C-14 Finding the Right Synergy from GLMs and Machine Learning

2010 CAS Annual Meeting

Claudine Modlin

November 8, 2010

Definitions

Parametric modeling

- Objective: build a predictive model
- User makes assumptions (e.g., distribution, model structure) and specifies preliminary list of explanatory variables
- User guides statistical method in order to effectively describe a particular response (e.g., claim frequency)
- Result is an algorithm, a set of parameters, and diagnostics
- Examples: minimum bias methods, linear regression, GLM

Definitions

Machine learning tools

- Objective: learn new things (which may help in building a model)
- Find patterns (often complex) in an unknown underlying distribution
- Tool may be supervised, unsupervised, or blend of the two
- Result might be a new variable, a tree, a grouping, a score, etc
- Examples: principal components analysis, decision trees, clustering, artificial neural networks

A confusing message …..

GLMs have "weaknesses," as evident by unexplained predictive power in the GLM residuals.

Therefore they need to be "corrected" via machine learning methods.

Before we jump to conclusions….

- Make sure your GLM is as good as it can be (i.e., follow best practices)
- Use machine learning methods to improve each stage of the GLM process

Before we jump to conclusions….

"All models are **wrong**, but some are **useful**"

– George E.P. Box

- What does "**useful**" imply other than reliably accurately predictive?
  - Easy to understand and communicate
  - Available in a timely manner
  - Capable of implementation

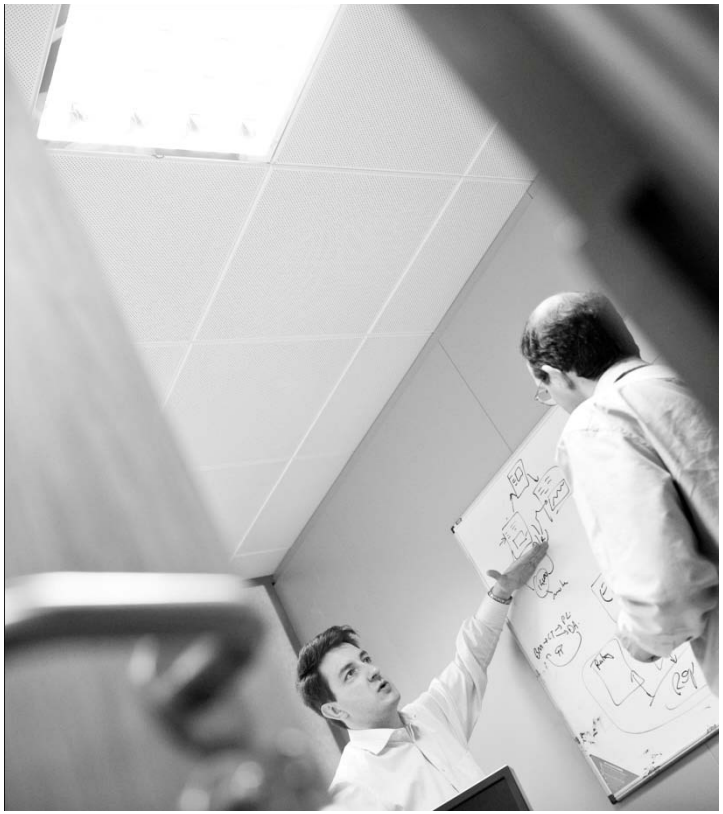
Agenda

- Kristi:
  - GLM best practices
  - Machine learning at every stage of GLM analysis
- Claudine:
  - Additional enhancements to GLM
  - Mining GLM residuals via machine learning

GLM enhancements

GLM enhancements

- Testing link function assumption
- Saddles for interaction detection



GLM enhancement

Test link function via Box-Cox investigation

$$E[Y_i] = \mu_i = g^{-1}(\Sigma X_{ij}.\beta_j + \xi_i) \quad Var[Y_i] = \phi.V(\mu_i)/\omega_i$$

Box-Cox link function defined as:

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

$\lambda = 1 \quad \Rightarrow g(x) = (x - 1) \quad \Rightarrow$ additive (with a base level shift)

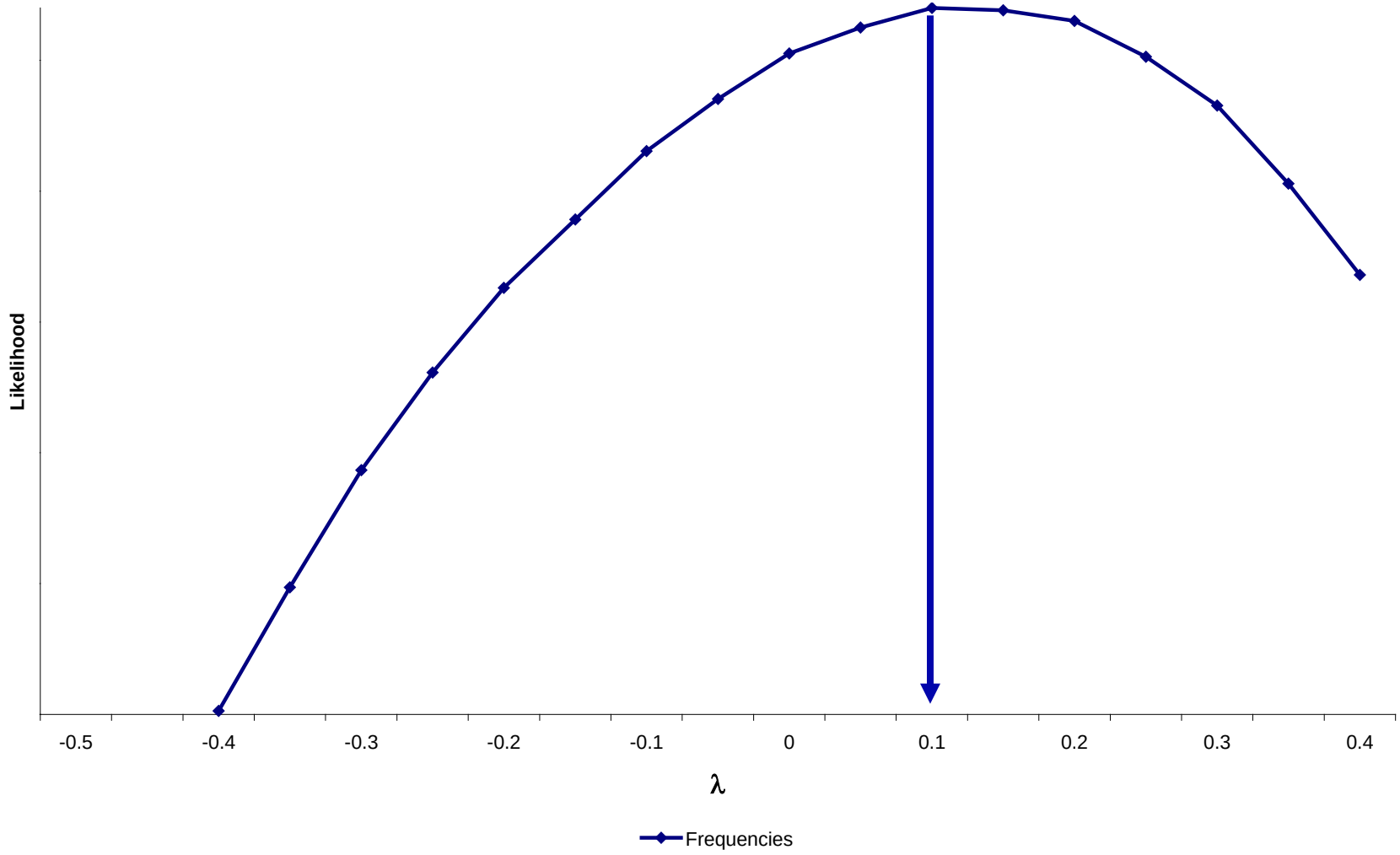
$\lambda \rightarrow 0 \quad \Rightarrow g(x) \rightarrow \ln(x) \quad \Rightarrow$ multiplicative (via l'Hôpital)

$\lambda = -1 \quad \Rightarrow g(x) = 1\text{-}1/x \quad \Rightarrow$ inverse (with a base level shift)

Test a range of values of λ and see which maximizes likelihood

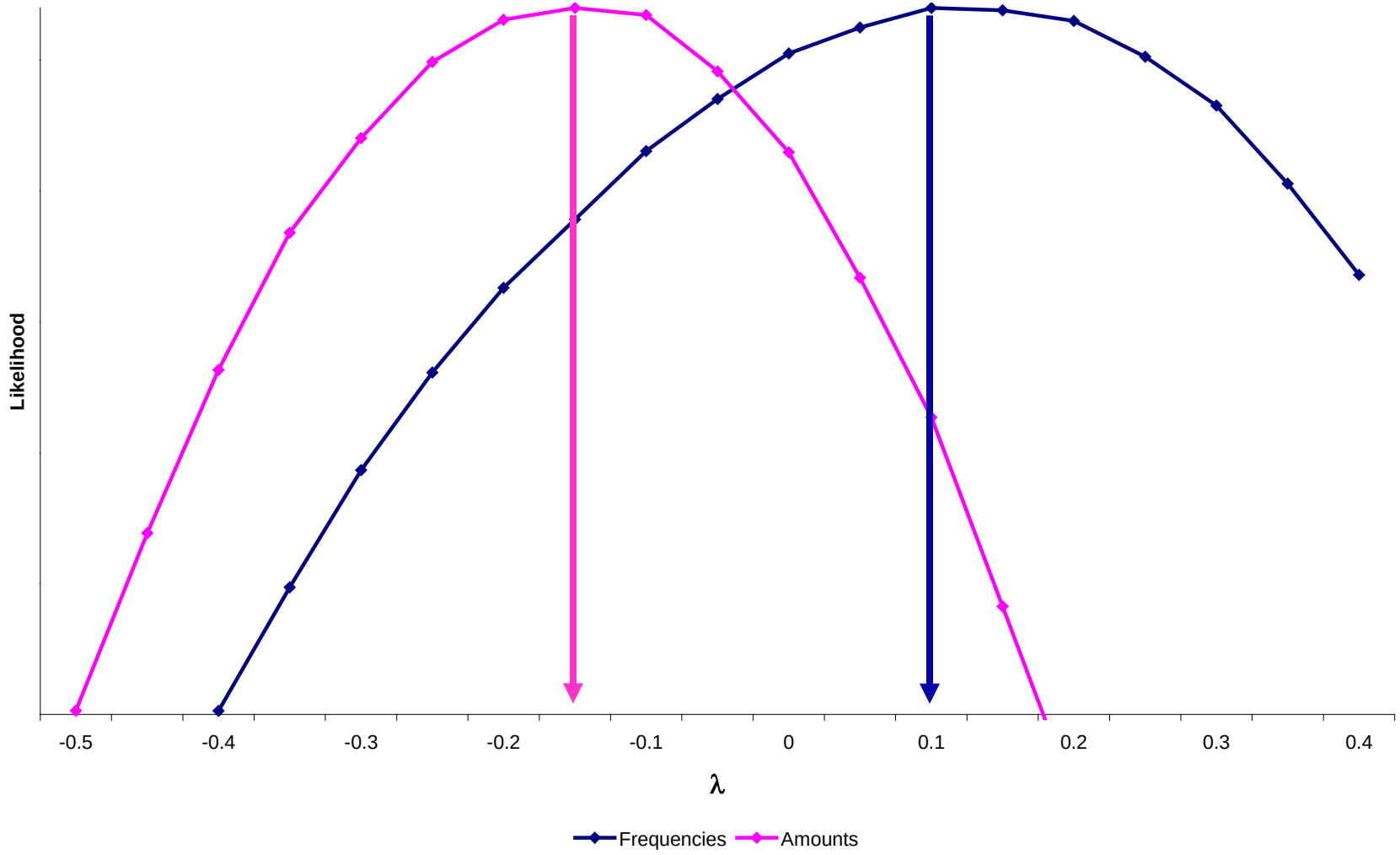
GLM enhancement

Test link function via Box-Cox investigation

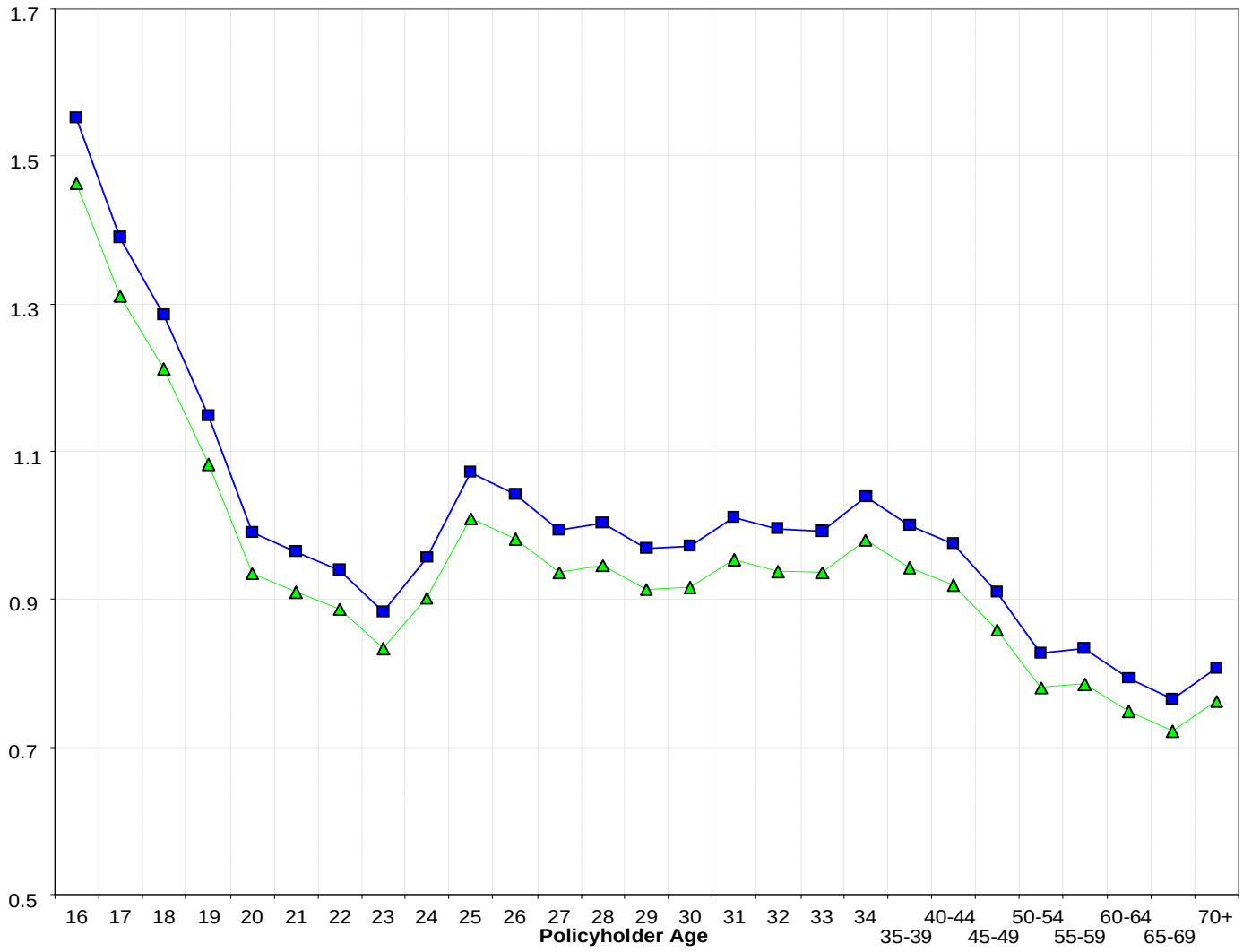


GLM enhancement

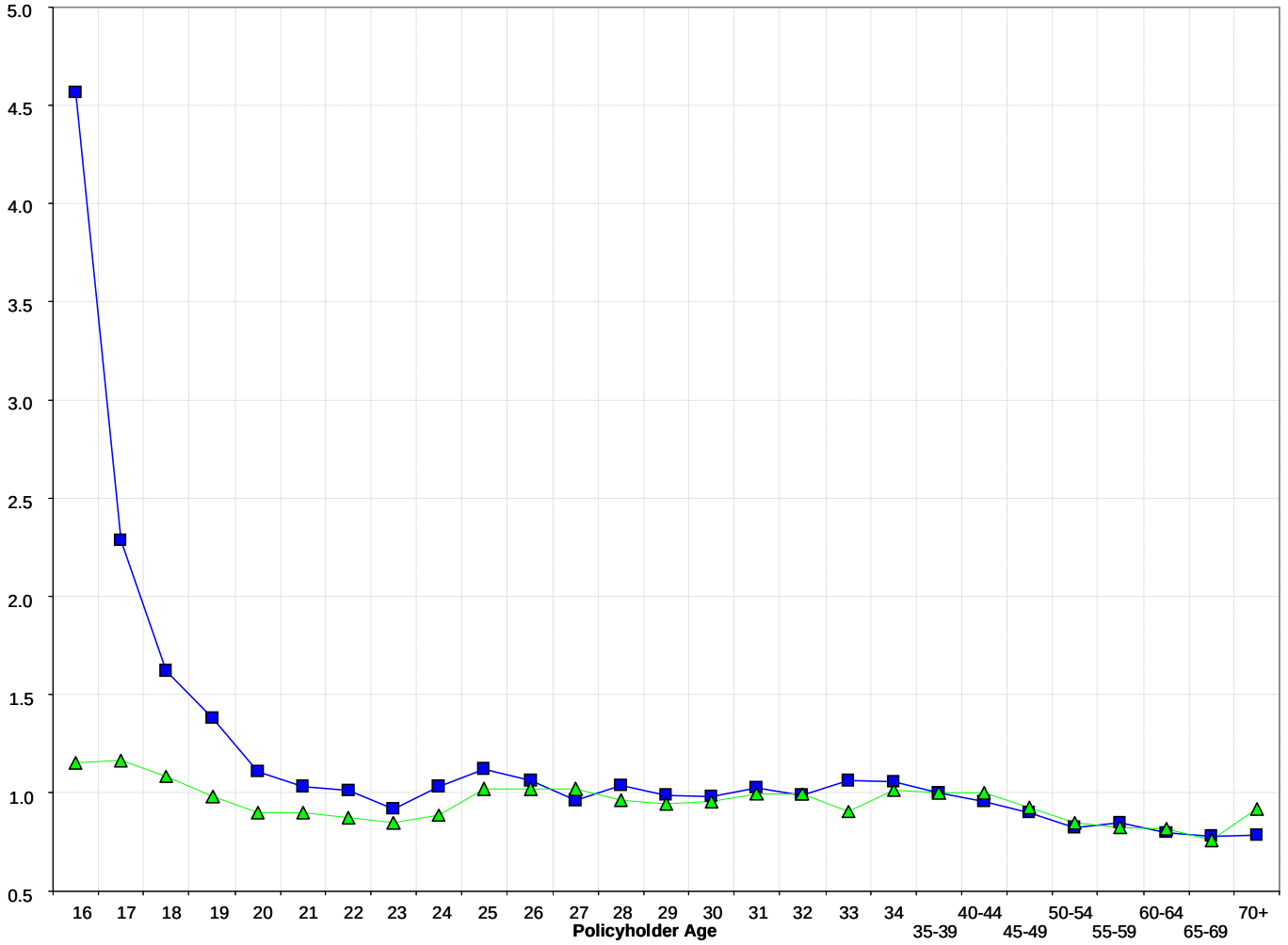
Test link function via Box-Cox investigation



