GLM III

CAS RPM 2019
Marcus Deckert
Brent Petzoldt

Antitrust Notice

- The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.
- Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.
- It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

AGENDA

JLT Re

- **Driver Averageing**
- **Modeling sparse claim types**
- **Product Evaluations**
- **The Tweedie Distribution**
- **Geographic risk**
- **“Quadrant Saddles”**

3

1 DRIVER AVERAGING

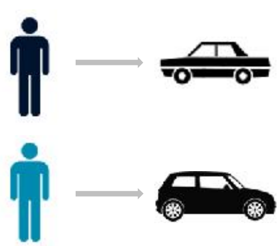
HOUSEHOLD AVERAGING

- Historically companies assigned operators to vehicles for the purpose of rating
- More recently driver averaging strategies have been deployed to capture the household
- Average may consider all drivers or a subset
  - This choice may affect other household composition factors
- Modeling data needs to mimic the transaction
- Types of averages
  - Straight vs. geometric average
  - Weighted average
  - Modified
  - Average/assigned hybrid

5

DRIVER AVERAGING VS DRIVER ASSIGNMENT
DRIVER ASSIGNMENT

- There is a one-to-one mapping of drivers to vehicles
- Assigned driver characteristics can be considered a vehicle characteristic
- Downstream tables do not need driver ID as a key
- Standard vehicle exposure is used

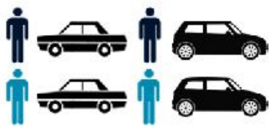


6

DRIVER AVERAGING VS DRIVER ASSIGNMENT
DRIVER AVERAGING

JLT Re

- There is a unique record for each driver-vehicle combination
- Characteristics of each driver is used for each combination
- Exposures for each vehicle are split amongst the number of drivers on the policy, i.e., annualized exposures / # drivers



7

MODEL DESIGN

JLT Re

- In all modeling projects, it is imperative that the data set up mimic the rating structure
- Consider the following example…

Vehicle	Operator	Vehicle Rate
V1	Dad	$500
V2	Mom	$450

Operator	Class Factor
Dad	0.80
Mom	0.85
Junior	2.80

- Assume Mom had a $1,000 claim while driving Dad's car

8

ASSIGNMENT

JLT Re

In driver assignment methodology, each record represents a single vehicle with one assigned operator

Veh	Op	Sym	MYR	Age	Sex	Type	Yths	Drvrs	Vehs	Exp	Clm	Losses	Prem
V1	Junior	17	2006	16	M	OO	1	3	2	1	1	1,000	1,400
V2	Mom	17	2005	43	F	PO	1	3	2	1	0	0	382

- Operator characteristics based on assigned operator
- Vehicle characteristics based on vehicle
- Policy characteristics "catch" other drivers
- Losses assigned to vehicle

9

STRAIGHT AVERAGE

- Straight average methodology:

$$h \quad \times \frac{(1 \quad + 2 \quad + 3 \quad)}{3}$$

- Which can be deconstructed::

$$h \quad \times \frac{(1 \quad)}{3}$$
$$h \quad \times \frac{(2 \quad)}{3}$$
$$h \quad \times \frac{(3 \quad)}{3}$$

10

STRAIGHT AVERAGE

In straight average methodology, each record represents a single vehicle and operator combination

Veh	Op	Sym	MYR	Age	Sex	Yths	Drvrs	Vehs	Exp	Clm	Loss	Prem
V1	Dad	17	2006	45	M	1	3	2	1/3	0	0	133
V1	Mom	17	2006	43	F	1	3	2	1/3	1	1,000	141
V1	Junior	17	2006	16	M	1	3	2	1/3	0	0	467
V2	Dad	17	2005	45	M	1	3	2	1/3	0	0	120
V2	Mom	17	2005	43	F	1	3	2	1/3	0	0	127
V2	Junior	17	2005	16	M	1	3	2	1/3	0	0	420

- Policy characteristics are same, but less predictive
- Driver exposure split amongst each vehicle
- Losses assigned to vehicle/operator combination
- iid is a major concern
- No clear solution for comprehensive coverage

11

GEOMETRIC AVERAGE

Geometric average methodology:

$$h \quad \times (1 \quad + 2 \quad + 3 \quad)^{/}$$

No direct decomposition

12

GEOMETRIC AVERAGE

Geometric methodology: each record represents a single vehicle

Veh	Sym	MYR	# Dads	# Moms	# Juniors	Exp	Clm	Loss	Prem
V1	17	2006	1/3	1/3	1/3	1	1	1,000	619.72
V2	17	2005	1/3	1/3	1/3	1	0	0	557.74

- Policy characteristics are same, but less predictive
- Predictors are translated to counts
- Losses assigned to vehicle
- More challenging to add operator interactions or variates

