GLM III: Advanced Modeling Strategy

2006 CAS Seminar on Predictive Modeling

WWW.WATSONWYATT.COM

Duncan Anderson MA FIA

Watson Wyatt Worldwide

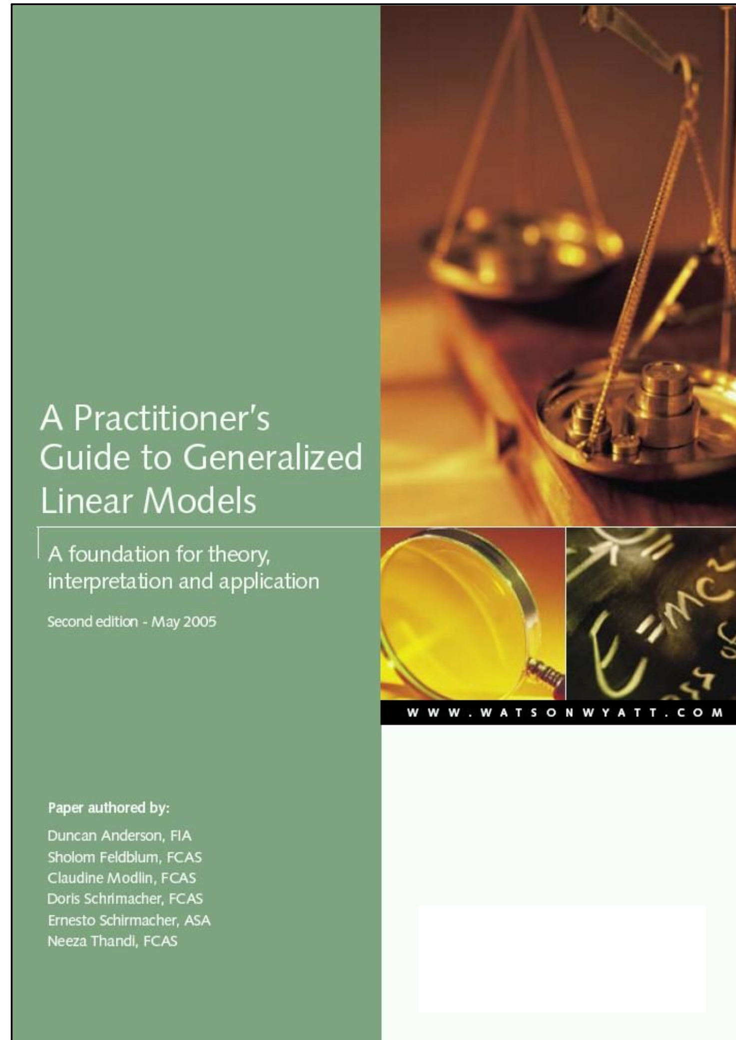
Watson Wyatt Worldwide

Agenda

- Introduction
- Testing the link function
- The Tweedie distribution
- Splines
- Reference models
- Aliasing / near aliasing
- Combining models across claim types
- Restricted models
- Model validation

"A Practitioner's Guide to GLMs"

- 2004 CAS Discussion Paper Program
- Discusses
  - testing the link function
  - the Tweedie distribution
  - aliasing / near aliasing
  - combining models across claim types
  - restricted models
- Copies available here



Generalized linear models

$$E[Y_i] = \mu_i = g^{-1}(\Sigma X_{ij}\beta_j + \xi_i)$$

$$Var[Y_i] = \phi.V(\mu_i)/\omega_i$$

- Consider all factors simultaneously
- Provide statistical diagnostics
- Allow for nature of random process
- Robust and transparent
- Increasingly a global industry standard

Generalized linear models

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

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

Generalized linear models

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

Link function

Offset term

Y-variate

Design matrix

Parameter estimates

Generalized linear models

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

Some function (user defined)

Observed thing (data)

Some matrix based on data (user defined)

Parameters to be estimated (the answer!)

Known effects

Generalized linear models

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

Scale parameter

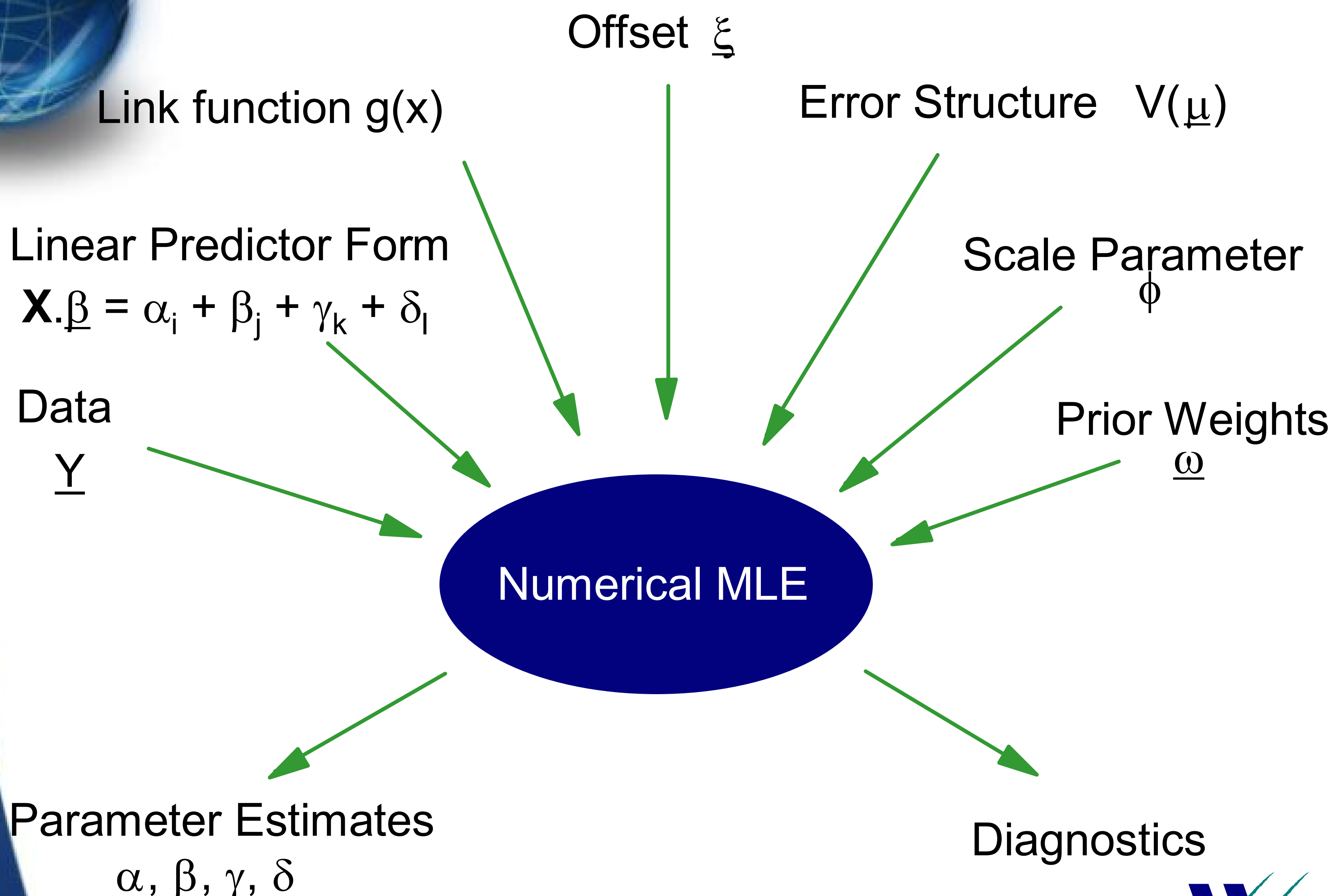
Variance function

Prior weights

- Usually assume exponential family, eg
- $\phi = \sigma^2$ (estimated), $V(x) = 1$ $\Rightarrow \mathrm{Var}[Y_i] = \sigma^2$ Normal
- $\phi = 1$ (specified), $V(x) = x$ $\Rightarrow \mathrm{Var}[Y_i] = \mu_i$ Poisson
- $\phi = k$ (estimated), $V(x) = x^2$ $\Rightarrow \mathrm{Var}[Y_i] = k\mu_i^2$ Gamma

Agenda

- Introduction
- Testing the link function
- The Tweedie distribution
- Splines
- Reference models
- Aliasing / near aliasing
- Combining models across claim types
- Restricted models
- Model validation



Offset $\underline{\xi}$
Link function g(x)
Error Structure V($\underline{\mu}$)
Linear Predictor Form
$\mathbf{X}.\underline{\beta} = \alpha_i + \beta_j + \gamma_k + \delta_l$
Scale Parameter
ϕ
Data
$\underline{Y}$
Prior Weights
$\underline{\omega}$
Numerical MLE
Parameter Estimates
$\alpha, \beta, \gamma, \delta$
Diagnostics

Model testing

- Use only those factors which are predictive
  - standard errors of parameter estimates
  - F tests / χ^2 tests on deviances
  - stepwise approach (helpful if used with care)
  - consistency over time
  - human intuition
- Make sure the model is reasonable
  - variance function: residual plots
    (histograms / Q-Q / residual vs fitted value etc)
  - outliers: leverage / Cook's distance
  - link function: Box-Cox

Box-Cox link function investigation

- GLM structure is

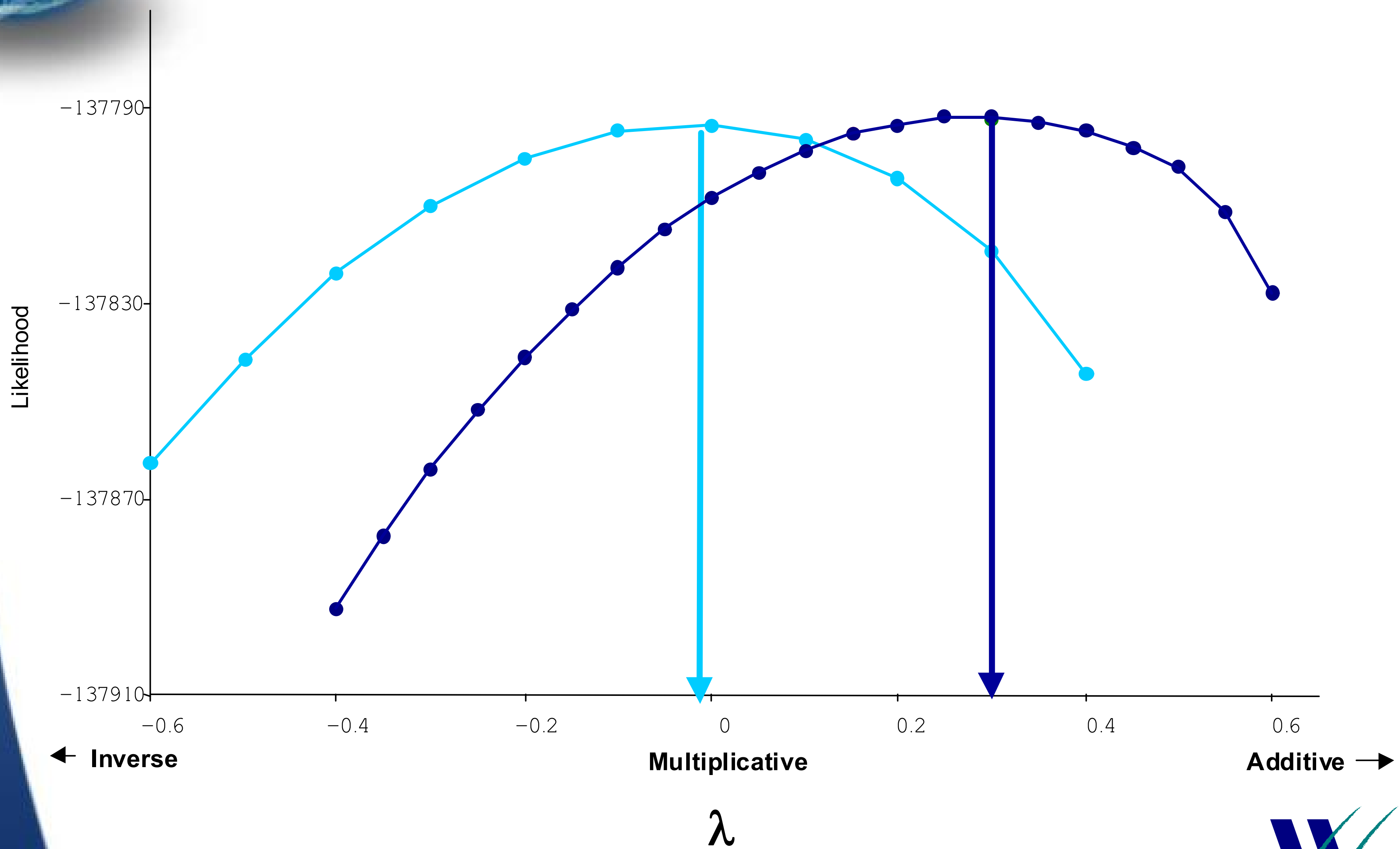
  $E[\underline{Y}] = \underline{\mu} = g^{-1}(\mathbf{X}.\underline{\beta} + \underline{\xi}) \qquad Var[\underline{Y}] = \phi.V(\underline{\mu}) / \underline{\omega}$

- Box Cox transforms defines

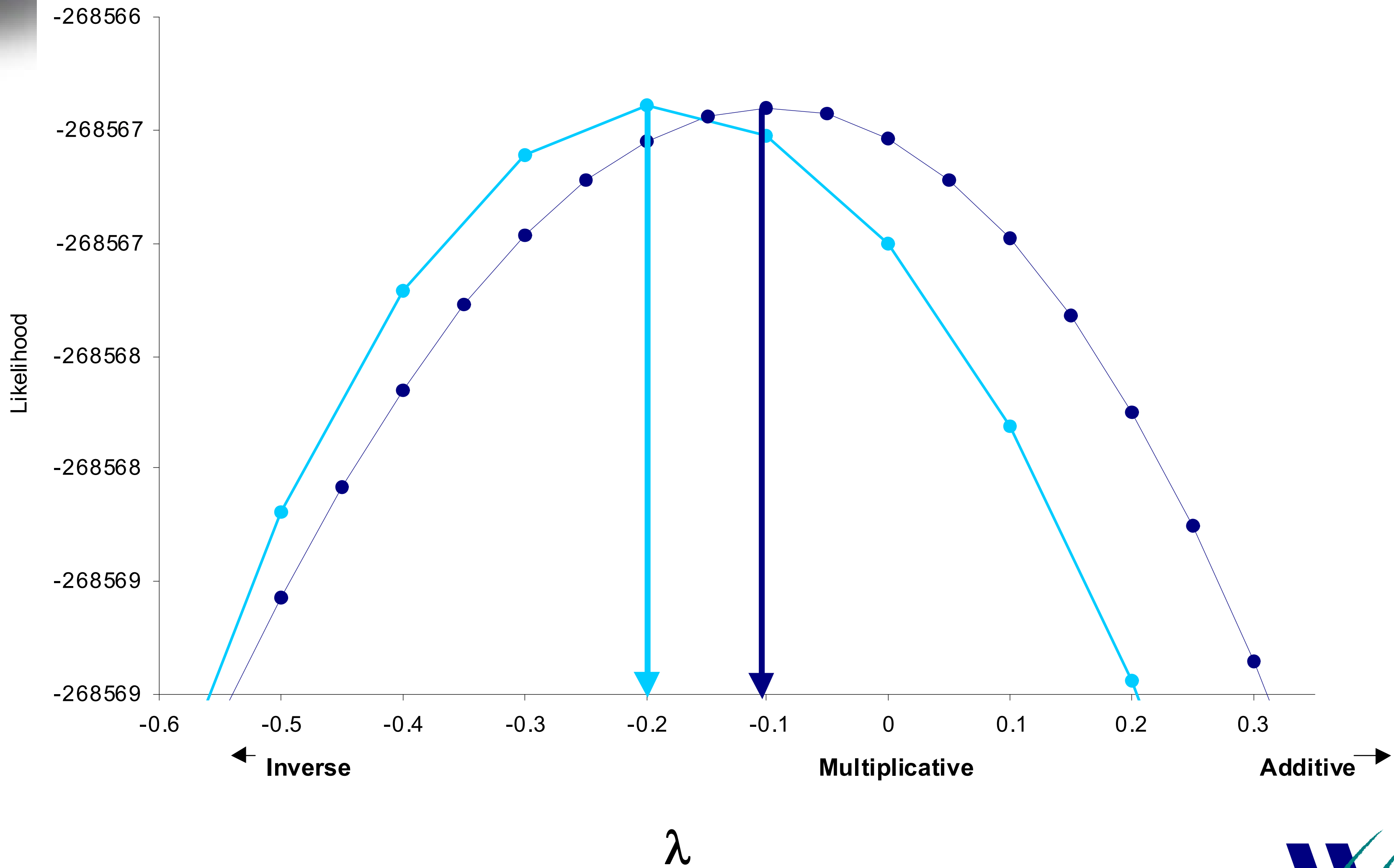
  $g(x) = (x^{\lambda} - 1) / \lambda$ for $\lambda \neq 0$, $\ln(x)$ for $\lambda = 0$

- $\lambda = 1 \Rightarrow g(x) = x - 1 \Rightarrow$ additive (with base level shift)
- $\lambda \to 0 \Rightarrow g(x) \to \ln(x) \Rightarrow$ multiplicative (via maths)
- $\lambda = -1 \Rightarrow g(x) = 1 - 1/x \Rightarrow$ inverse (with base level shift)
- Try different values of λ and measure goodness of fit to see which fits experience best

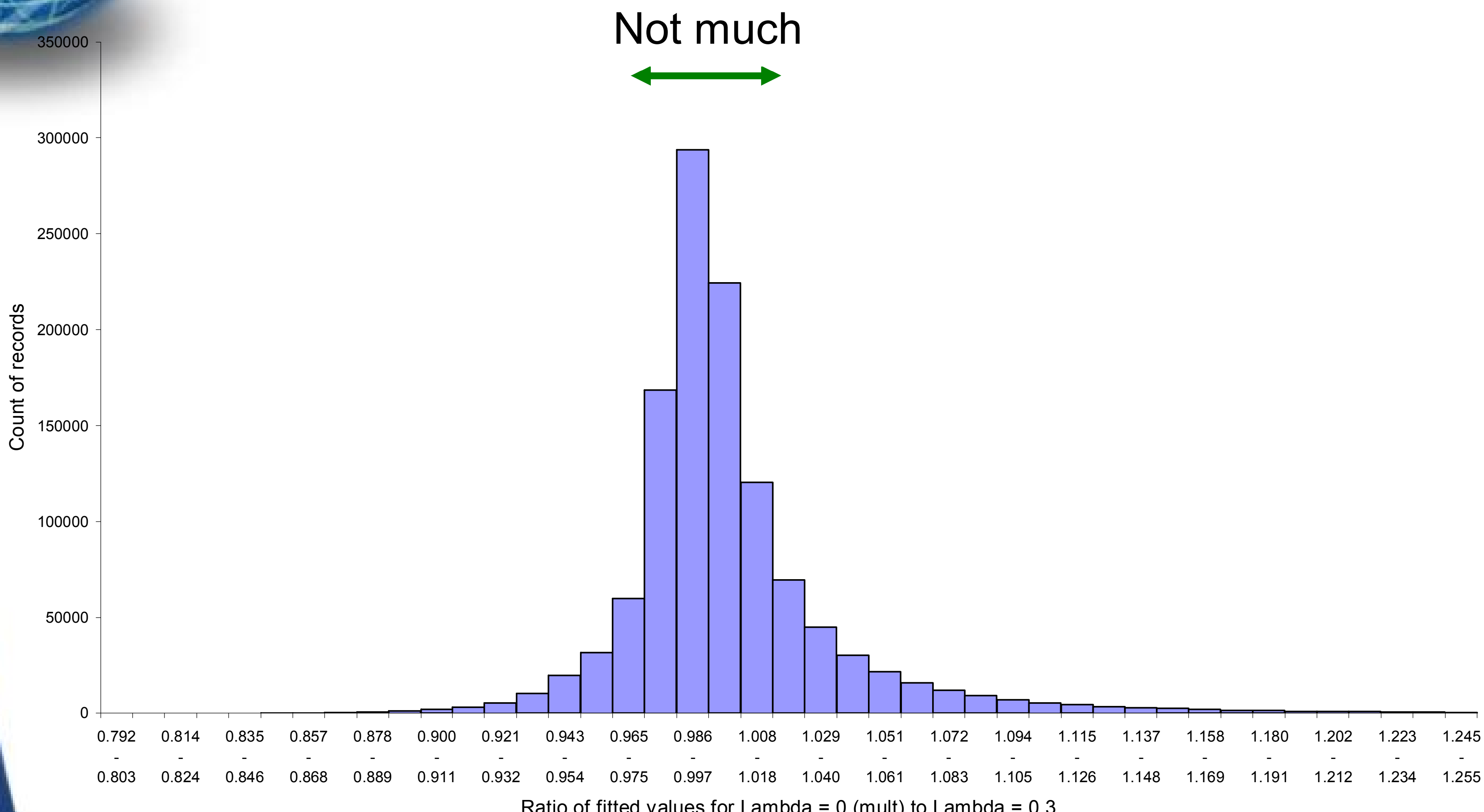
Box-Cox link function investigation
Auto third party property damage frequencies



Box-Cox link function investigation
Auto third party property damage average amounts



Box-Cox link function investigation
Comparing fitted values of different link functions

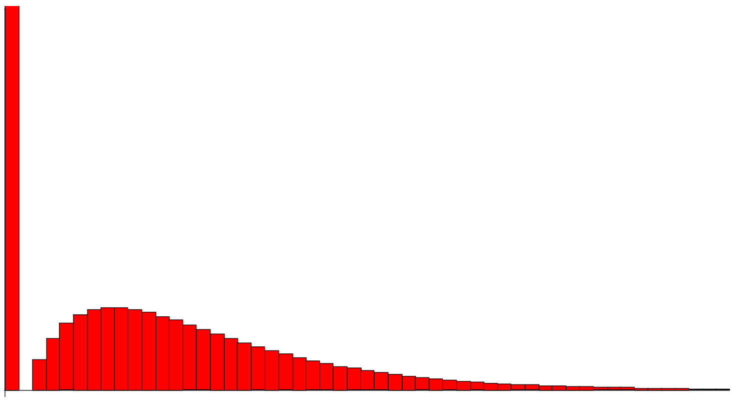


Agenda

- Introduction
- Testing the link function
- The Tweedie distribution
- Splines
- Reference models
- Aliasing / near aliasing
- Combining models across claim types
- Restricted models
- Model validation

Tweedie distributions

- Incurred losses have a point mass at zero and then a continuous distribution
- Poisson and gamma not suited to this



- Tweedie distribution can have point mass at zero and parameters which can alter the shape to be like Poisson and gamma above zero

$$f_Y(y;\theta,\lambda,\alpha) = \sum_{n=1}^{\infty} \frac{\left\{(\lambda\omega)^{1-\alpha}\kappa_\alpha(-1/y)\right\}^n}{\Gamma(-n\alpha)n!\,y}.\exp\{\lambda\omega[\theta_0 y - \kappa_\alpha(\theta_0)]\} \quad \text{for } y>0$$

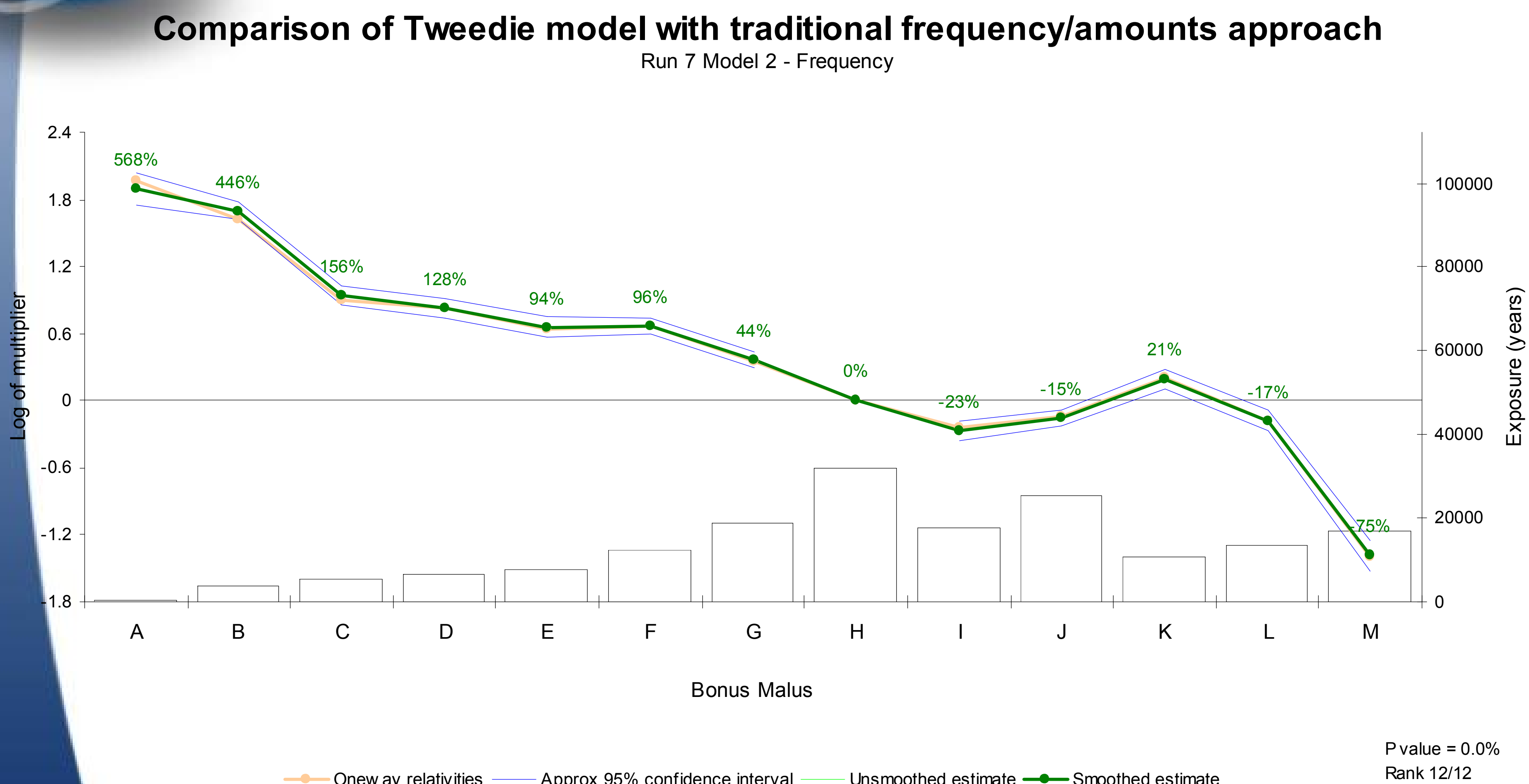
$$p(Y=0) = \exp\{-\lambda\omega\kappa_\alpha(\theta_0)\}$$

Tweedie distributions

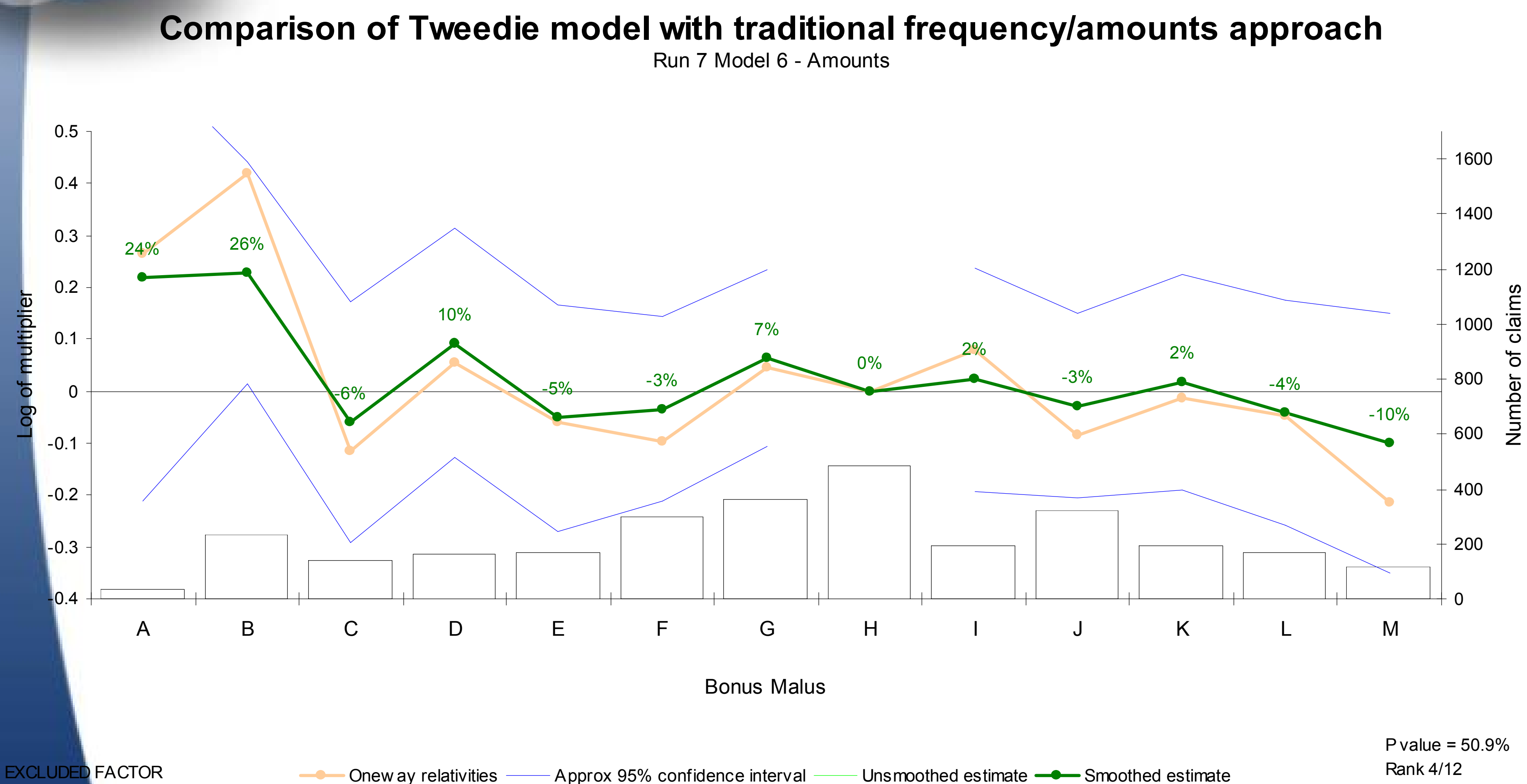
$$\text{Tweedie: } \phi = k,\ V(x) = x^p \Rightarrow \mathrm{Var}[\underline{Y}] = k\underline{\mu}^p$$