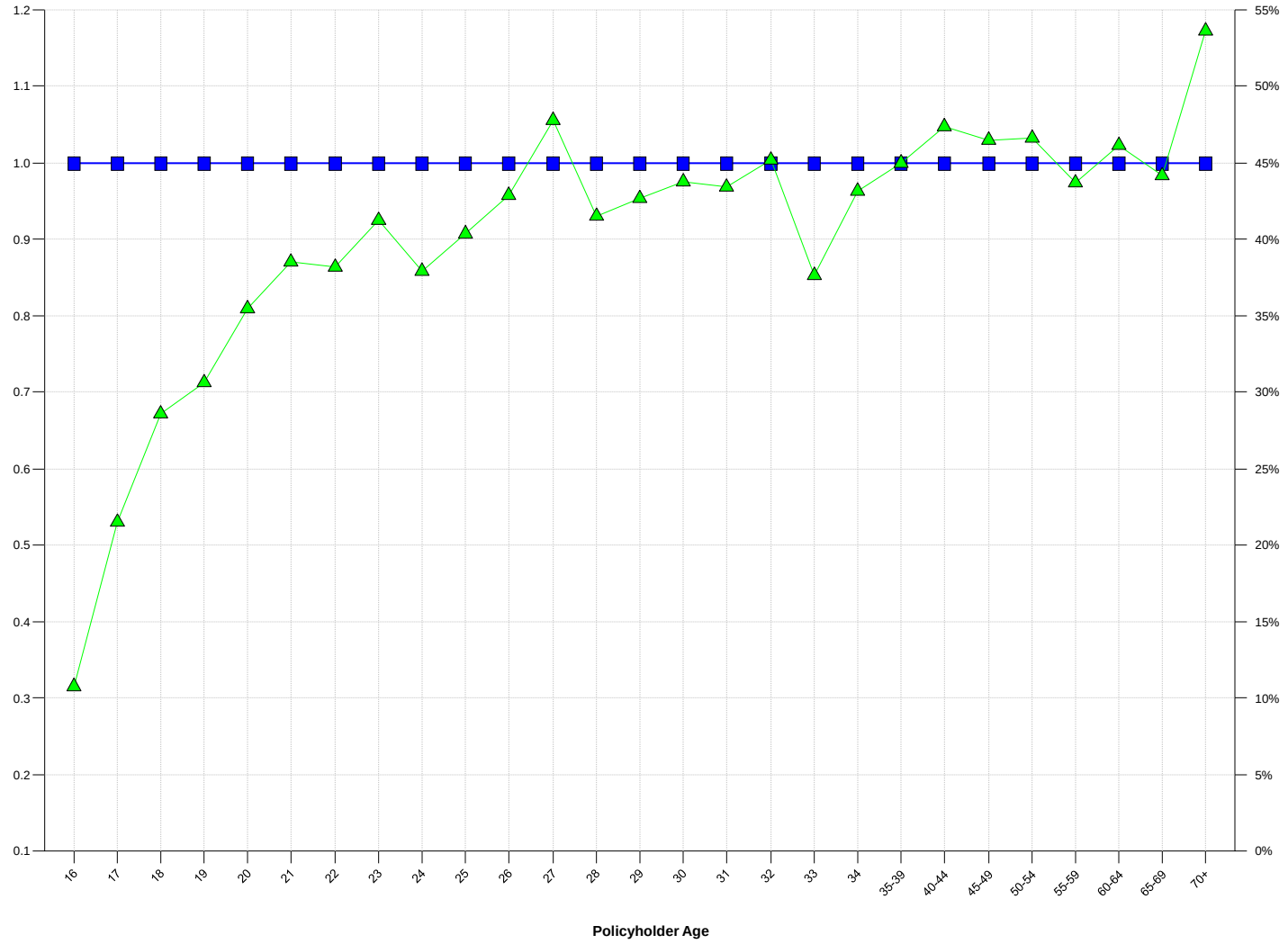
Interactions



Interactions



Interactions



GLM enhancement

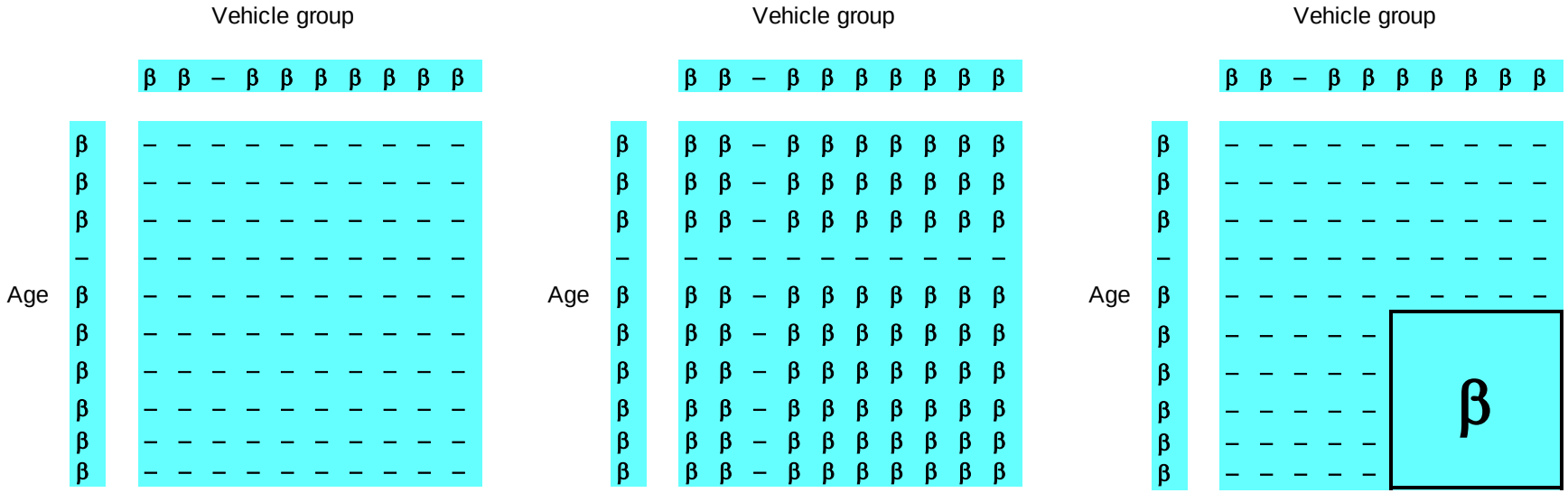
Interactions

Why are interactions present?

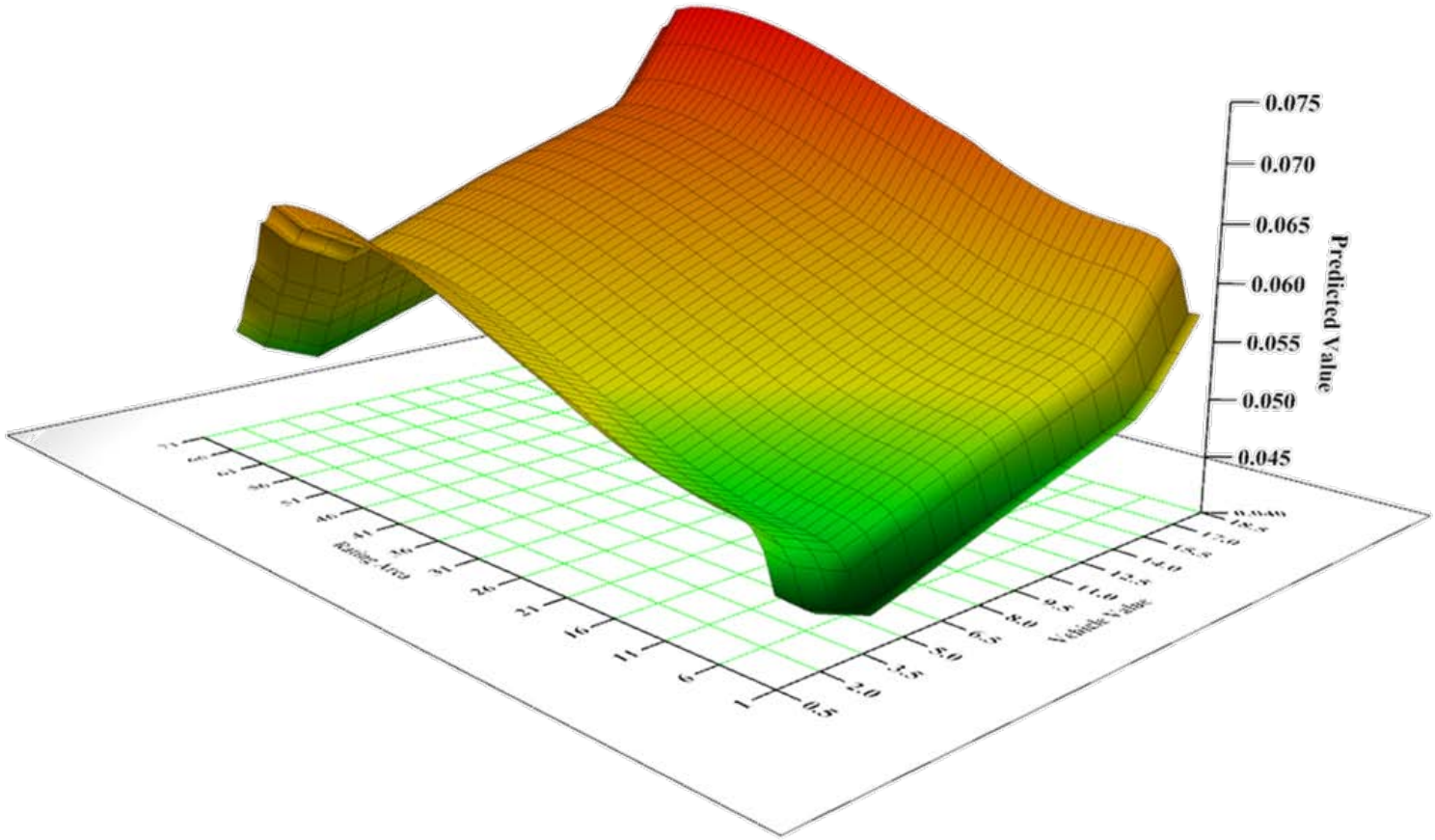
1. Because that's how the factors behave
2. Because multiplicative models can go wrong at the edges
   - 1.5 * 1.4 * 1.7 * 1.5 * 1.8 * 1.5 * 1.8 = 26!

GLM enhancement

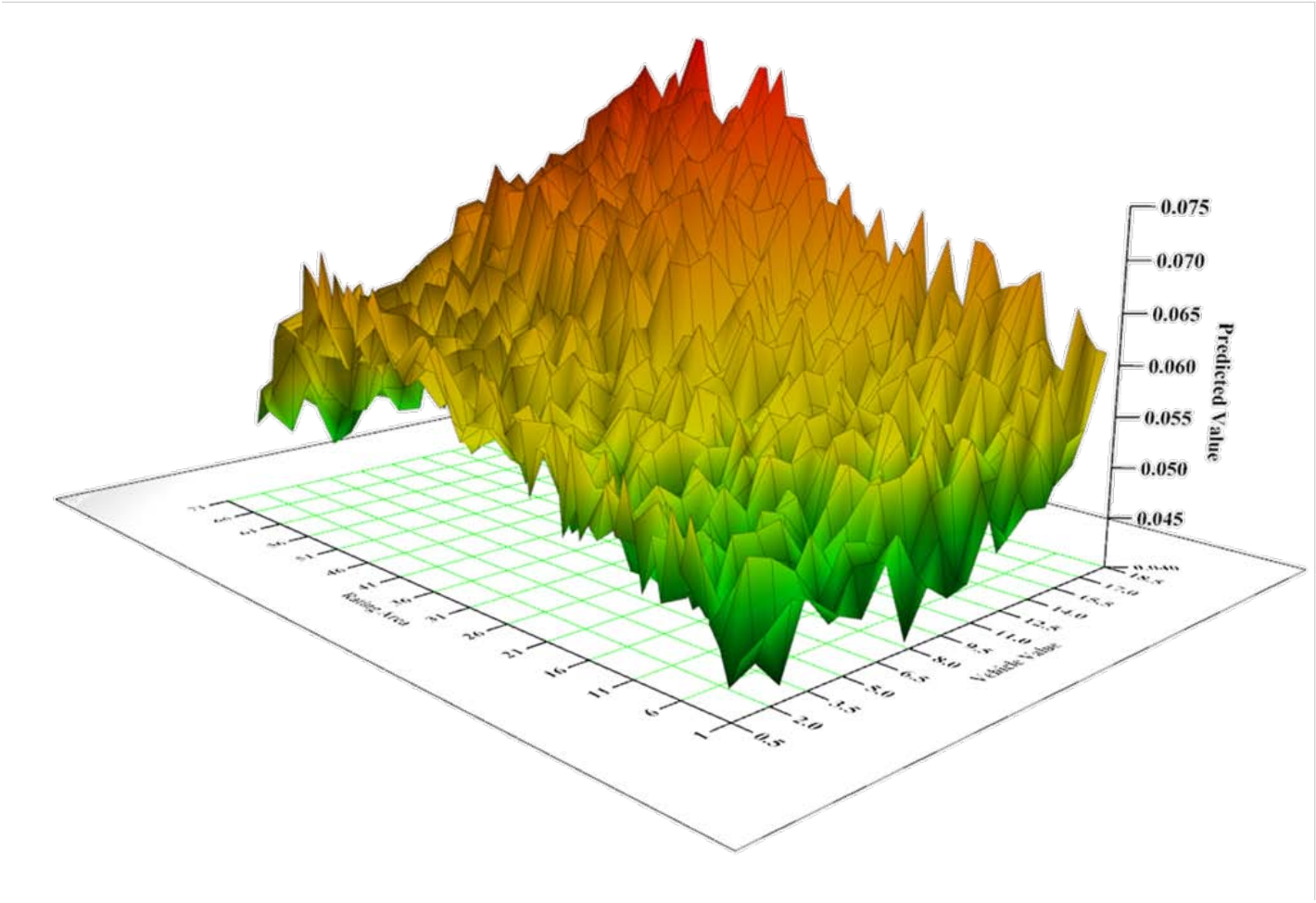
Interaction detection within GLMs - Saddles