WEIGHTED AVERAGE

Weighted average methodology for a straight average approach

Veh	Op	Sym	MYR	Age	Sex	Type	Yths	Drvrs	Vehs	Exp	Clm	Loss	Prem
V1	Dad	17	2006	45	M	PO	1	3	3	1/3	0	0	133
V1	Mom	17	2006	43	F	OC	1	3	3	1/3	1	1,000	141
V1	Junior	17	2006	16	M	OC	1	3	3	1/3	0	0	467
V2	Dad	17	2005	45	M	OC	1	3	3	1/3	0	0	120
V2	Mom	17	2005	43	F	PO	1	3	3	1/3	0	0	127
V2	Junior	17	2005	16	M	OC	1	3	3	1/3	0	0	420

- Creates a relationship between the vehicle and the operator
- Uses the model to determine the weights
- More accurate since it uses more information…if correct

$$h \qquad 1 \qquad \times \frac{(1 \qquad * \quad + 2 \qquad * \quad + 3 \qquad * \quad)}{3}$$

2 MODELING SPARSE CLAIM TYPES

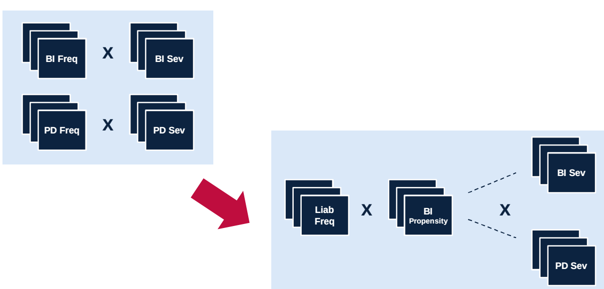
AMPLIFICATION OF THE BI SIGNAL USING PD EXPERIENCE

JLT Re

- **Fit straight to BI**
- **Use PD model as a guide in free fitting BI**
- **Use PD model structure**
- **Offset PD relativities onto BI data as starting point**
- **BI/PD proportion model:**
  - BI frequency = BI/PD proportion * PD frequency

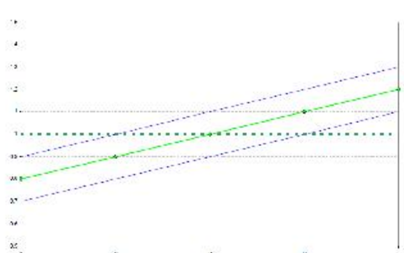
PROPORTION MODEL

JLT Re



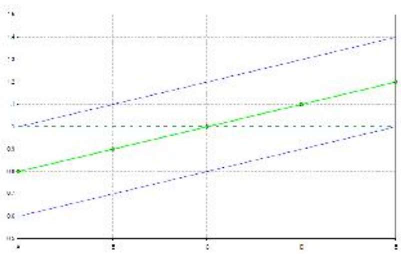
REFERENCE MODELS

JLT Re



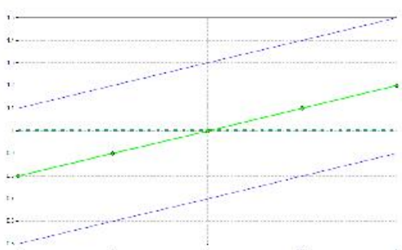
REFERENCE MODELS

JLT Re



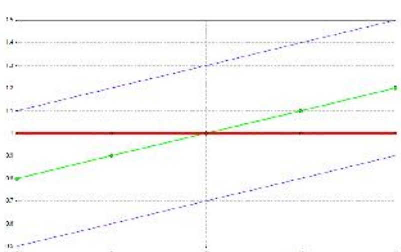
REFERENCE MODELS

JLT Re



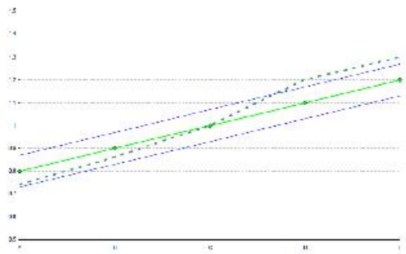
REFERENCE MODELS

JLT Re



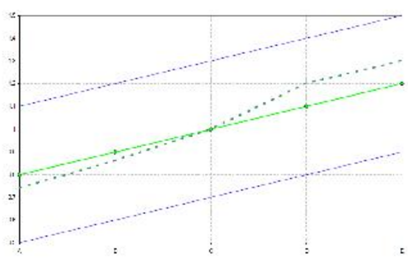
REFERENCE MODELS

JLT Re



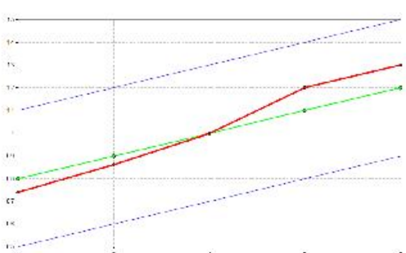
REFERENCE MODELS

JLT Re



REFERENCE MODELS

JLT Re



REFERENCE MODELS

JLT Re

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

Offset Term

- When modeling BI, set PD fitted values to be offset term
- GLM will seek effects over and above assumed PD effect

