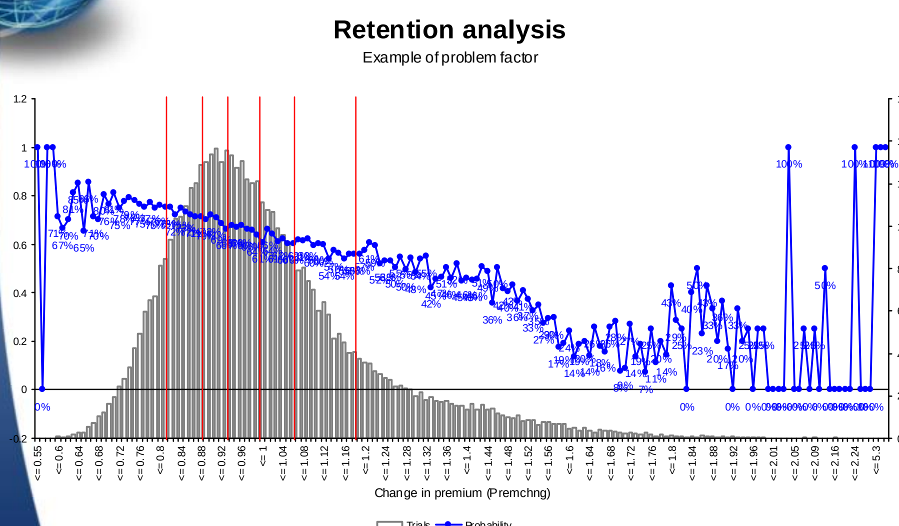


Copyright © Watson Wyatt Worldwide. All rights reserved.