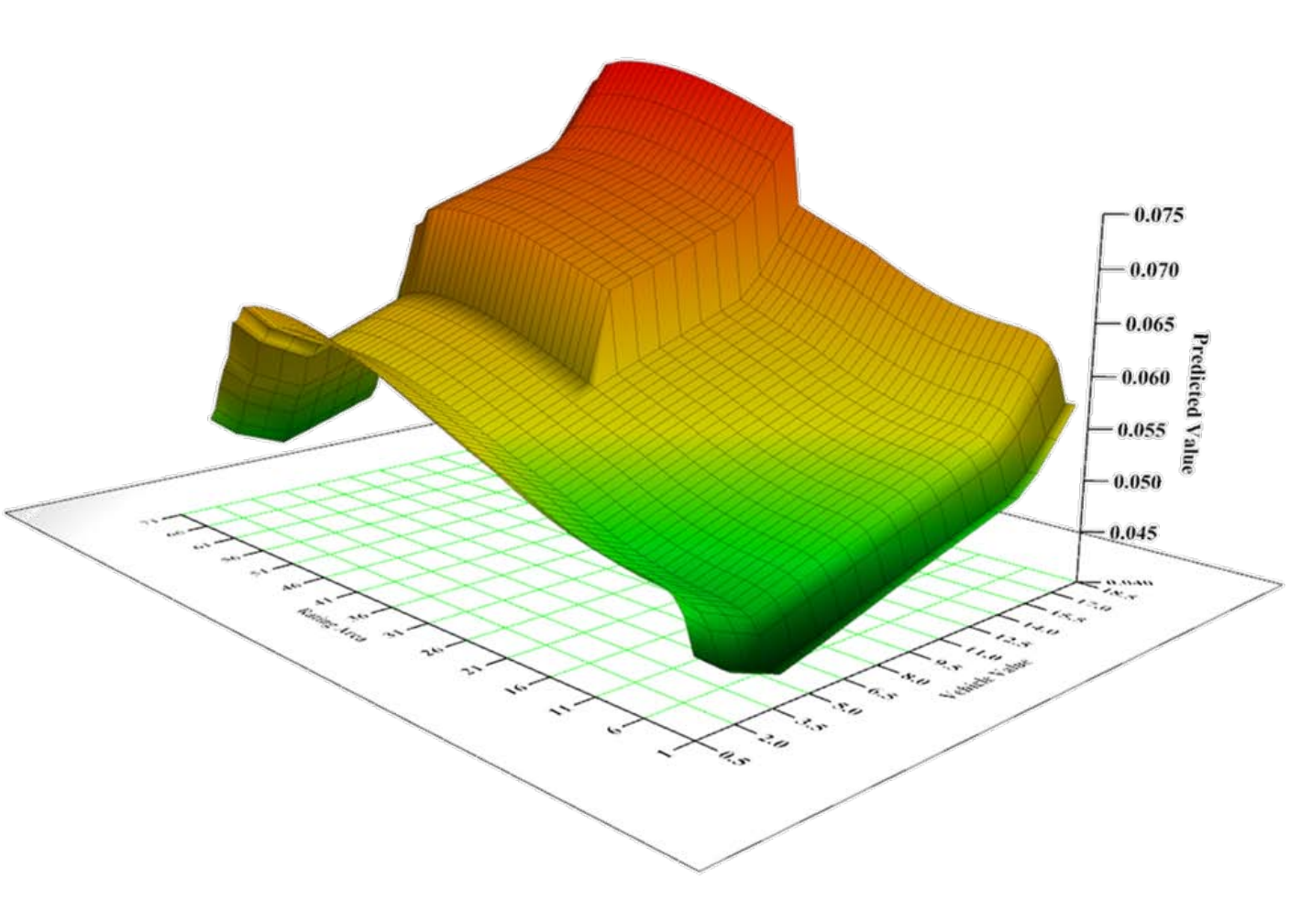
Example



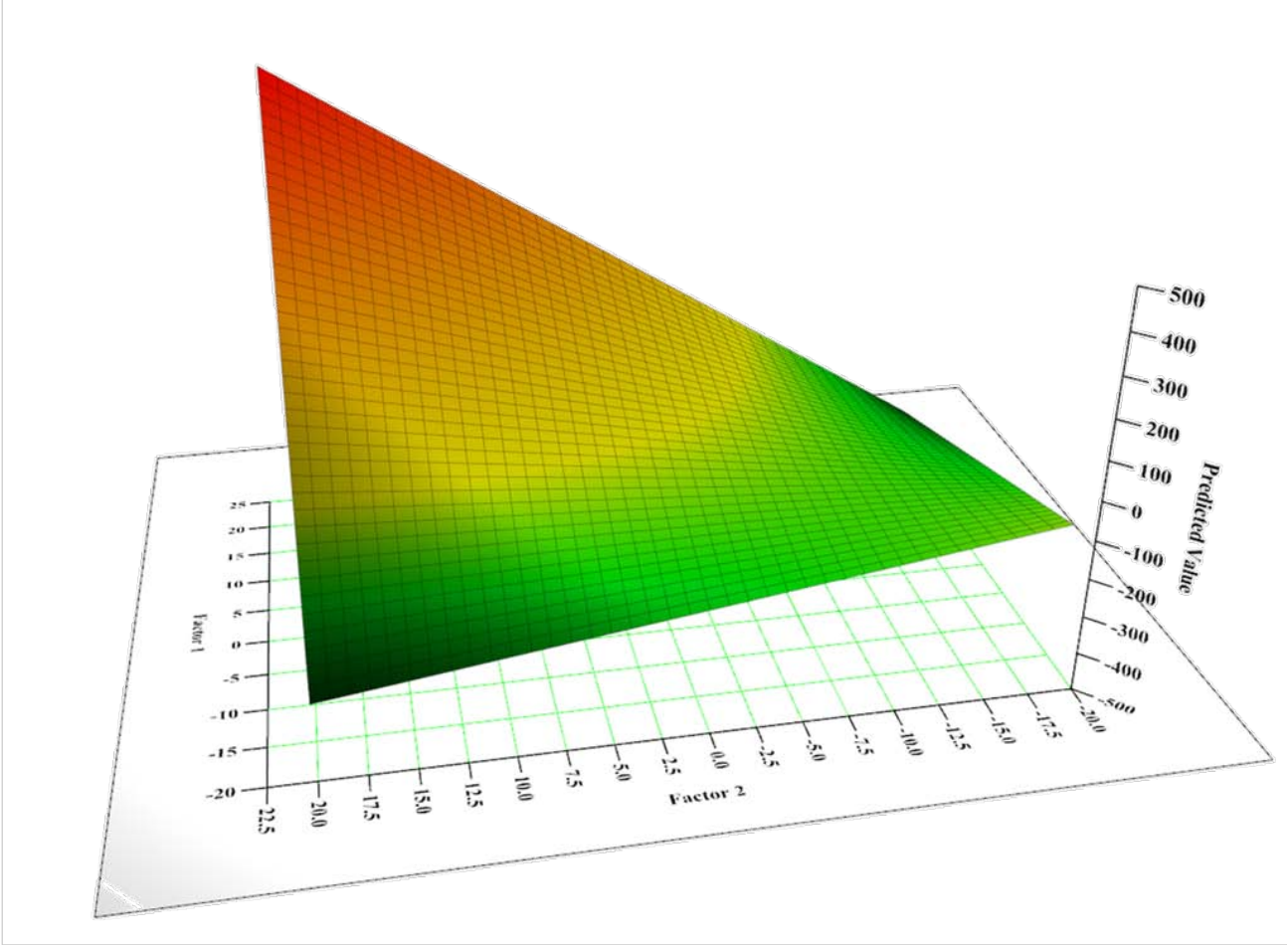
Example



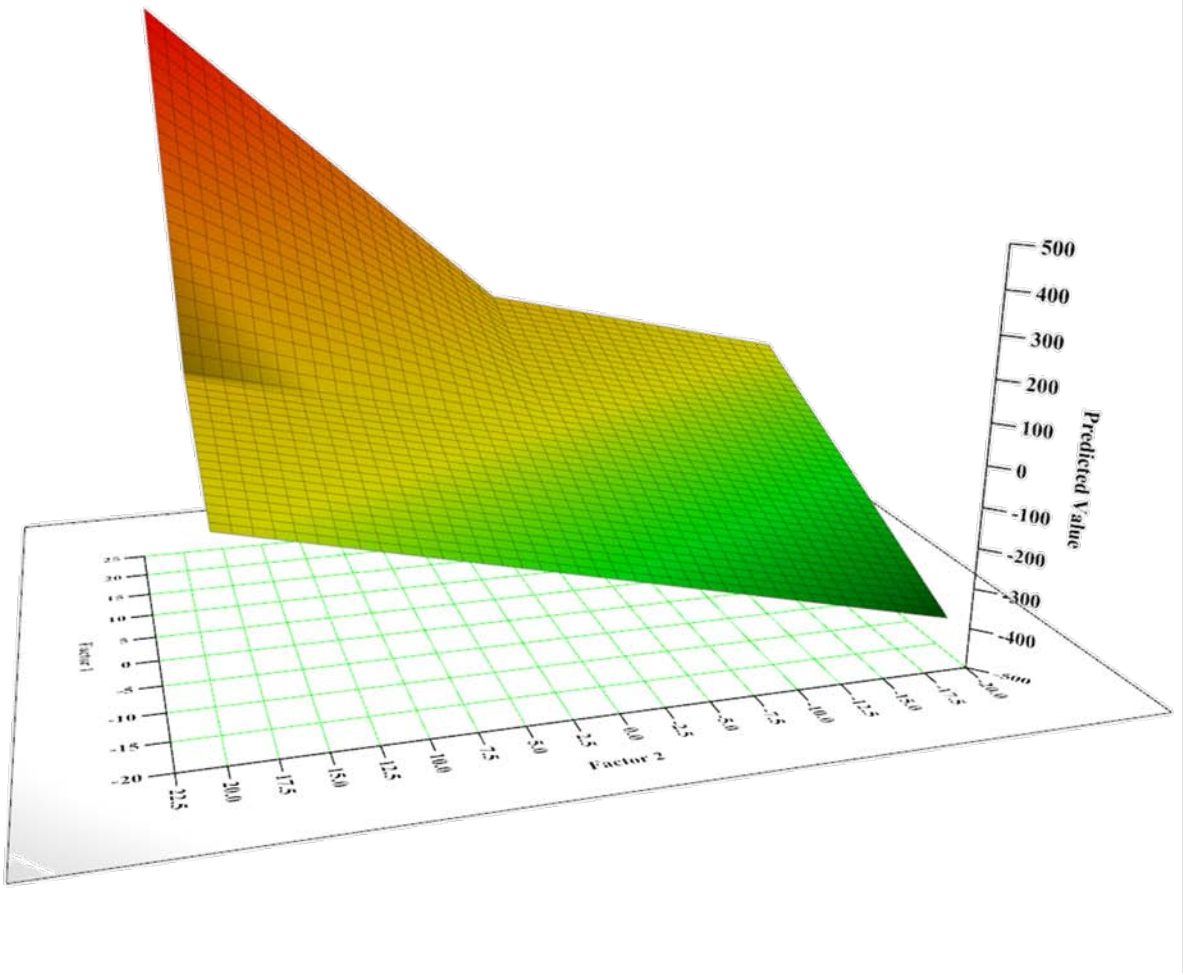
Example



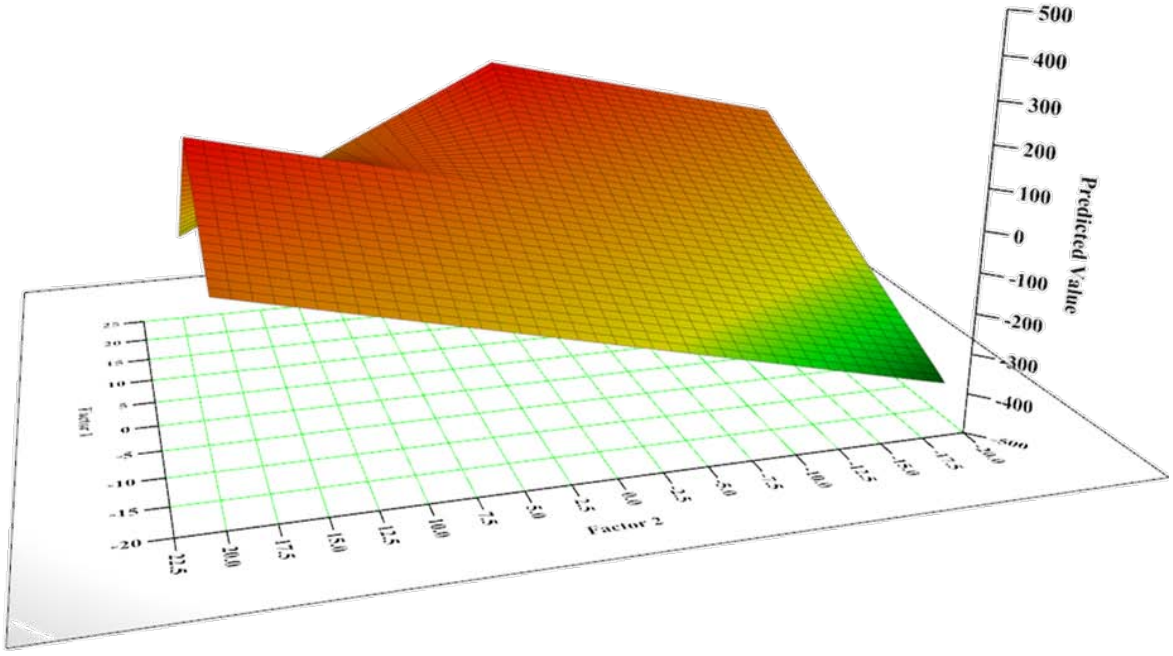
Saddles



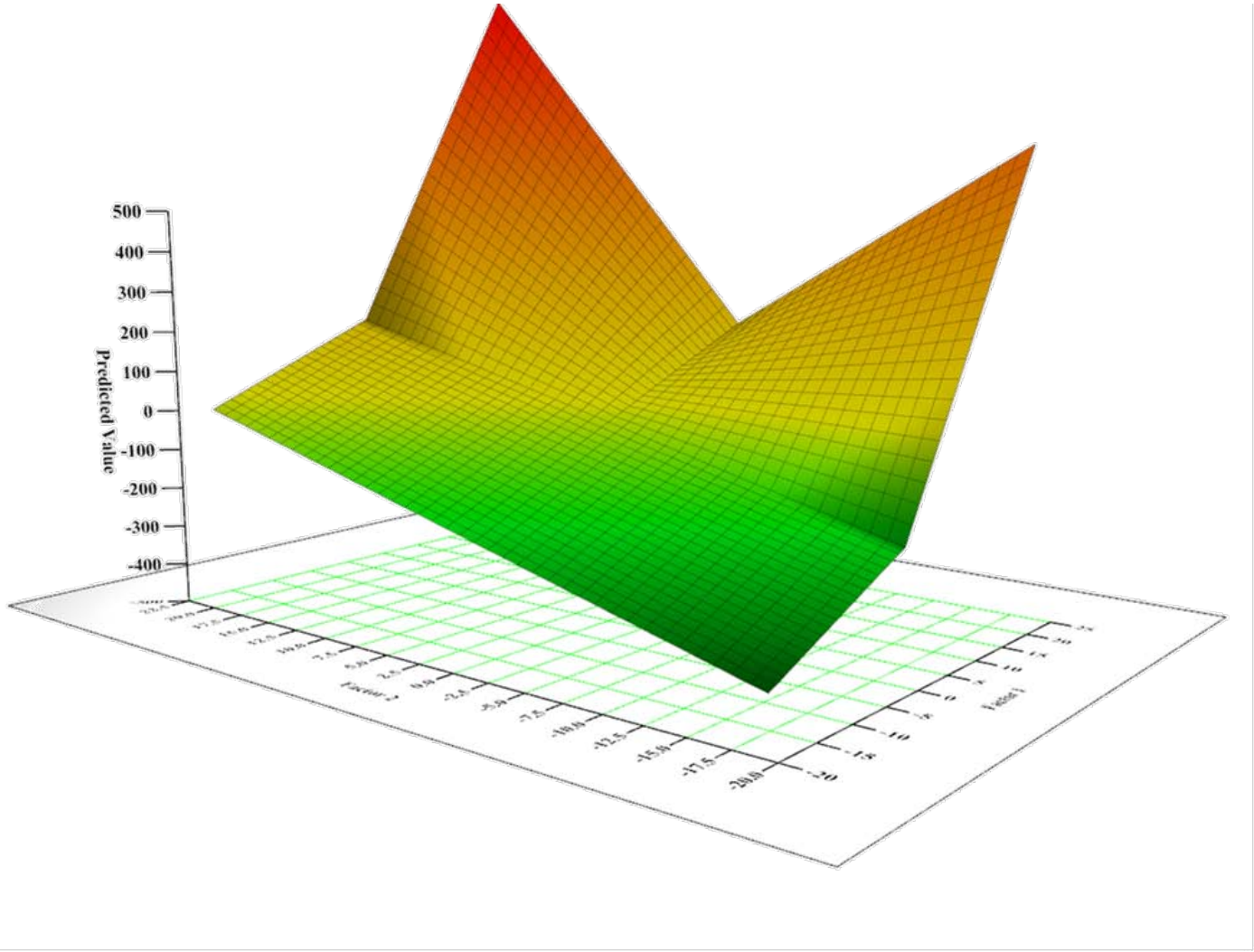
Saddles



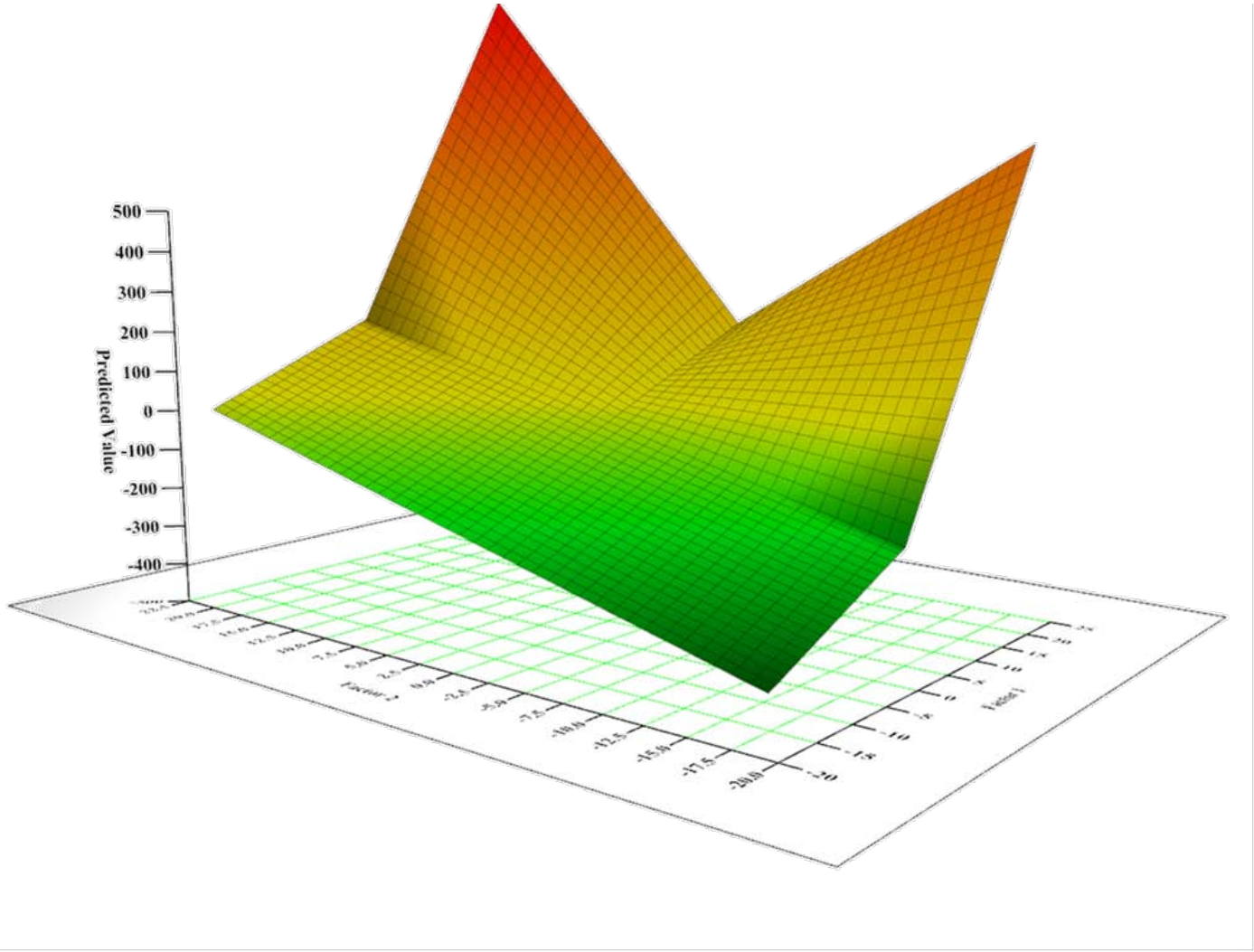
Saddles



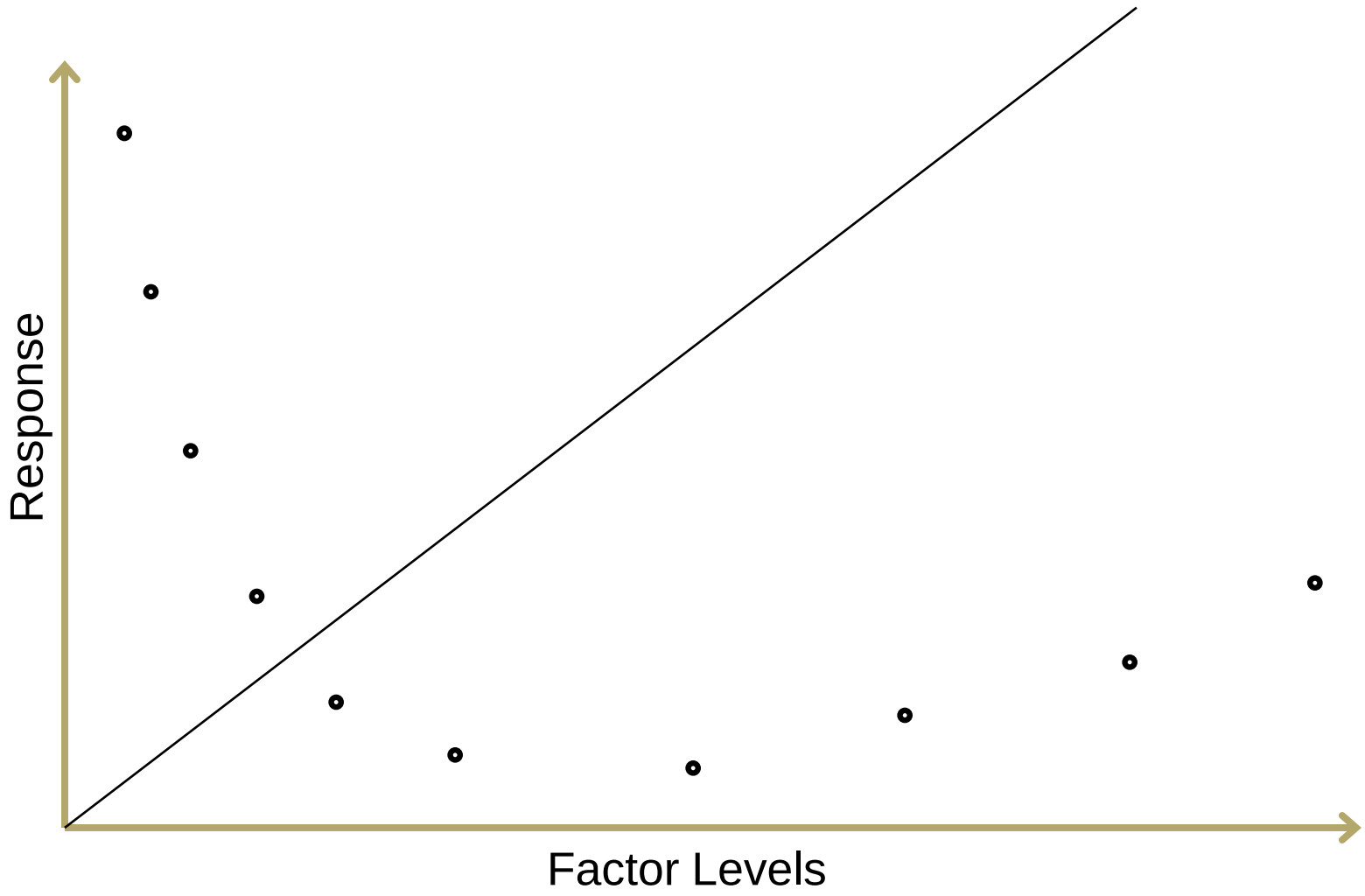