25

EXPERIMENT

JLT Re

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

Real large company

10% random sample

Pretend small company

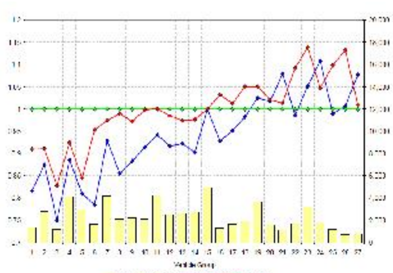
(2) Traditional GLM on BI claims on the "small company"

(3) Propensity reference model on BI claims of PD claims

26

EXAMPLE RESULT

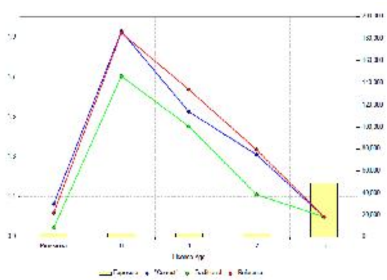
JLT Re



27

EXAMPLE RESULT

JLT Re



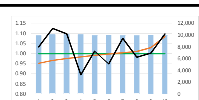
28

JLT Re

3 MODEL & PRODUCT EVALUATION

DOUBLE LIFT CHARTS & RESIDUAL ANALYSIS

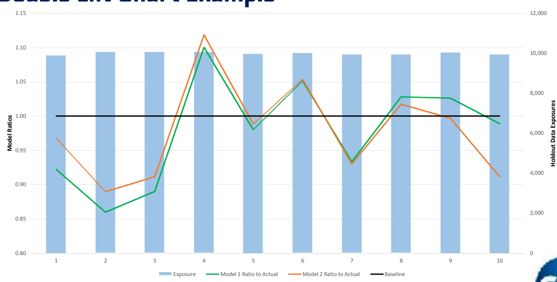
Double Lift Charts



- Double lift charts allow for performance comparison between two models on the holdout dataset.
- The x-axis metric used is the “M ratio”, (model 1 prediction / model 2 prediction).
- Visual inspection can be used by counting the number of points on the chart where a model “wins”, or a calculation can be used. As with all evaluation techniques some judgement is needed.

CAS

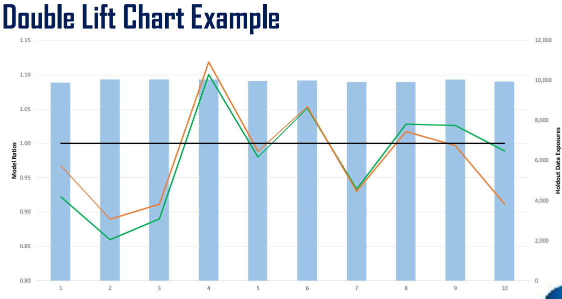
Double Lift Chart Example



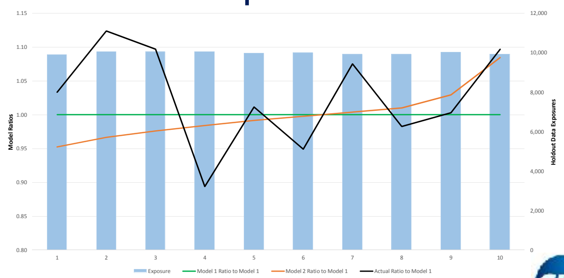
Double Lift Chart Data

Decile	Exposure	Actual	Model 1 Prediction	Model 2 Prediction	M Ratio	Model 1 Ratio to Actual	Model 2 Ratio to Actual	Baseline	Model 1 ABS Error	Model 2 ABS Error
1	9,900	1,176,932	1,085,213	1,139,291	0.95	0.92	0.97	1.00	0.08	0.03
2	10,060	1,424,253	1,224,576	1,267,062	0.97	0.86	0.89	1.00	0.14	0.11
3	10,049	1,299,990	1,156,955	1,185,253	0.98	0.89	0.91	1.00	0.11	0.09
4	10,062	1,076,663	1,184,883	1,204,219	0.98	1.10	1.12	1.00	0.10	0.12
5	9,980	1,176,702	1,153,482	1,163,320	0.99	0.98	0.99	1.00	0.02	0.01
6	10,001	1,130,289	1,188,551	1,190,952	1.00	1.05	1.05	1.00	0.05	0.05
7	9,924	1,292,249	1,206,454	1,201,915	1.00	0.93	0.93	1.00	0.07	0.07
8	9,919	1,225,162	1,259,580	1,246,783	1.01	1.03	1.02	1.00	0.03	0.02
9	10,032	1,221,396	1,253,443	1,217,445	1.03	1.03	1.00	1.00	0.03	0.00
10	9,937	1,347,950	1,332,839	1,228,693	1.08	0.99	0.91	1.00	0.01	0.09
									0.63	**0.59**