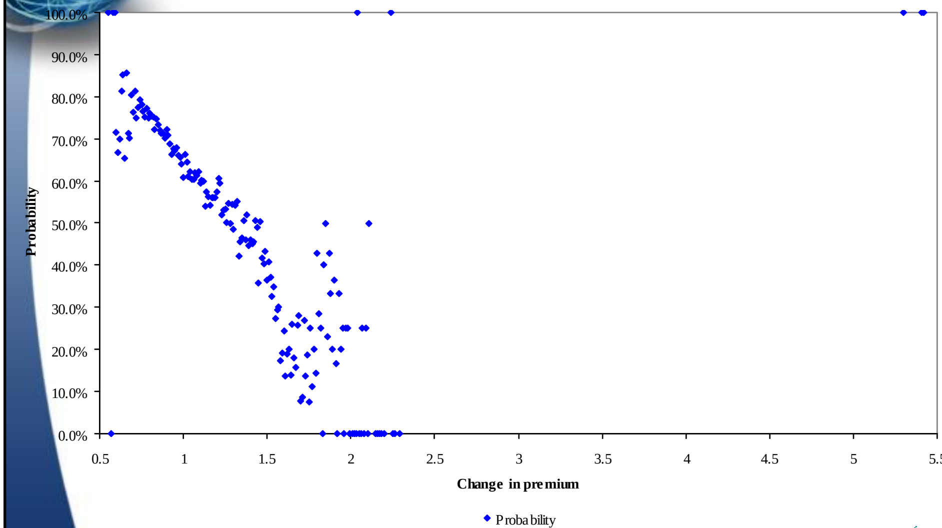
Detailed oneway



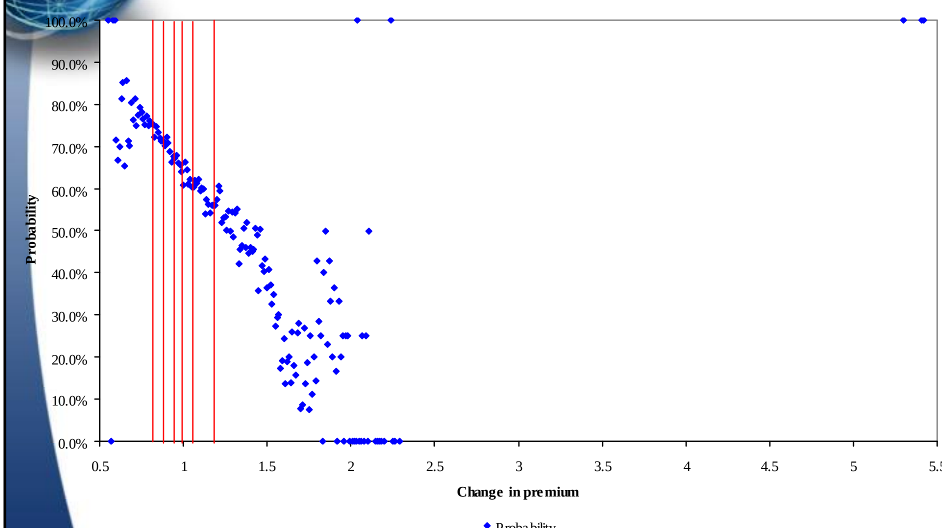
Detailed oneway



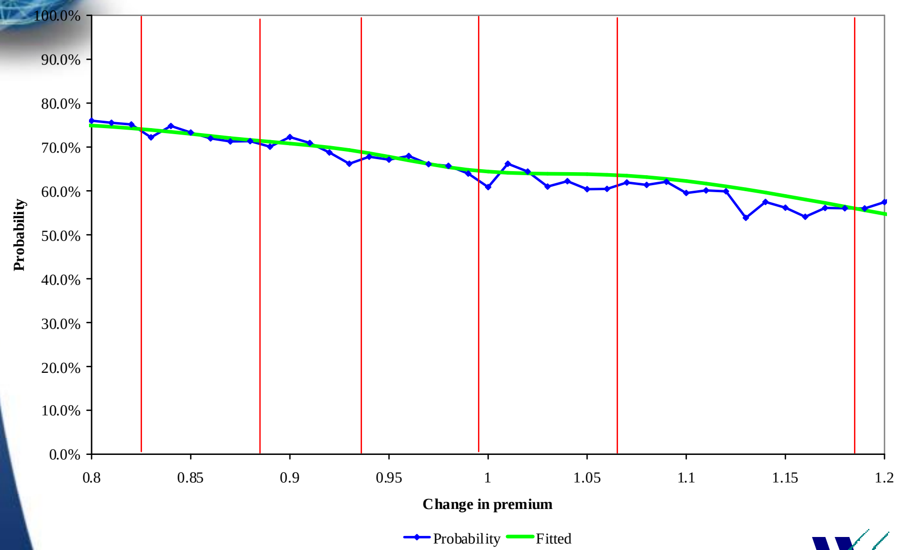
X-Y plot



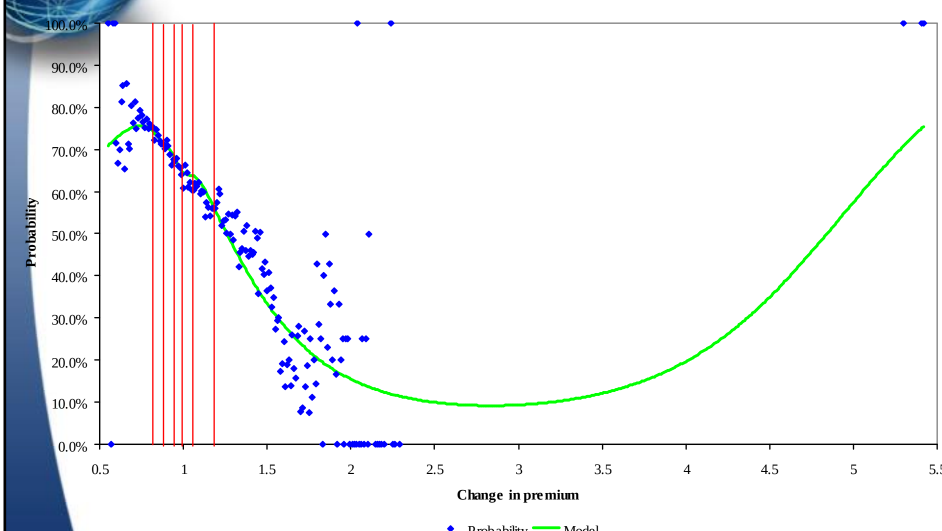
X-Y plot



Model results



Model results



Agenda

- Variates, Polynomials and Splines
- Continuous interactions
- Practical problems with continuous variables
- Over-dispersed Poisson