Saddles



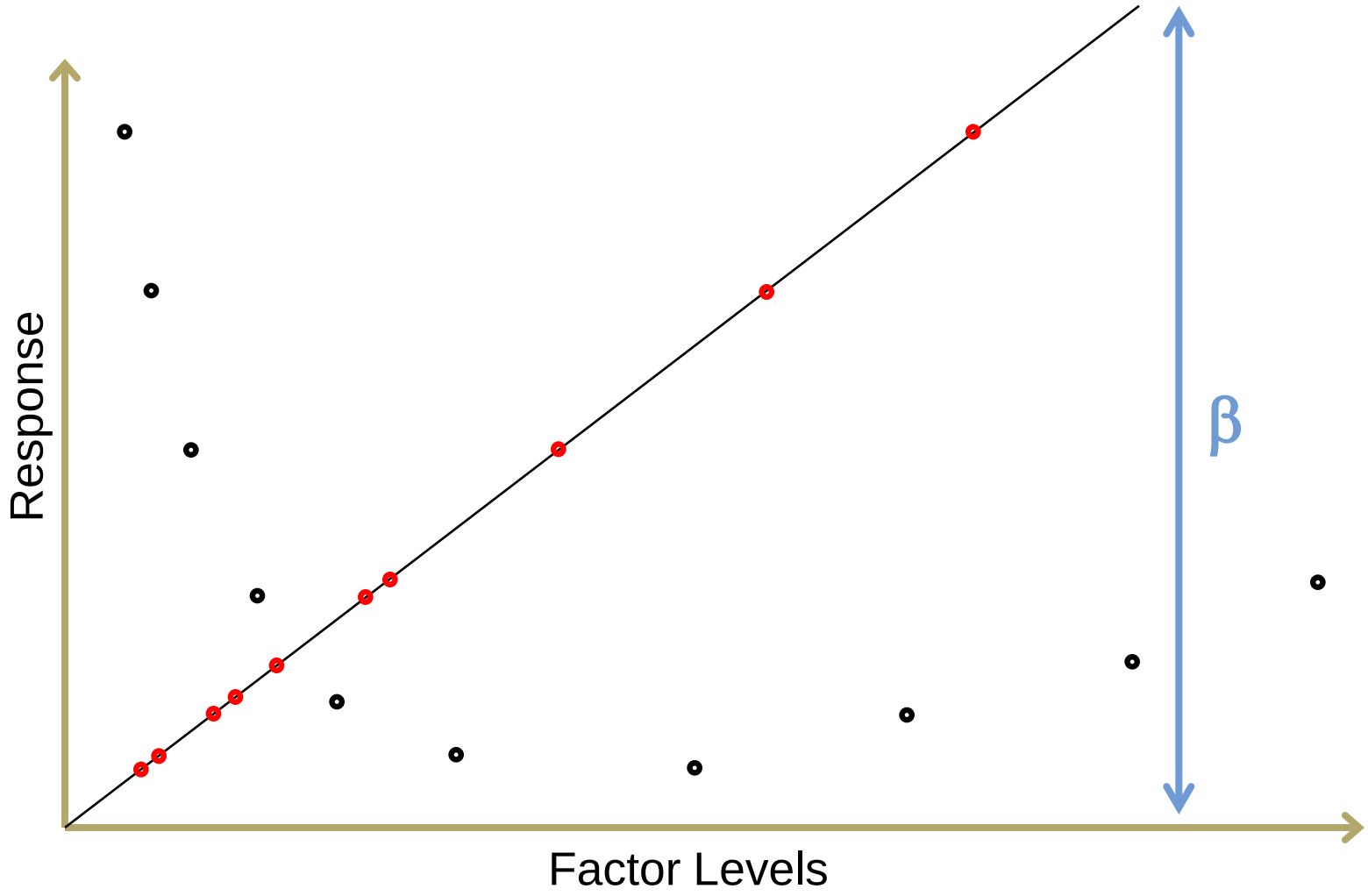
Saddles



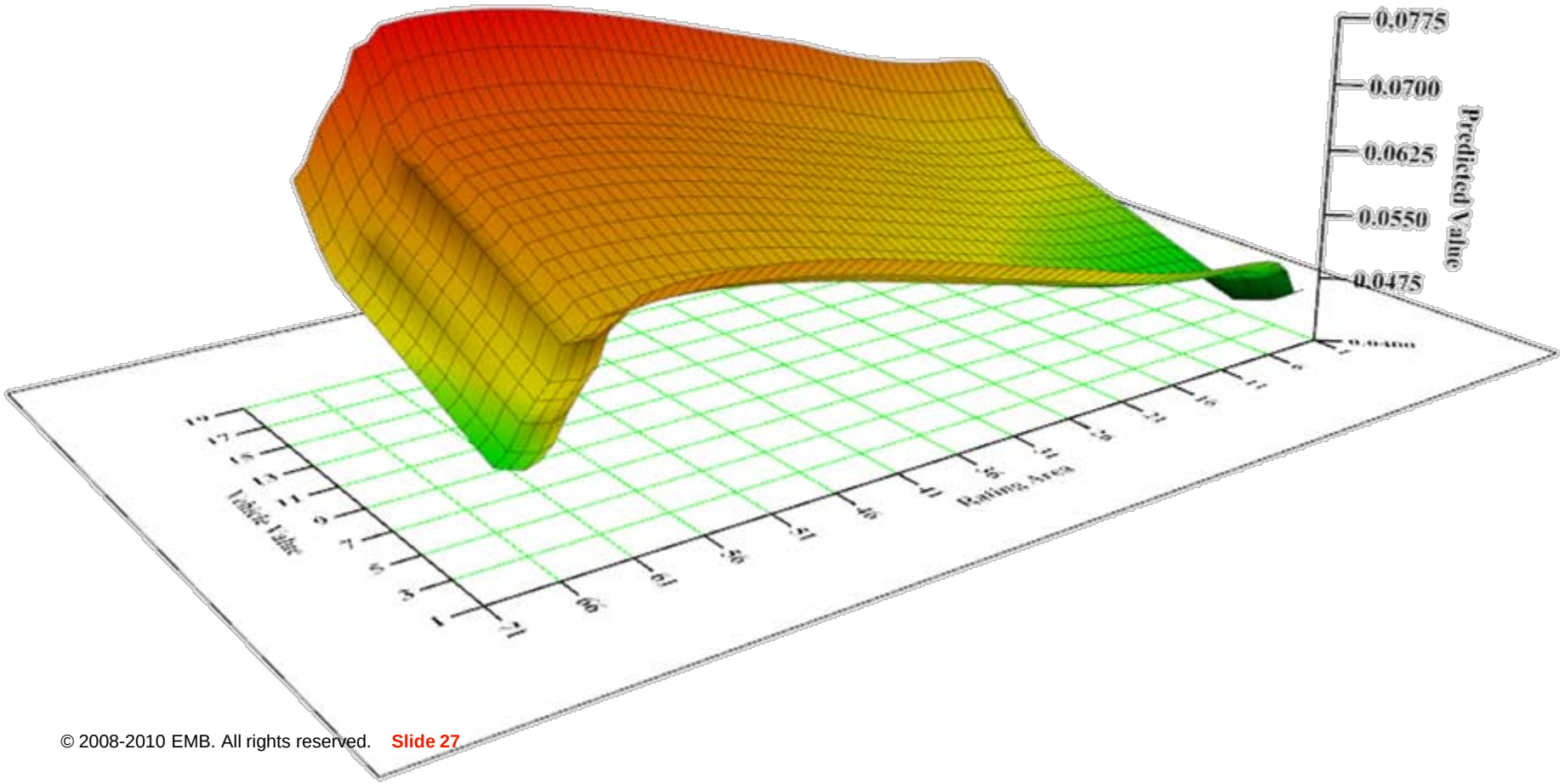
Transforming categorical and non-linear responses into single parameter variates



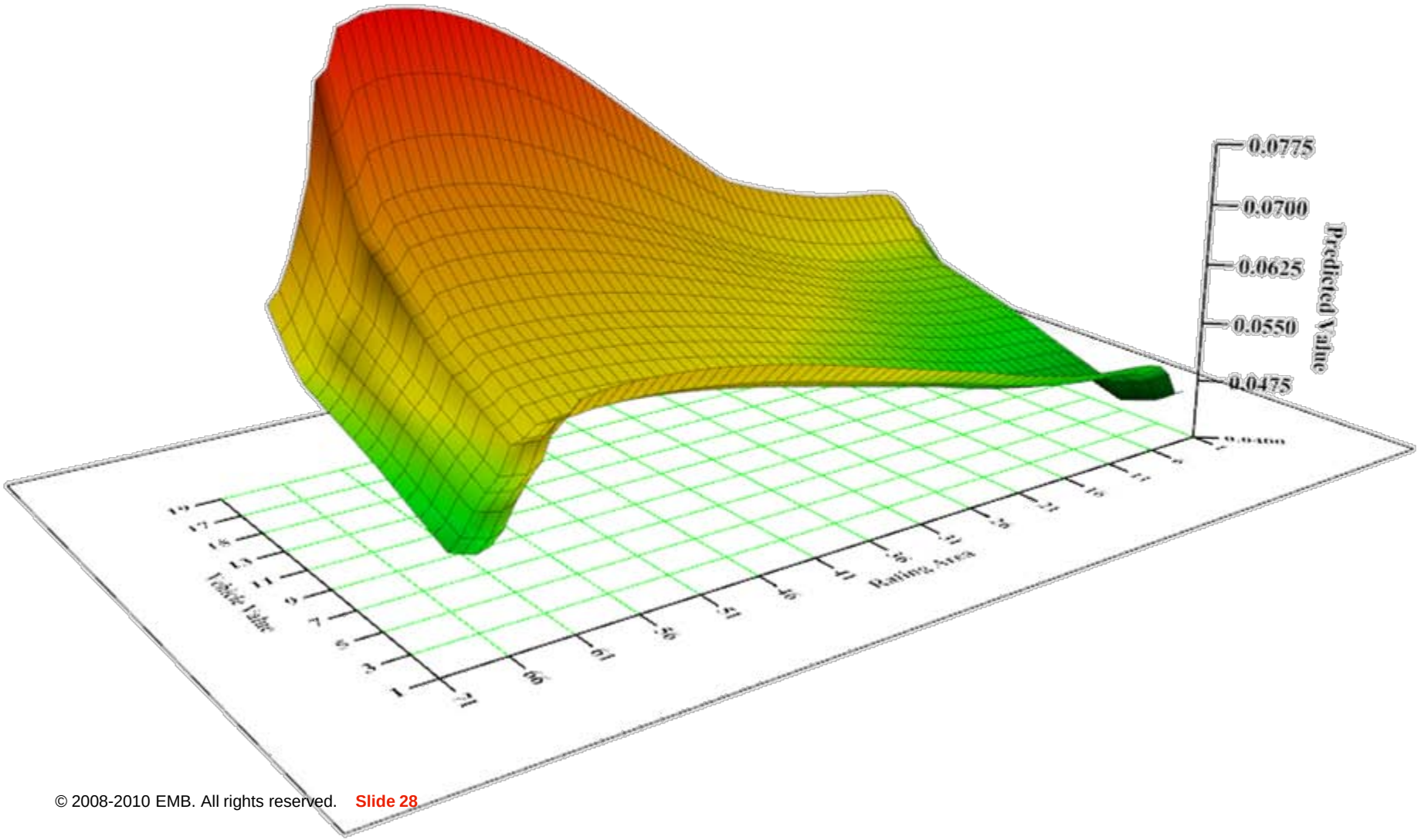
Transforming categorical and non-linear responses into single parameter variates