- p=1 corresponds to Poisson, p=2 to gamma
- Defines a valid distribution for $p<0$, $1<p<2$, $p>2$
- Can be considered as Poisson/gamma process for $1<p<2$
- Need to estimate both k and p when fitting models
  - often estimate a where $p = (2-a)/(1-a)$
- Typical values of p for insurance incurred claims around, or just under, 1.5

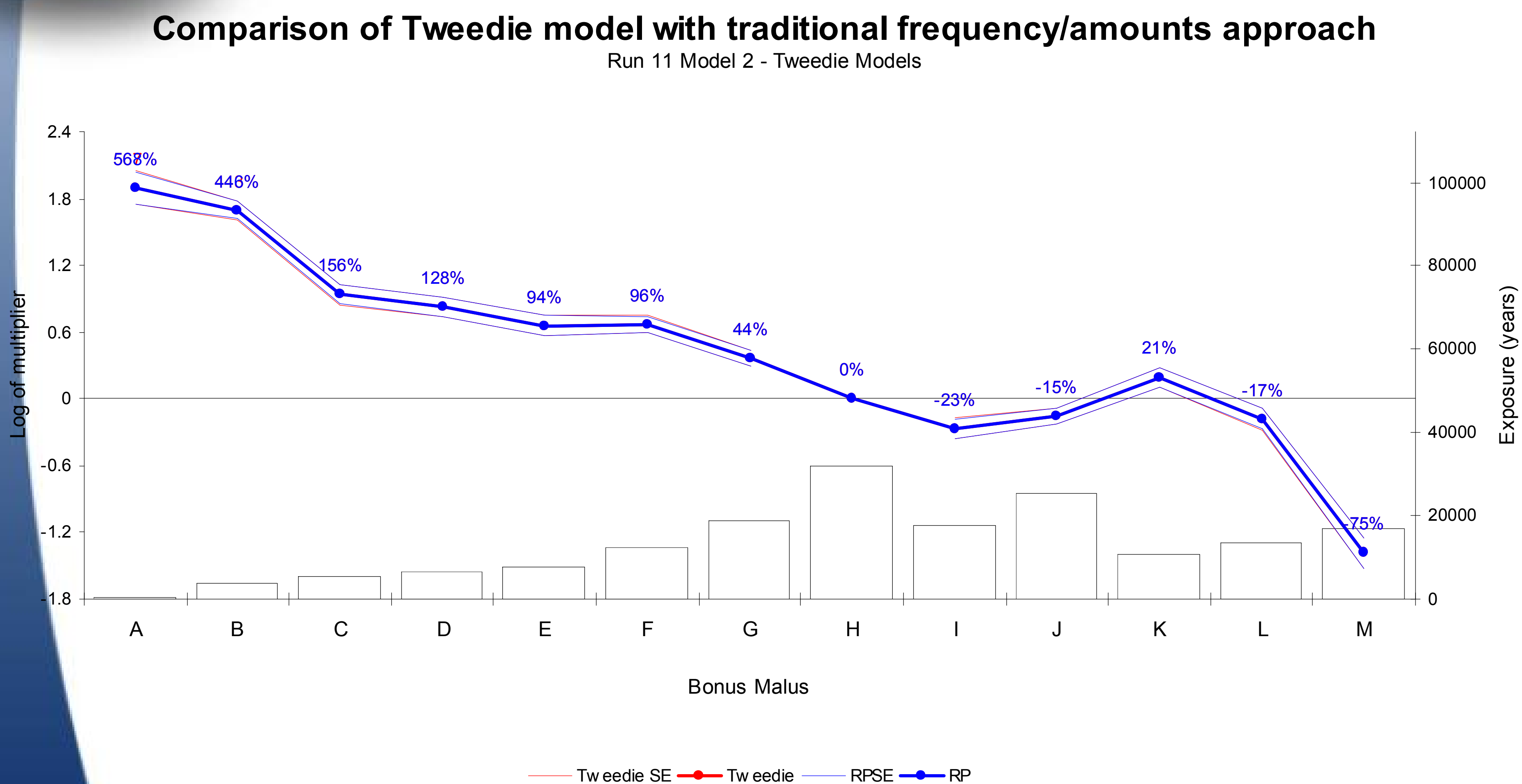
Example 1: frequency



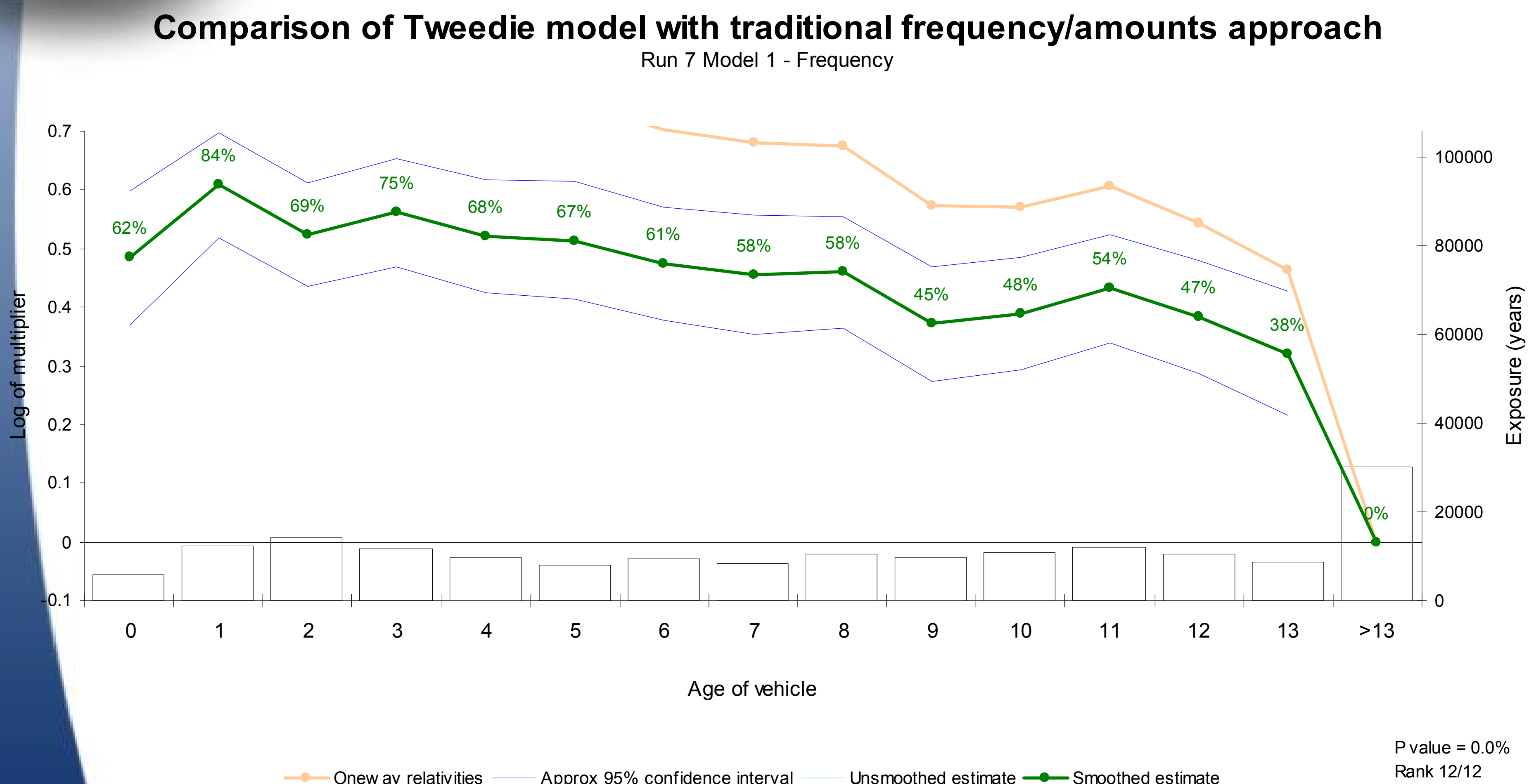
Example 1: amounts



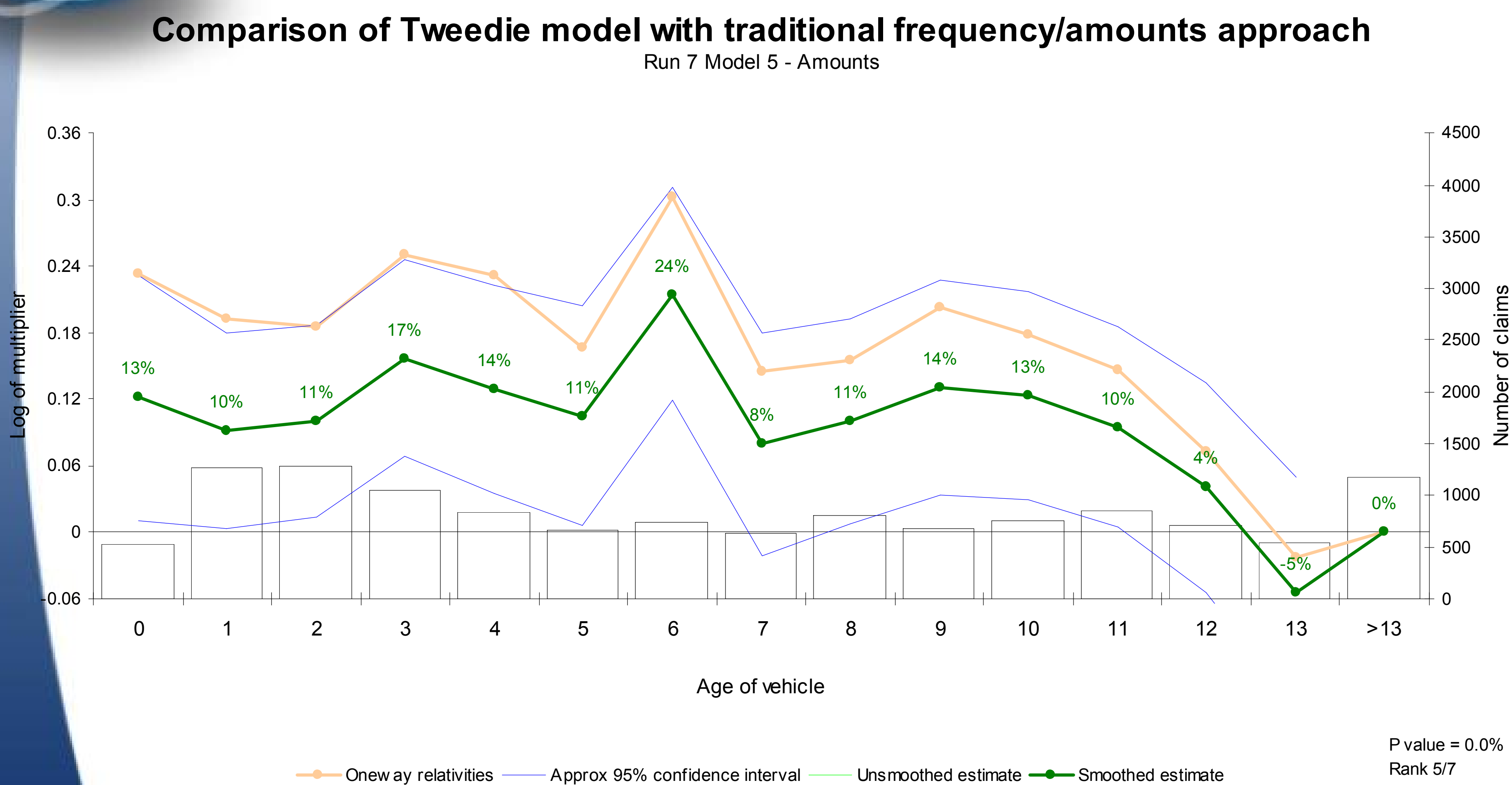
Example 1: traditional RP vs Tweedie



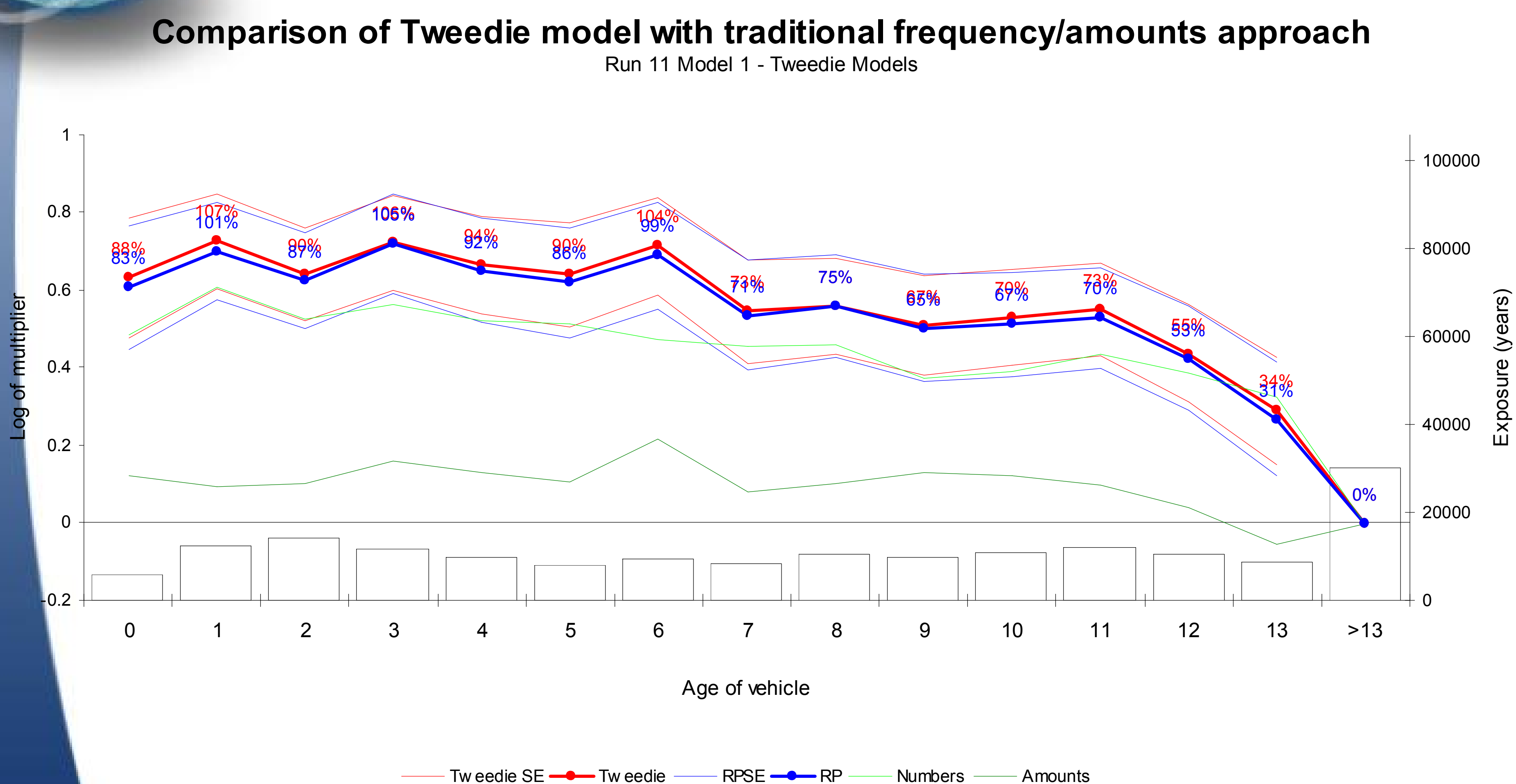
Example 2: frequency



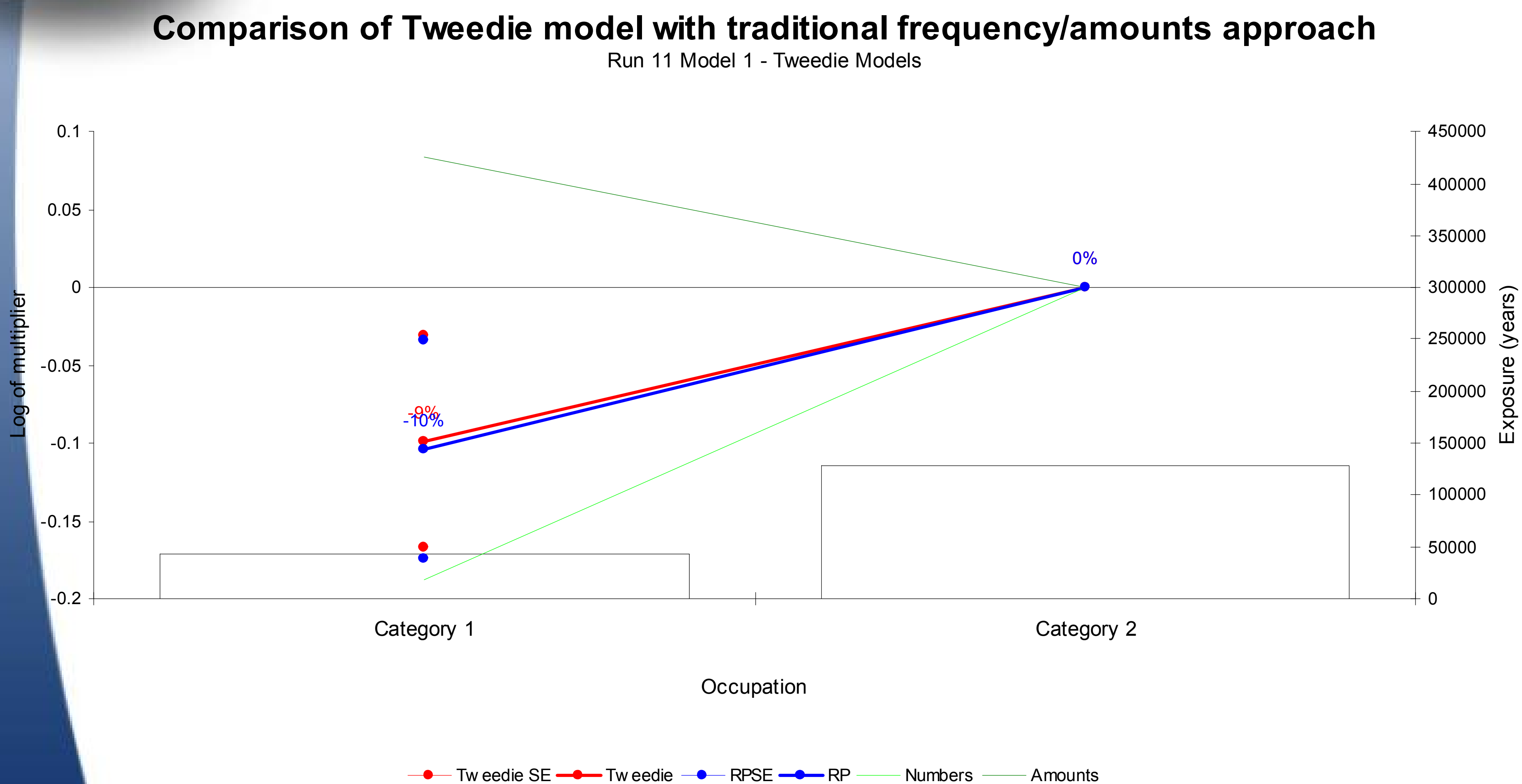
Example 2: amounts



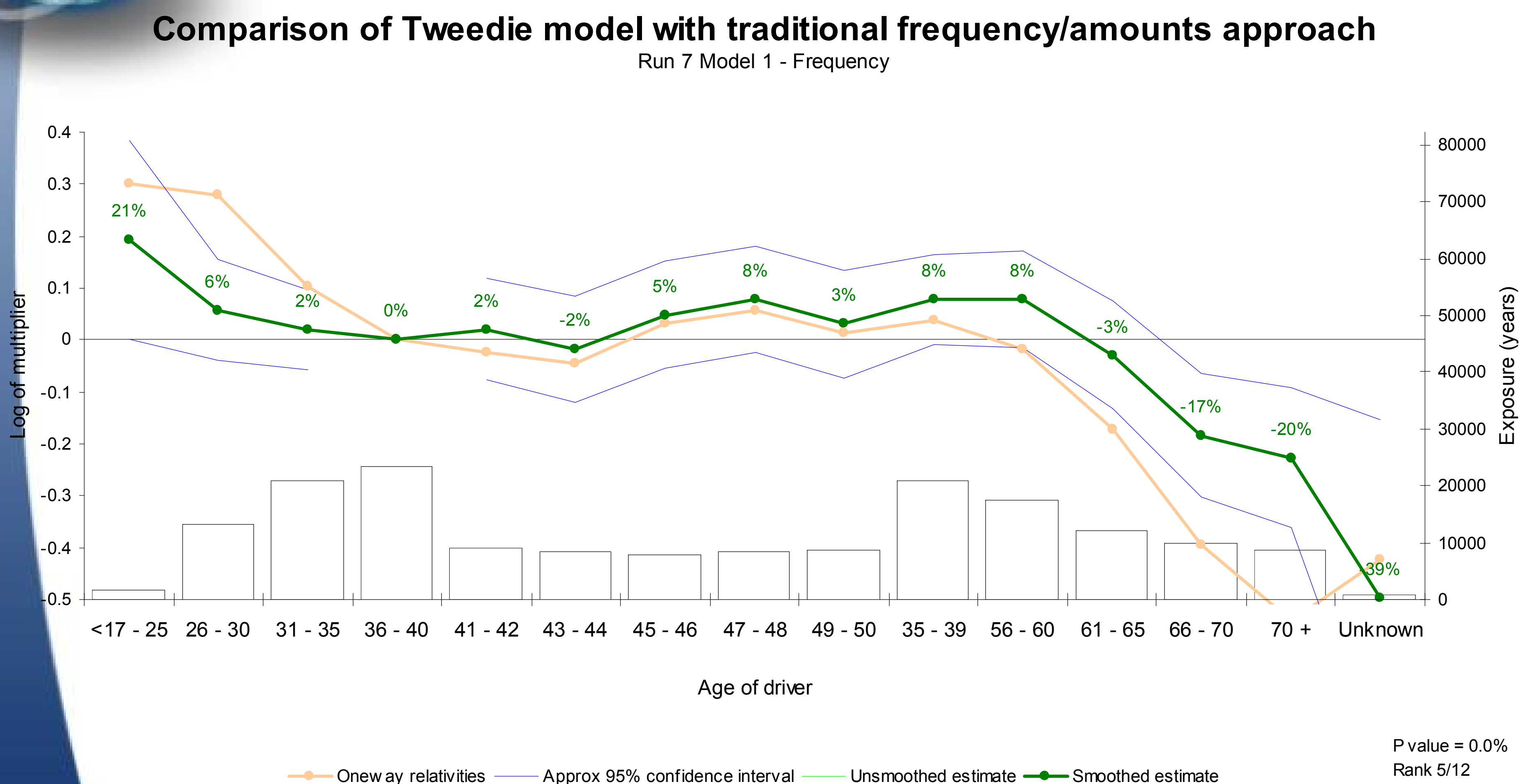
Example 2: traditional RP vs Tweedie



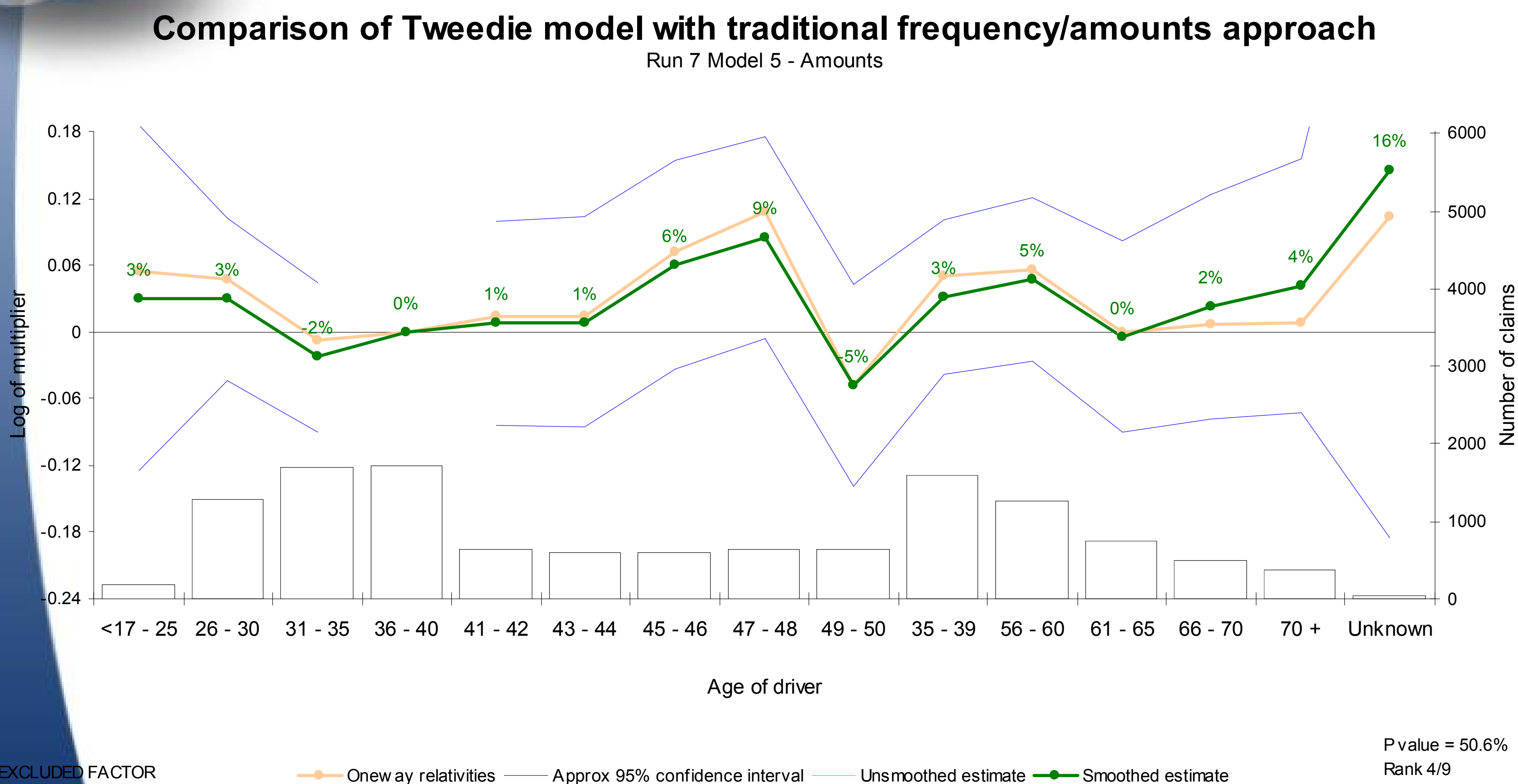
Example 3: traditional RP vs Tweedie



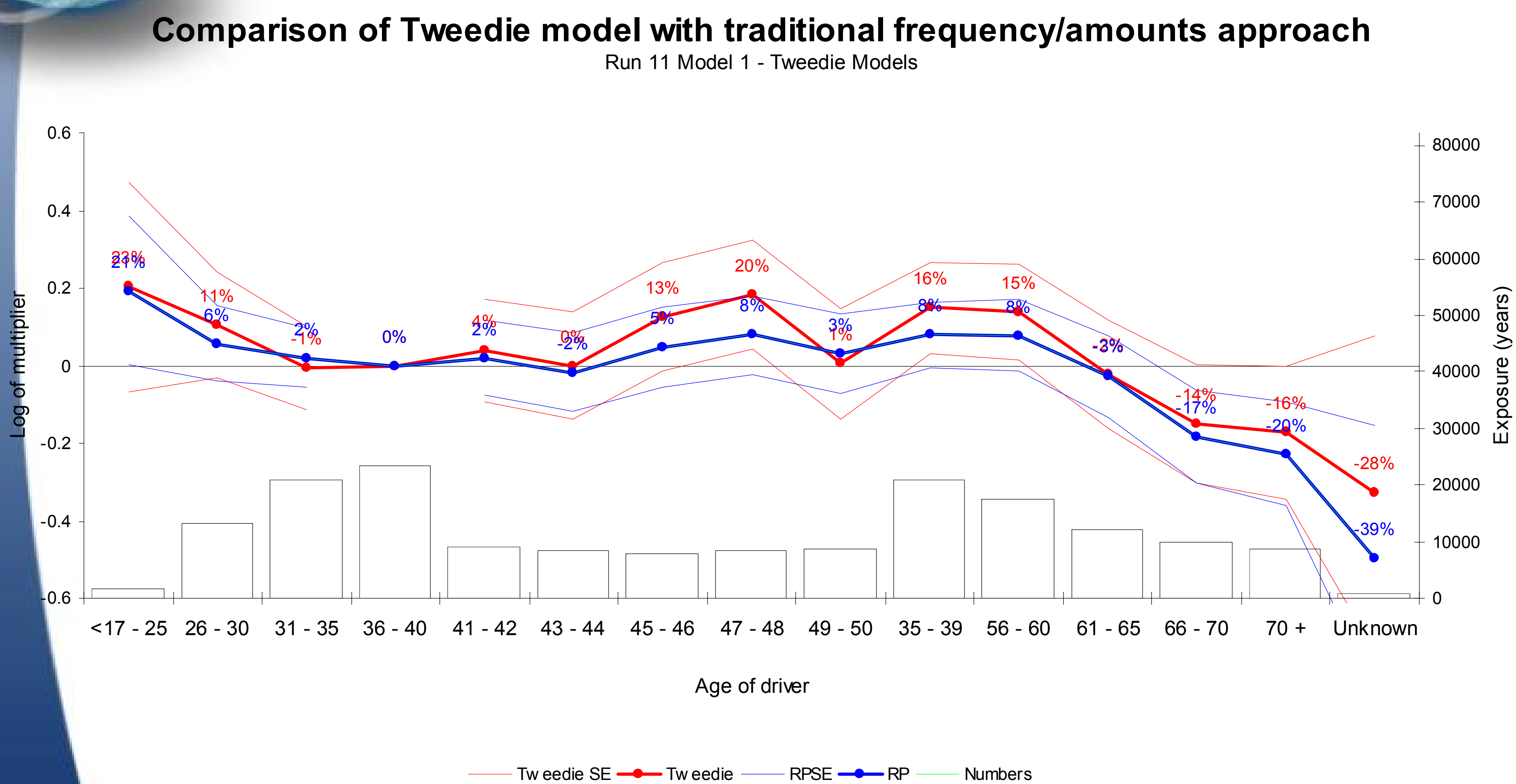
Example 4: frequency



Example 4: amounts



Example 4: traditional RP vs Tweedie

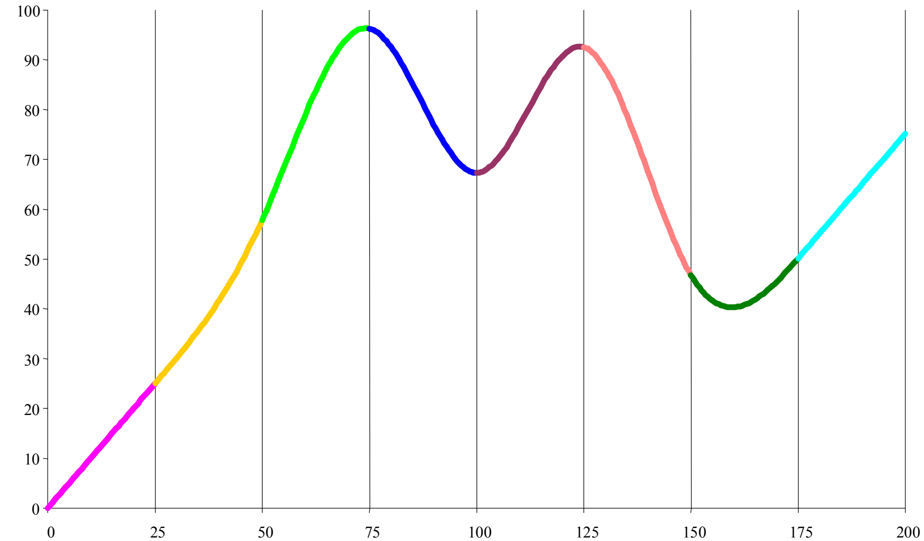


Agenda

- Introduction
- Testing the link function
- The Tweedie distribution
- Splines
- Reference models
- Aliasing / near aliasing
- Combining models across claim types
- Restricted models
- Model validation