Interdependence of claim events

- Sometimes claim events are not individually independent
- A real life example:
  - Health insurer recorded data such that individual PMI claim payments could not be matched to a particular single medical event
  - this meant that each transactional claim payment was recorded as a new claim regardless of the event
  - claim numbers therefore appear in multiples per policy in the data
- This invalidates assumptions underlying the GLM framework

Mathematical implications

- Poisson model used for numbers assumes

$$E[\underline{Y}] = \underline{\mu} \qquad Var[\underline{Y}] = \underline{\mu}$$

- Replacing every one claim with K claims gives

$$E[\underline{Y}] = K.\underline{\mu} \qquad Var[\underline{Y}] = K^2.\underline{\mu}$$

- But the Poisson GLM modelling process applies the Poisson assumptions which are, in this case wrong!

$$E[\underline{Y}] = K.\underline{\mu} \qquad Var[\underline{Y}] = K.\underline{\mu}$$

Mathematical implications

- Fitting a Poisson GLM to claims data that is not independent does not effect the parameter estimates
- But it does effect the standard errors!

Generalised linear models

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

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

scale parameter

- Inclusion of a scale parameter adjusts the variance assumed in the model

Estimating the scale parameter

- Deviance scale for Poisson

$$\phi = D / (n - p)$$

- Pearson scale

$$\chi^2 = \Sigma\ \omega_i\ (Y_i - \mu_i)^2 / V(\mu_i)$$

$$\phi = \chi^2 / (n - p)$$

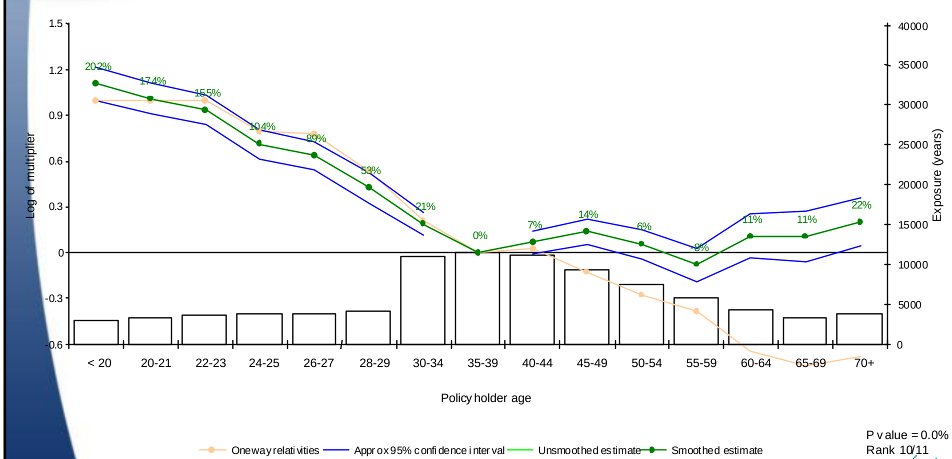
Theoretical case study

- Fitted numbers model to the TPPD claims numbers
- Data was adjusted by
  - multiplying exposure by 1000
  - multiplying claims numbers by 10
- Tried fitting models
  - Poisson
  - overdispersed Poisson (with Deviance scale)
  - overdispersed Poisson (with Pearson scale)