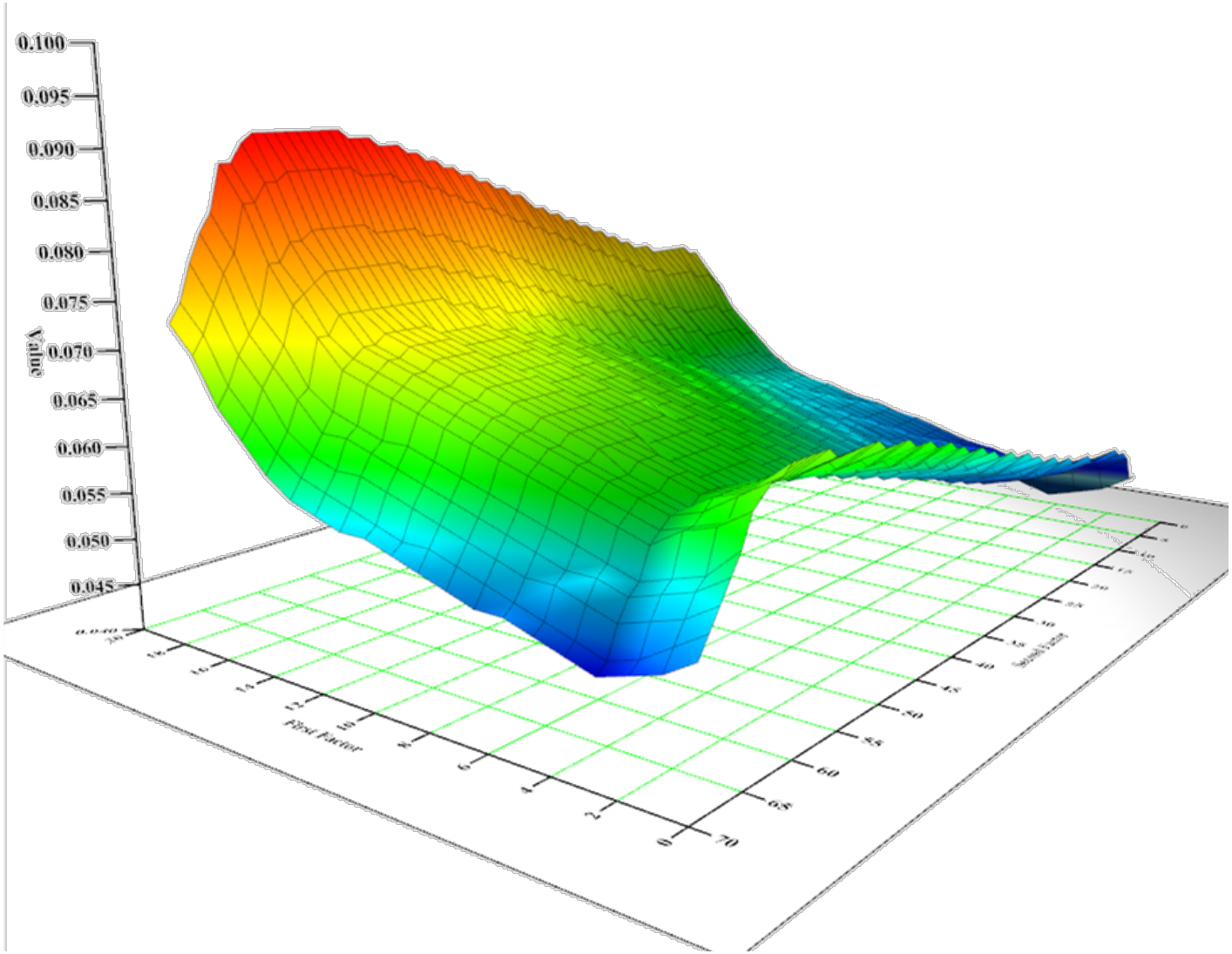
Saddles



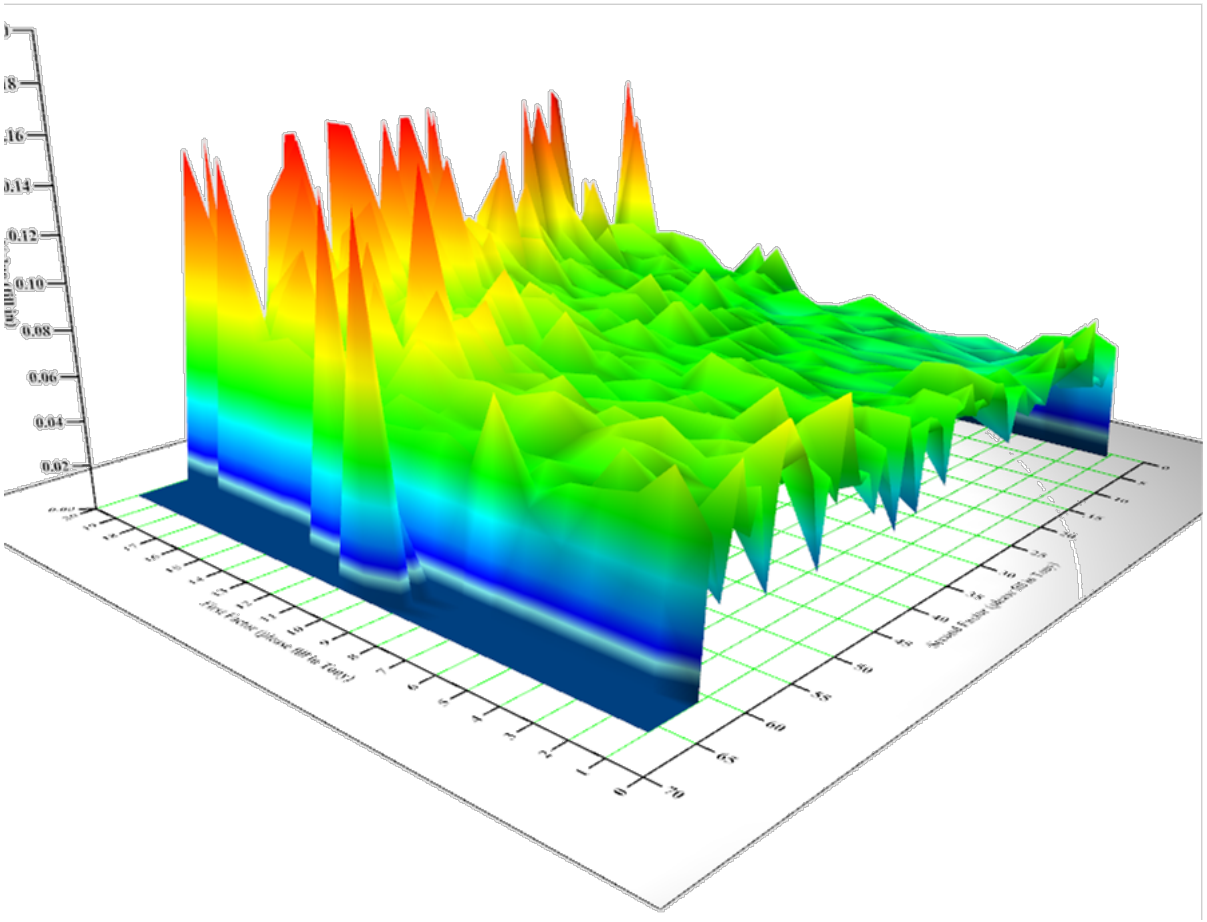
Saddles



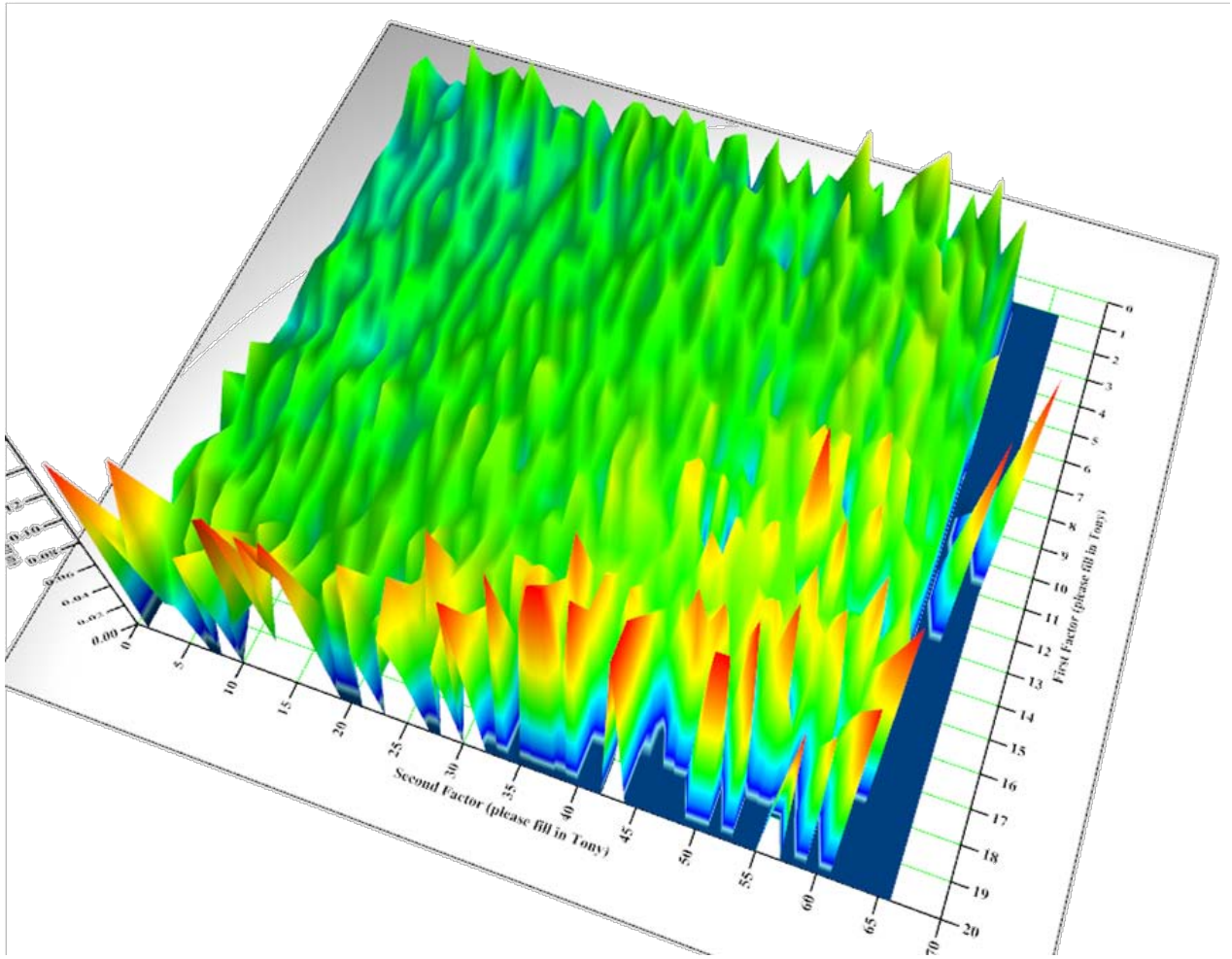
Saddles example: no interaction



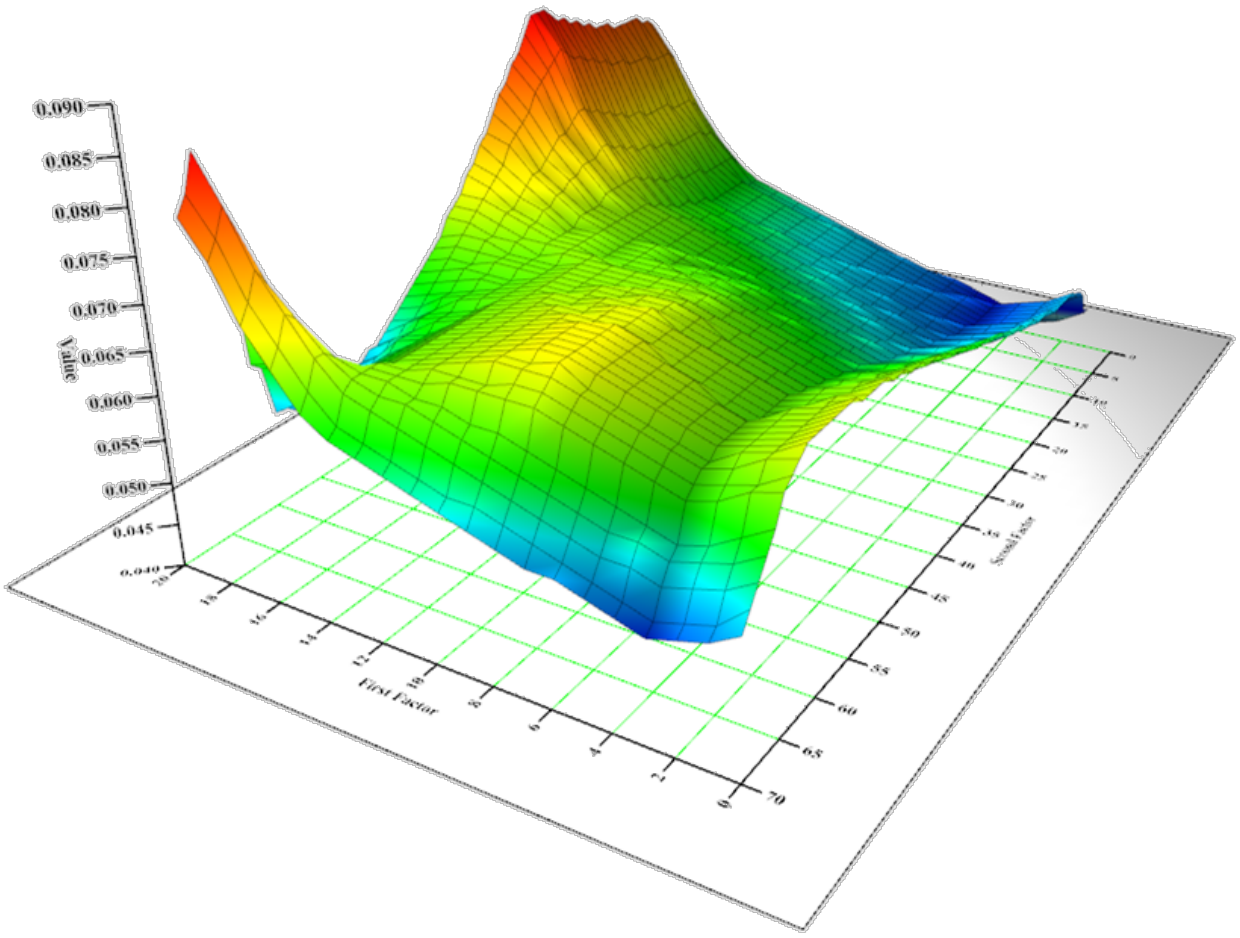
Saddles example: unsimplified interaction

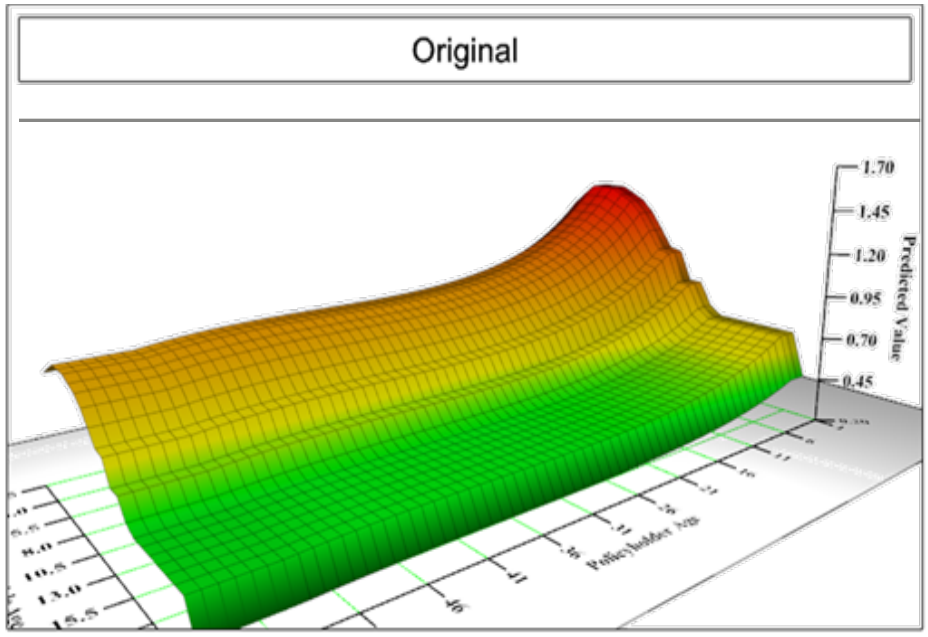


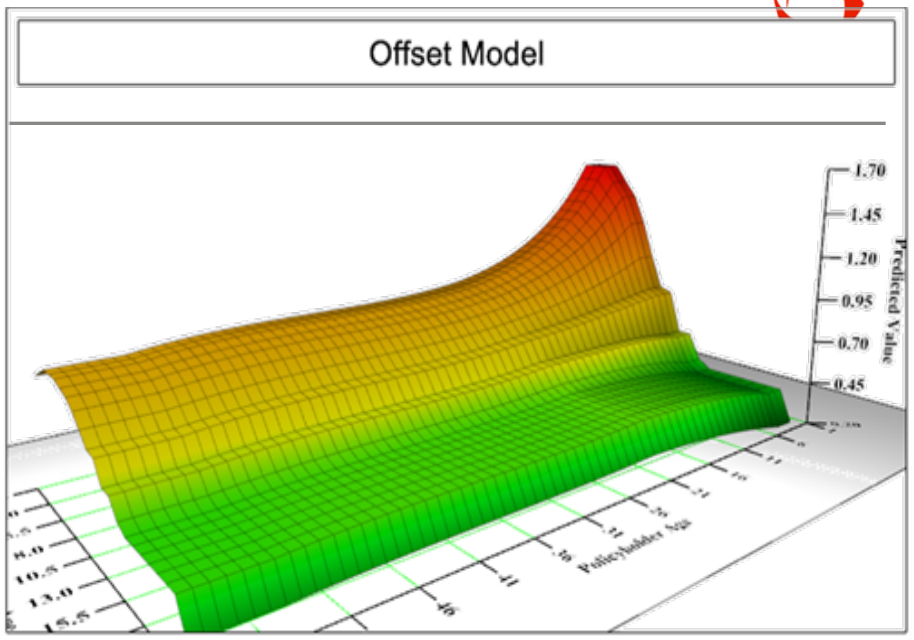
Saddles example: unsimplified interaction