Spline definition

- A series of polynomial functions, with each function defined over a short interval



- Intervals are defined by k+2 knots
  - two exterior knots at extremes of data
  - variable number (k) of interior knots
- At each interior knot the two functions must join "smoothly"

Cubic splines

- Each polynomial is a cubic
  - $a + bx + cx^2 + dx^3$
- "Smoothness" at interior knots is defined as:
  - continuous
  - continuous first derivative
  - continuous second derivative

Regression splines

- The position of the knots is specified by the user
- Standard GLMs can be used by careful definition of variates

- Pros
  – fits easily into existing structures
  – no complex re-sampling needed
- Cons
  – position of knots can effect final answer

Smoothing splines

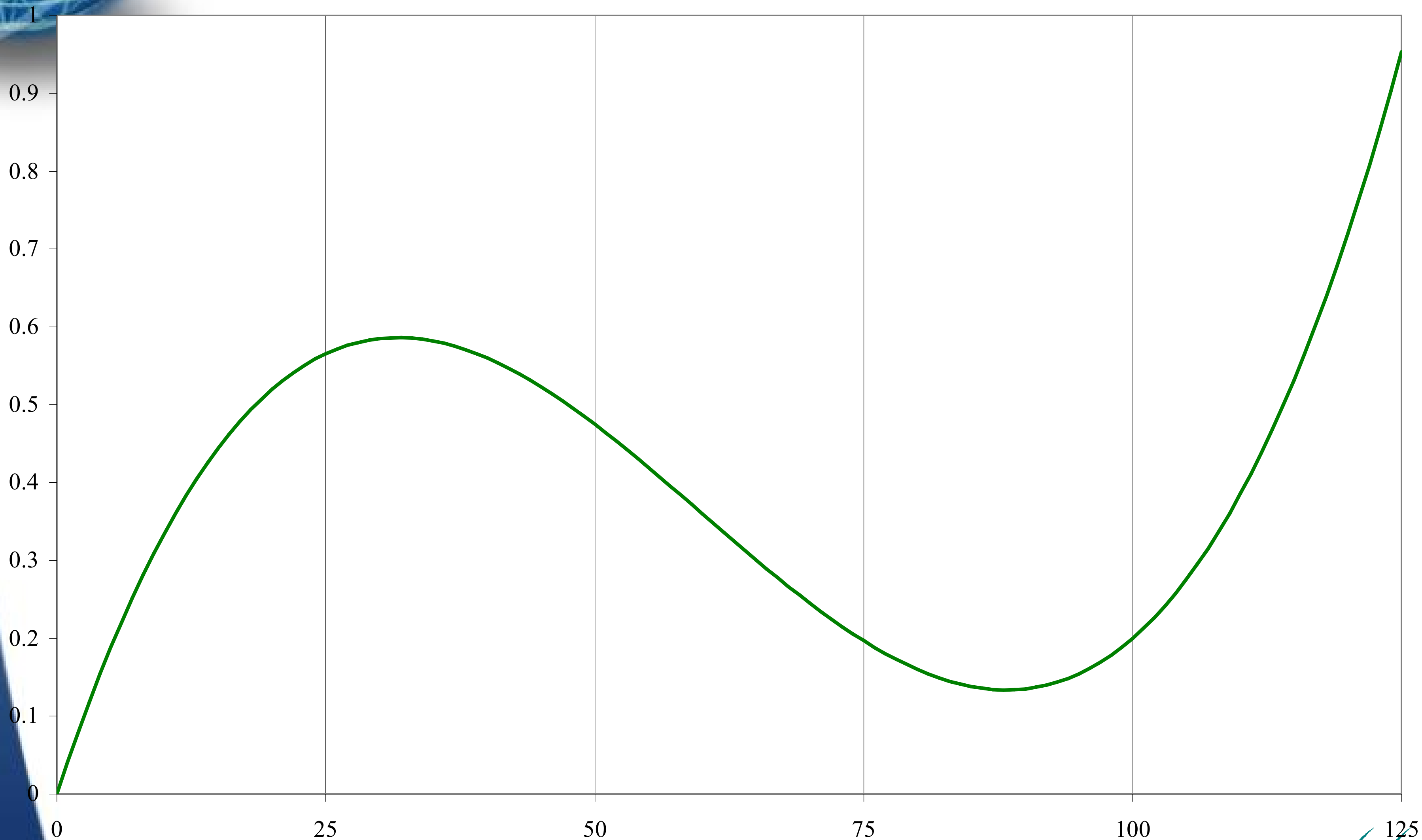
- One knot at each unique data value
- Additional curvature penalty prevents over fitting
- Curvature penalty selected by repeatedly sampling subsets and optimising generalised goodness of fit measure such as AIC

- Pros
  - allows data to guide final result
- Cons
  - 100s of knots required
  - optimisation process is time-consuming
  - difficult to produce new fitted values

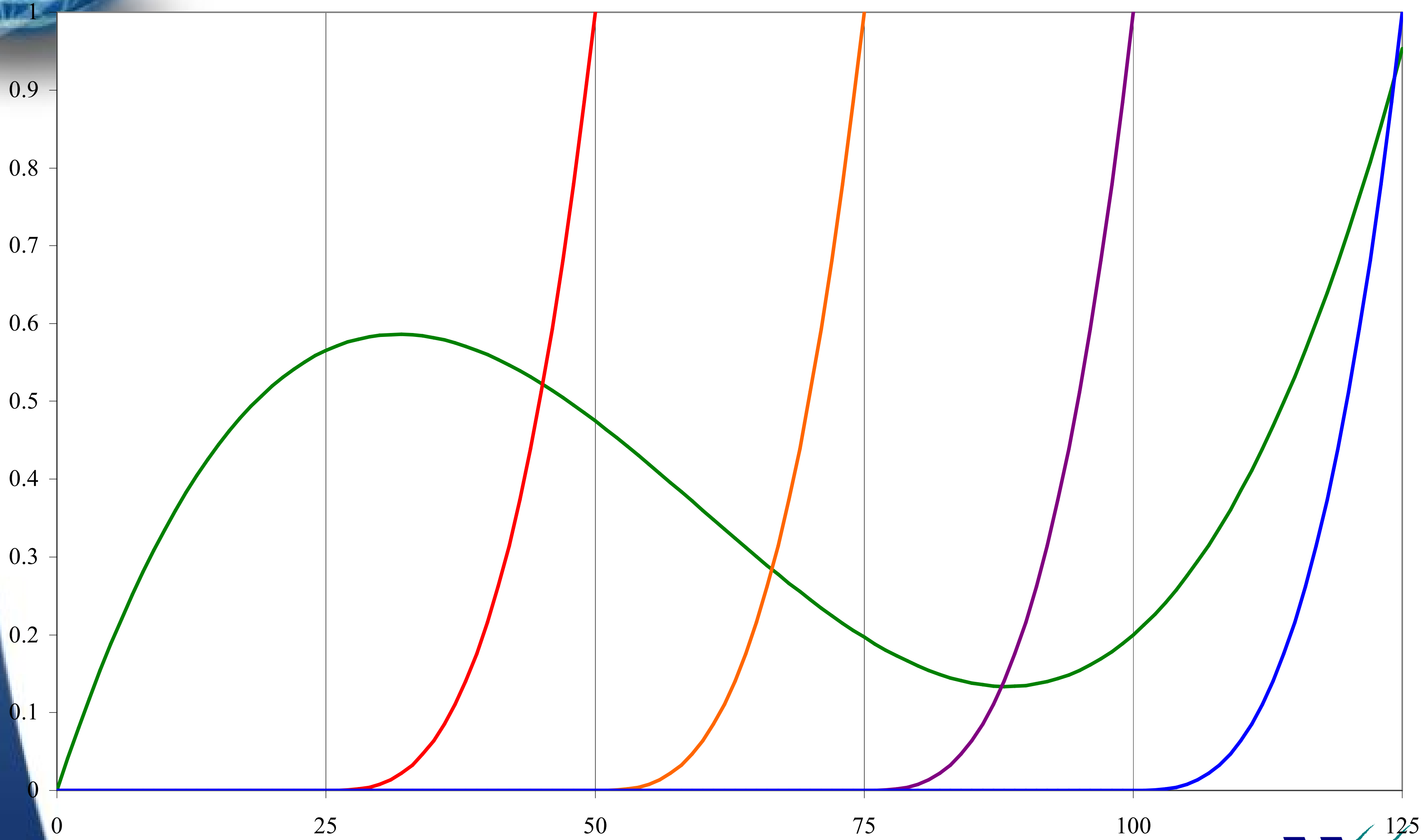
"Easy" regression splines

- Fit a cubic over the whole range
  - simply define x, x^2 and x^3 as variates and include in the model
- Fit additional cubic "correction" variates for each interval, defined as
  - 0 if $x<k_r$
  - $((x - k_r)/(k_{r+1} - k_r))^3$ otherwise

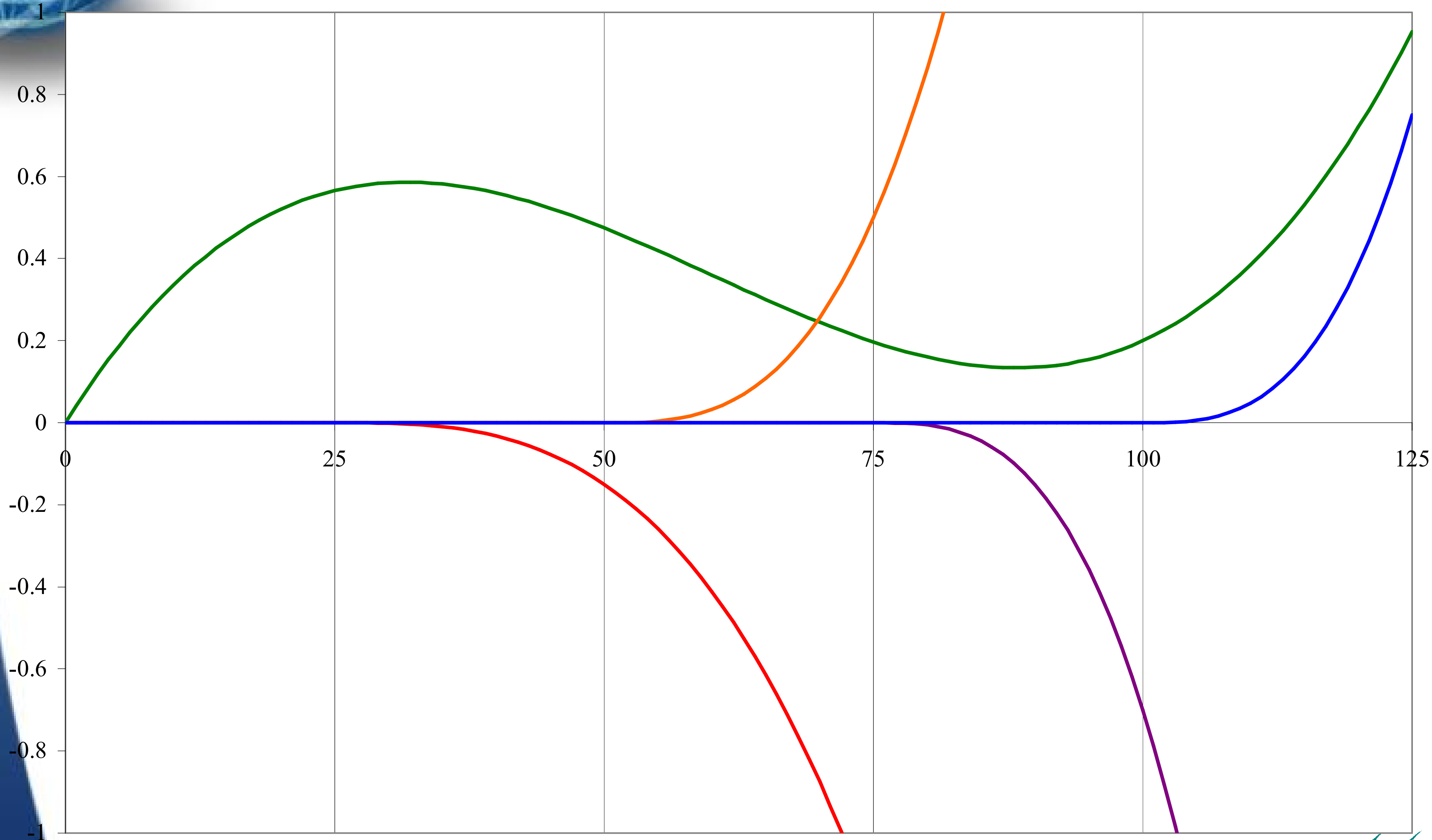
"Easy" regression splines



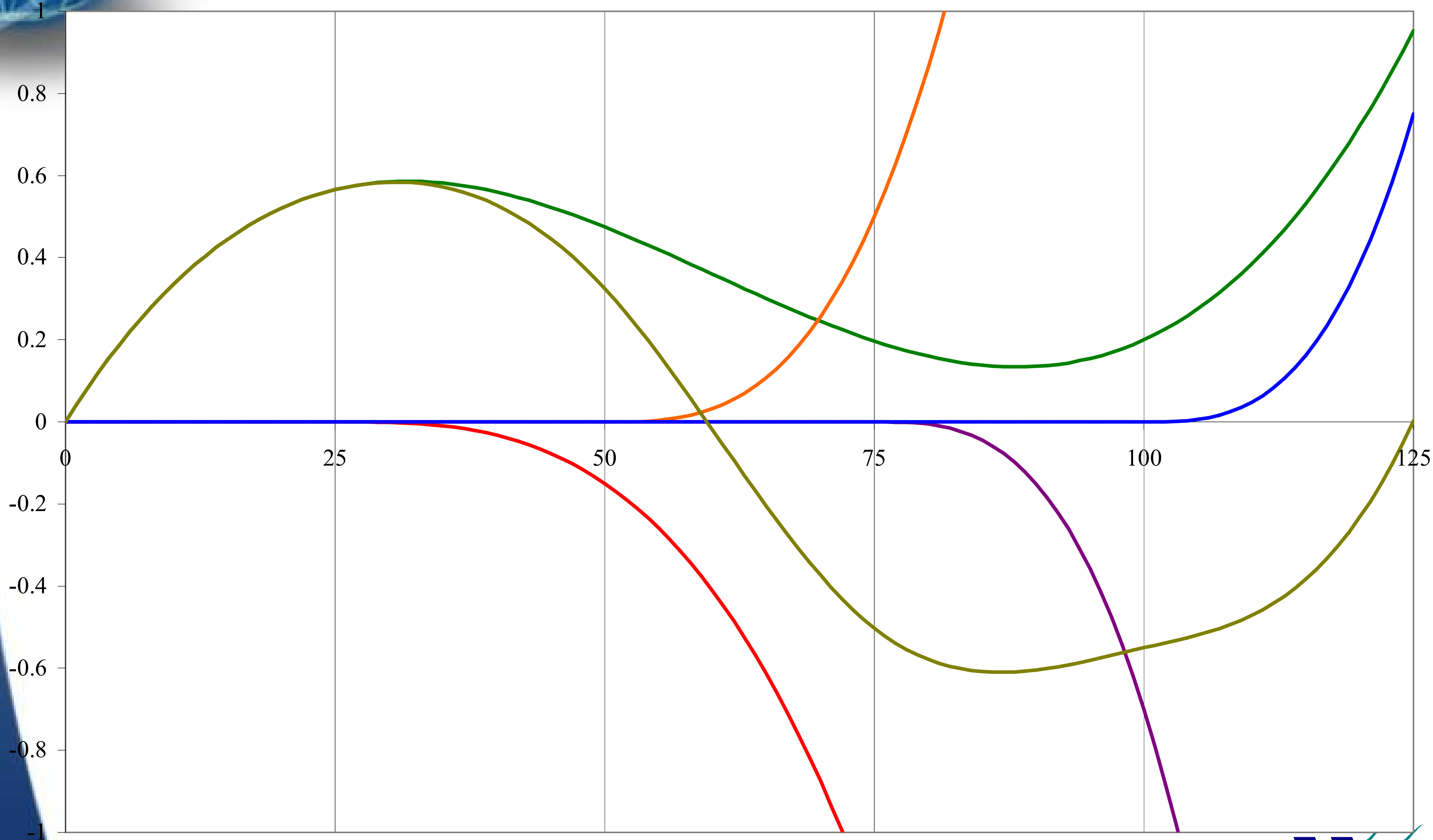
"Easy" regression splines



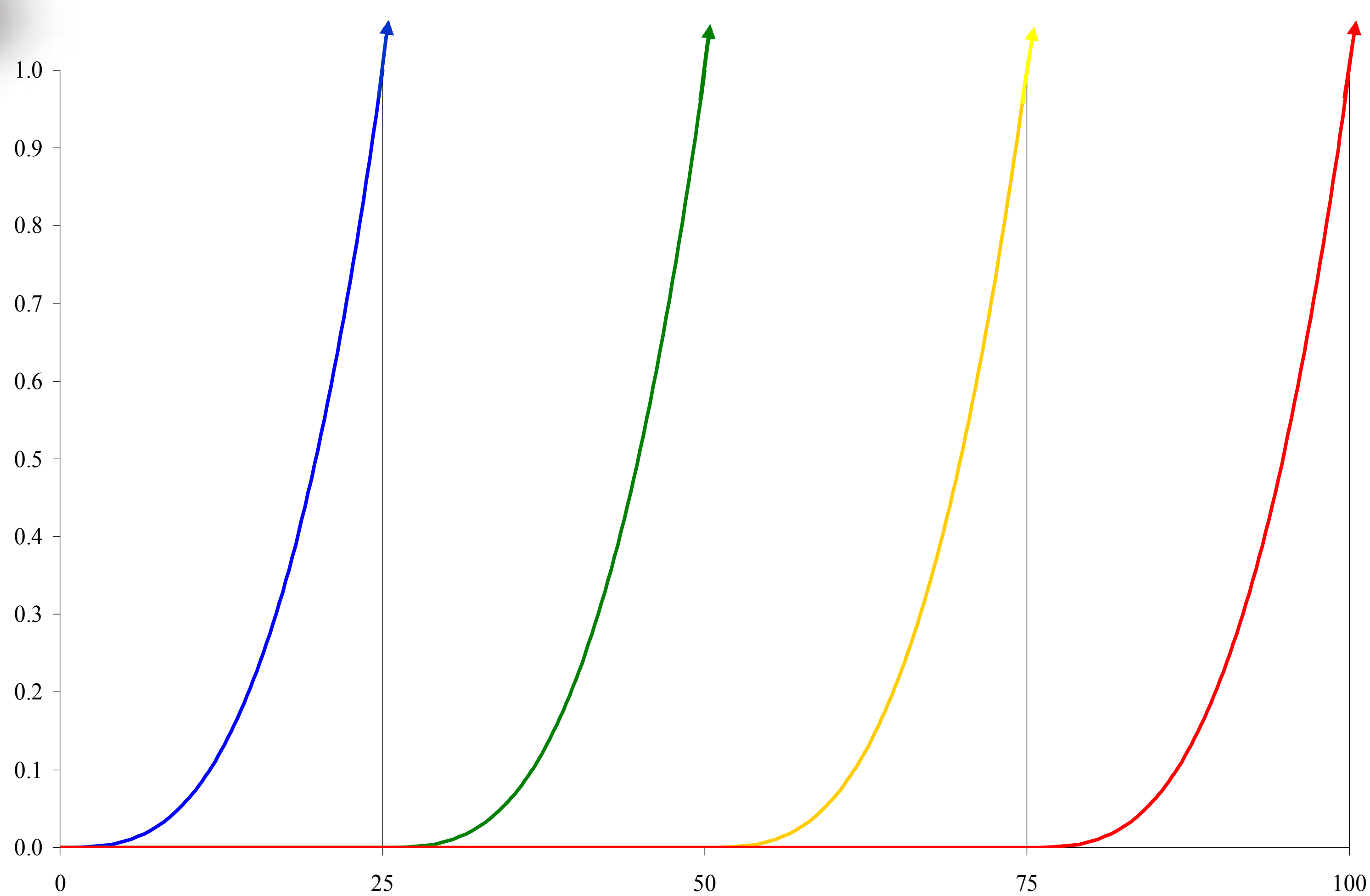
"Easy" regression splines



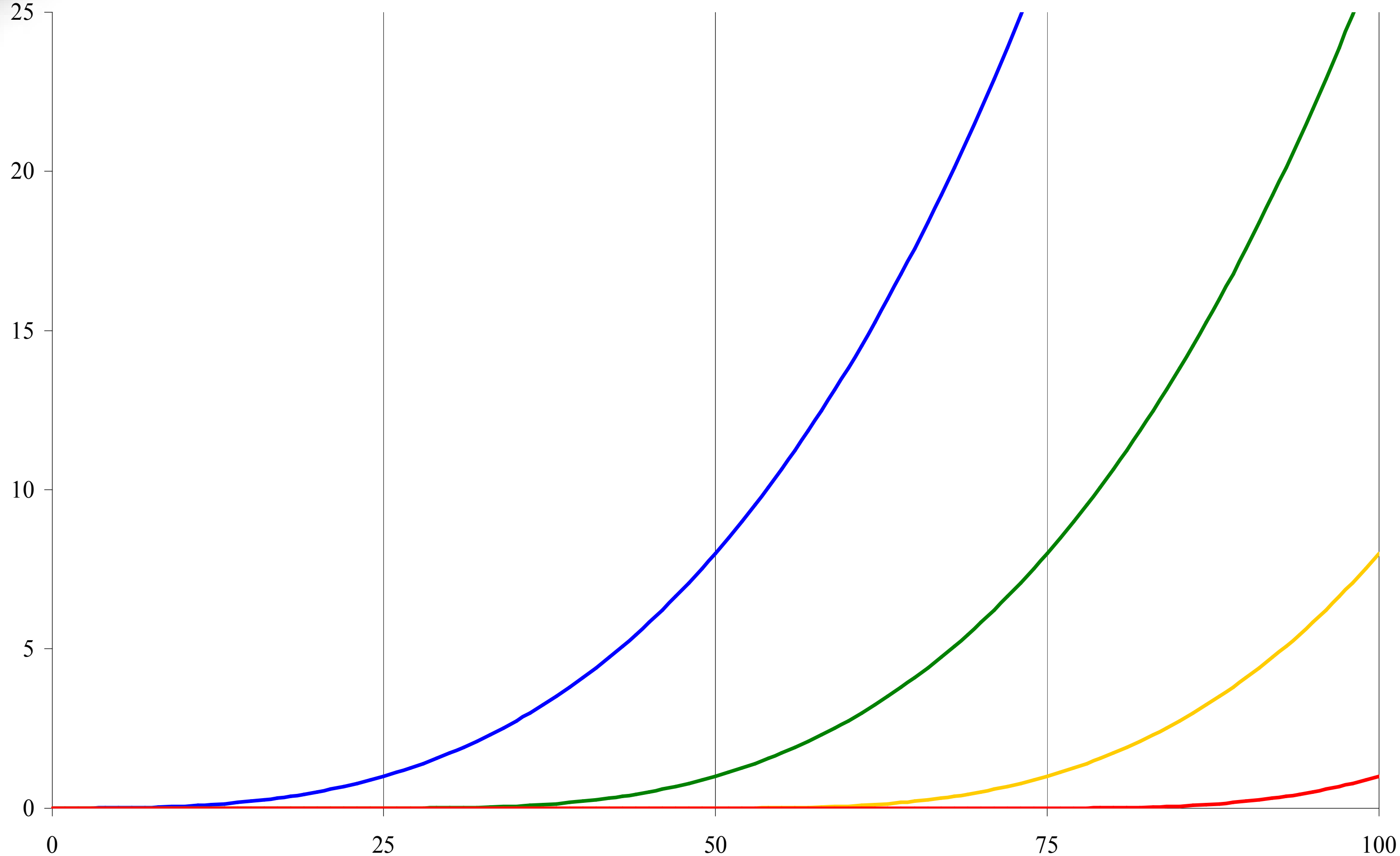
"Easy" regression splines



"Easy" regression splines



"Easy" regression splines

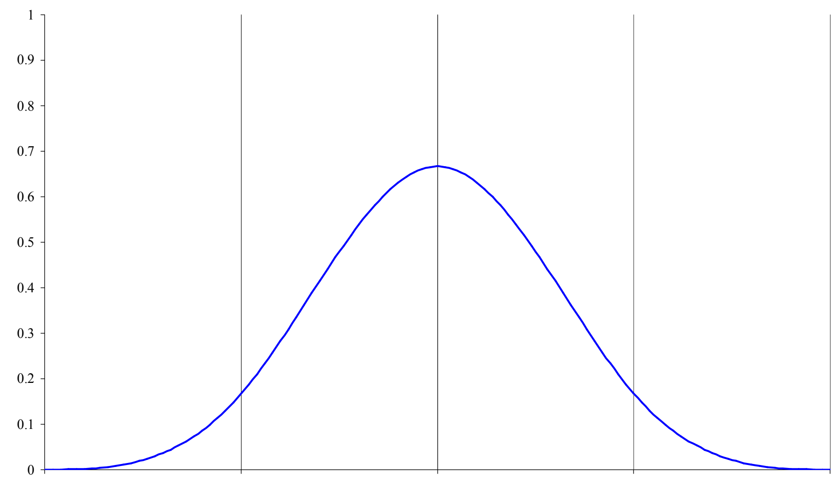


"Easy" regression splines

- "Correction" variates get large quickly
- In practice GLM process can struggle with these large numbers
- Alternate basis is clearly desirable so that:
  - underlying variate remains small for all possible values of x
  - easy to impose additional edge constraints (linear or constant extrapolation is desirable)

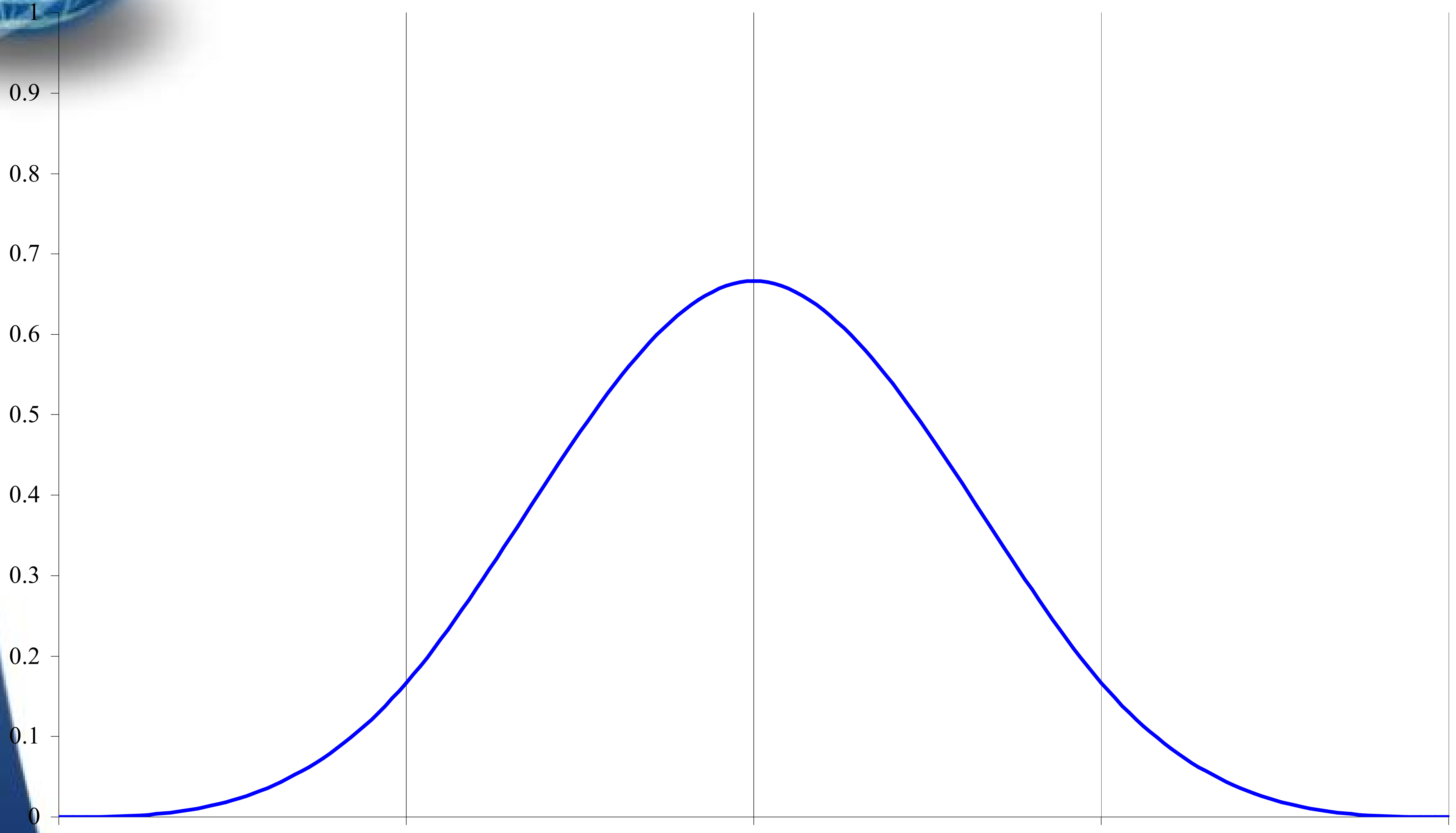
B-Splines

- Set of basis functions usually covering four segments (defined by five knots)
- Each function is itself a cubic spline