Double Lift Chart Example



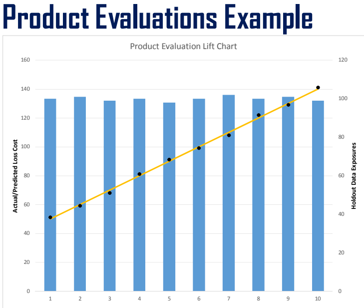
Double Lift Chart Example Alternate



Product Evaluations

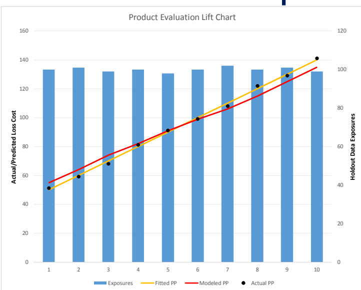
- **Double lift charts** and **residual models** can both be used to accurately assess if a product is worth purchasing.
- The **double lift** chart method allows you to directly compare models with and without the product to evaluate if the improvement from the product is significant.
- **Residual models** are an alternative and involve:
  1. Fitting the best model without the product
  2. Sending the vendor the holdout data including your prediction, the actual target variable, and independent variables required to calculate vendor's prediction
  3. Vendor calculates their model prediction on the holdout data
  4. Vendor builds standard decile lift chart on their predictions, and also plots your average prediction for each decile
- Both of these techniques intend to eliminate the effects of correlation between the vendor's product and the predictor variables already in our model. If we simply look at the one-way analysis from their product, we may be misled into believing it will provide great benefit when it really only provides marginal lift beyond our existing models.

Product Evaluations Example



- Fitted tracks with actual
- Large amount of lift
- Let's buy it!

Product Evaluations Example



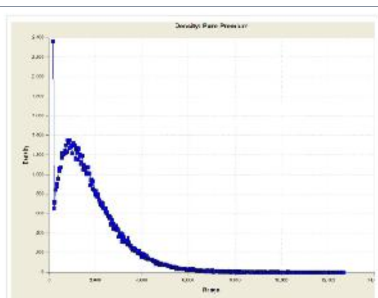
- Fitted tracks with actual
- Large amount of lift
- Wait a minute; our model can segment risk, too
- This product may be helpful, but it's strongly correlated with our current information sources and the impact will be limited

JLT Re

4 THE TWEEDIE DISTRIBUTION

TWEEDIE GLMS

JLT Re

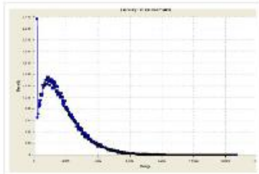


39

TWEEDIE GLMS

JLT Re

Raw pure premiums
- Incurred losses have a point mass at zero and then a continuous distribution
- Poisson and gamma not appropriate here
- Tweedie distribution has
  - Point mass at zero
  - A parameter which changes shape above zero



$$(\; ; \; , \lambda, \;) = \sum \frac{\{(\quad) \quad (\quad -)\}}{(\quad) \, !} \exp\{ \; [\quad - \kappa \; (\quad)]\} \text{ for y>0}$$

$$(\quad = 0) = \exp\{- \quad \kappa \; (\quad)\}$$

FORMULIZATION OF GLMS

JLT Re

Observed Response	Most Appropriate Link Function	Most Appropriate Error Structure	Variance Function
-	-	Normal	μ^0
Claim Frequency	Log	Poisson	μ^1
Claim Severity	Log	Gamma	μ^2
Claim Severity	Log	Inverse Gaussian	μ^3
Raw Pure Premium	Log	Tweedie	μ^T
Retention Rate	Logit	Binomial	μ (1-μ)
Conversion Rate	Logit	Binomial	μ (1-μ)

41

FORMULIZATION OF GLMS

JLT Re

- **More formally:**

$$(\quad) = \frac{(\hat{})}{}$$

Variance function

Prior weights

Scale parameter

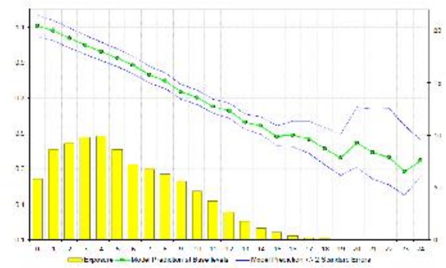
- **Tweedie's Variance function:**
  - p=1 Poisson
  - p=2 Gamma
  - 1<p<2 Poisson/Gamma process
- **Other concerns**
  - Need to estimate both & p when fitting models
  - Typically p ~= 1.5 for incurred claims