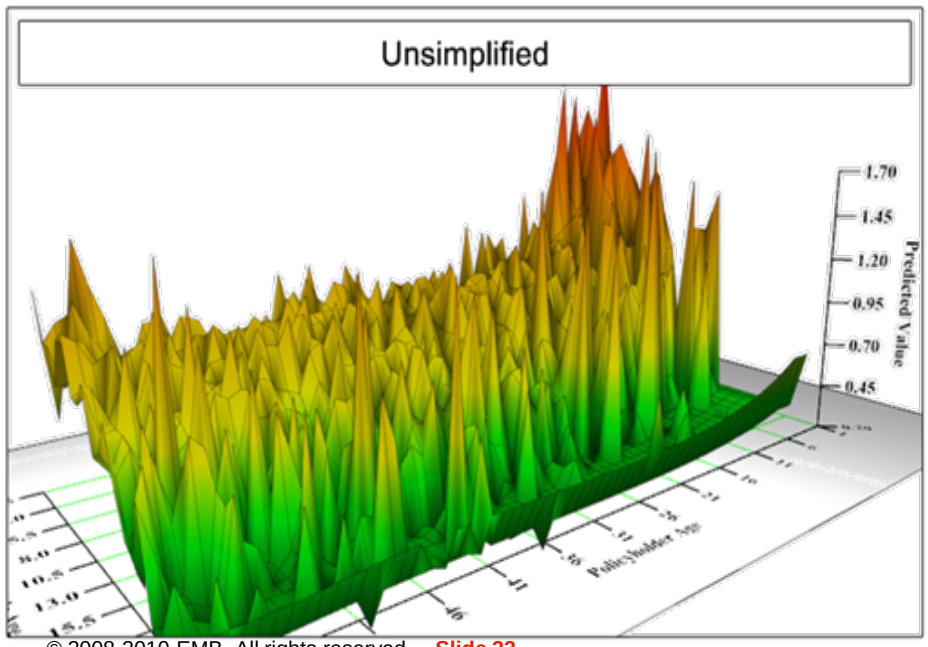


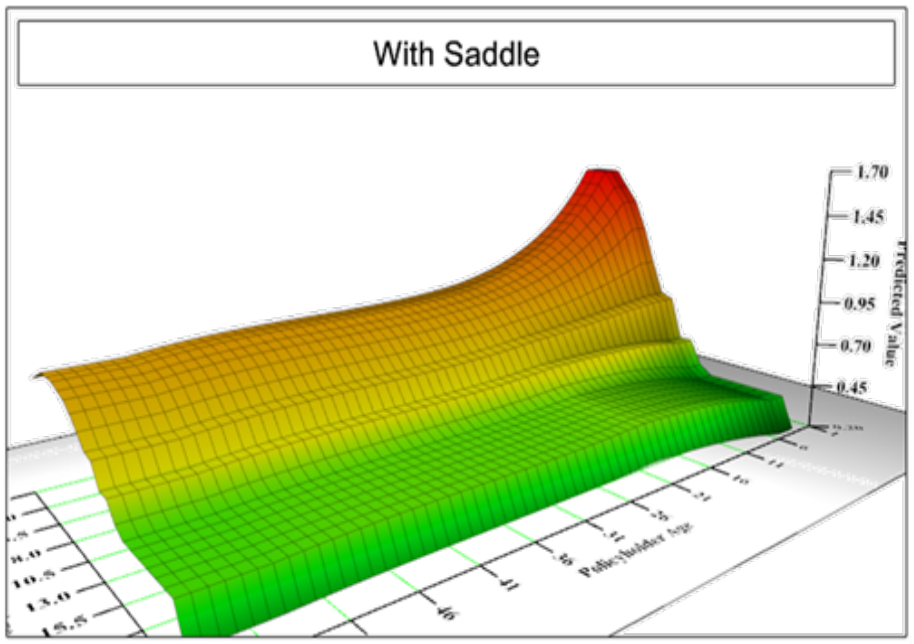
Saddles example: quadrant interaction







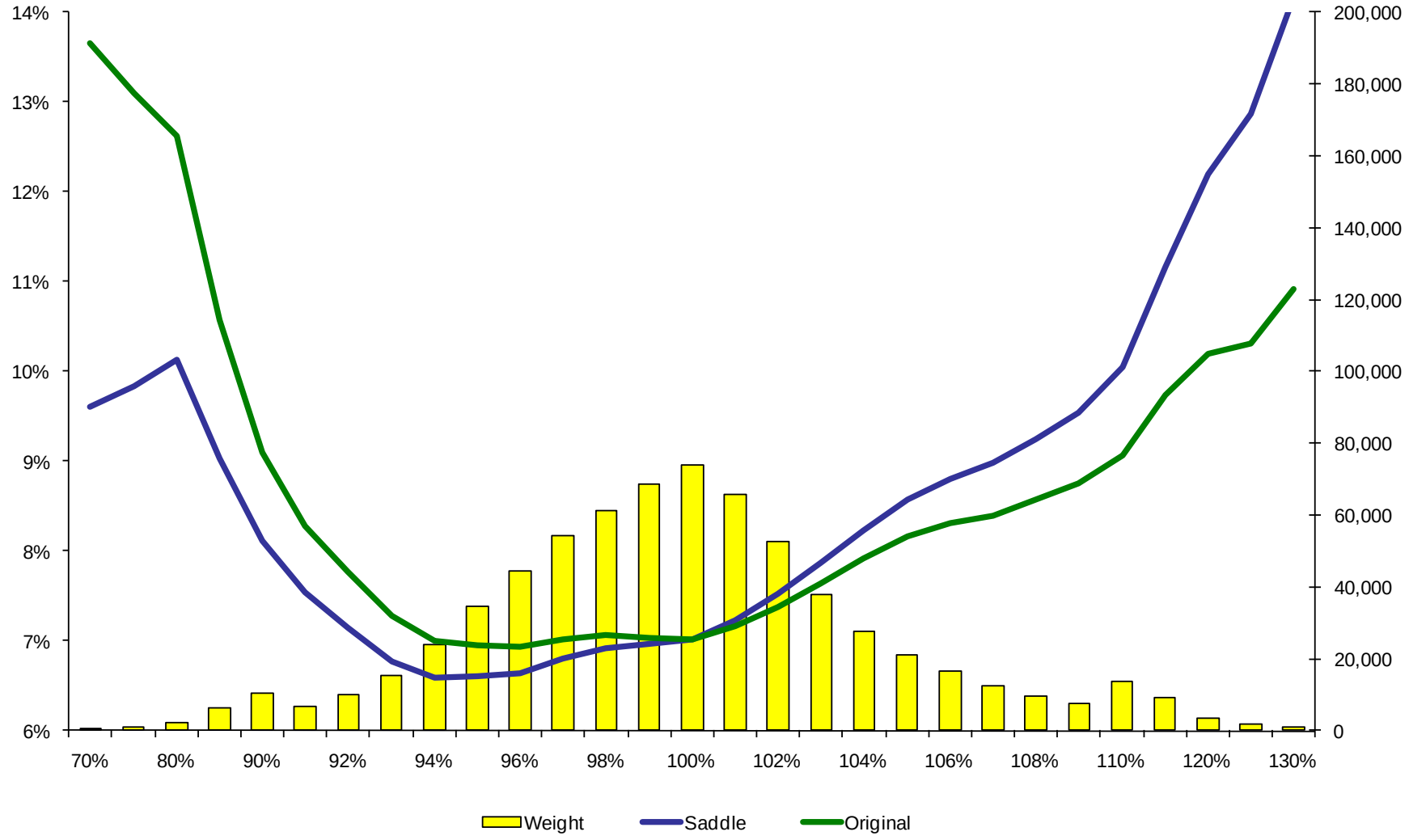




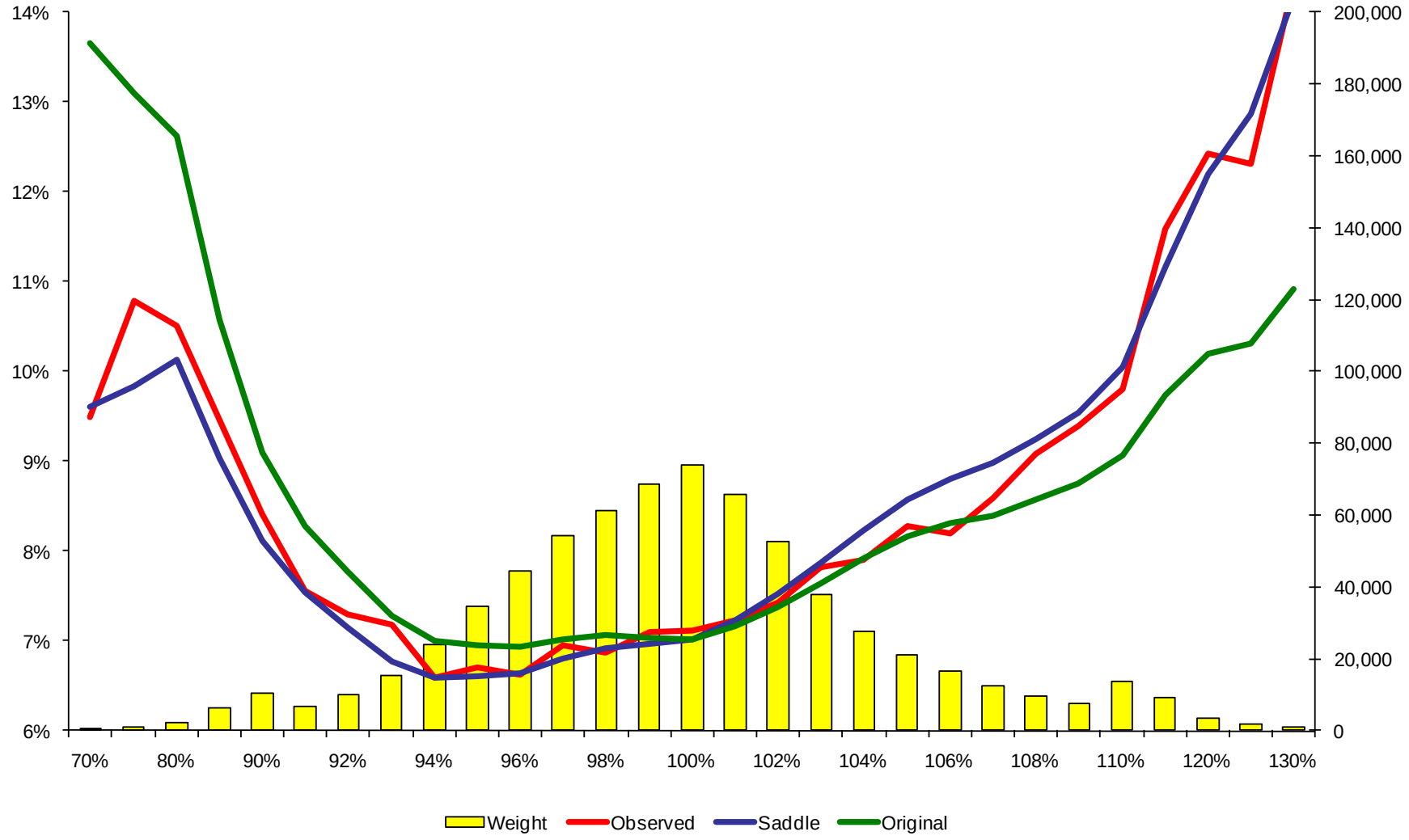
Higher dimensions

- Why stop with 2 dimensions
- Fast and parsimonious way of detecting complex signals and model corrections
- Can be used to guide GLM refinement or used in own right
- Underwriting rules

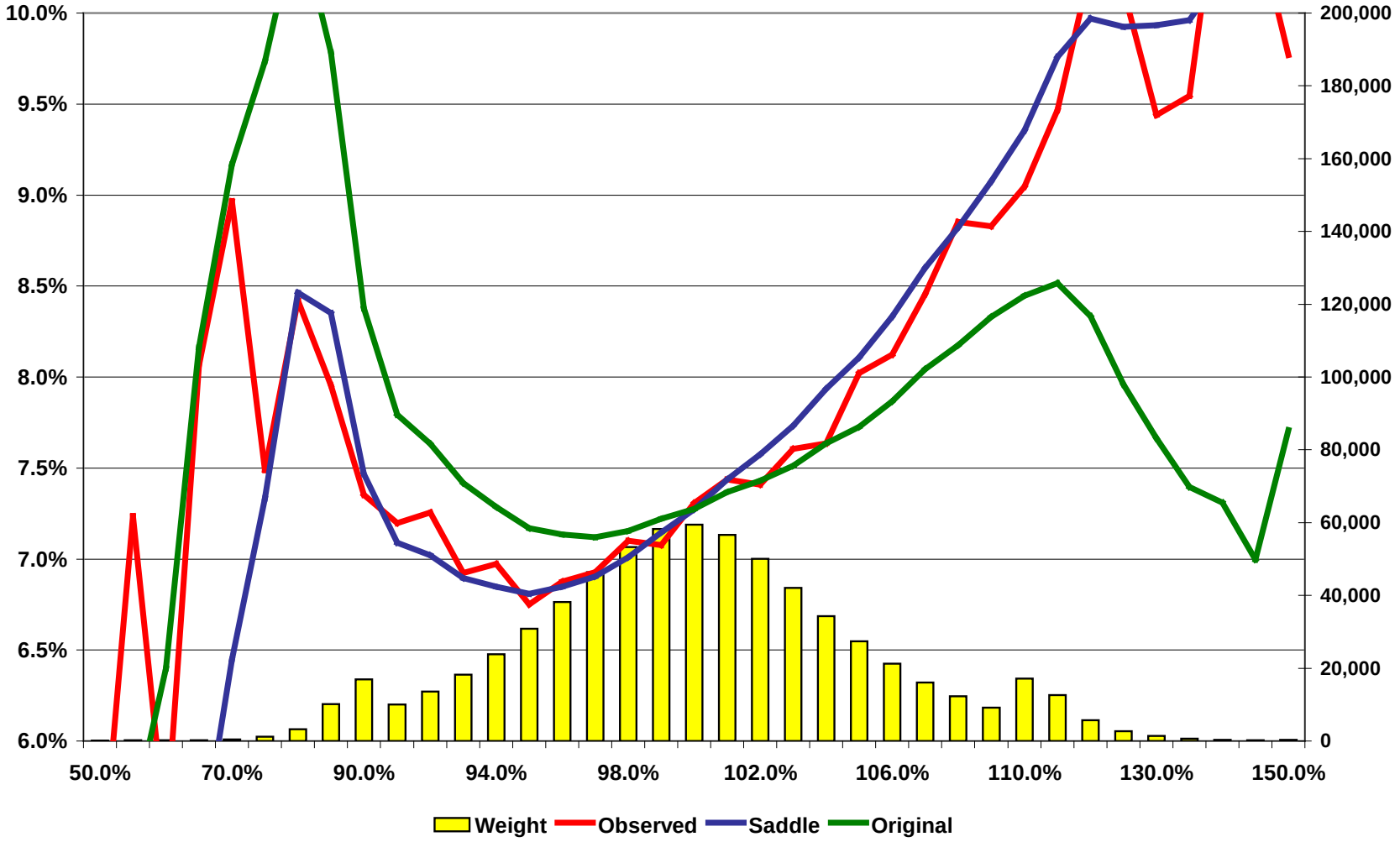
Saddles - model comparison



Saddles - model comparison



Saddles - model comparison



GLM residuals

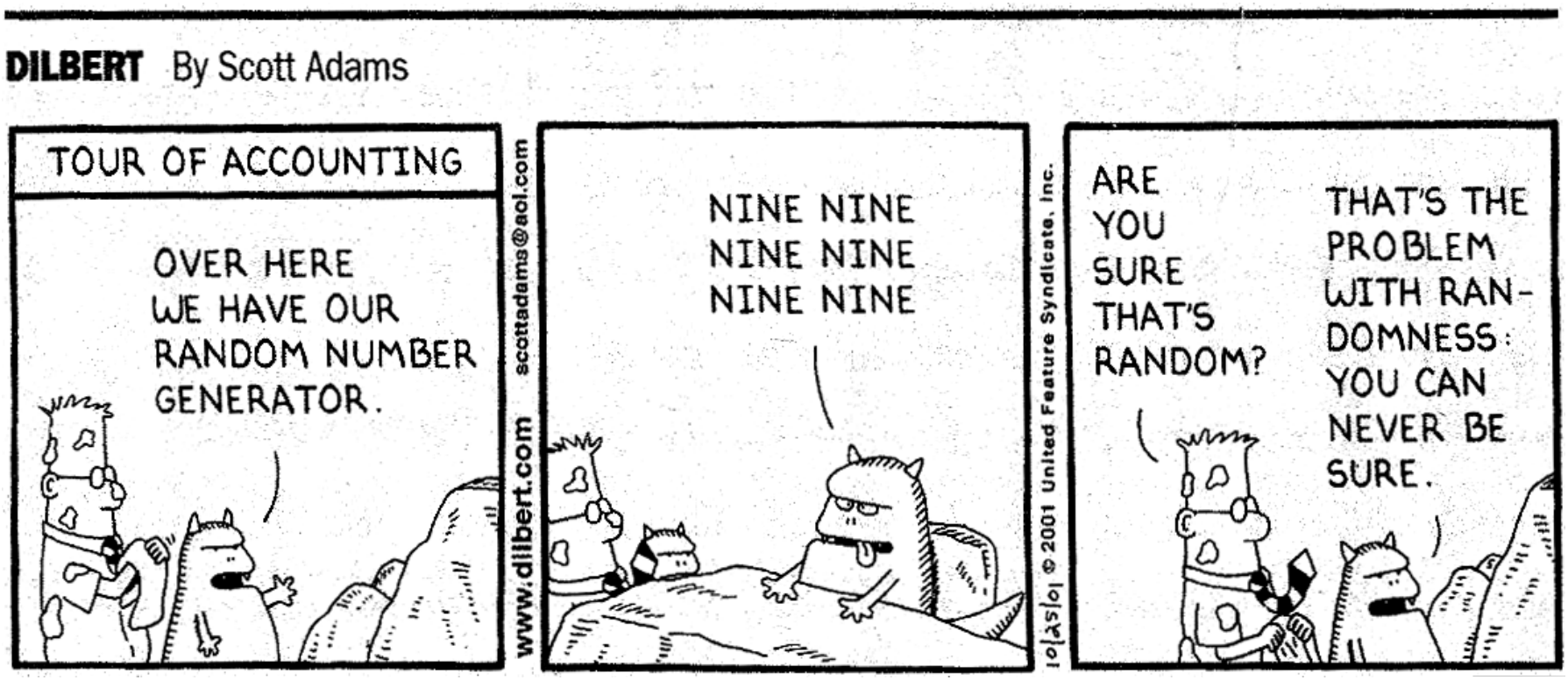
What if there is still unexplained power in the GLM residuals – and why?

- Limited list of explanatory variables
- Missing interactions
- Poor decisions in factor selection
- Other?