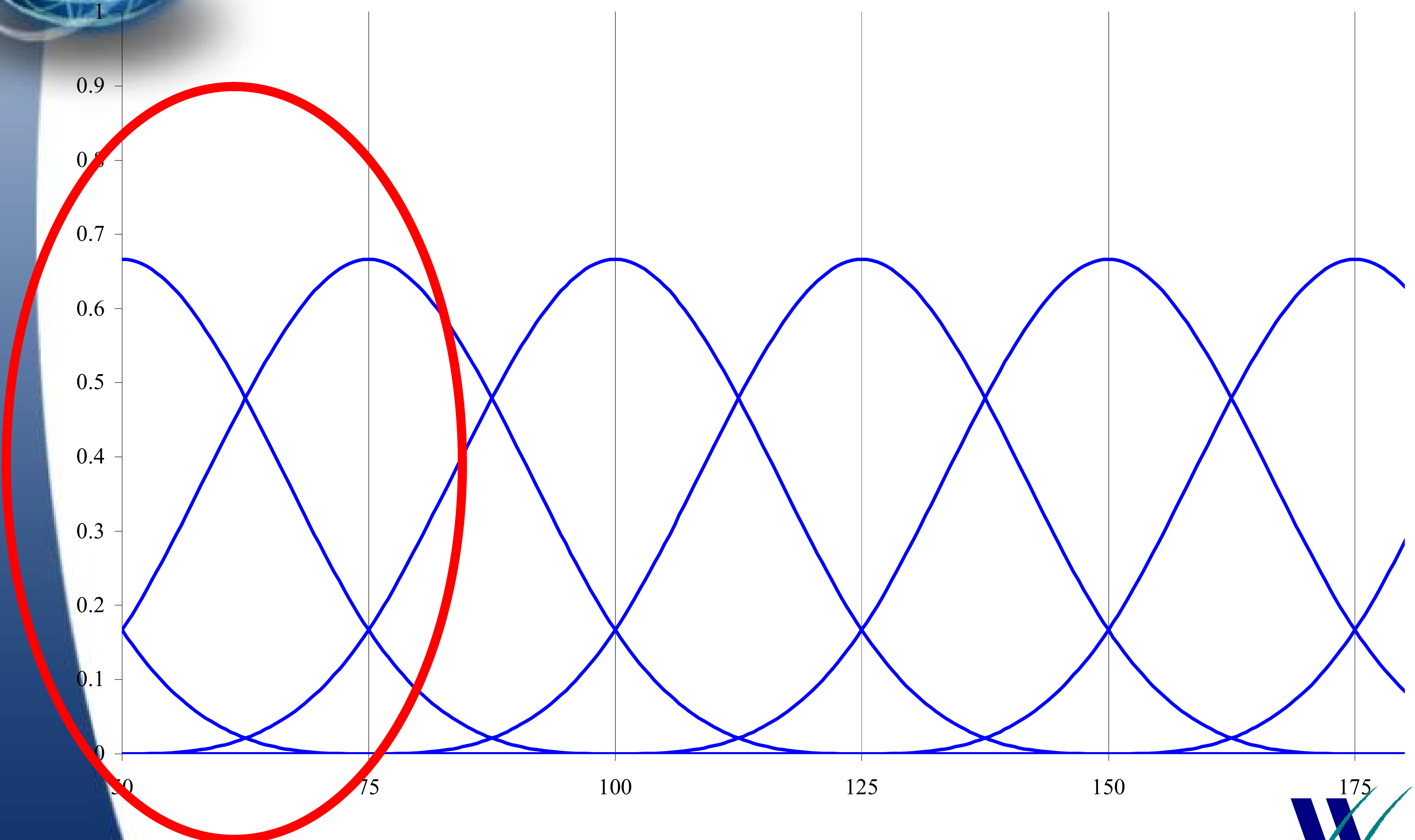


- Each basis function has the same shape, except for the three basis functions at each extreme which occupy fewer than four segments