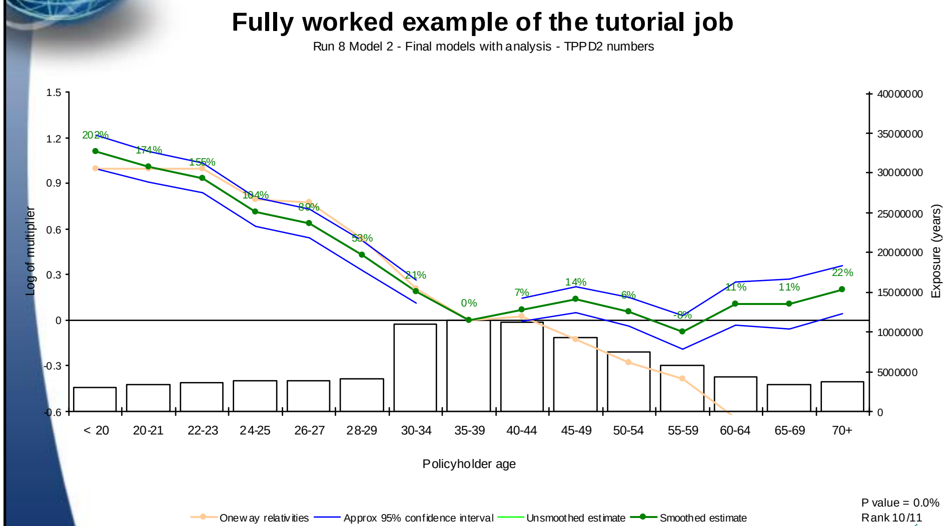
The correct answer

Fully worked example of the tutorial job

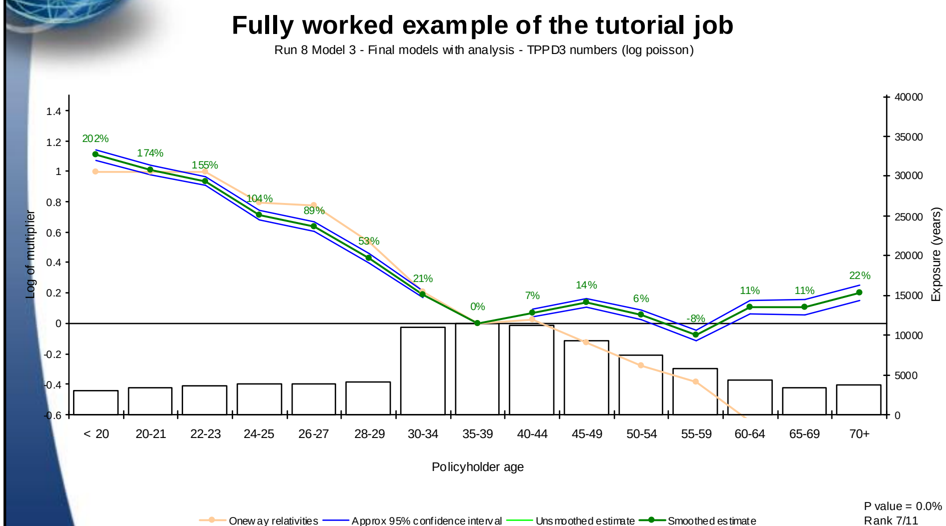
Run 8 Model 1 - Final models with analysis - TPPD numbers



Exposure * 1,000



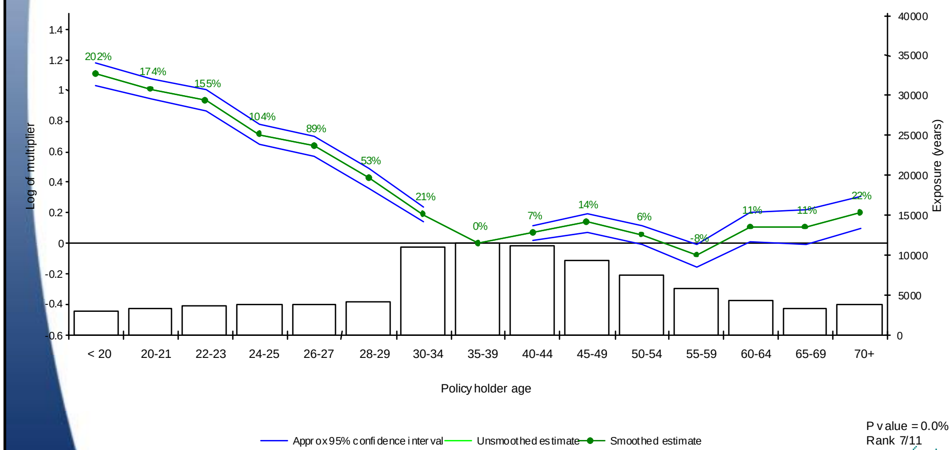
Num * 10



Overdispersed Poisson (deviance)