Mining GLM residuals

- Supervised machine learning tools can mine residuals from GLM and develop algorithms that group risks with similar residuals
- Results can form basis of a single correction factor to the GLM
- Potential disadvantages of this approach
  - Hard to distinguish signal from noise in the residual when no basis for evaluating residual
  - Prone to overfitting
  - Difficult to understand and explain effect on model, which can lead to implementation issues

Mining GLM residuals



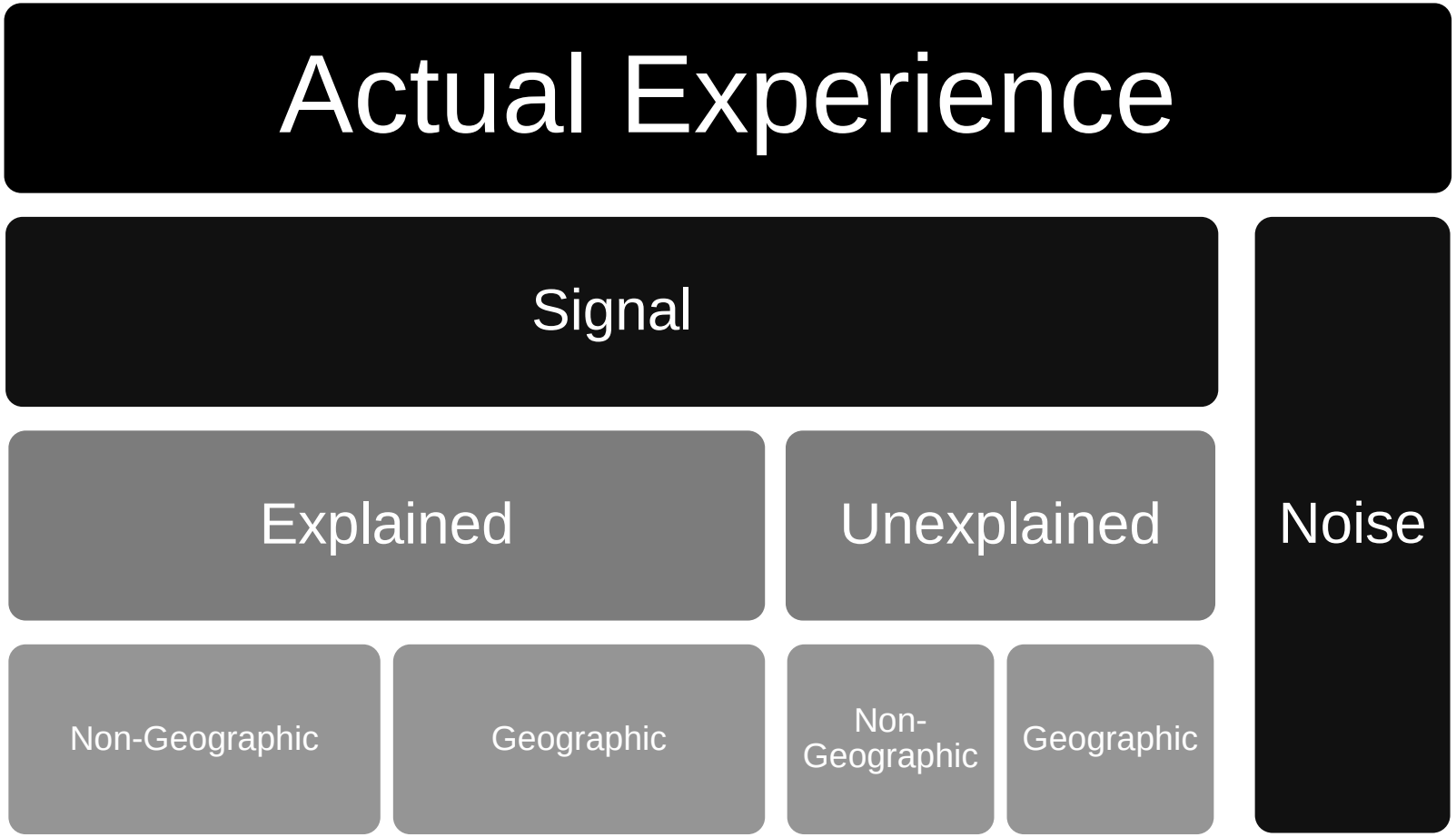
An alternative approach to mine GLM residuals

- Identify additional signal in residual that can be attributed to a particular high-dimension factor – for example,
  - Geography (zip code)
  - Vehicle (VIN)
  - Worker compensation SIC code
  - Any factor requiring a large number of small units as building blocks – and many building blocks have little or no claims experience
- EMB uses a Bayesian-based data mining method that utilizes the signal in the residuals to “correct” the GLM results for that high-dimension factor
- This type of focused correction factor is easier to control and understand

Mining GLM residuals in controlled manner

Geography example

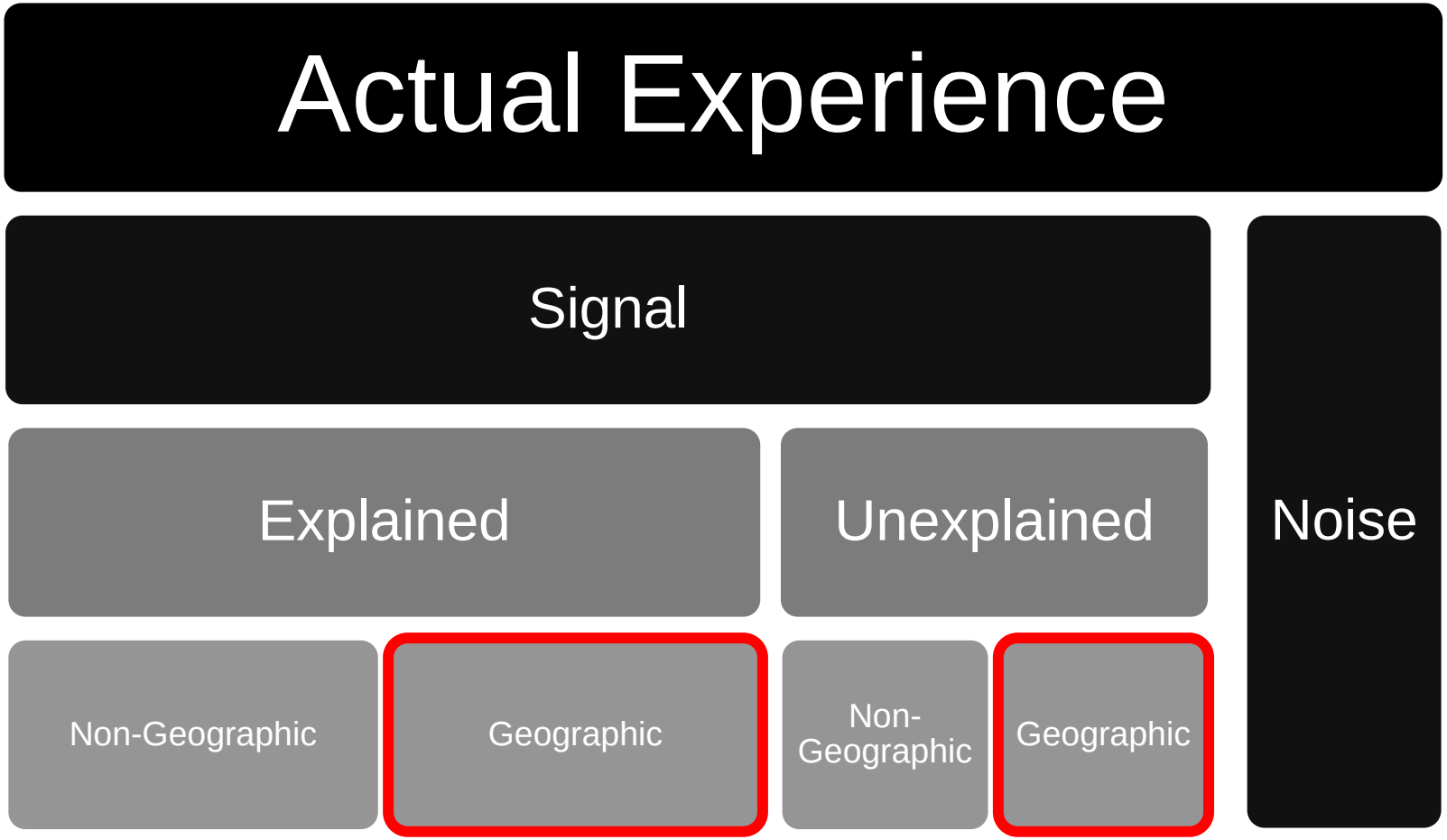
Goal is to remove the “noise” and find the “signal”



Mining GLM residuals in controlled manner

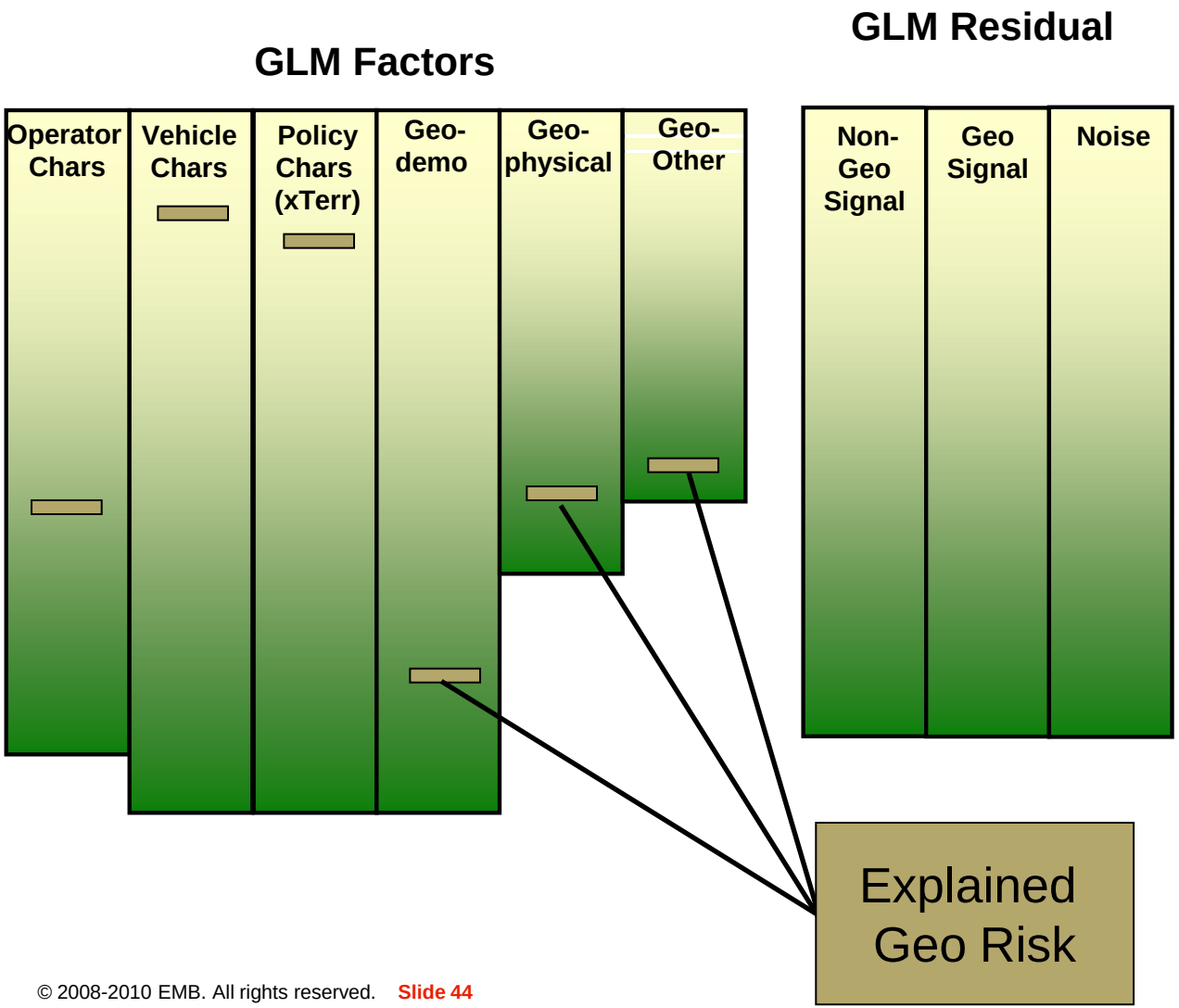
Geography example

Goal is to find the geographic signal



Mining GLM residuals in controlled manner

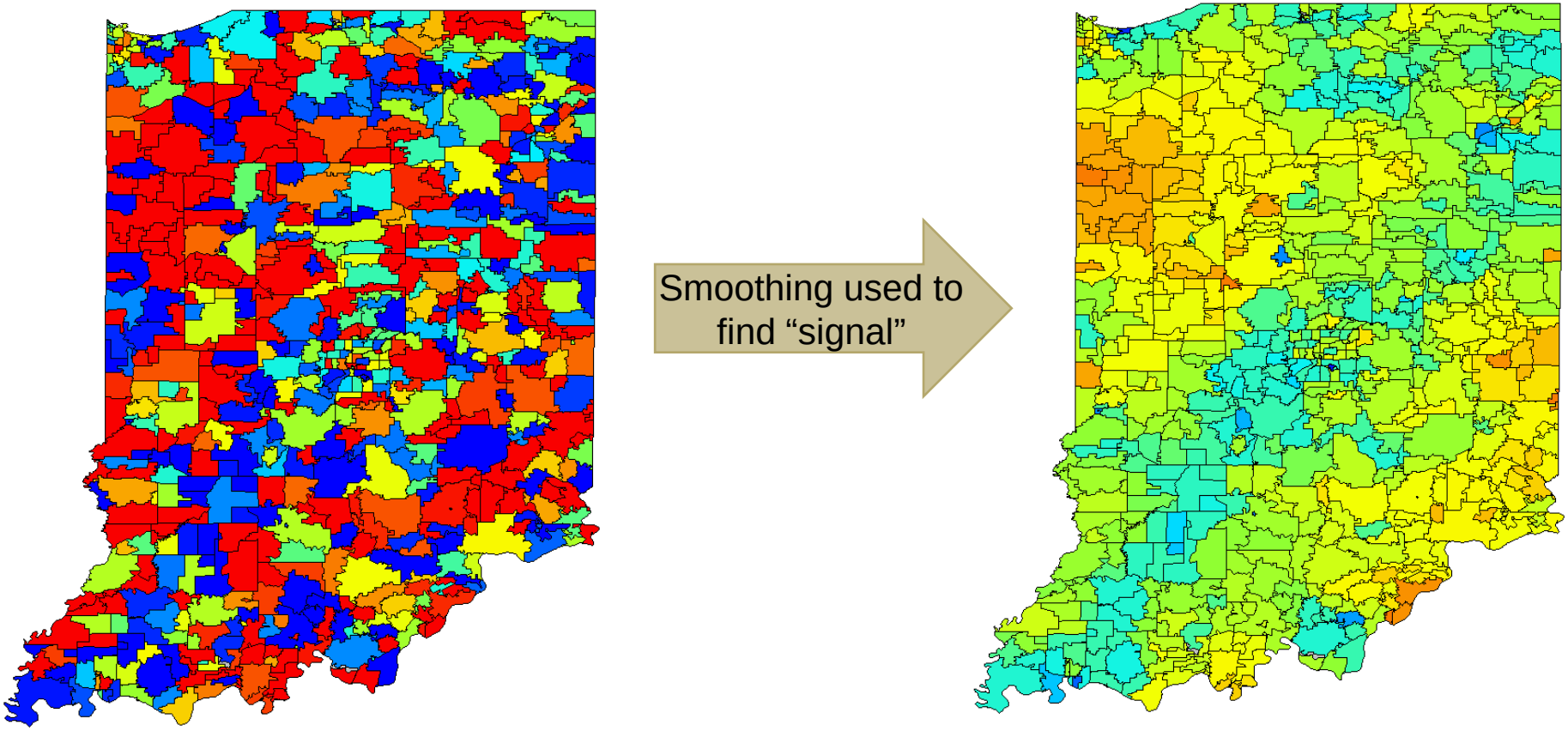
Geography example



Mining GLM residuals in controlled manner

Geography example

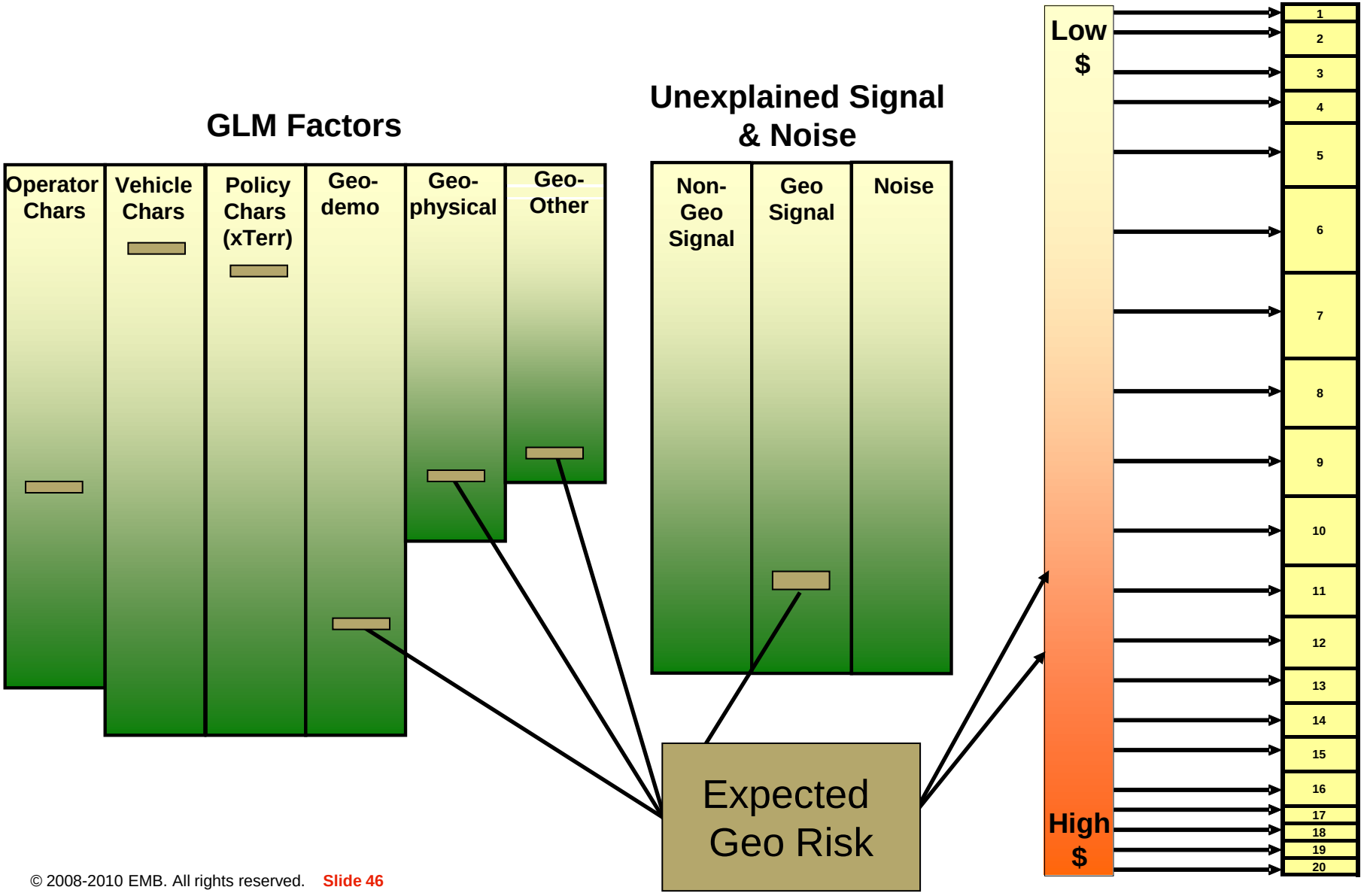
- Check the residuals to determine if there is any unexplained systematic effect