B-Splines



B-Splines



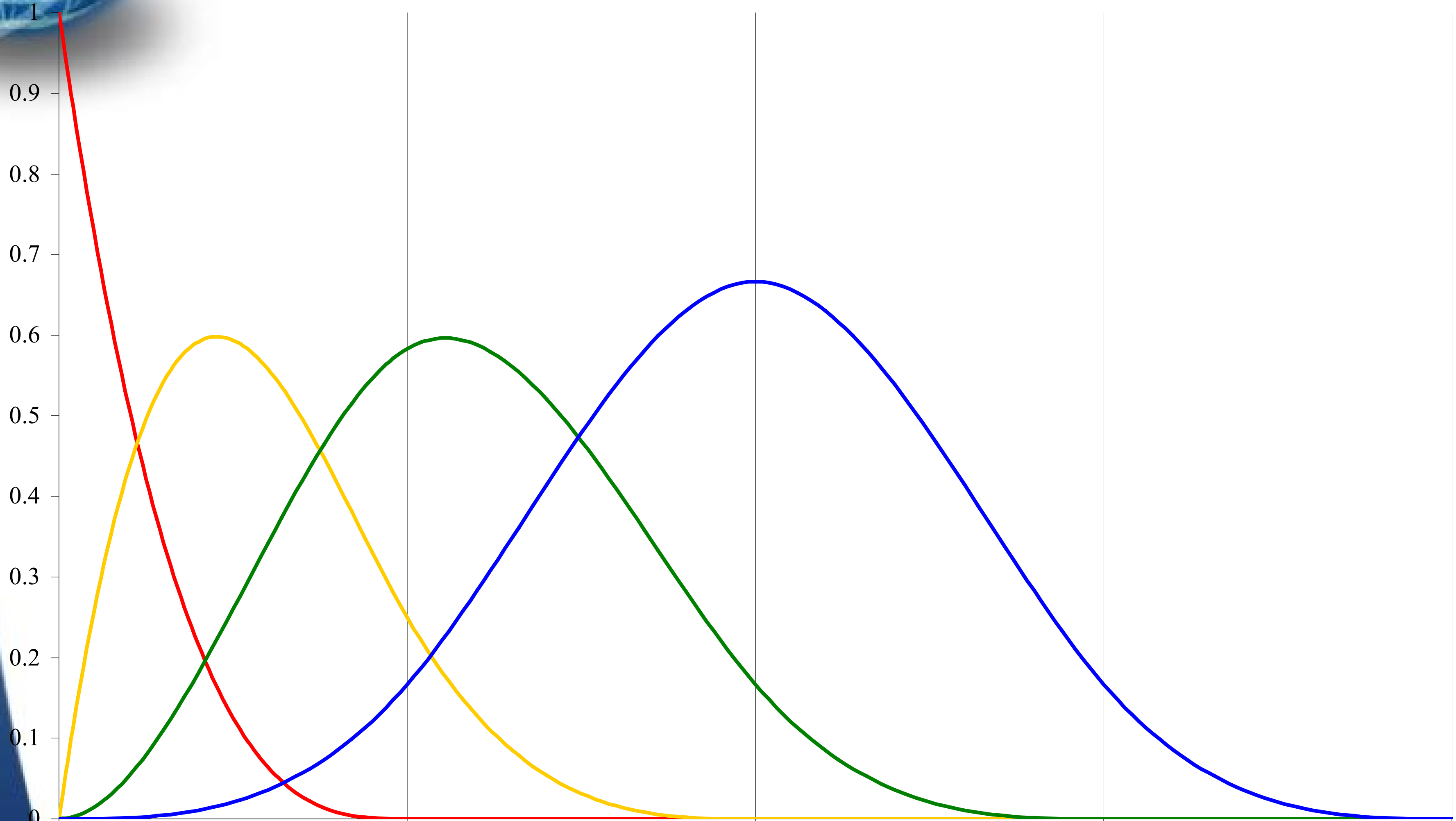
B-Splines

B-Splines

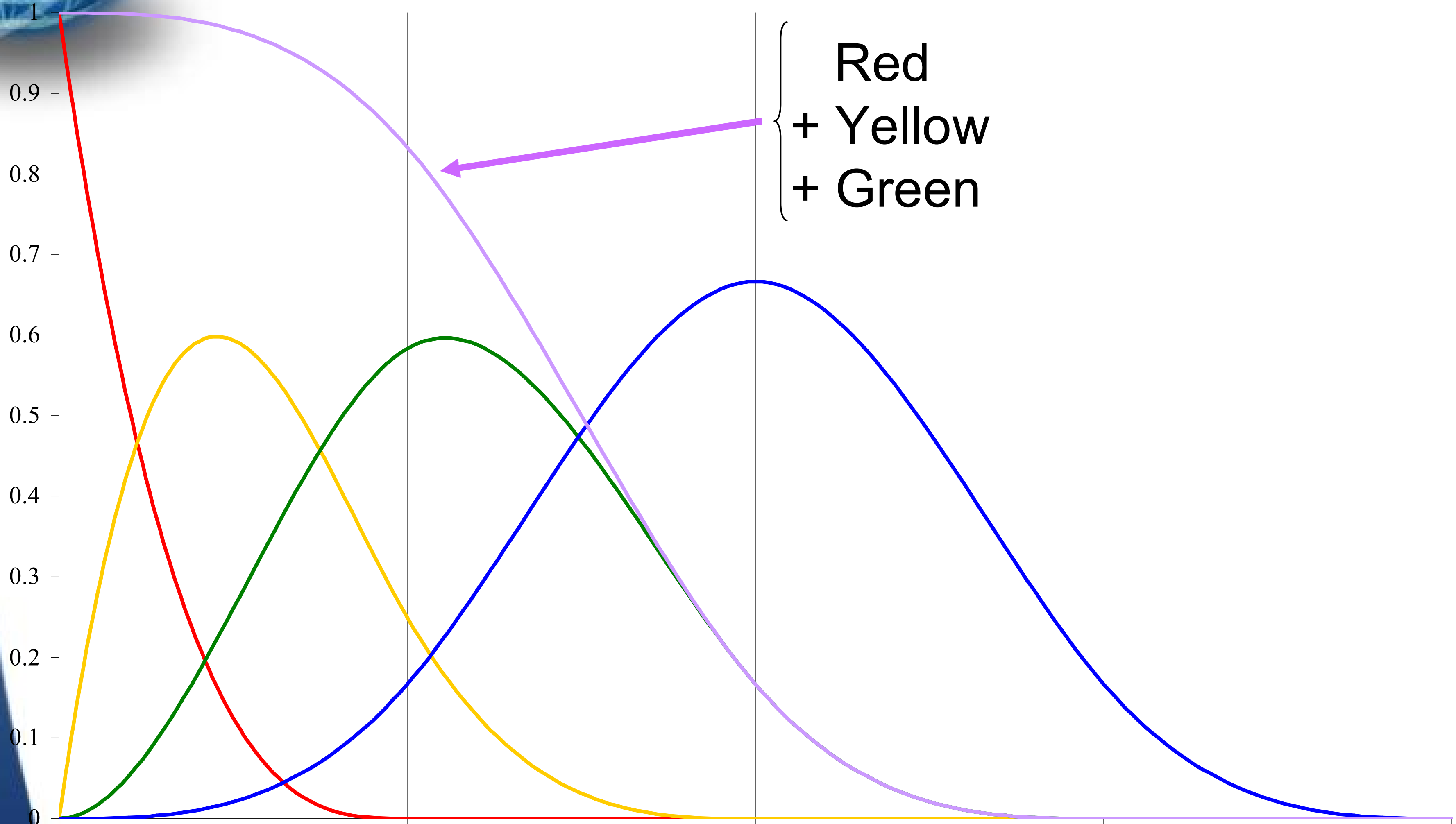
B-Splines

B-Splines

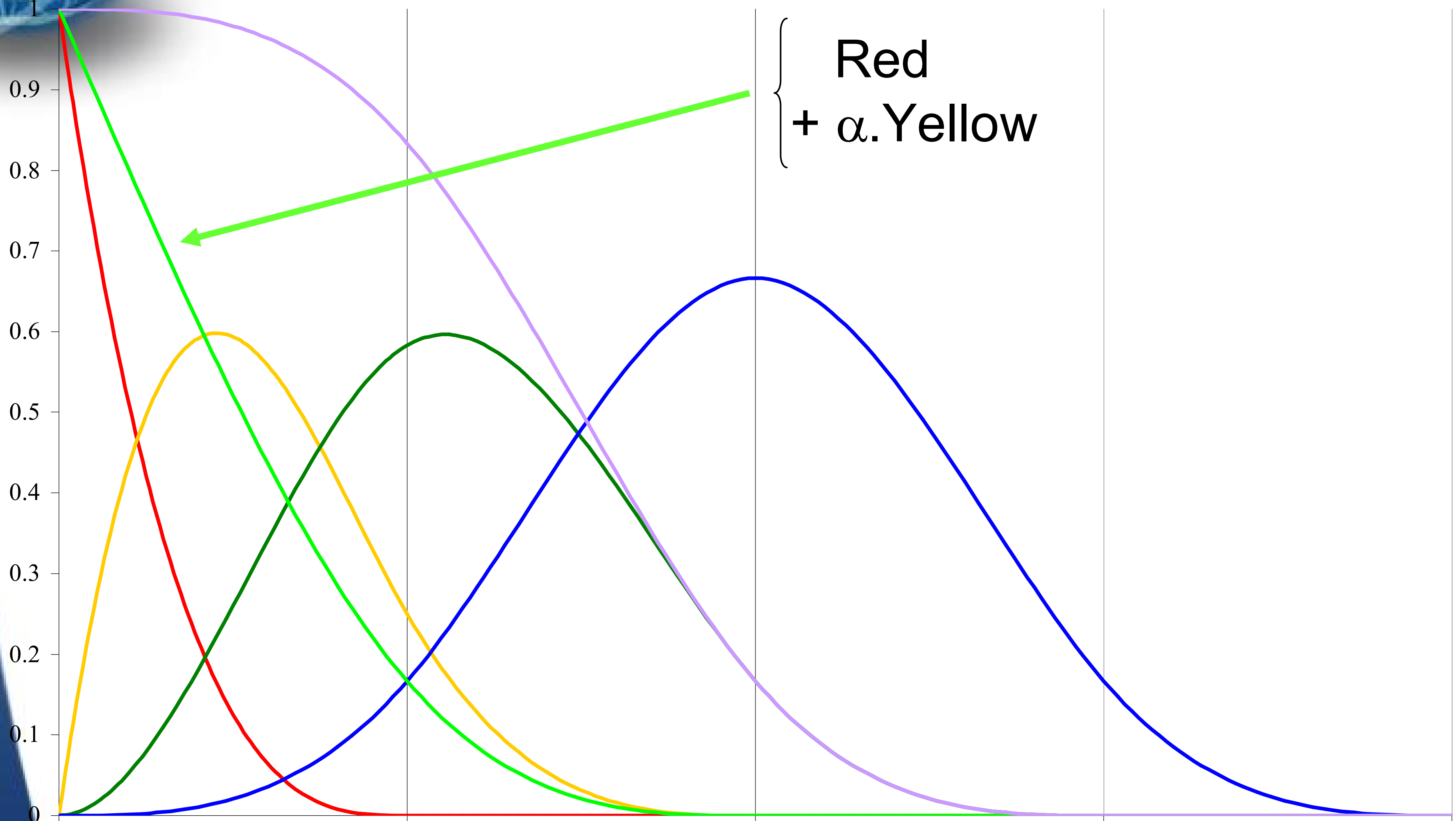
B-Splines



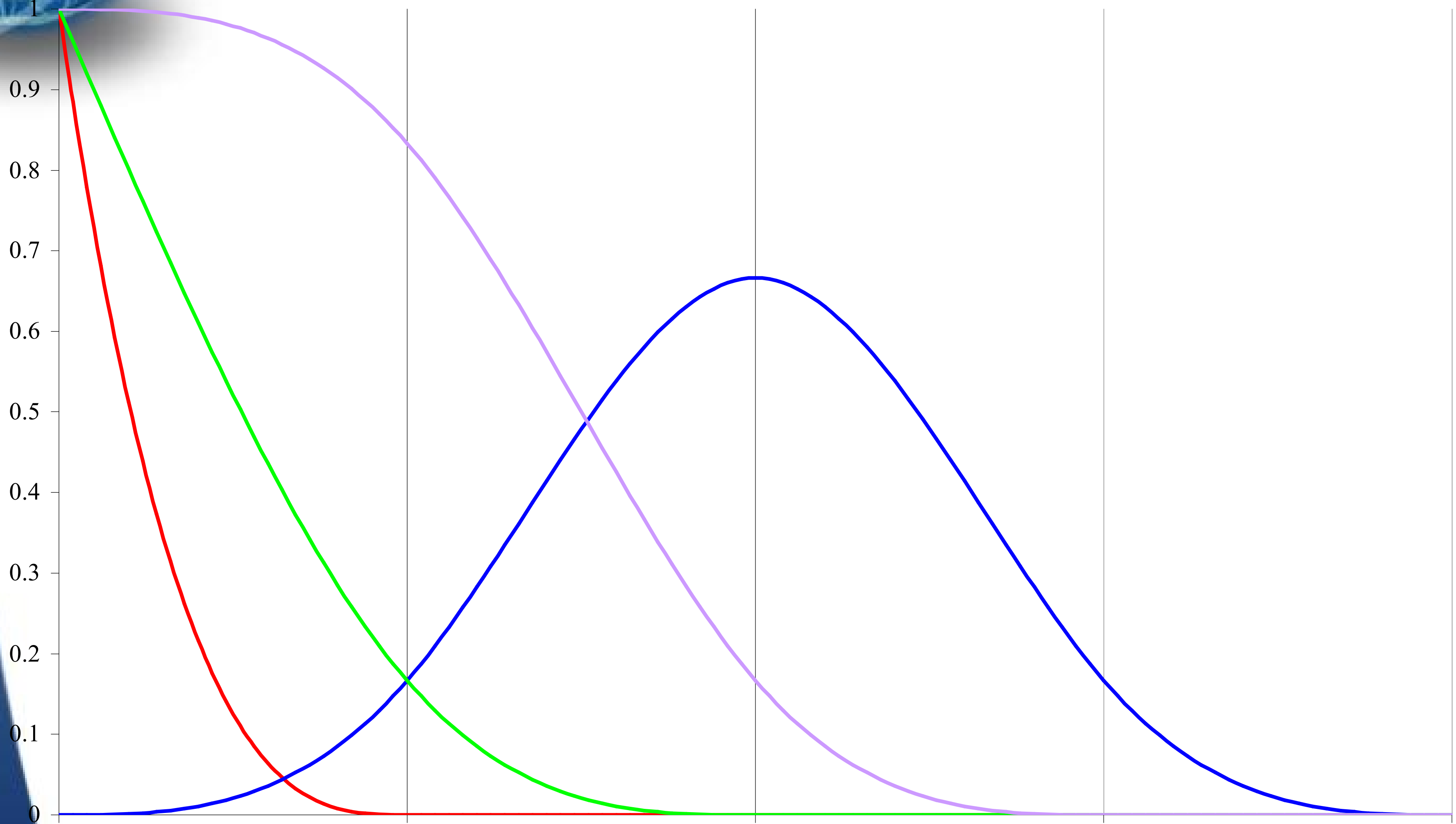
B-Splines



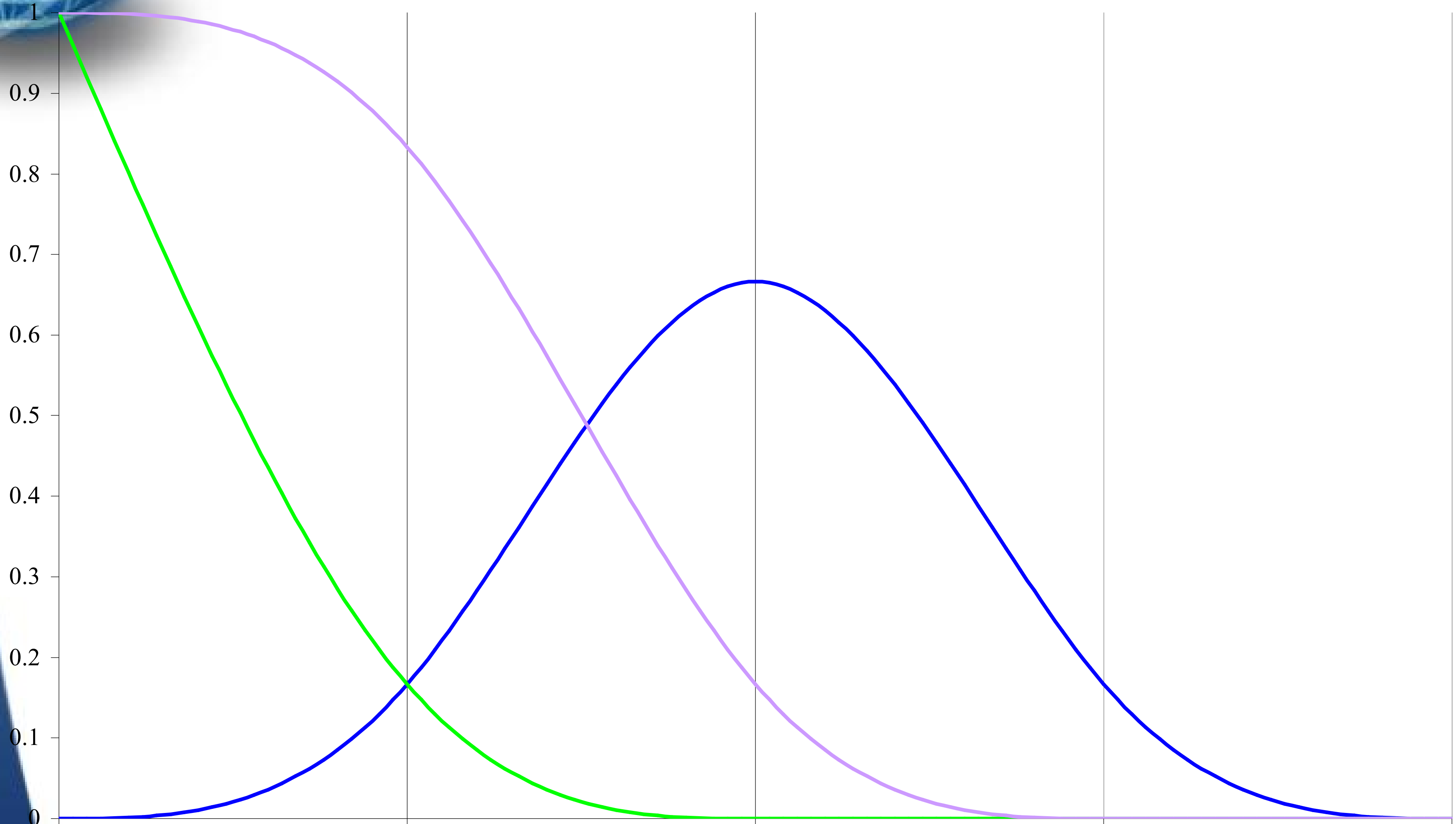
B-Splines



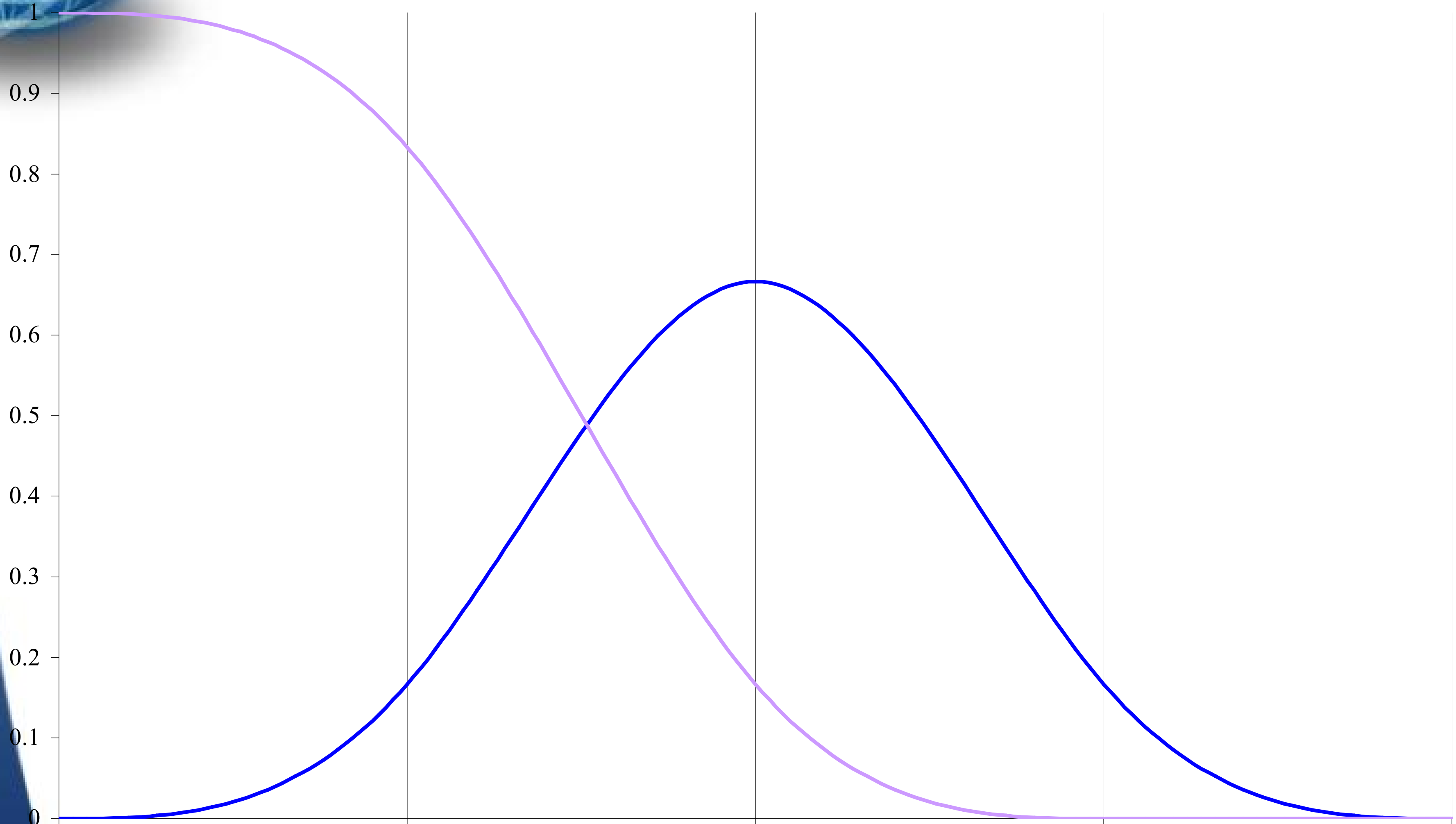
B-Splines – quadratic extrapolation



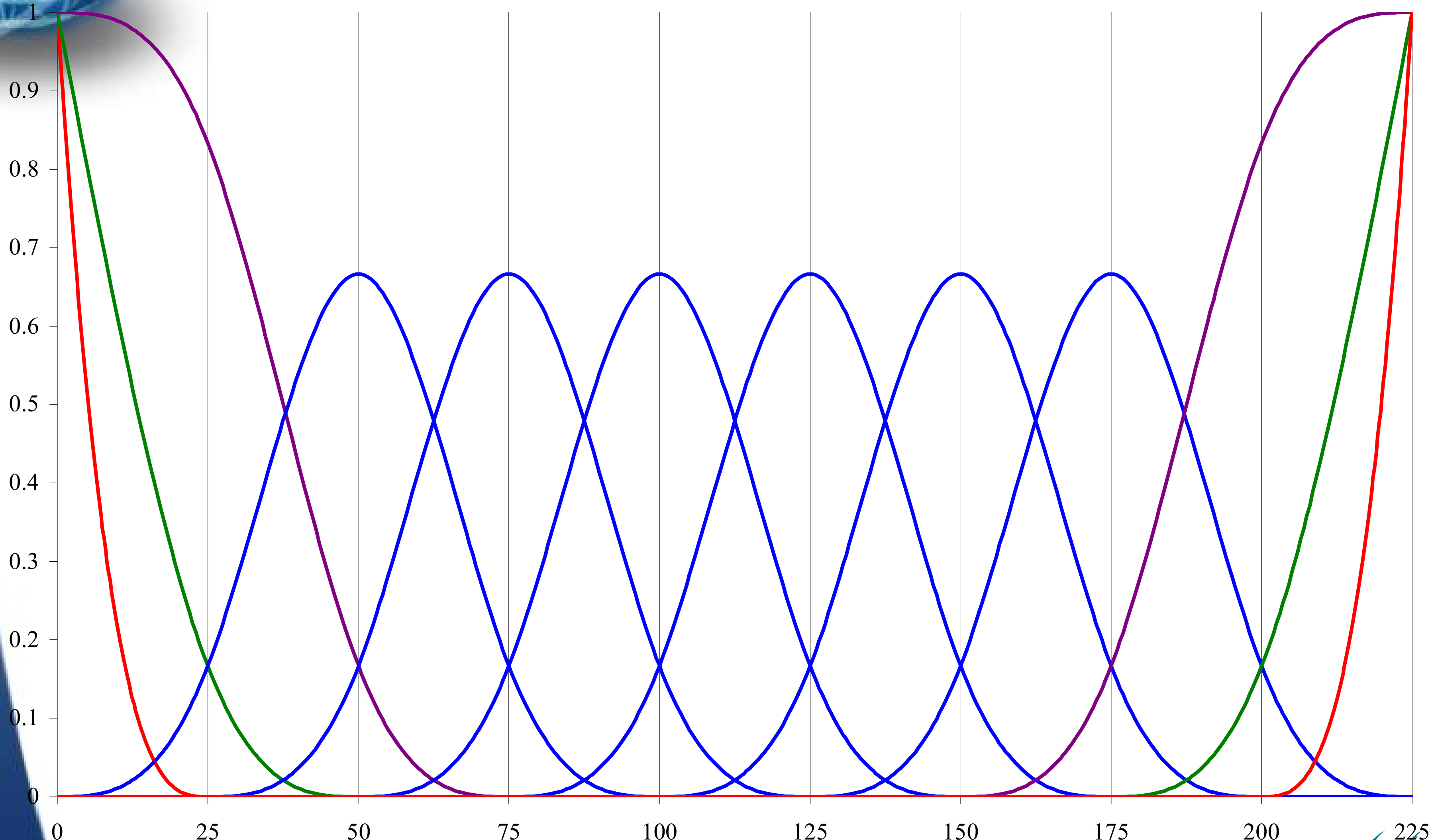
B-Splines – linear extrapolation



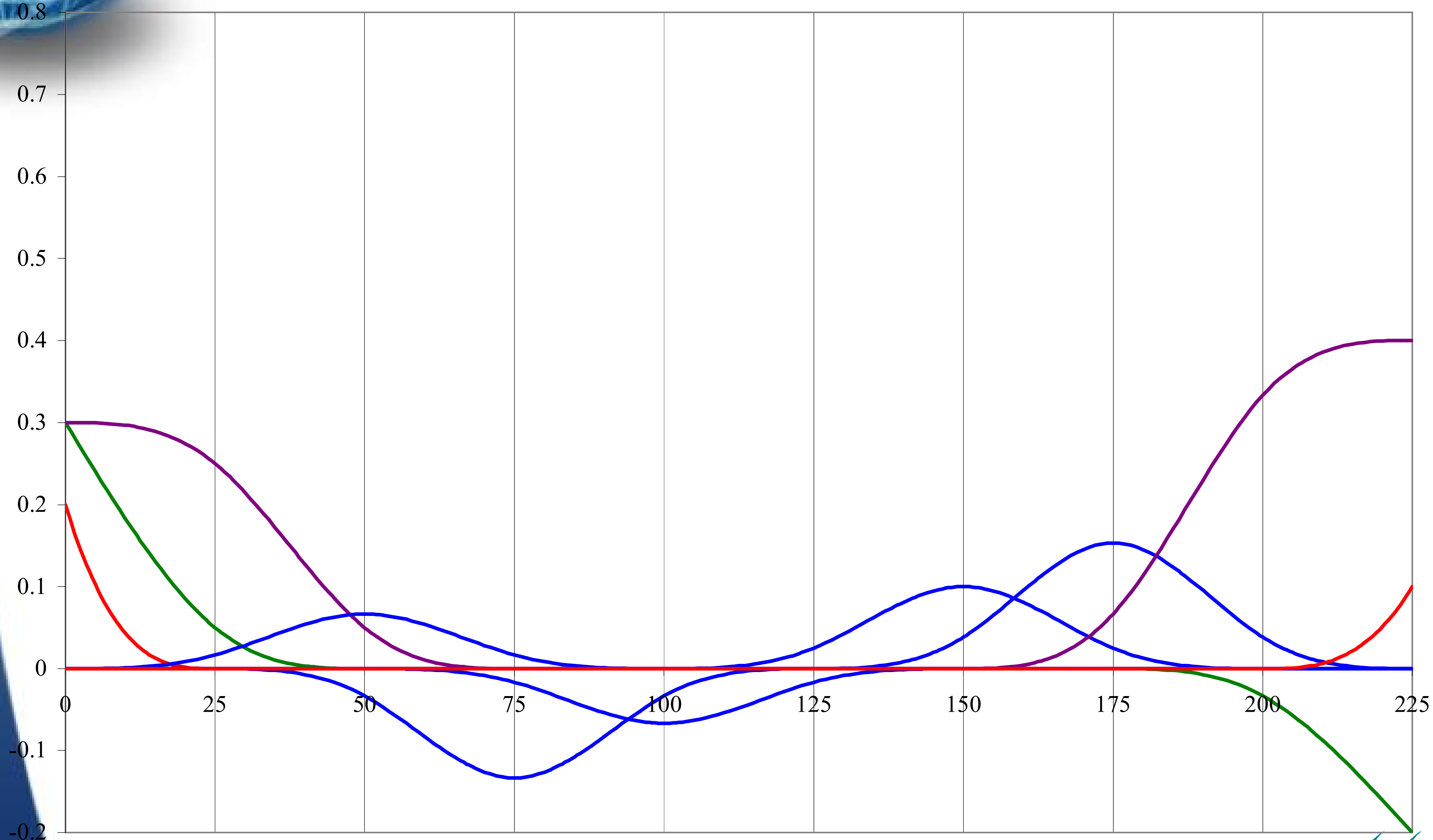
B-Splines – constant extrapolation



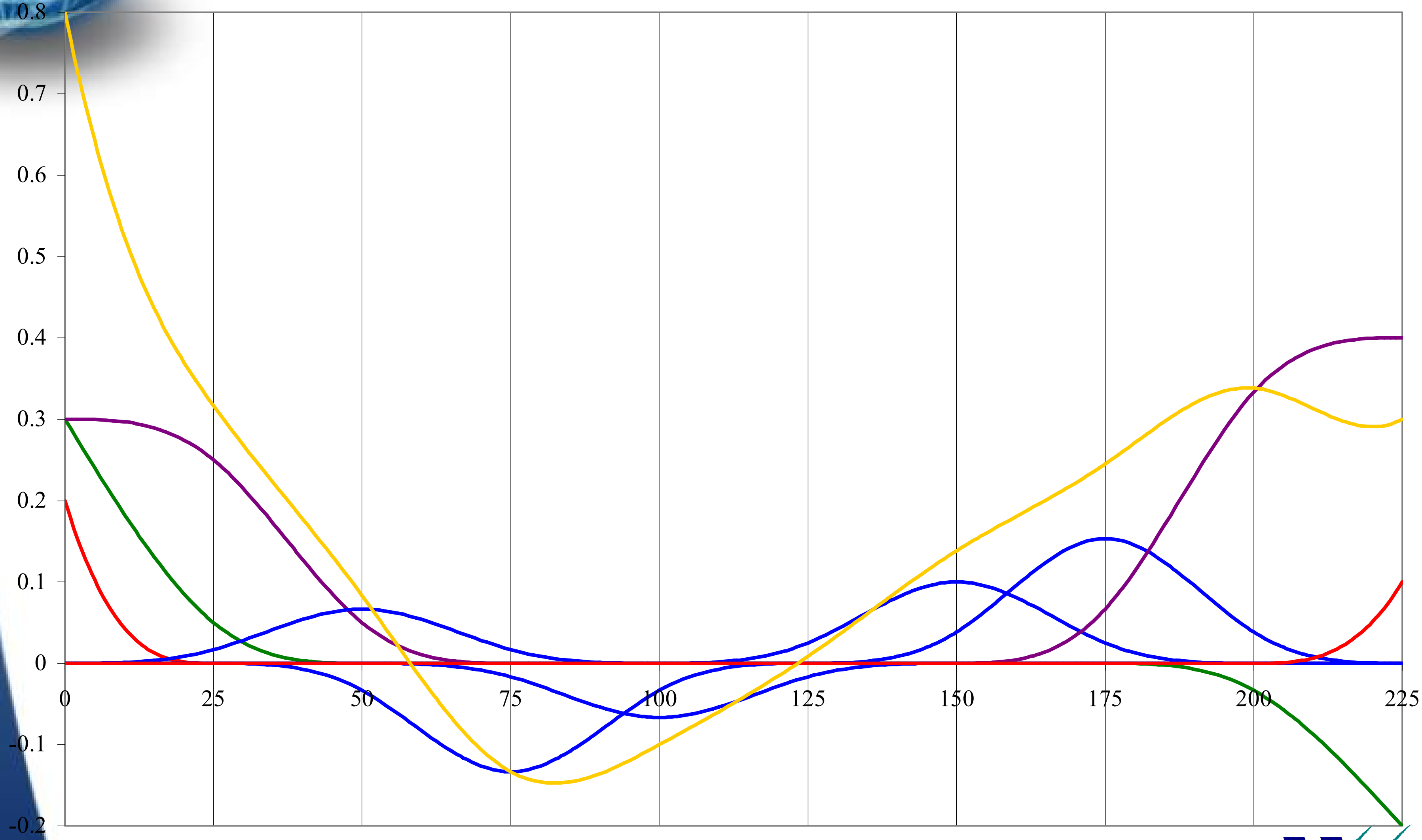
B-Splines - example



B-Splines – example



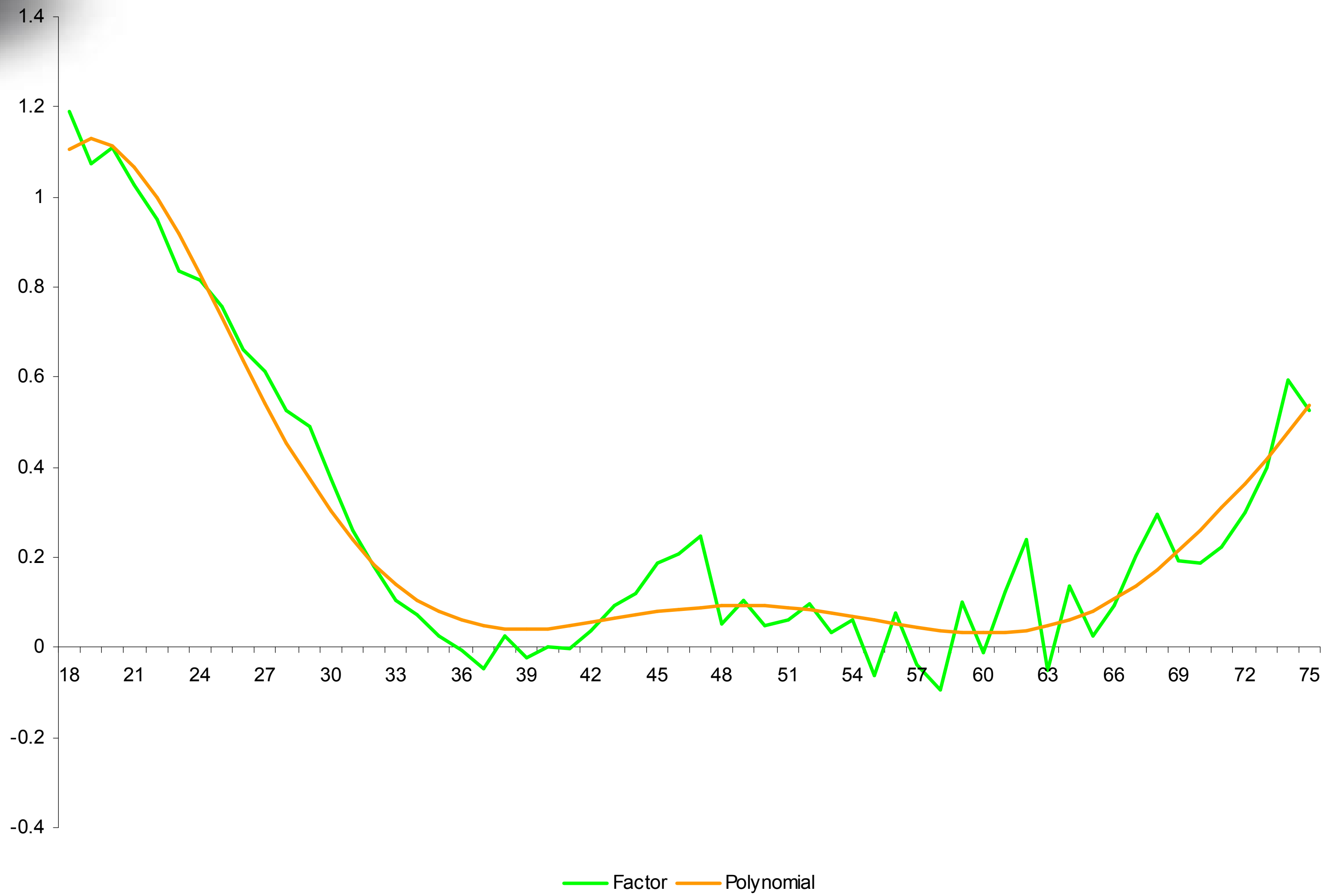
B-Splines - example



Example

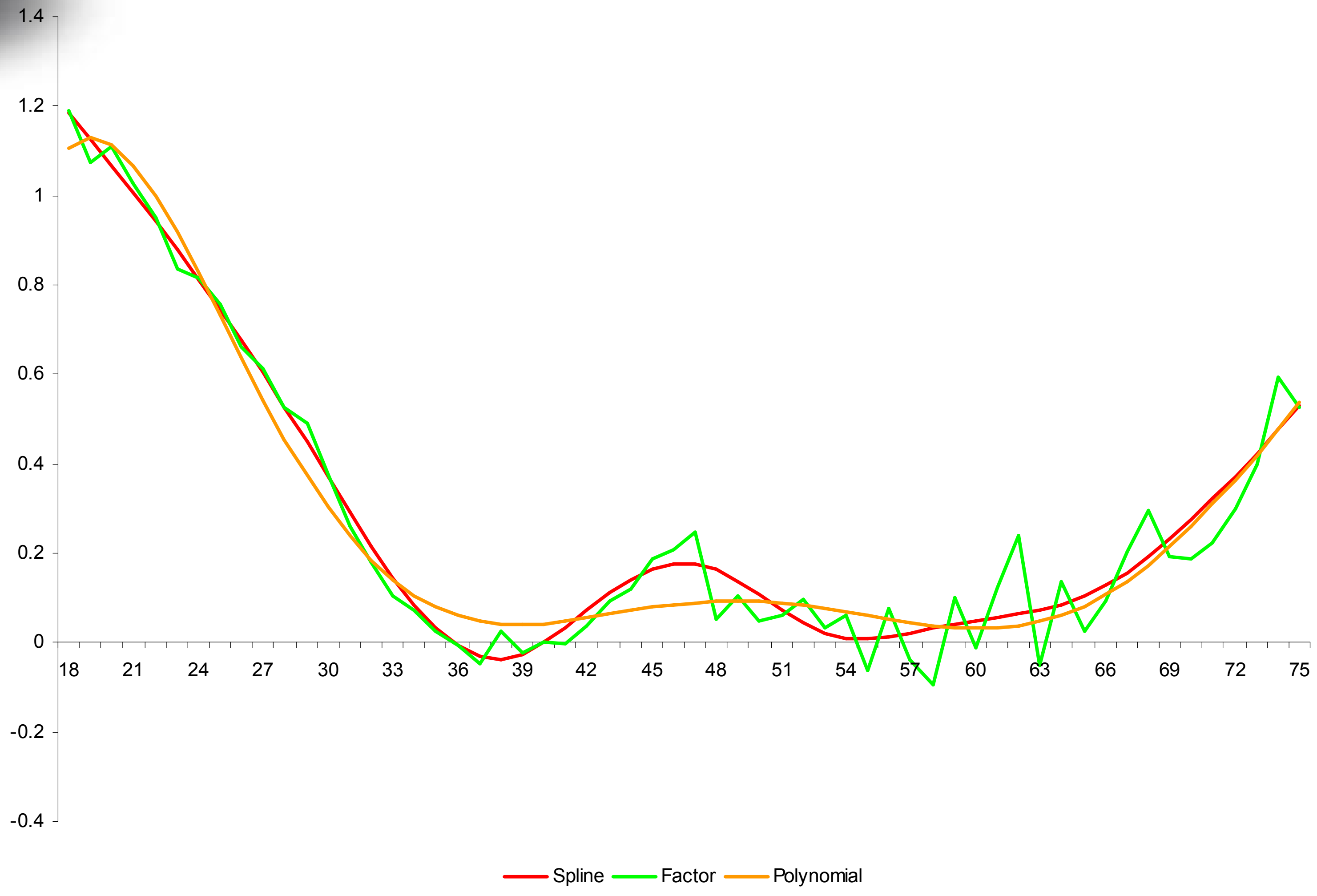
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
18
21
24
27
30
33
36
39
42
45
48
51
54
57
60
63
66
69
72
75
Factor

Example



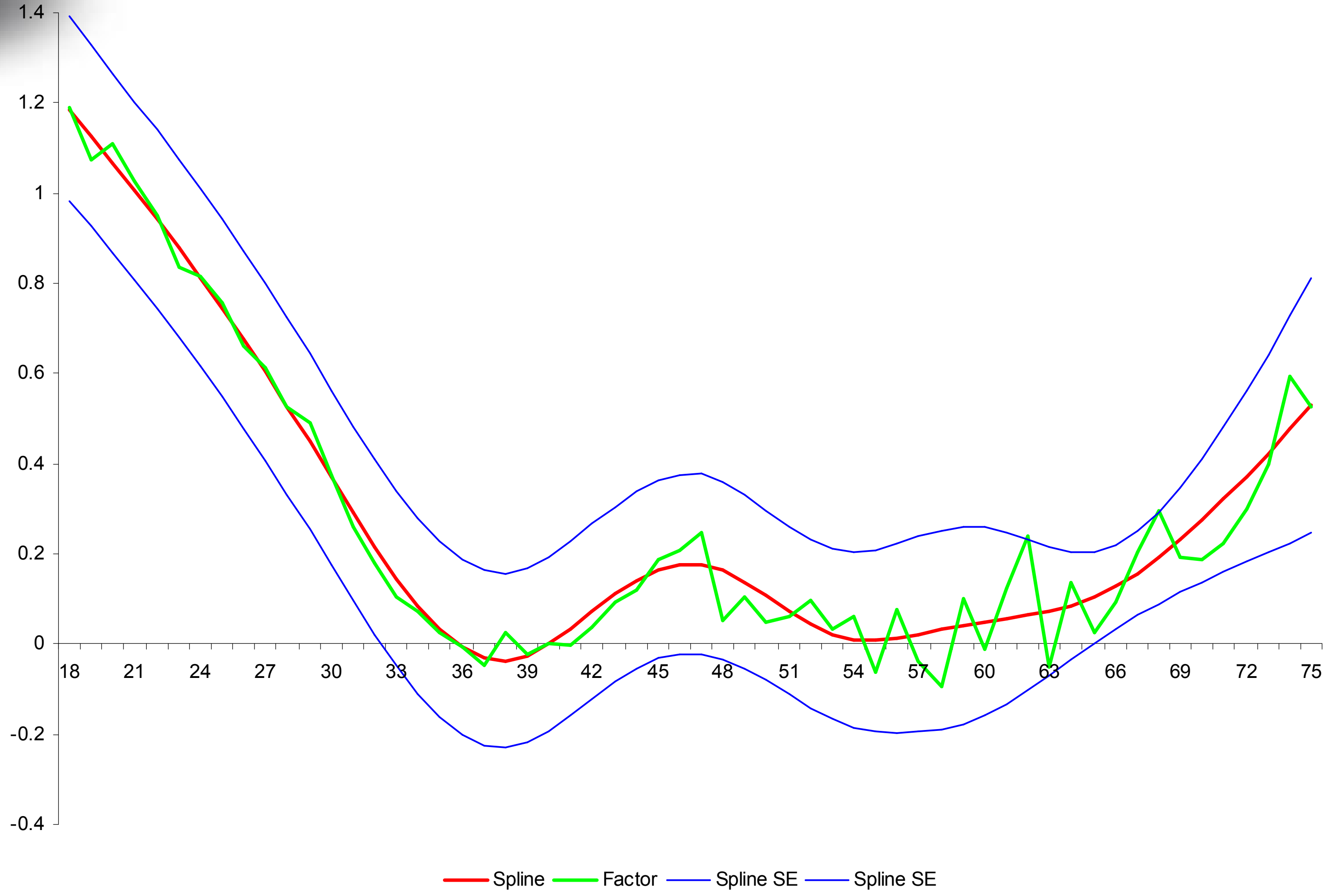
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
18 21 24 27 30 33 36 39 42 45 48 51 54 57 60 63 66 69 72 75
Factor
Polynomial

Example



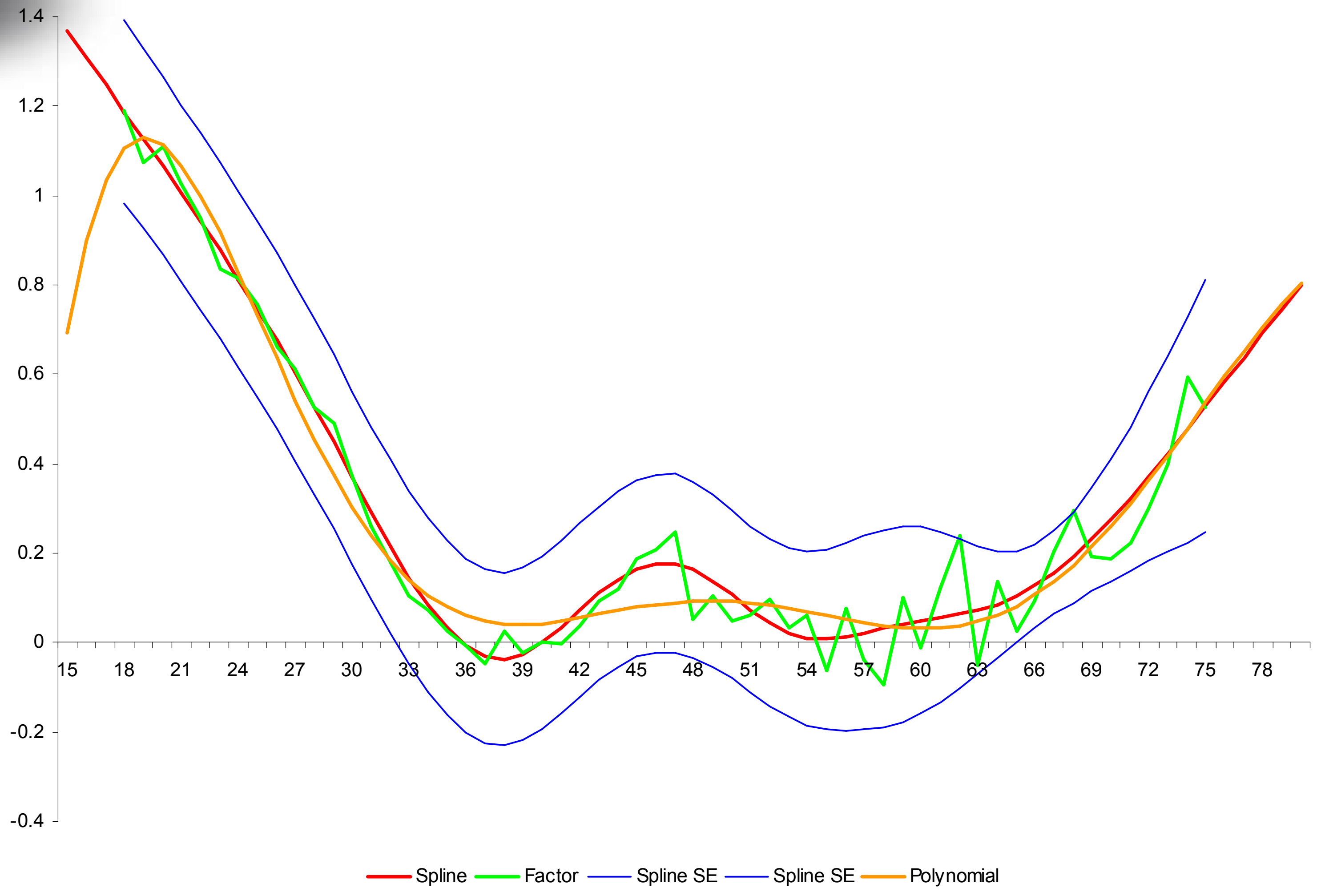
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
18
21
24
27
30
33
36
39
42
45
48
51
54
57
60
63
66
69
72
75
Spline
Factor
Polynomial

Example



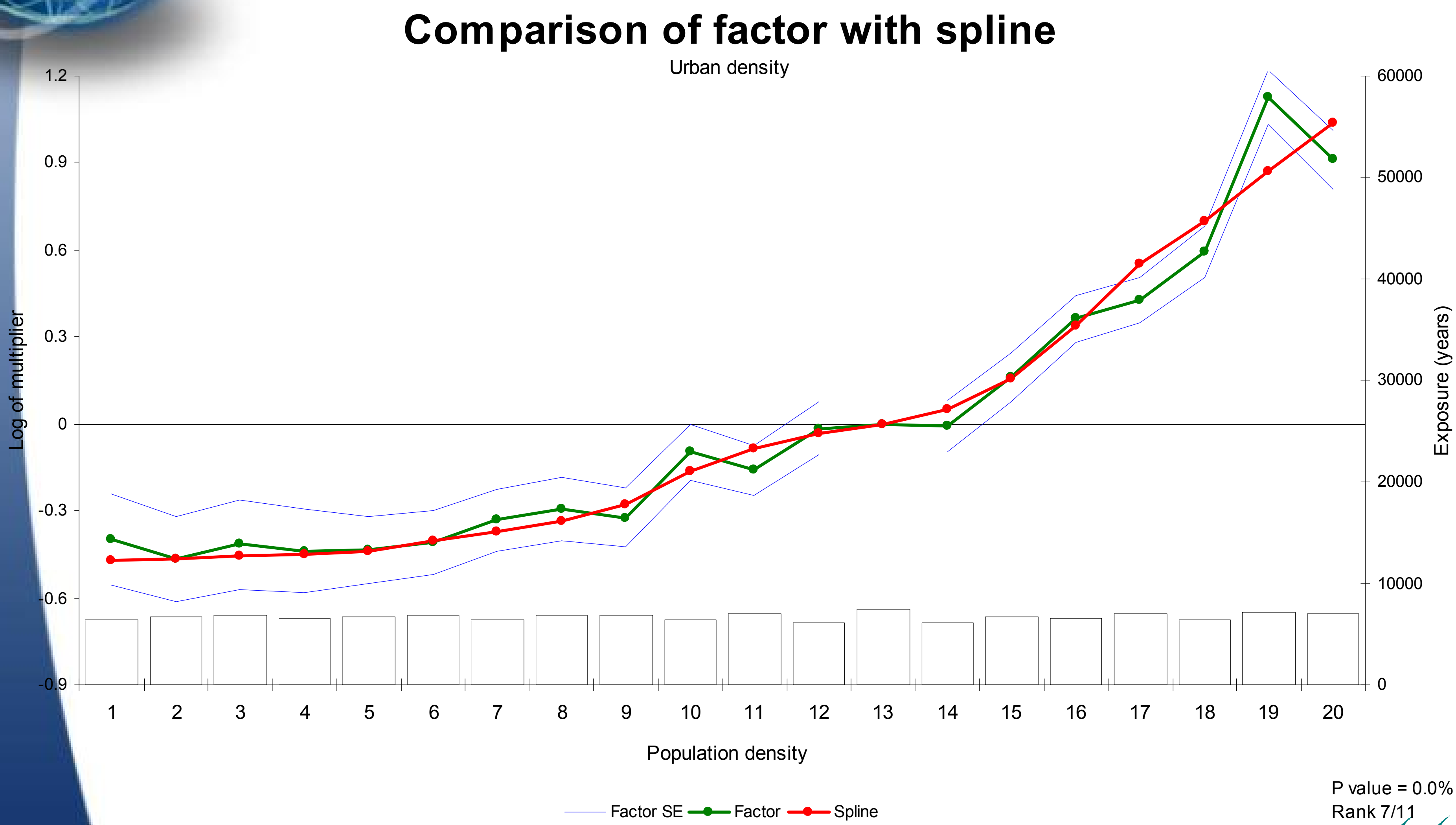
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
18
21
24
27
30
33
36
39
42
45
48
51
54
57
60
63
66
69
72
75
Spline
Factor
Spline SE
Spline SE