42

EXAMPLE 1

JLT Re

Vehicle Age - Frequency

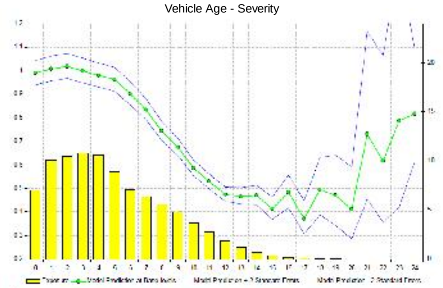


43

EXAMPLE 1

JLT Re

Vehicle Age - Severity

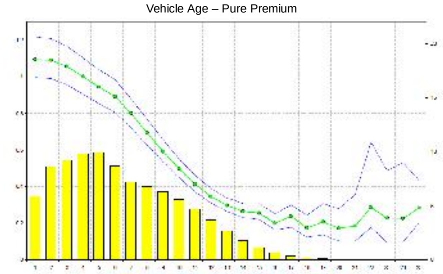


44

EXAMPLE 1

JLT Re

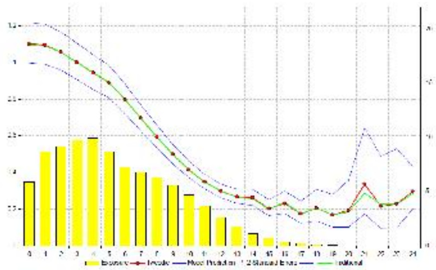
Vehicle Age – Pure Premium



45

EXAMPLE 1

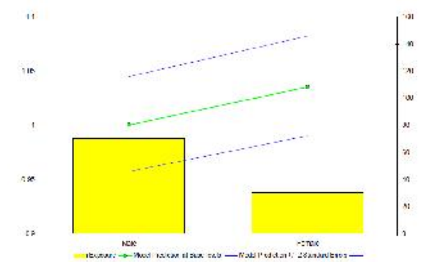
Vehicle Age – Pure Premium



46

EXAMPLE 2

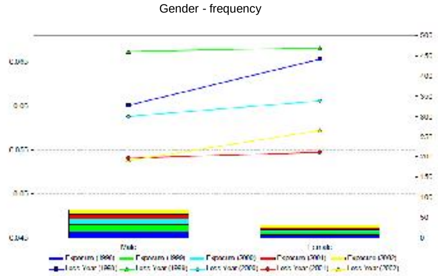
Gender - frequency



47

EXAMPLE 2

Gender - frequency

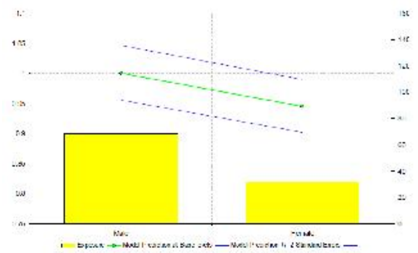


48

EXAMPLE 2

JLT Re

Gender - severity

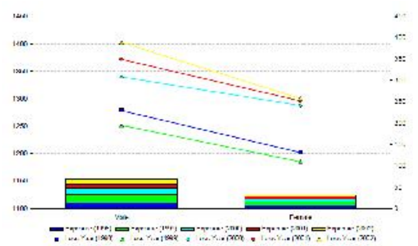


49

EXAMPLE 2

JLT Re

Gender - severity

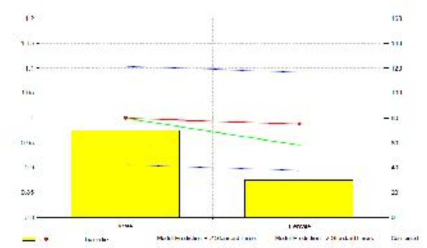


50

EXAMPLE 2

JLT Re

Gender – pure premium



51

TWEEDIE GLMS

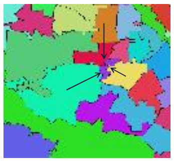
- Helpful when it's important to fit to loss cost directly
- Similar results to frequency/severity traditional approach if frequency and severity effects are clearly weak or clearly strong
- Distorted by large insignificant effects
- Removes understanding of what is driving results
- Smoothing harder

52

5 GEOGRAPHIC RISK

TERRITORIAL BOUNDARY/RELATIVITY ANALYSIS

- Location is critical as a major risk driver and accounts for a substantial portion of the variation in insurance risk
- Two elements:
  - Segmentation of the risk (territorial boundaries)
  - Quantification of the risk (territorial relativities)
- Historically, the market focus has been on relativities
  - Initial boundaries typically based on limited data, anecdotal evidence, competitors, bureaus, and judgment
  - Regular reviews of relativities, while merely tweaking the boundaries when necessary

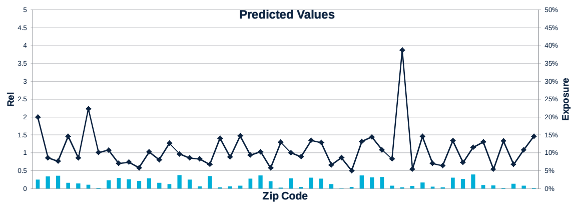


54

HIGH DIMENSIONAL CATEGORICAL VARIABLES
STANDARD DIMENSION REDUCTION TECHNIQUES FALL SHORT

JLT Re

- Grouping difficult to evaluate
- Cannot "order" geographic units, so curves not an option

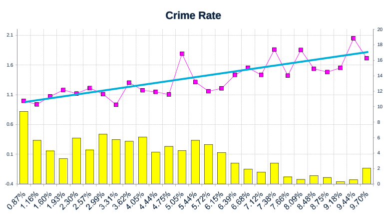


55

SOLUTION 1: USE PROXIES

JLT Re

- Proxies attach at the code level
- High-dimensional, but ordered; so we can fit curves
- Geo-demographics such as:
  - Population density
  - Crime rate

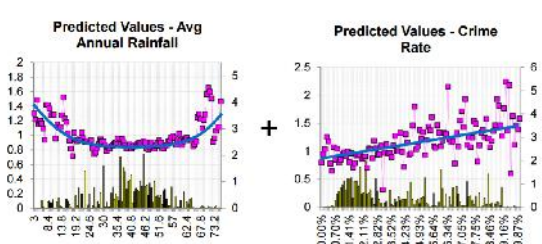


56

PROBLEM WITH PROXIES ONLY
HOW TO DETERMINE RIGHT PROXIES (OR COMBINATIONS THEREOF) HAVE BEEN USED?