- “Correction factors” applied to geo estimates to determine best estimate

Mining GLM residuals in controlled manner

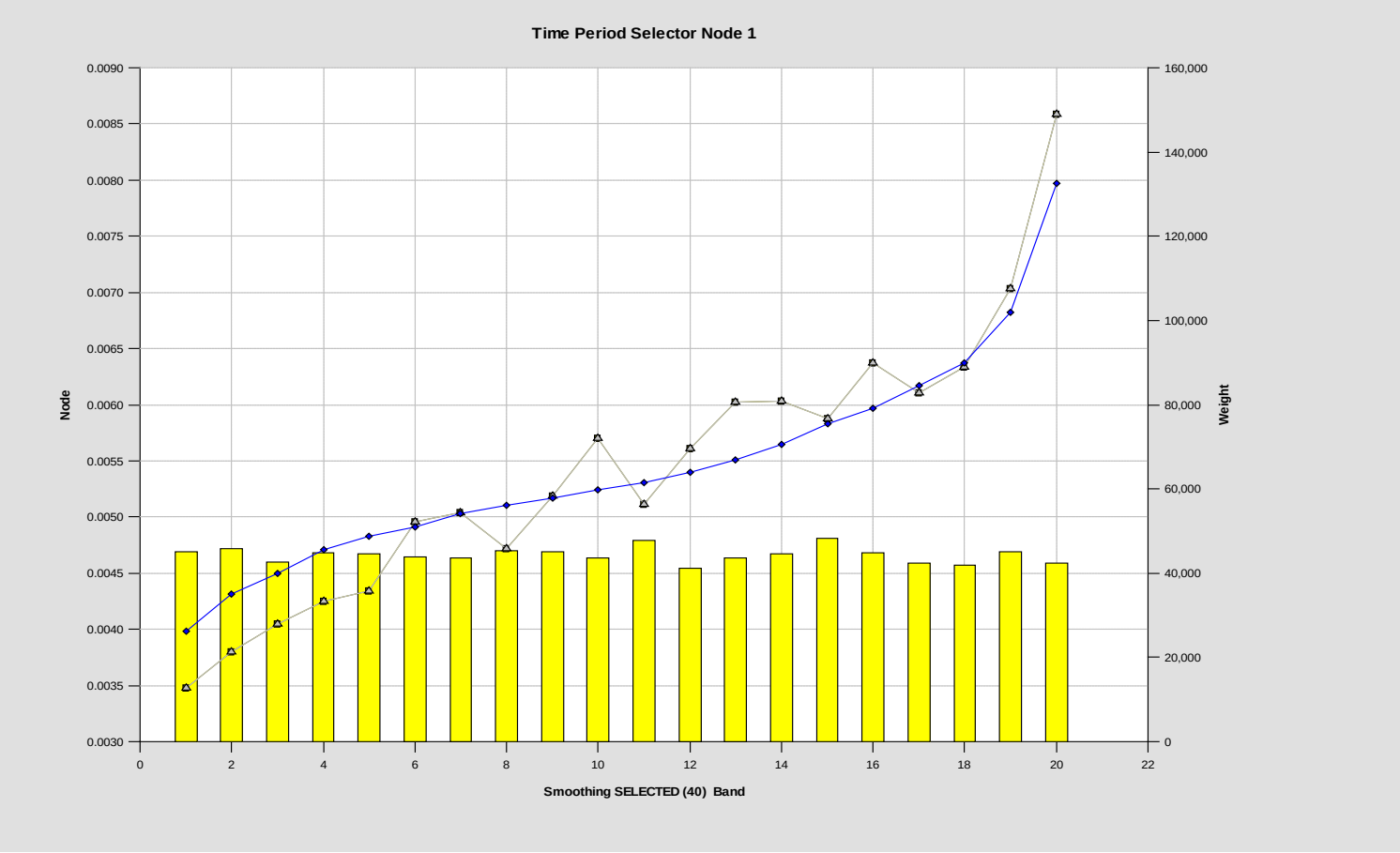
Geography example



Mining GLM residuals in controlled manner

Geography example

Assess whether new territorial groupings follow observed data well (ideally on hold-out data)



Mining GLM residuals in controlled manner

Vehicle example

Standardized fitted value from GLM

Spatially corrected standardized fitted value

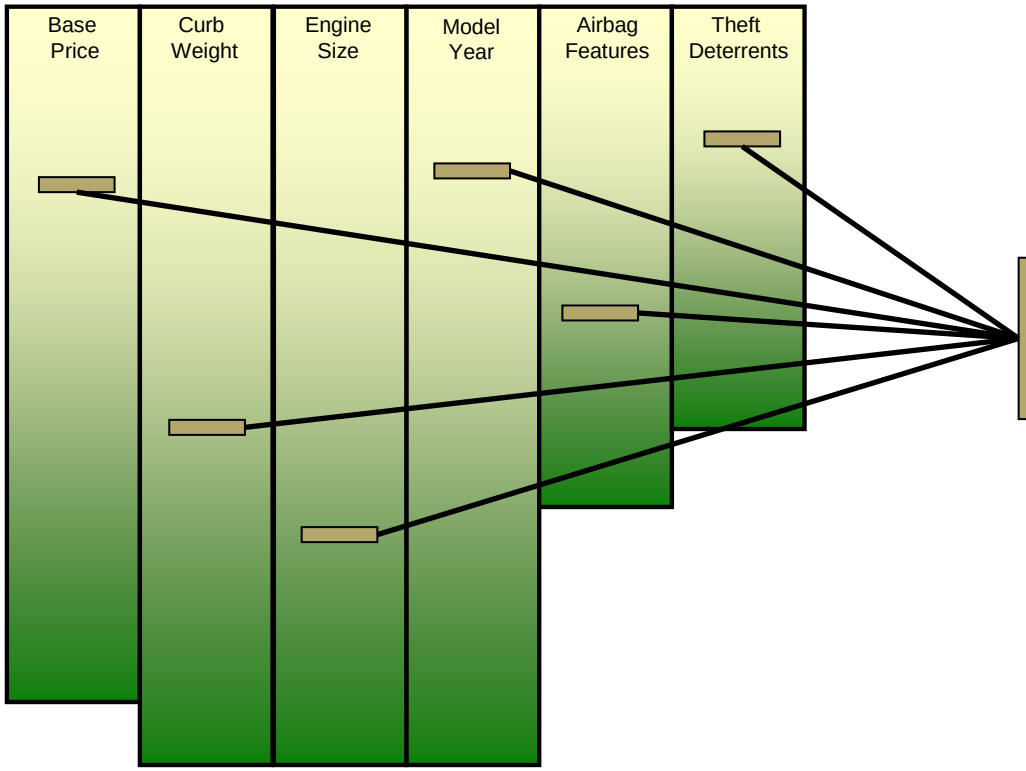
Clustered to form symbols

Relativities calculated for each symbol

Mining GLM residuals in controlled manner

Vehicle example

Initial Estimator:



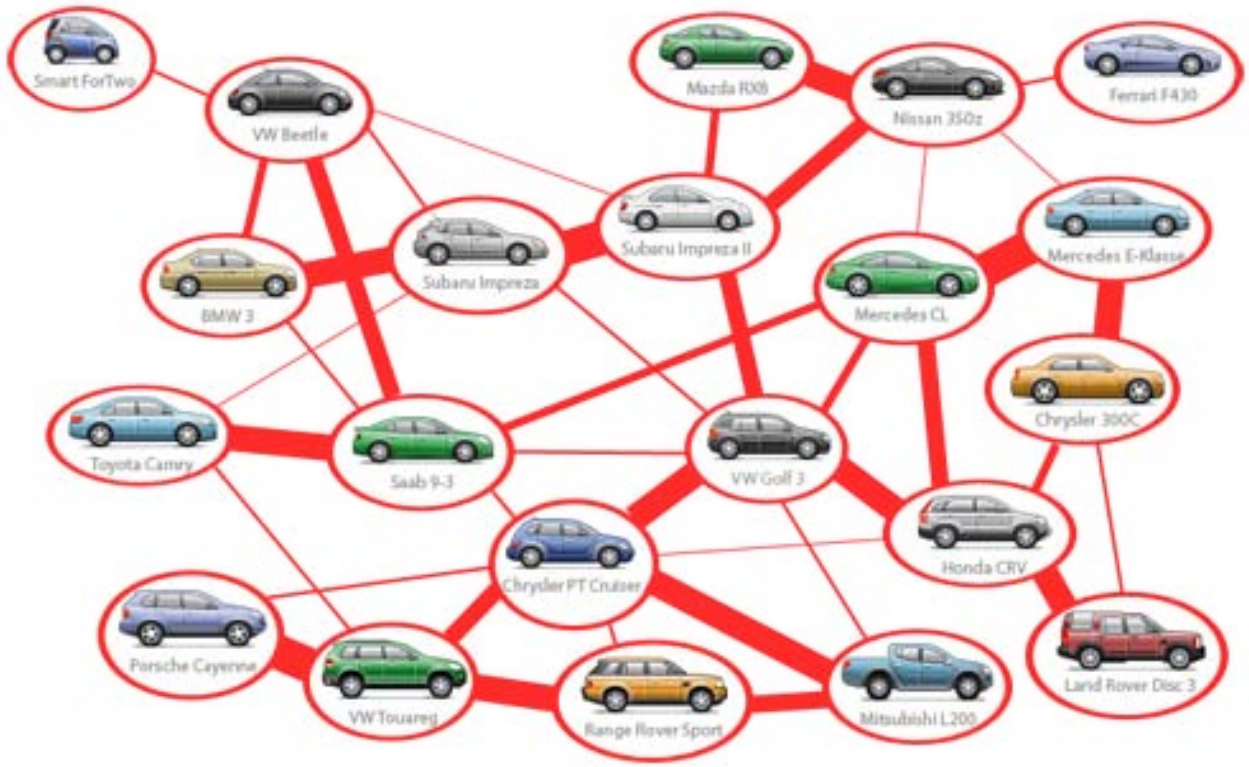
Vehicle Component Relativities

- Vehicle characteristics are used as proxies for VIN
- These are standardized for non-vehicle factors (e.g., age of driver)

Mining GLM residuals in controlled manner

Vehicle example

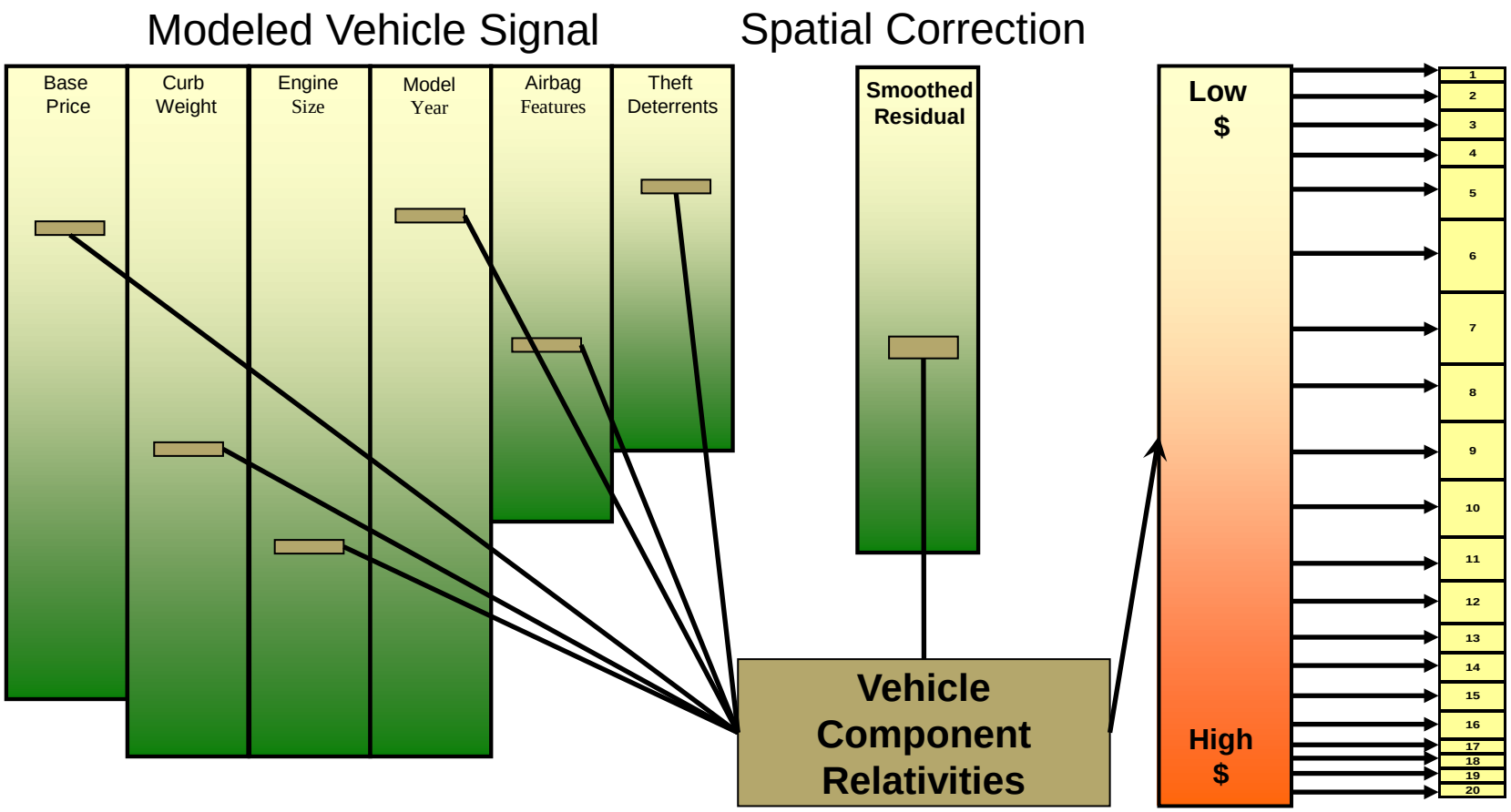
Smooth residuals across “neighbor” vehicles



Mining GLM residuals in controlled manner

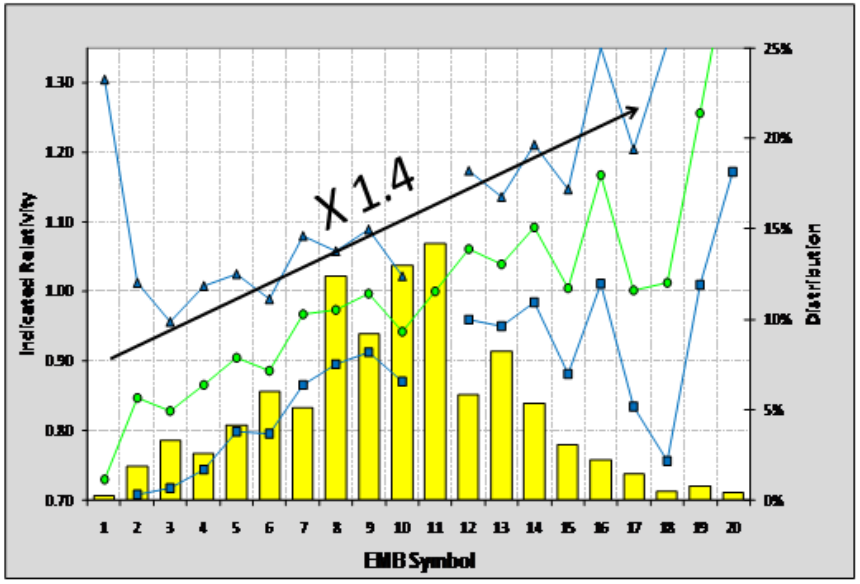
Vehicle example

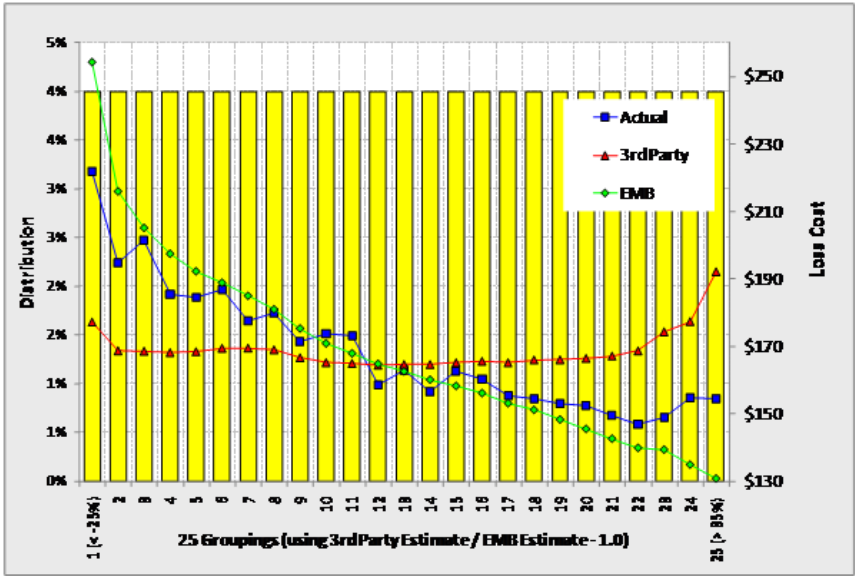
Vehicle estimator is clustered into new symbols



Mining GLM residuals in controlled manner

Vehicle example





Technique has proven very successful based on proper hold-out sampling validation

Summary

- Model building tools build models. Machine learning tools explore data.
- GLMs are a powerful and practical multivariate method for insurance analysis, particularly ratemaking.
- Model-building in general can be improved by following best practices and enhancements.
- Machine learning tools can improve the GLM process at every stage: data preparation, variable reduction, interaction detection, variable simplification, model validation.
- Data mining methods can squeeze additional predictive power out of GLM residuals. Rather than mining residuals on a broad basis, consider mining residuals and correcting a particular high-dimension factor
  - easier to control
  - easier to understand

Contact us

EMB
12235 El Camino Real
Suite 150
San Diego, California
92130

T +1 (858) 793-1425
F +1 (858) 793-1589
www.emb.com