Example



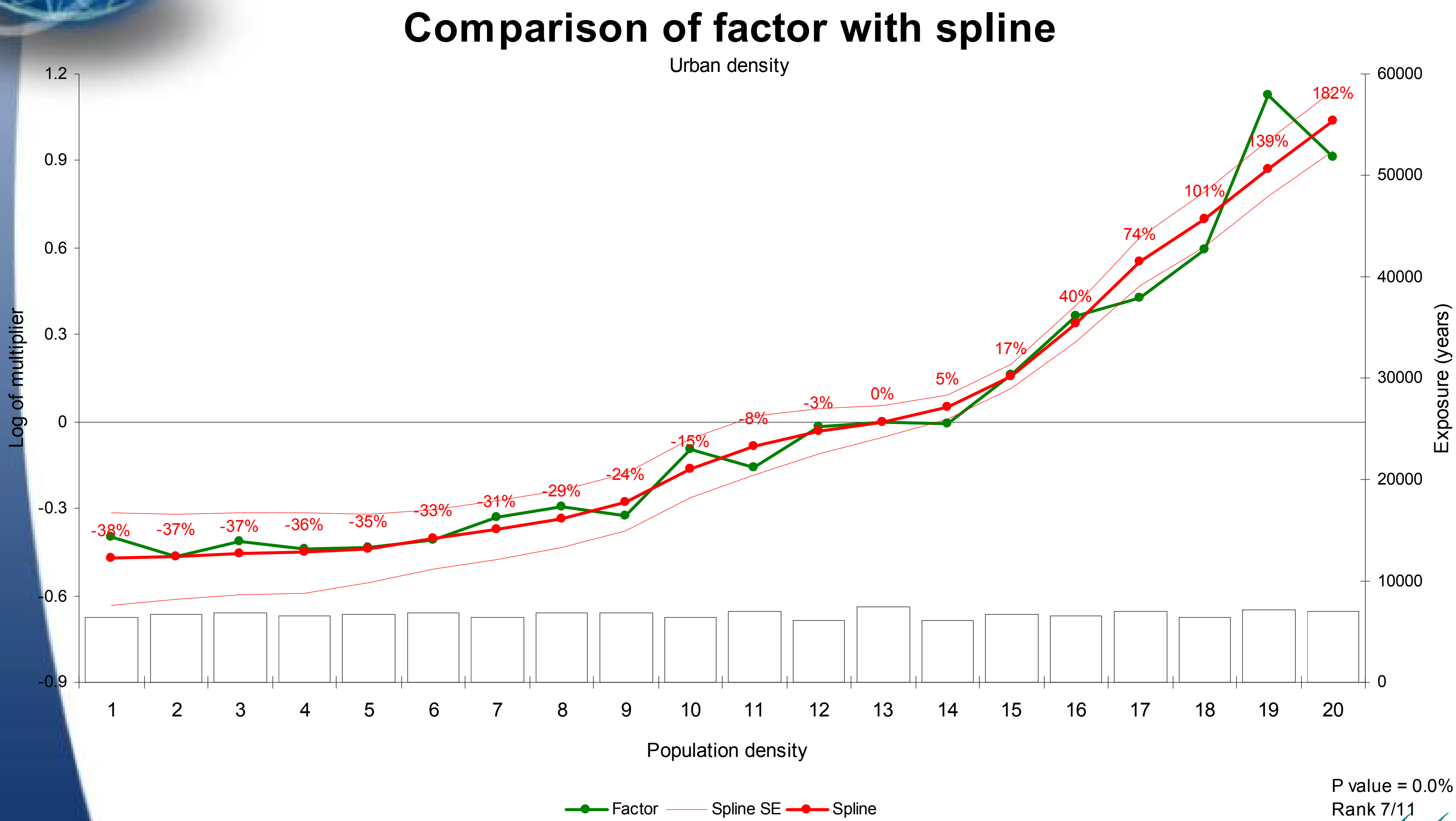
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
15 18 21 24 27 30 33 36 39 42 45 48 51 54 57 60 63 66 69 72 75 78
Spline
Factor
Spline SE
Spline SE
Polynomial

Further example



P value = 0.0%
Rank 7/11

Further example



Knot placement

- Position of knots is important
- Equal width
  - B-splines symmetric
  - Knots may not fall on turning points
- Equal exposure
  - Concentrates knots in high volume segments
  - Can be poor fit at edges
- By eye
  - Can place knots near known turning points
  - Subjective

Splines

- Practical way of modelling continuous variables
- Often better than polynomials
- Increases complexity, therefore best used
  - when it is important that rates vary continuously with a variable
  - when modeling elasticity to be used in price optimization analyses

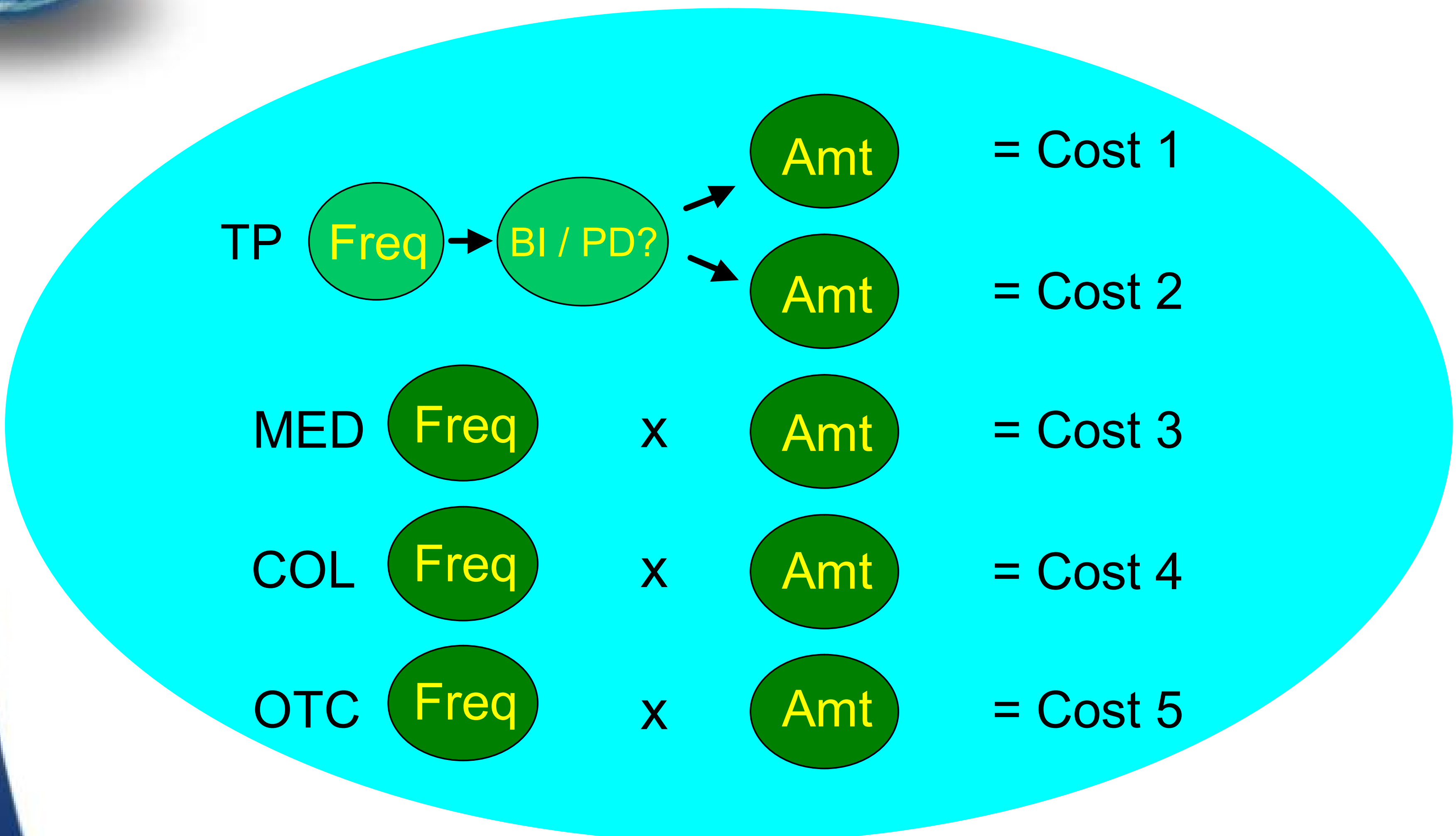
Agenda

- Introduction
- Testing the link function
- The Tweedie distribution
- Splines
- Reference models
- Aliasing / near aliasing
- Combining models across claim types
- Restricted models
- Model validation

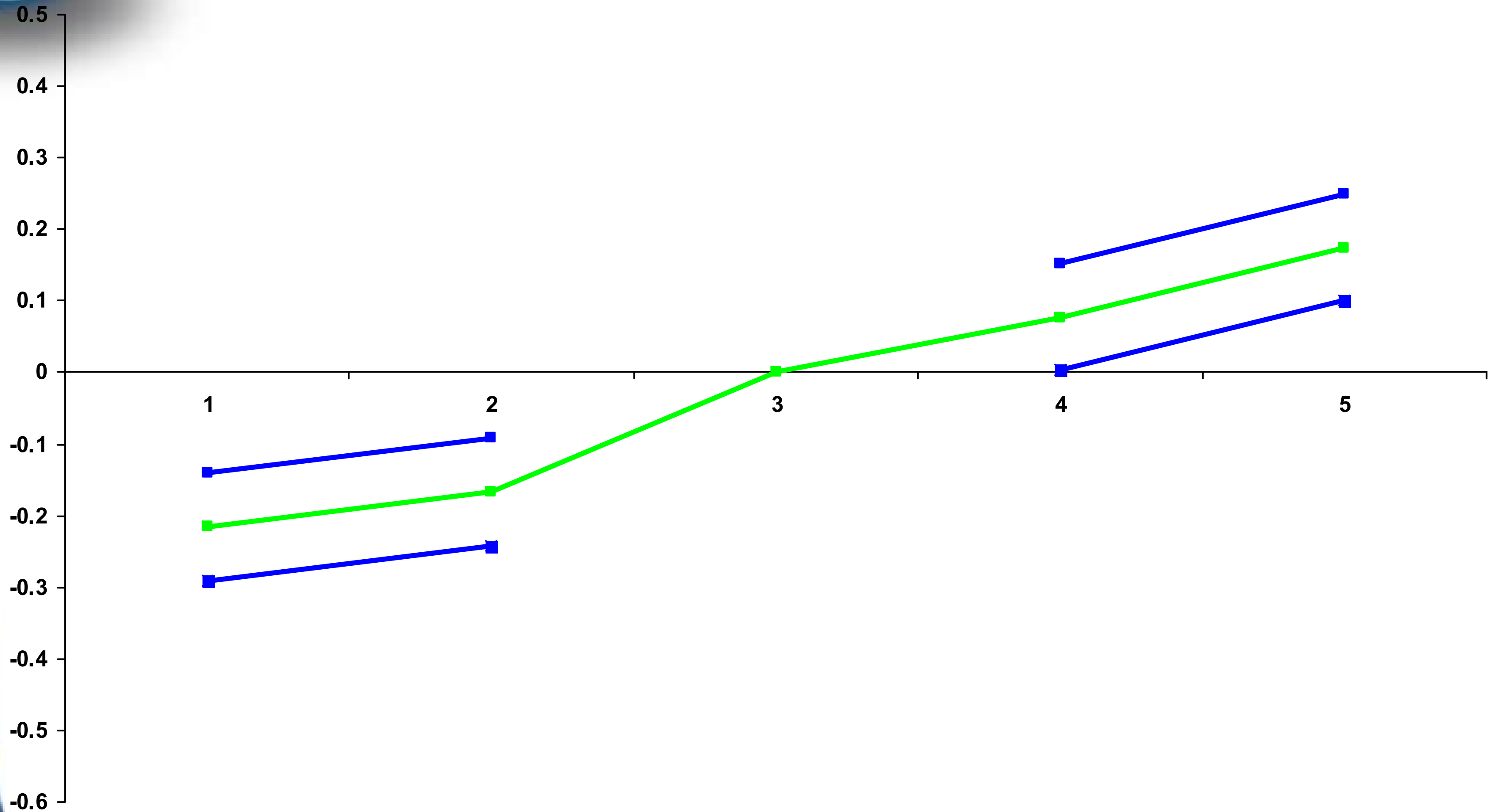
Standard approach

BI	Freq	x	Amt	= Cost 1
PD	Freq	x	Amt	= Cost 2
MED	Freq	x	Amt	= Cost 3
COL	Freq	x	Amt	= Cost 4
OTC	Freq	x	Amt	= Cost 5