JLT Re

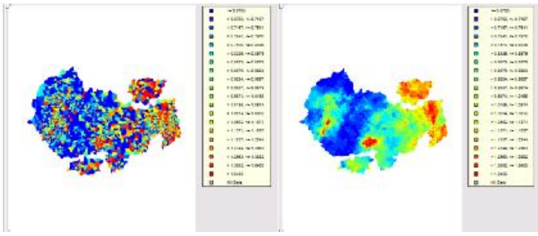
- How to determine the right proxies (or combinations thereof) have been used?



57

SOLUTION 2: USE PROXIES WITH SPATIAL CORRECTION

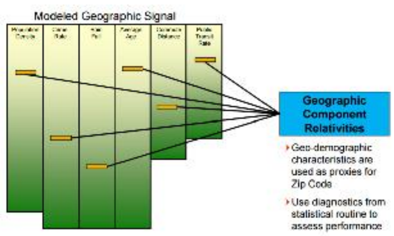
1. Include proxies in GLM
2. Then apply geo-spatial smoothing



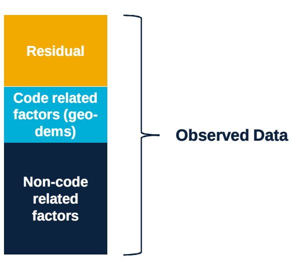
GEOGRAPHIC ESTIMATOR

Initial Estimator:

- Component models built using geographic proxies



SPATIAL CORRECTION APPROACH



SPATIAL CORRECTION APPROACH

JLT Re

Spatial smoothing extracts additional signal

Effect modeled in GLM

Residual

Code related factors (geo-dems)

Non-code related factors

Total geographical segmentation

61

SPATIAL CORRECTION APPROACH

JLT Re

Residual

Code related factors (geo-dems)

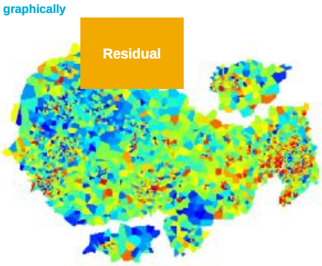
Non-code related factors

62

MAPPING THE RESIDUALS

JLT Re

- View the residuals graphically

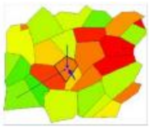


63

SPATIAL SMOOTHING METHODS

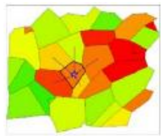
JLT Re

- Uses knowledge of surrounding areas to enhance estimates of the underlying risk in each area based on the "Principle of locality"



Distance-based

- Simpler to implement and interpret
- Does not consider natural boundaries such as rivers
- May over-smooth urban areas and under-smooth rural
- Best peril uses: windstorm



Adjacency-based

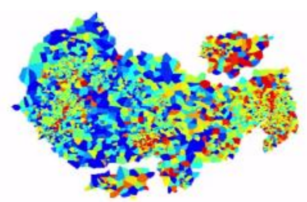
- Distribution assumptions about claims process can be incorporated
- Distance can be built in
- Considers natural boundaries
- Potential lines: auto, HO theft

64

SMOOTHING THE RESIDUALS

JLT Re

- View the residuals graphically
- Are there any patterns?

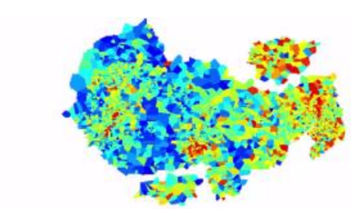


65

SMOOTHING THE RESIDUALS

JLT Re

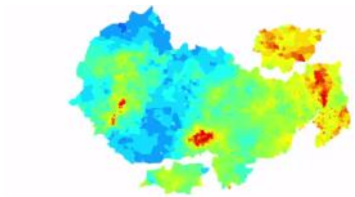
- View the residuals graphically
- Are there any patterns?



66

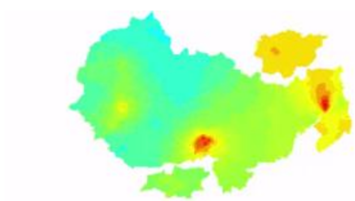
SMOOTHING THE RESIDUALS

- View the residuals graphically
- Are there any patterns?