Fully worked example of the tutorial job

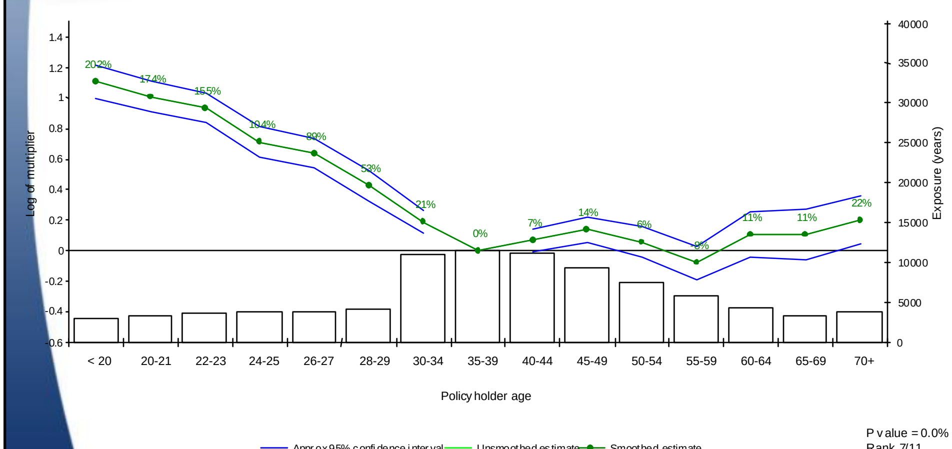
Run 8 Model 4 - Final models with analysis - TPPD3 numbers (log over-dispersed deviance poisson)



Overdispersed Poisson (Pearson)

Fully worked example of the tutorial job

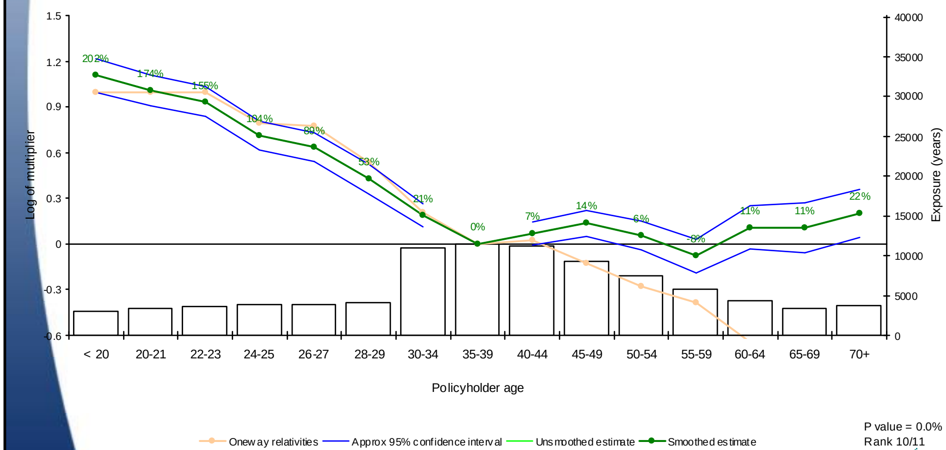
Run 8 Model 5 - Final models with analysis - TPPD3 numbers (log over-dispersed pearson poisson)



The correct answer

Fully worked example of the tutorial job

Run 8 Model 1 - Final models with analysis - TPPD numbers



Practical issues in model design

CAS Special Interest Seminar on Predictive Modeling

Boston, October 2006

James Tanser

Watson Wyatt Worldwide