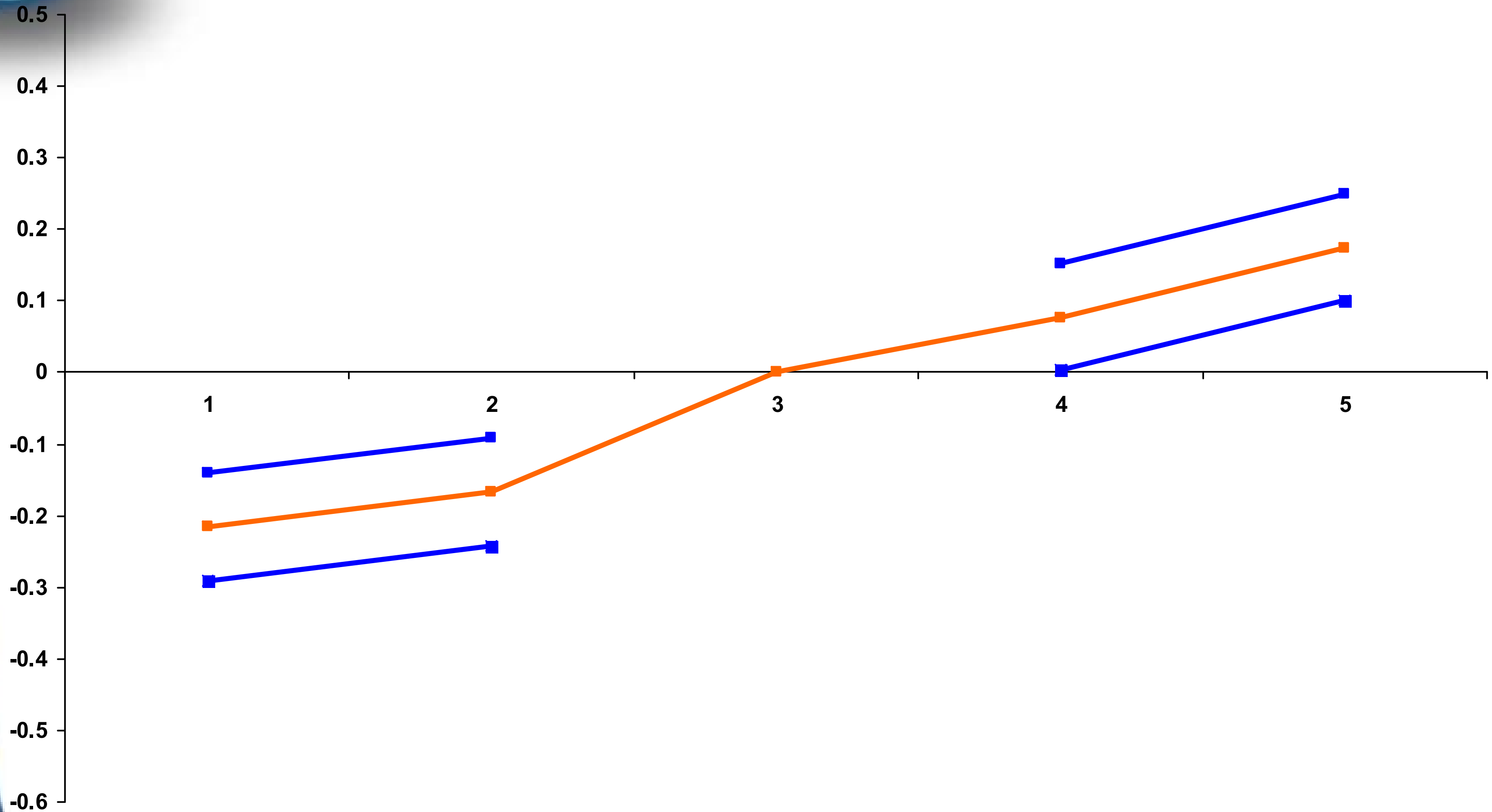
Binomial reference models



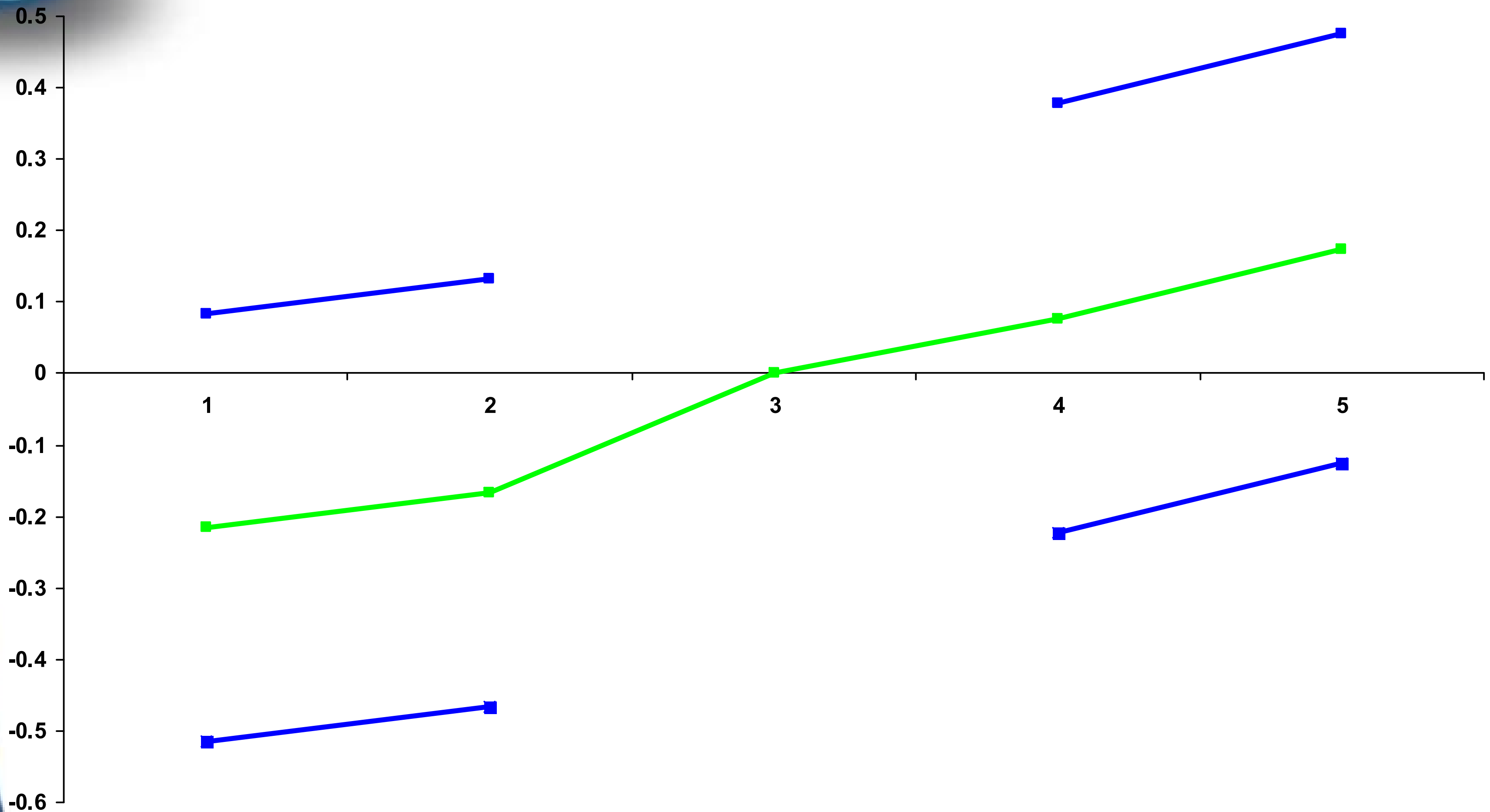
Offset reference model



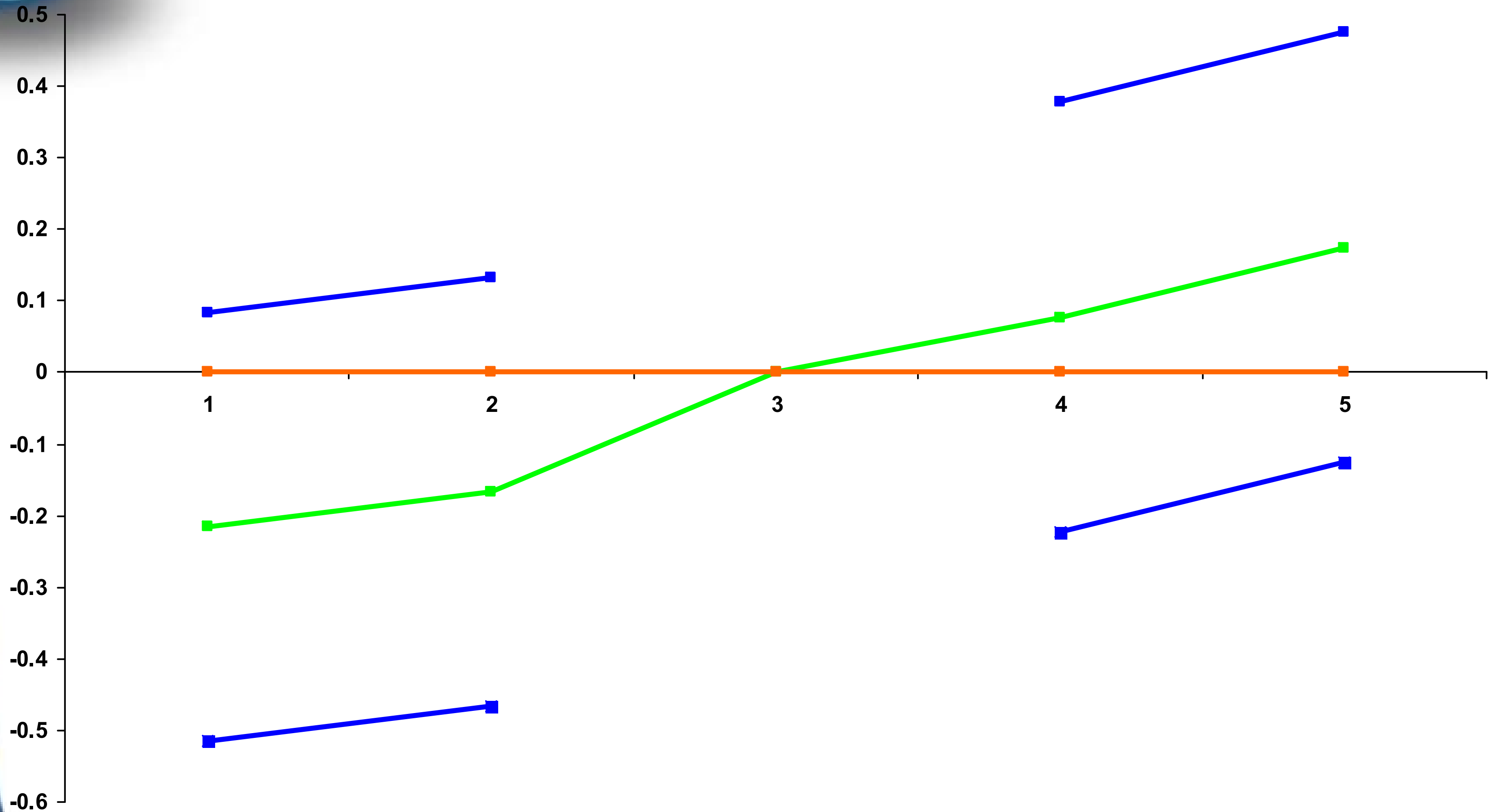
Offset reference model



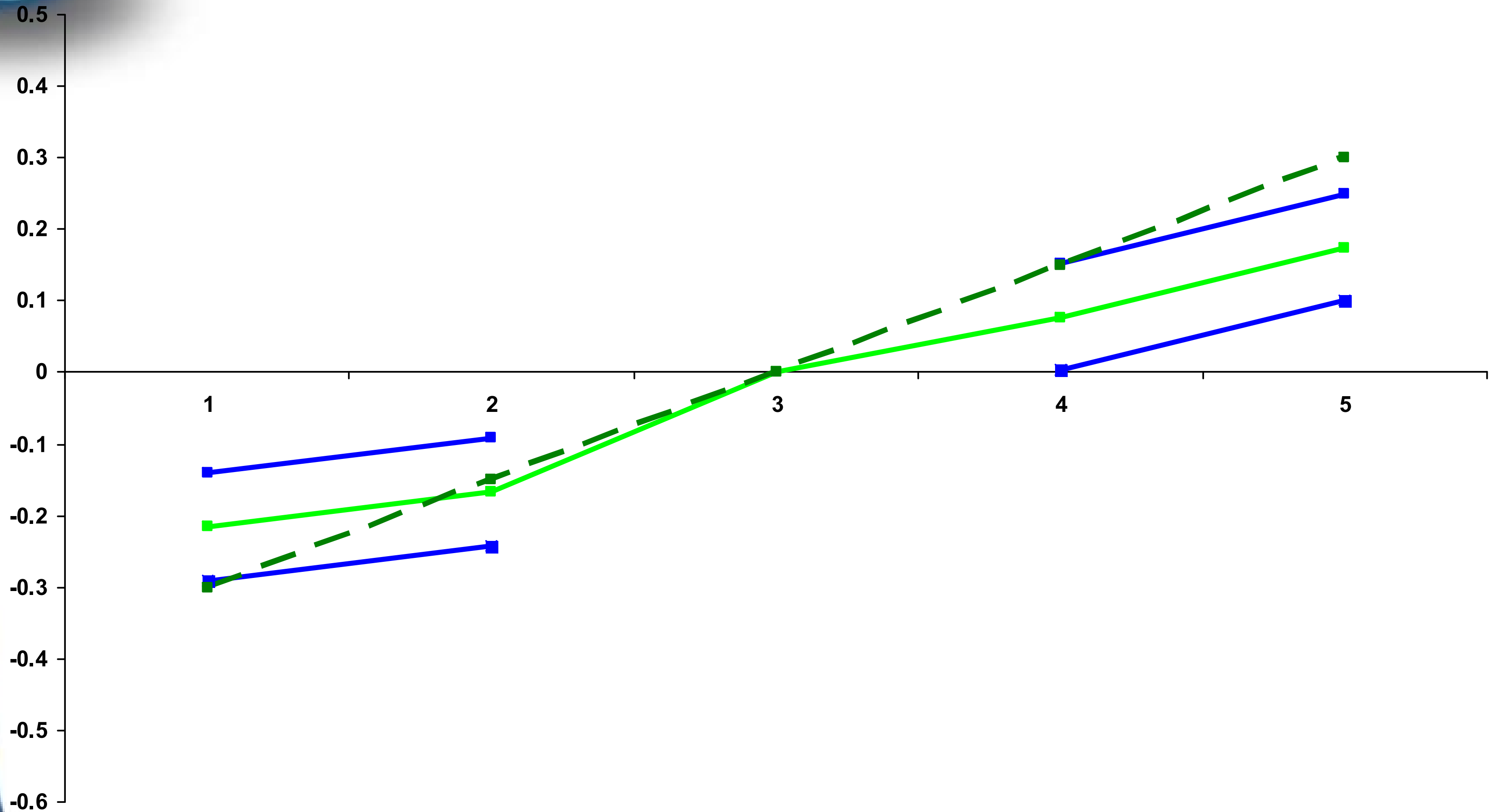
Offset reference model



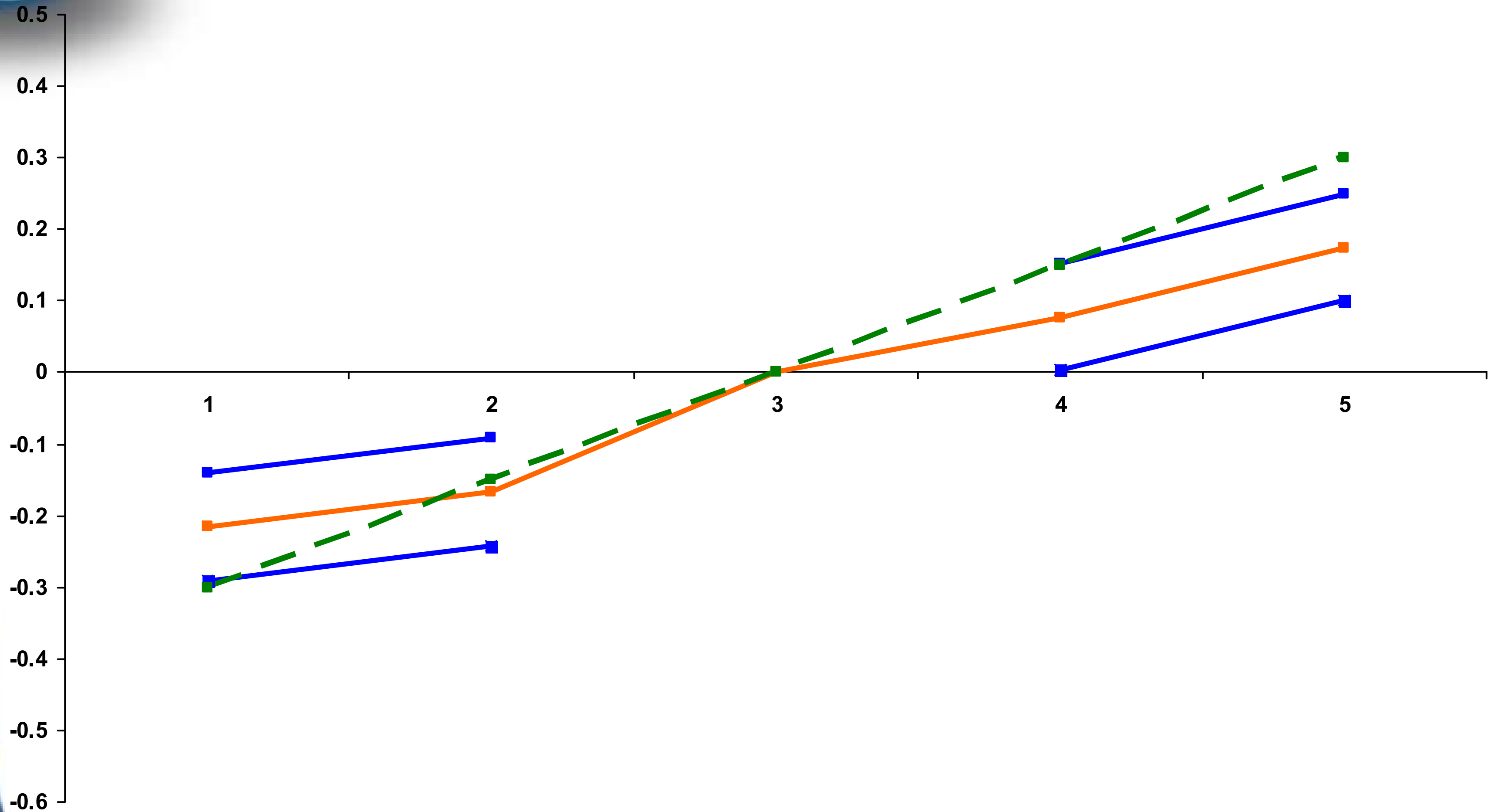
Offset reference model



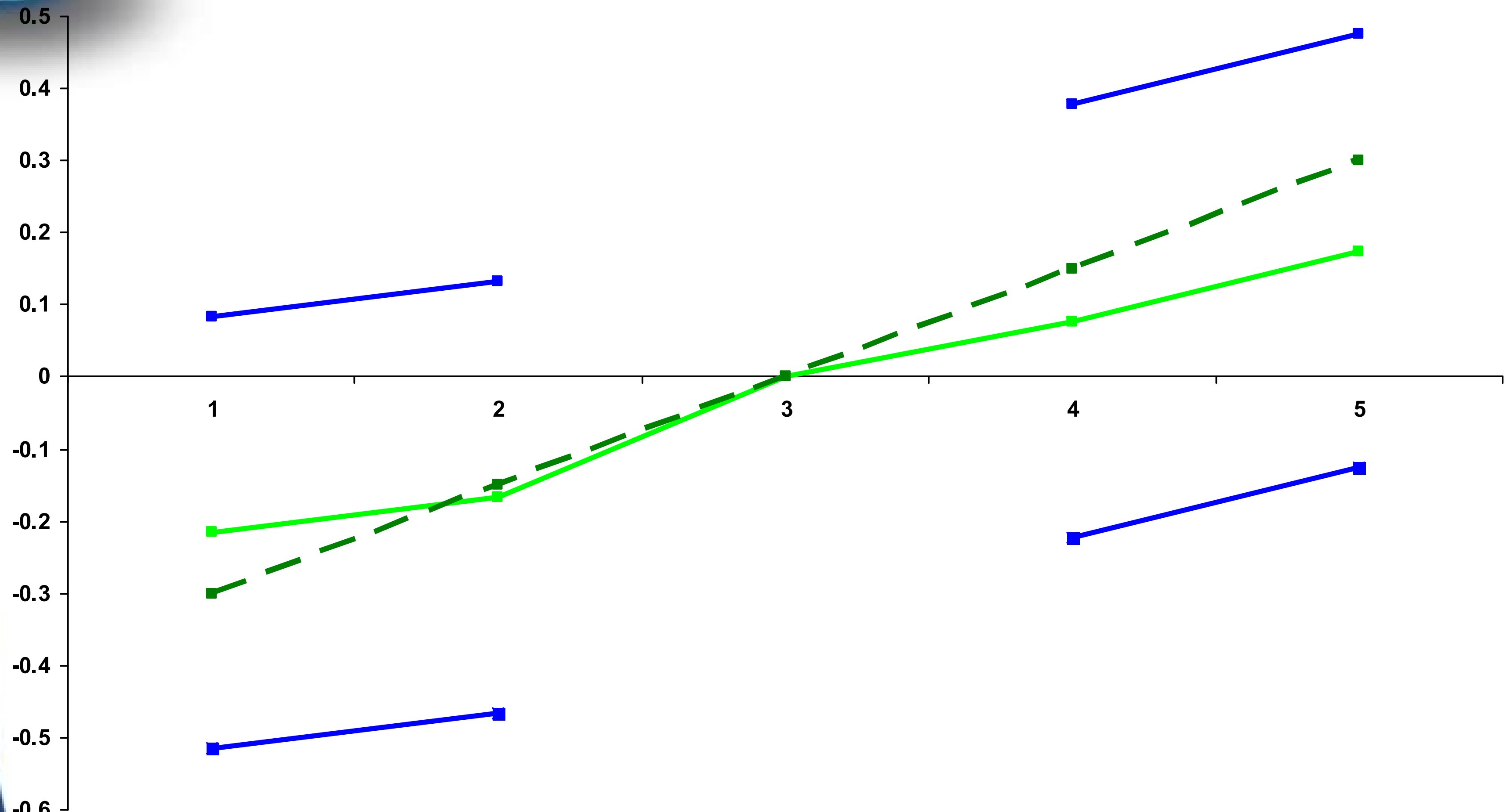
Offset reference model



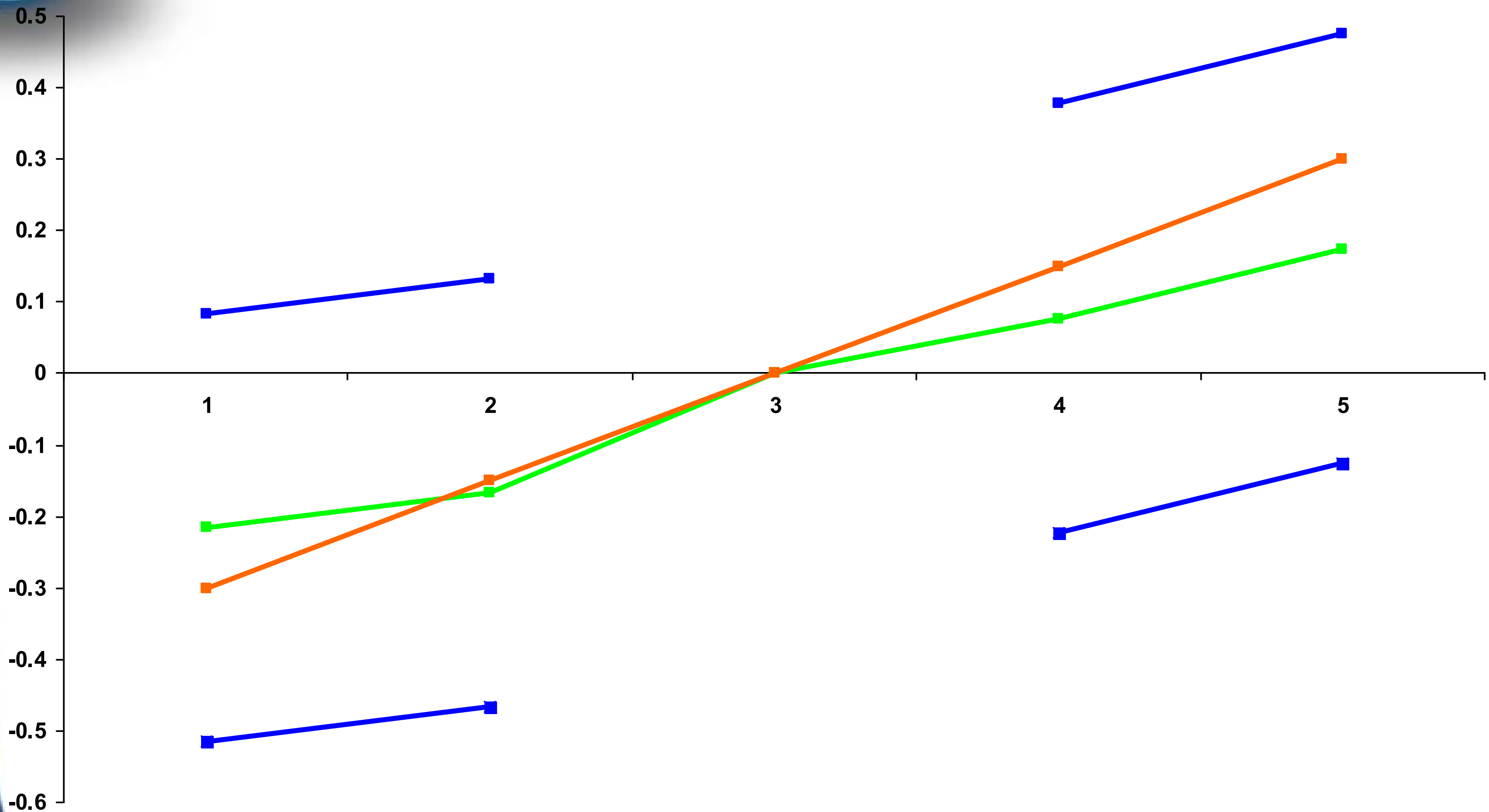
Offset reference model



Offset reference model

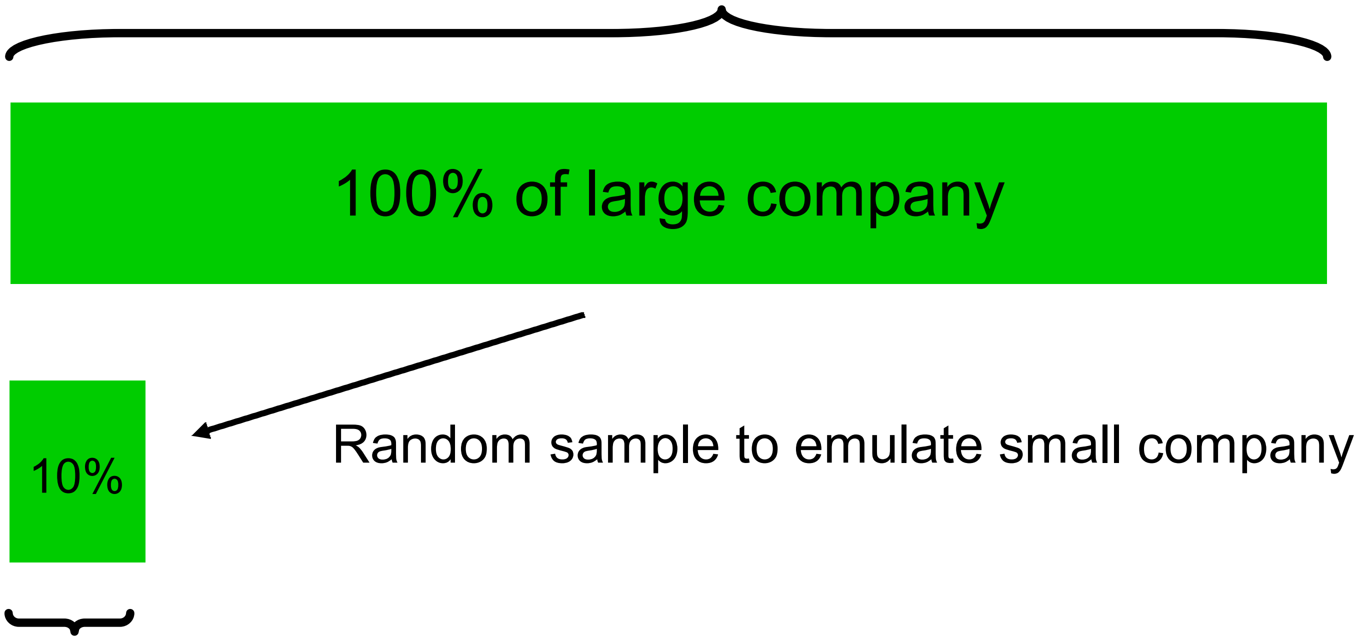


Offset reference model



Testing the reference model approach

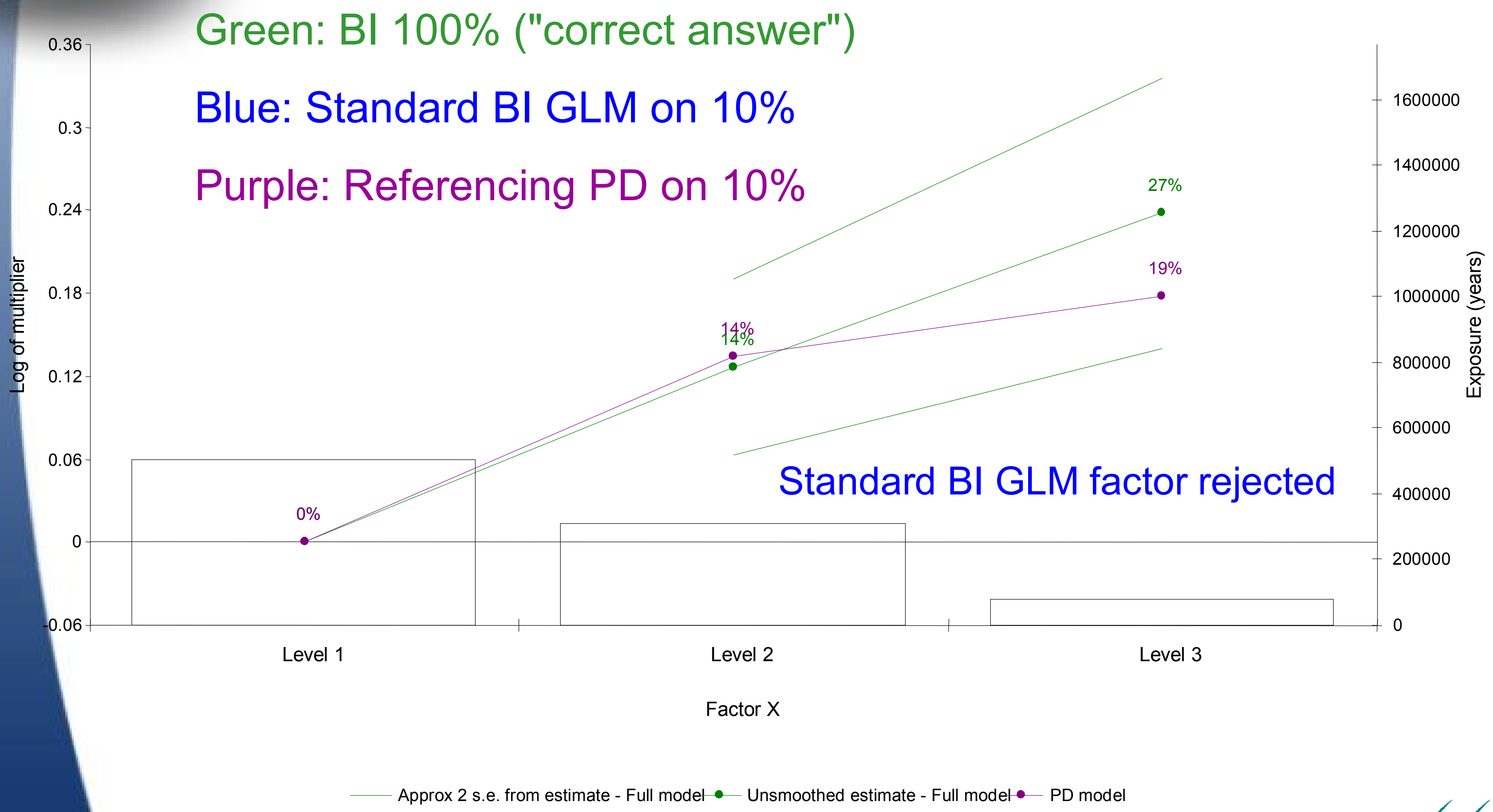
(1) Fit to BI claims on all data - the "correct answer"



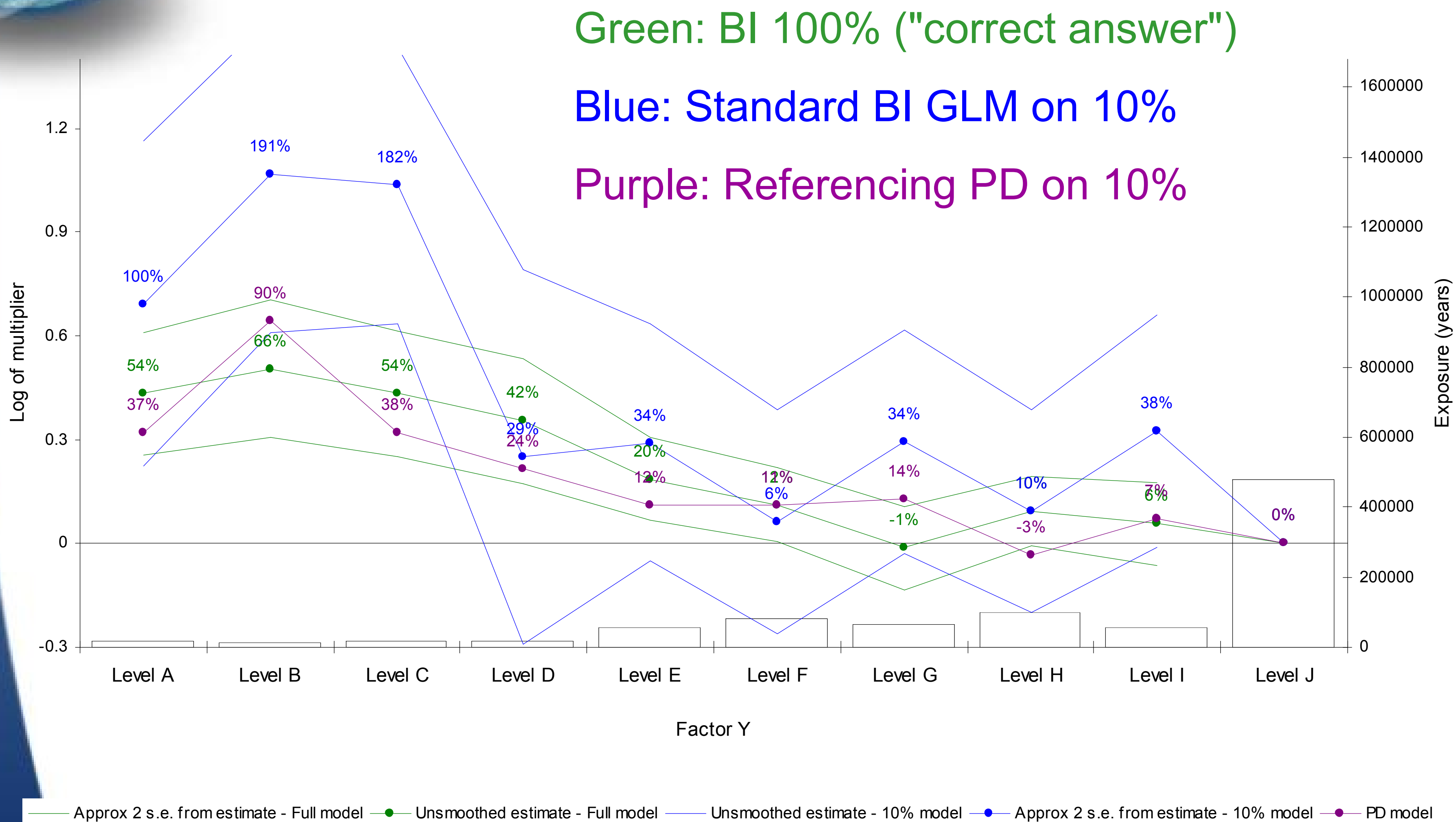
(2) Model BI claims with standard approach

(3) Model BI claims referencing PD experience on this small sample

Example of reference model method working



Example of reference model method working



Agenda

- Introduction
- Testing the link function
- The Tweedie distribution
- Splines
- Reference models
- Aliasing / near aliasing
- Combining models across claim types
- Restricted models
- Model validation

Aliasing and "near aliasing"

- Aliasing
  - the removal of unwanted redundant parameters
- Intrinsic aliasing
  - occurs by the design of the model
- Extrinsic aliasing
  - occurs "accidentally" as a result of the data

Example

- Suppose we wanted a model of the form:

$$\underline{\mu} = \alpha + \beta_1 \text{ if } \underline{\text{age}} < 30$$

$$+ \beta_2 \text{ if } \underline{\text{age}} \text{ 30 - 40}$$

$$+ \beta_3 \text{ if } \underline{\text{age}} > 40$$

$$+ \gamma_1 \text{ if } \underline{\text{sex}} \text{ male}$$

$$+ \gamma_2 \text{ if } \underline{\text{sex}} \text{ female}$$

Form of X.$\underline{\beta}$ in this case

		Age			Sex	
		<30	30-40	>40	M	F
1	1	0	1	0	1	0
2	1	1	0	0	1	0
3	1	1	0	0	0	1
4	1	0	0	1	1	0
5	1	0	1	0	0	1
	......	......	......	......	......	......
	......	......	......	......	......	......

$$\cdot \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \gamma_1 \\ \gamma_2 \end{pmatrix}$$

Example

- Suppose we wanted a model of the form:

$$\underline{\mu} = \alpha + \beta_1 \text{ if } \underline{\text{age}} < 30$$

$$+ \beta_2 \text{ if } \underline{\text{age}} \ 30 - 40$$

"Base levels"

$$+ \beta_3 \text{ if } \underline{\text{age}} > 40$$

$$+ \gamma_1 \text{ if } \underline{\text{sex}} \text{ male}$$

$$+ \gamma_2 \text{ if } \underline{\text{sex}} \text{ female}$$

X.β having adjusted for base levels

		Age			Sex	
		<30	30-40	>40	M	F
1	1	0	1	0	1	0
2	1	1	0	0	1	0
3	1	1	0	0	0	1
4	1	0	0	1	1	0
5	1	0	1	0	0	1
	....	....	....	....	....	....
	....	....	....	....	....	....

$$\cdot \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \gamma_1 \\ \gamma_2 \end{pmatrix}$$

X.$\underline{\beta}$ having adjusted for base levels

		Age		Sex
		<30	>40	F
1	1	0	0	0
2	1	1	0	0
3	1	1	0	1
4	1	0	1	0
5	1	0	0	1
	....			
	....			

$$\cdot \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_3 \\ \gamma_2 \end{pmatrix}$$

Intrinsic aliasing

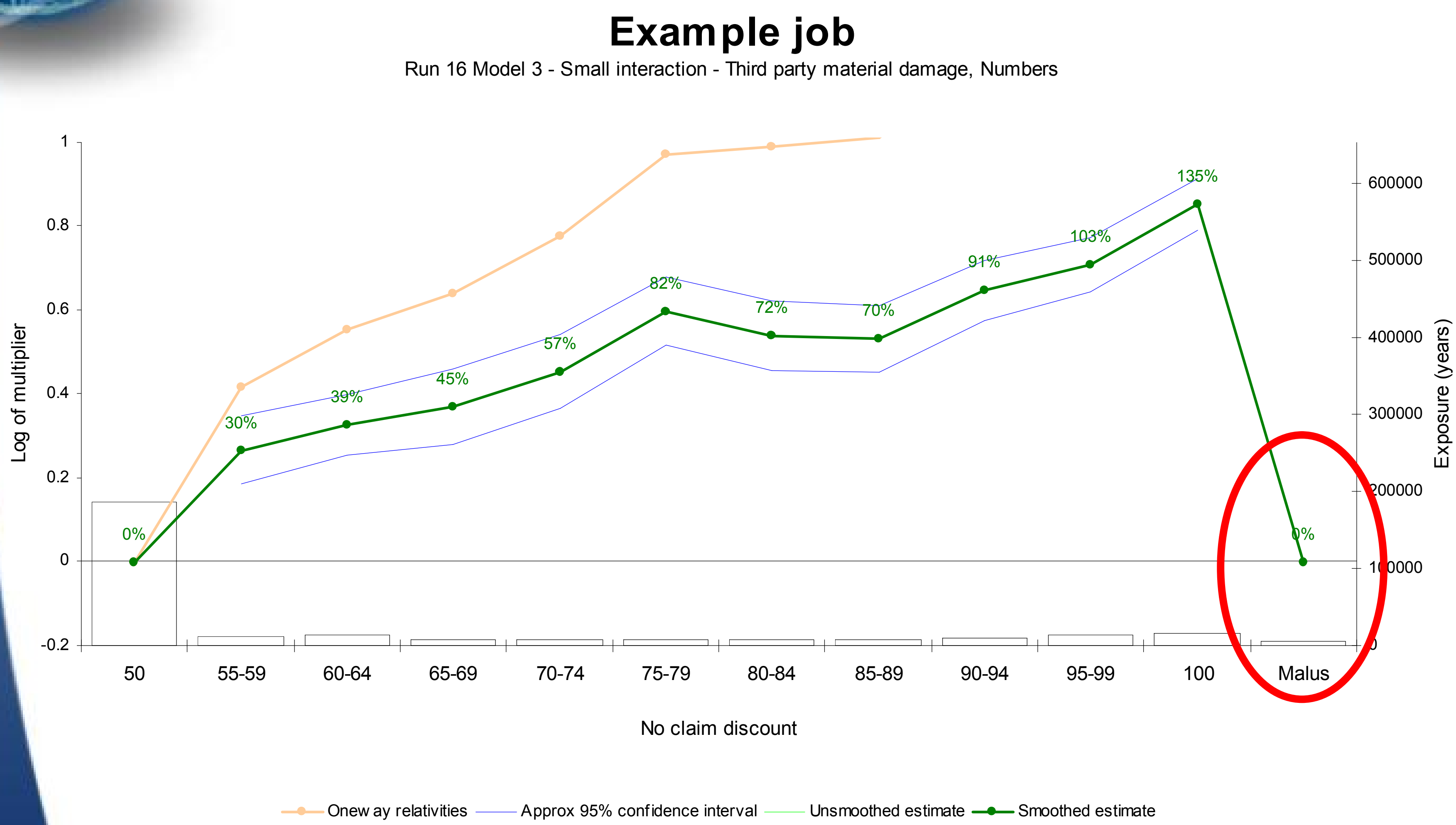
Extrinsic aliasing

- If a perfect correlation exists, one factor can alias levels of another
- Eg if doors declared first:

Exposure: # Doors → / Colour ↓	2	3	4 (Selected base)	5	Unknown
Red (Selected base)	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
Unknown (Further aliasing)	0	0	0	0	3,242

- This is the only reason the order of declaration can matter (fitted values are unaffected)

Extrinsic aliasing



"Near aliasing"

- If two factors are almost perfectly, but not quite aliased, convergence problems can result and/or results can become hard to interpret

Exposure: # Doors → / Colour ↓	2	3	4 (Selected base)	5	Unknown
Red (Selected base)	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

- Eg if the 2 black, unknown doors policies had no claims, GLM would try to estimate a very large negative number for unknown doors, and a very large positive number for unknown colour

"Near aliasing" - solution

1. Spot it
2. Fix the data!

Exposure: # Doors → Colour ↓	2	3	4	5	Unknown
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

Agenda

- Introduction
- Testing the link function
- The Tweedie distribution
- Splines
- Reference models
- Aliasing / near aliasing
- Combining models across claim types
- Restricted models
- Model validation

Combining claim elements - I

BI Freq x Amt = Cost 1

PD Freq x Amt = Cost 2

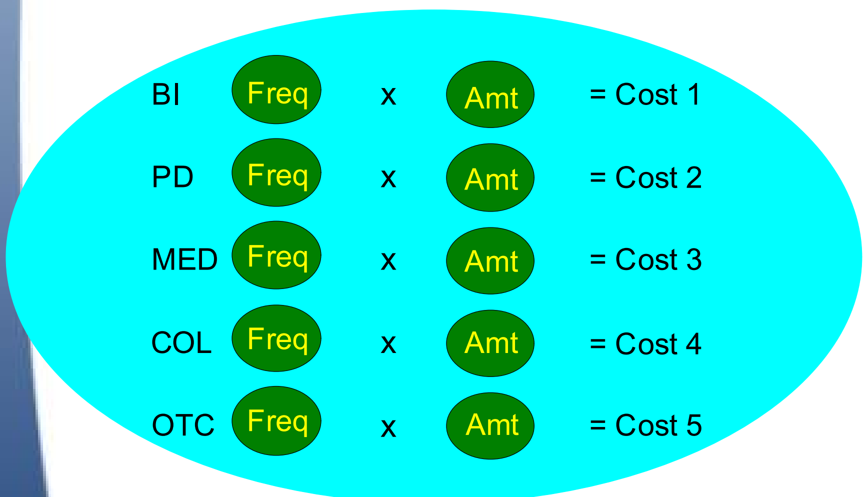
MED Freq x Amt = Cost 3

COL Freq x Amt = Cost 4

OTC Freq x Amt = Cost 5

- Multiply factors for frequencies and amounts
- Calculate risk premium as sum of claim elements

Combining claim elements - II



- Consider current exposure
- Calculate expected frequency and amount for each claim type for each record
- Combine to give expected total cost of claims for each record
- Fit model to this expected value

Calculation of risk premium

		TPPD Numbers	TPPD Amounts	TPBI Numbers	TPBI Amounts
Intercept		32%	£1000	12%	£4860
Sex	Male	1.000	1.000	1.000	1.000
	Female	0.750	1.200	0.667	0.900
Area	Town	1.000	1.000	1.000	1.000
	Country	1.250	0.700	0.750	0.833

Policy	Sex	Area	WWNUM1	WWAMT1	WWNUM2	WWAMT2	WWCC1	WWCC2	WWRSKPRM
…	…	…	…	…	…	…	…	…	…
82155654	M	T	32%	1000	12%	4860	320	583.20	903.20
82168746	F	T	24%	1200	8%	4374	288	349.92	637.92
82179481	M	C	40%	700	9%	4050	280	364.50	644.50
82186845	F	C	30%	840	6%	3645	252	218.70	470.70
…	…	…	…	…	…	…	…	…	…

Risk premium standard errors

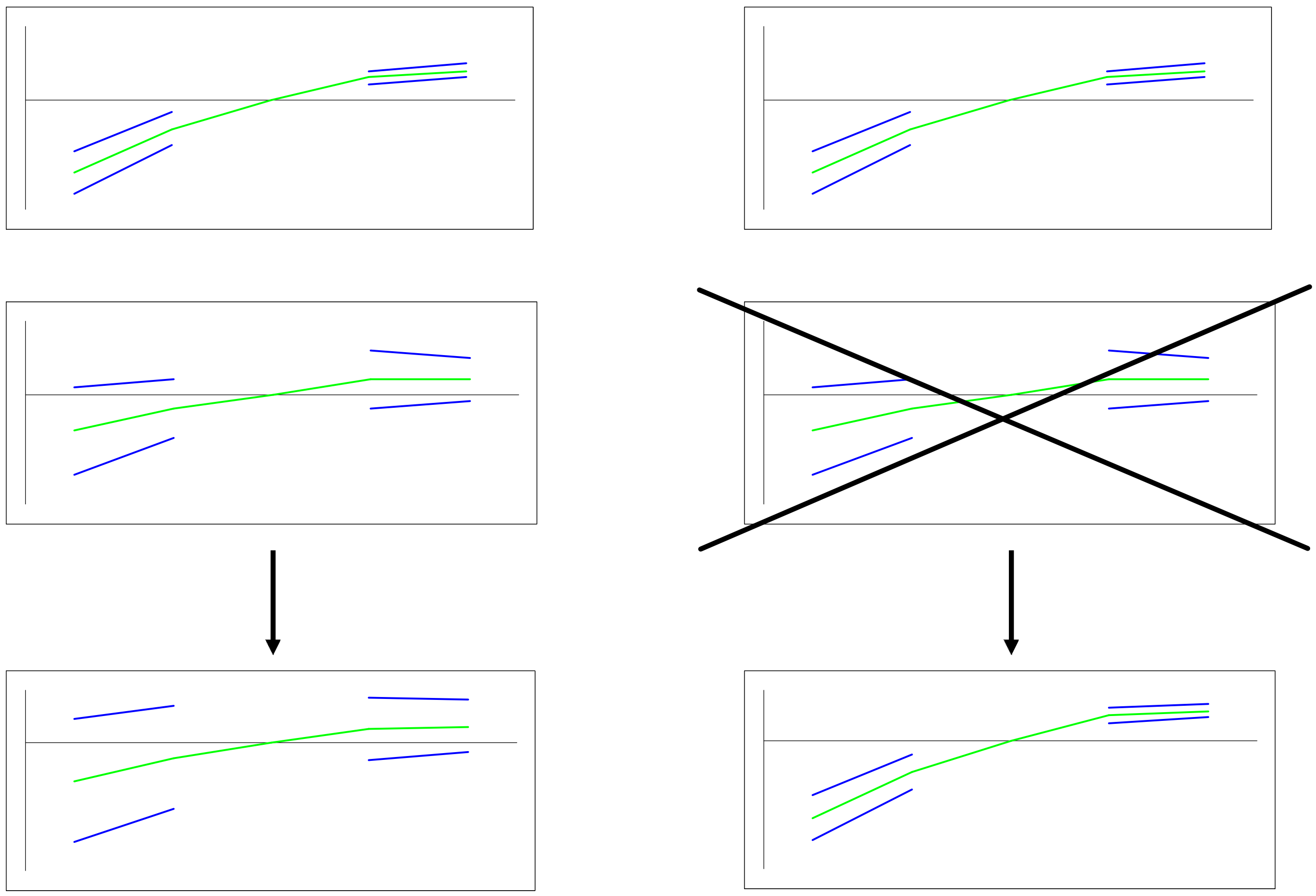
- Risk premium model standard errors are small owing to the smoothness of the expected value
- It is possible to approximate standard error of risk premium parameter estimates based on standard errors of parameter estimates in underlying models
- Care needed in interpreting such approximations since they do not reflect model error, eg deciding to exclude a marginal factor

Risk premium standard errors - failings

Numbers

Amounts

Risk premium



Agenda

- Introduction
- Testing the link function
- The Tweedie distribution
- Splines
- Reference models
- Aliasing / near aliasing
- Combining models across claim types
- Restricted models
- Model validation

Restricted models

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

Offset

- Offset term used for known effects, eg exposure in a numbers model
- Can also be used to constrain model (eg claim free years / payment frequency / amount of cover)
- Other factors adjusted to compensate

Restricted models

		Age		Sex
		<30	>40	F
1	1	0	0	0
2	1	1	0	0
3	1	1	0	1
4	1	0	1	0
5	1	0	0	1
	....	....	....	....
	....	....	....	....