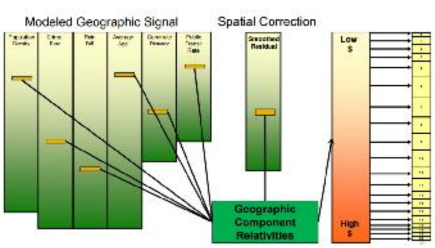
SMOOTHING THE RESIDUALS

- View the residuals graphically
- Are there any patterns?



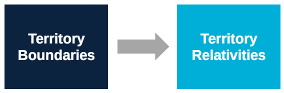
TERRITORIES

Clustering
- Cumulative geographic signal clustered into territories



DETERMINING TERRITORIAL RELATIVES

JLT Re



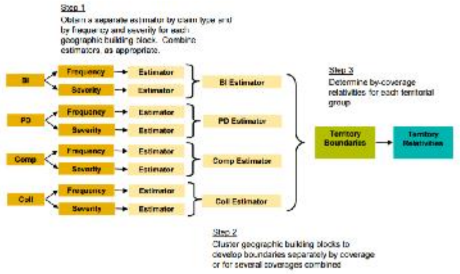
- GLM model fit using data grouped by new territorial boundaries
- Test relativities using standard GLM tests
  - Predictive in GLM
  - Consistent over time
- Refine boundaries/relativities as appropriate
  - Incorporate rules-based restrictions
  - Apply actuarial knowledge
  - Investigate neighboring territories with very different relativities

70

TERRITORY RATING - OVERVIEW

JLT Re

- Accurate estimation of underlying risk associated with geography is a three stage process



71

SUMMARY

JLT Re

- Territory is a major driver of risk, thus it is critical that companies review boundaries and relativities regularly
- Issues exist that create special challenges with regards to territorial analysis
  - High-dimensionality
  - Heavily correlated
- Territory boundary analysis requires a range of different approaches and tools (as there are different loss drivers)
- Diagnostics needed to ensure best model possible

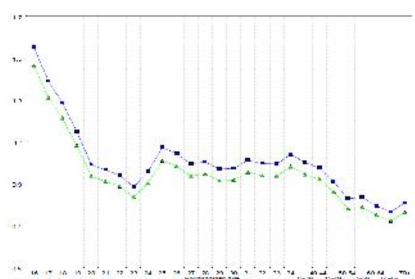
72

JLT Re

6 QUADRANT SADDLES

INTERACTIONS

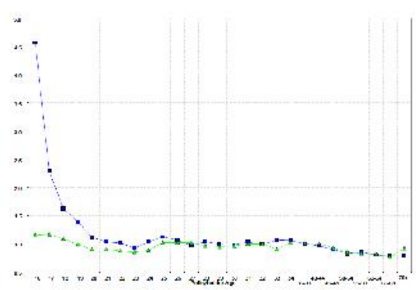
JLT Re



74

INTERACTIONS

JLT Re



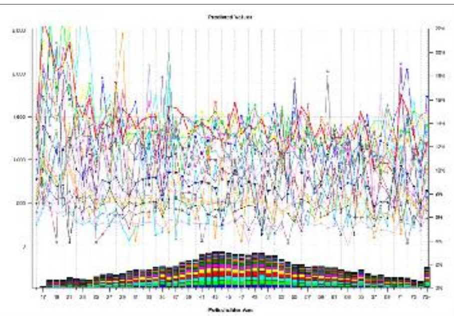
75

WHY ARE INTERACTIONS PRESENT?

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

76

INTERACTIONS



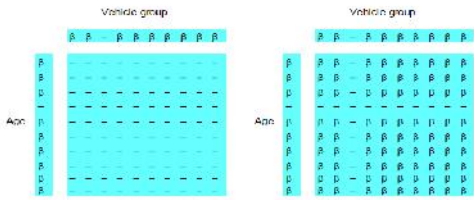
77

INTERACTIONS

78

INTERACTIONS

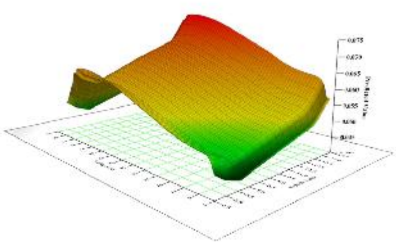
JLT Re



79

EXAMPLE

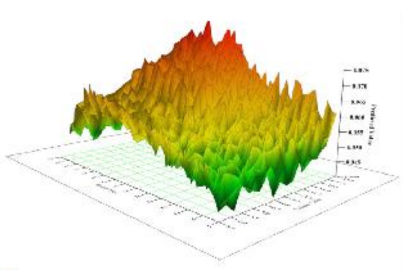
JLT Re



80

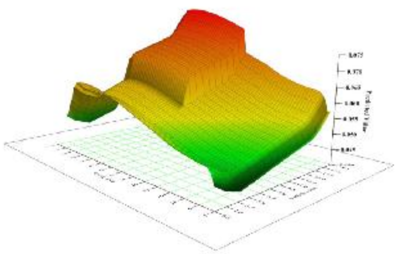
EXAMPLE

JLT Re



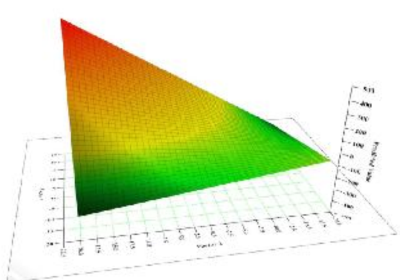
81

EXAMPLE

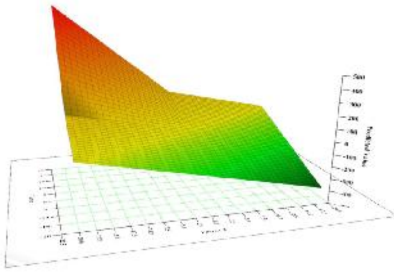


INTERACTIONS

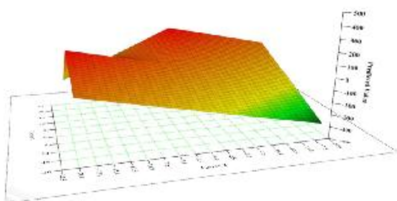
SADDLES



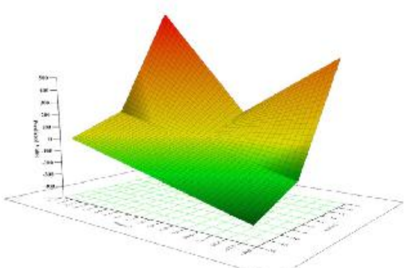
SADDLES



SADDLES

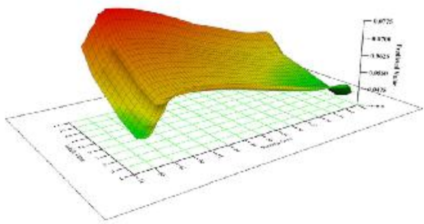


SADDLES



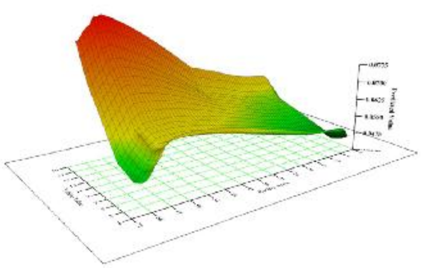
SADDLES

JLT Re



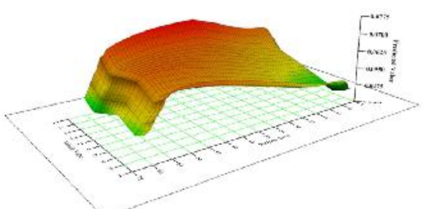
SADDLES

JLT Re

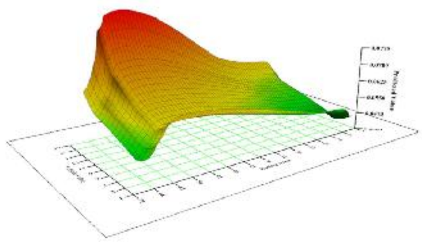


SADDLES

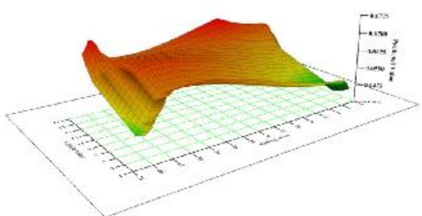
JLT Re



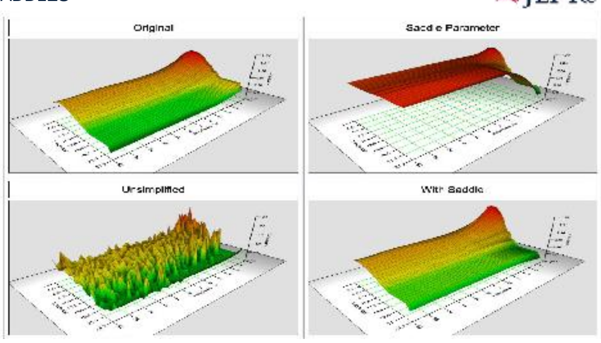
SADDLES



SADDLES

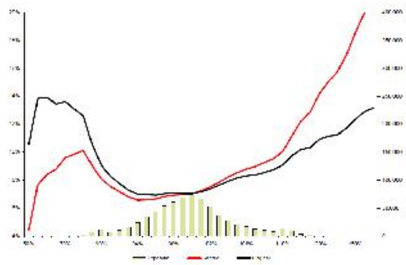


SADDLES



SADDLES- MODEL COMPARISON
AUTO FREQUENCY – OUT OF SAMPLE

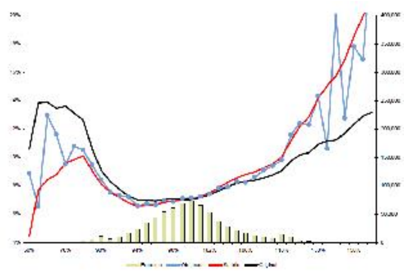
JLT Re



94

SADDLES- MODEL COMPARISON
AUTO FREQUENCY – OUT OF SAMPLE

JLT Re



95