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix} \cdot \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_3 \\ \gamma_2 \end{pmatrix}$$

Restricted models

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

$$\begin{array}{c} \\ \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ \\ \\ \end{array} \begin{pmatrix} & \text{Age} & & \text{Sex} \\ & <30 & >40 & F \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix} . \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_3 \\ \gamma_2 \end{pmatrix}$$

Restricted models

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

		Age <30	Age >40
1	1	0	0
2	1	1	0
3	1	1	0
4	1	0	1
5	1	0	0

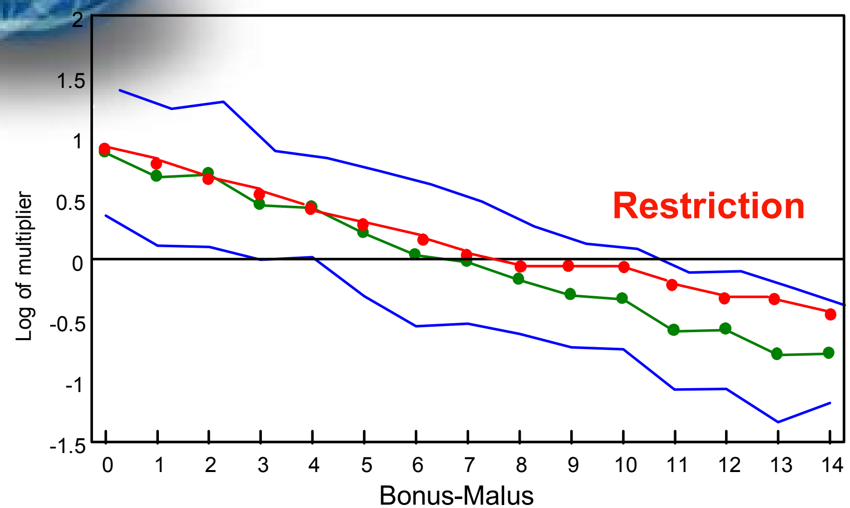
$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ \dots & \dots & \dots \\ \dots & \dots & \dots \end{pmatrix} . \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0.1 \\ 0 \\ 0.1 \\ \dots \\ \dots \end{pmatrix}$$

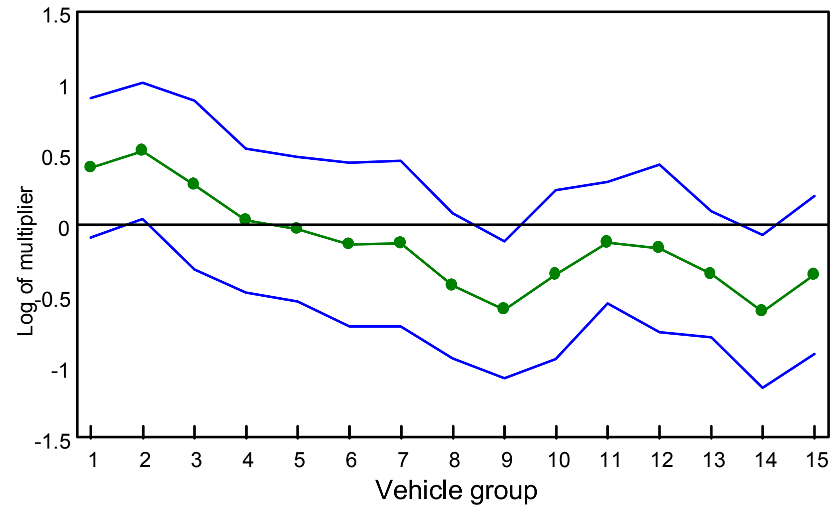
Restricted models

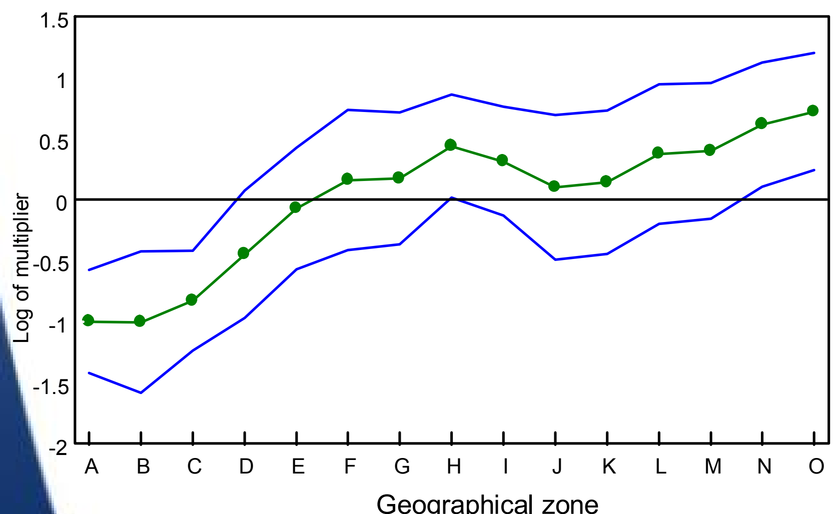
		Age		Sex
		<30	>40	F
1	1	0	0	0
2	1	1	0	0
3	1	1	0	1
4	1	0	1	0
5	1	0	0	1
	......	......	......	......
	......	......	......	......

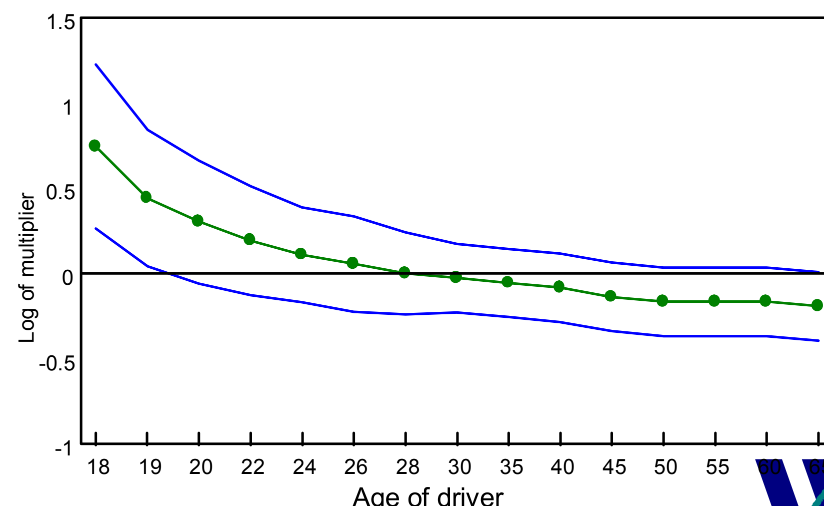
$$\cdot \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_3 \\ \mathbf{0.1} \end{pmatrix}$$

Restricted models

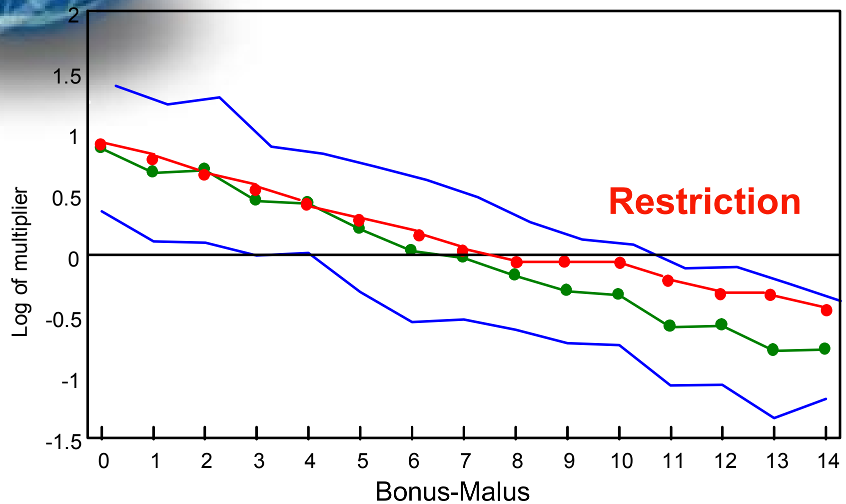


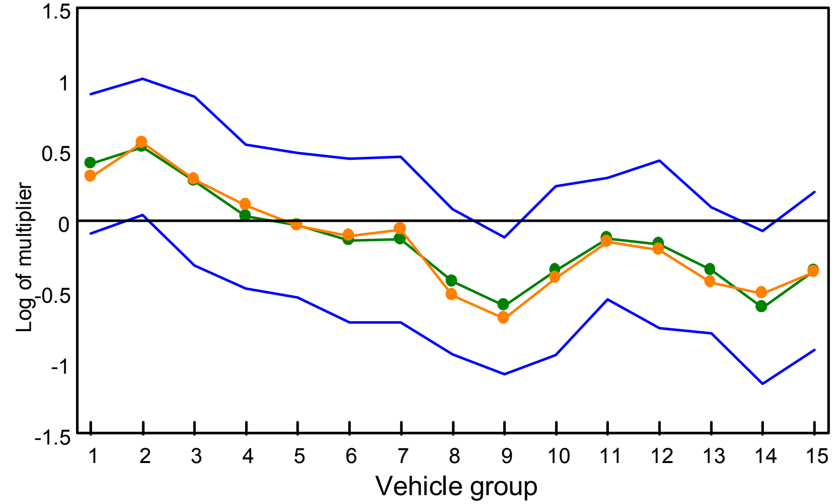


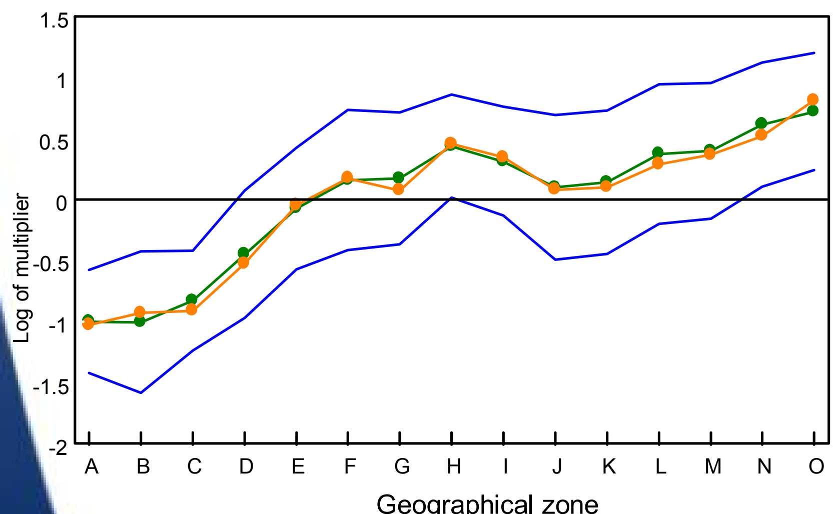


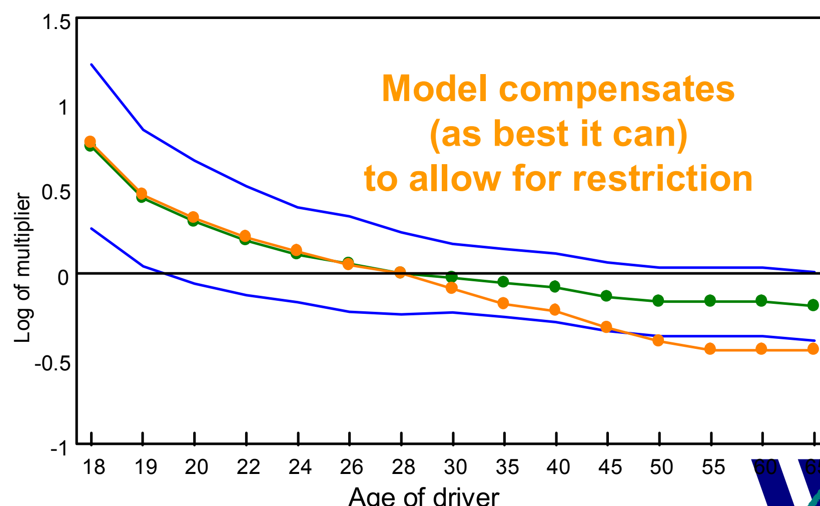


Restricted models



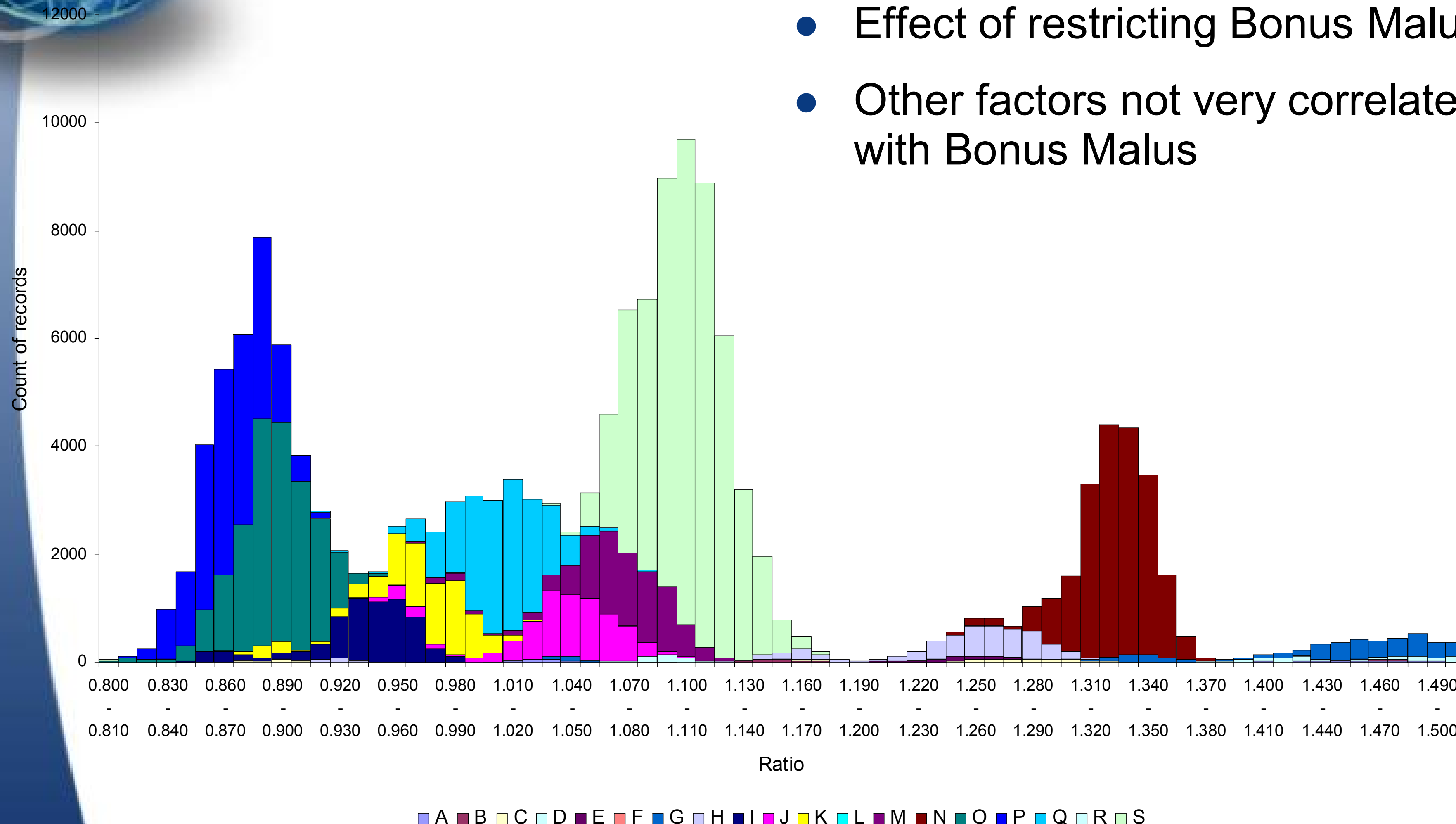




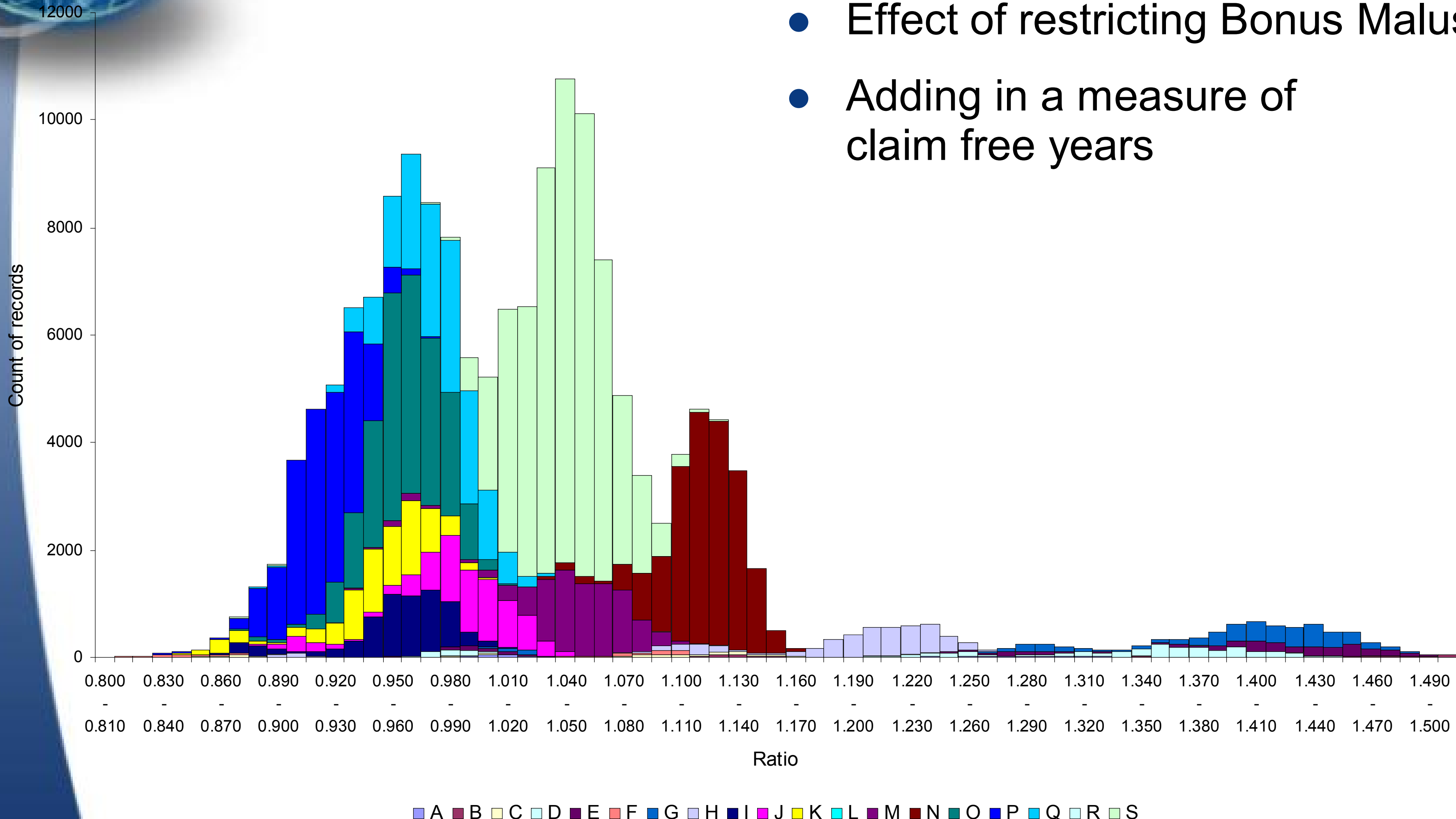


Testing the effectiveness of restrictions

- Effect of restricting Bonus Malus
- Other factors not very correlated with Bonus Malus



Testing the effectiveness of restrictions



- Effect of restricting Bonus Malus
- Adding in a measure of claim free years

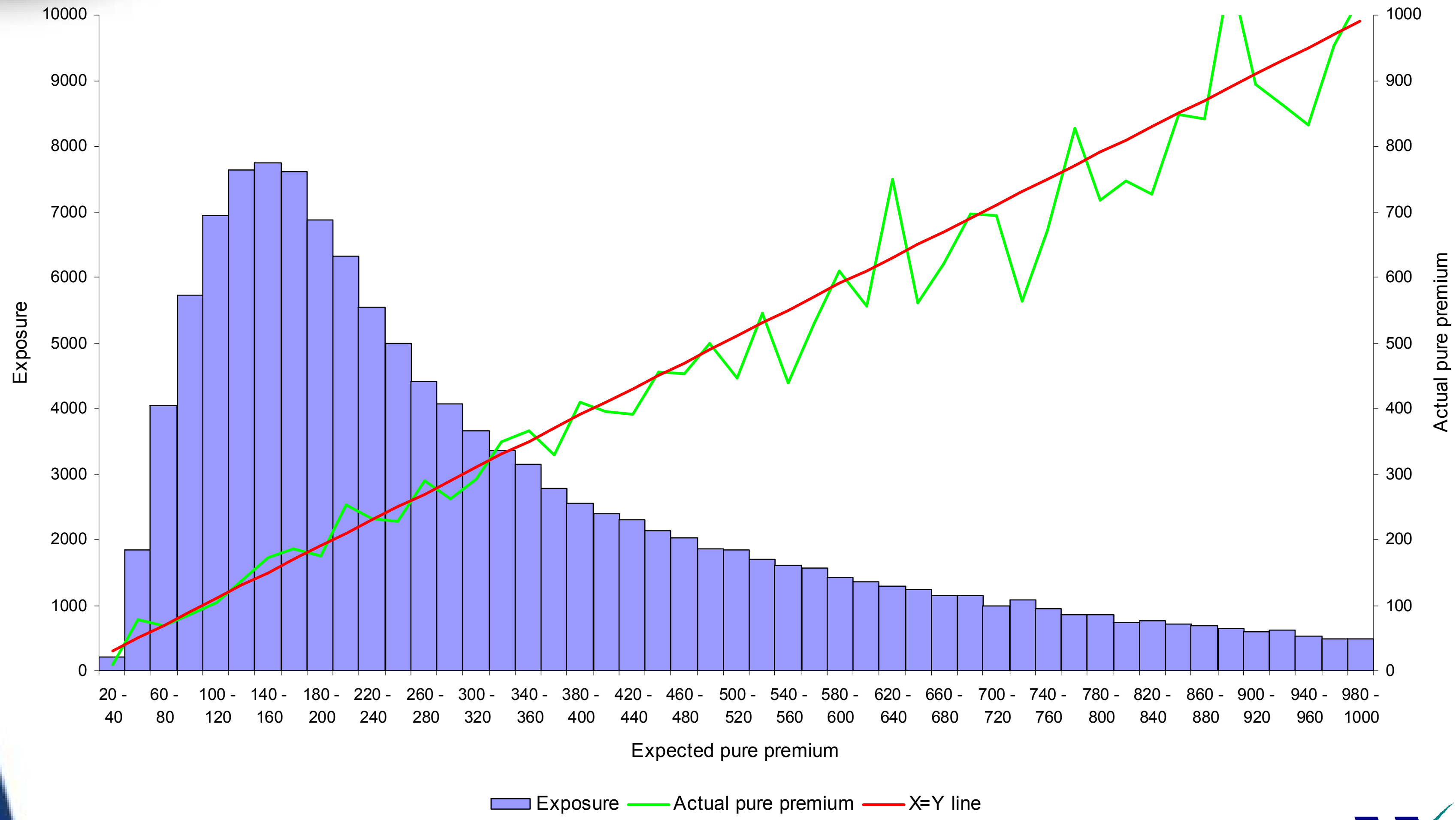
Restrictions

- Only use to "get around" restrictions
- A commercial smoothing is a commercial smoothing
- Apply at risk premium stage

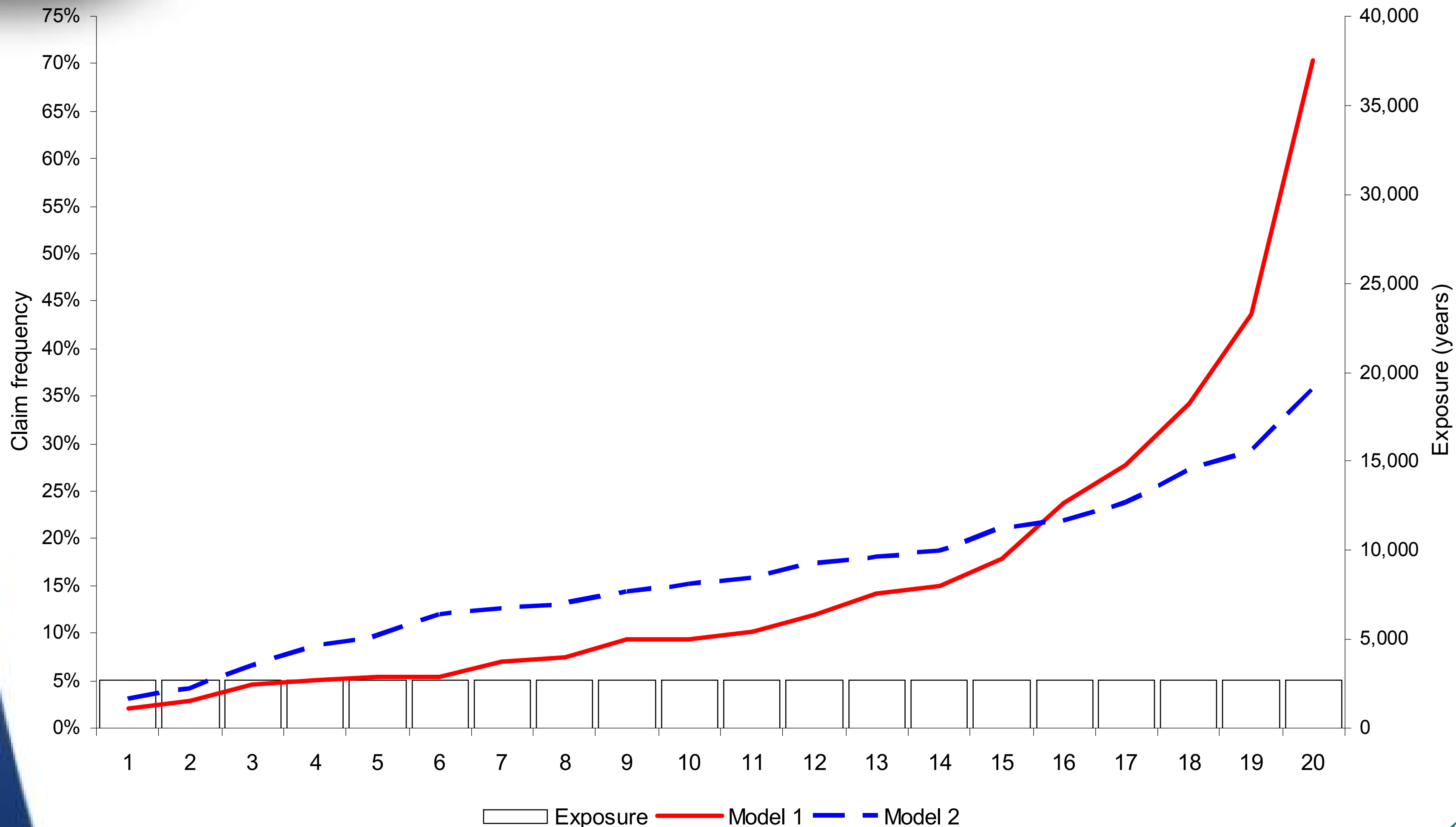
Agenda

- Introduction
- Testing the link function
- The Tweedie distribution
- Splines
- Reference models
- Aliasing / near aliasing
- Combining models across claim types
- Restricted models
- Model validation

Model validation



Lift curves



75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Claim frequency
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
Exposure (years)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Exposure
Model 1
Model 2

GLM III: Advanced Modeling Strategy

WWW.WATSONWYATT.COM

2006 CAS Seminar on Predictive Modeling

Duncan Anderson MA FIA

Watson Wyatt Worldwide

Watson Wyatt Worldwide