NATIONAL RECONNAISSANCE OFFICE

(U) A Method for Regression Analysis on Sparse Datasets

Daniel Barkmeyer
NRO CAAG
June 2015

SUPRA ET ULTRA

Background

- Traditional regression analysis, e.g. Zero-Bias Minimum Percent Error (ZMPE), can run into problems when important independent variables are known only for some datapoints (sparsely populated)
  - Omit data for which not all drivers tested are known, or
  - Do not test as drivers those data fields that are not fully populated

- NRO CAAG's Commercial-like Acquisition Program Study (CAPS)* ameliorated this issue
  - Empirically-derived scoring term based on known drivers
  - Scores independent of unknown drivers
  - Regression determines contribution of drivers to score, and coefficients expressing DV as a function of score
  - Linear regression only

OBJECTIVE: Apply score-based regression to power-form functions with multiplicative error terms

* Alvarado, W., Barkmeyer, D., and Burgess, E. "Commercial-like Acquisitions: Practices and Costs." Journal of Cost Analysis and Parametrics, V3, Issue 1.

Regression on Sparse Datasets

- Advantage – retain **explanatory power** of sparsely-populated drivers, **degrees of freedom** in regressions derived from sparsely-populated datasets
- For a given independent variable *n*, if x_n is unknown for a datapoint, the influence of *n* is removed from the score for that datapoint
  - Datapoint can be retained in the regression as long as some x_n are known
  - Allows all partially-populated datapoints to inform regression

Data Point	Cost	Weight	Operating Wavelength	Mobile (1) or Stationary (0)	Operational (1) or Experimental (0)
1	$ 18	154	250	0	0
2	$ 95	650			1
3	$ 54		450	0	
4	$ 52	310	500		1
5	$ 68	776	450	0	0
6	$ 165	490	505	1	1
7	$ 307	900	800		0
8	$ 60	100		1	1
9	$ 123	281	550	1	0
10	$ 82	200	380	1	

ZMPE Regression
Include
~~Omit~~
~~Omit~~
~~Omit~~
Include
Include
~~Omit~~
~~Omit~~
Include
~~Omit~~

Scoring Regression	
Include	S = f(W,λ,Mob,Op)
Include	S = f(W,Op)
Include	S = f(λ,Mob)
Include	S = f(W,λ,Op)
Include	S = f(W,λ,Mob,Op)
Include	S = f(W,λ,Mob,Op)
Include	S = f(W,λ,Op)
Include	S = f(W,Mob,Op)
Include	S = f(W,λ,Mob,Op)
Include	S = f(W,λ,Mob)

With 4 drivers, **0** DOF

With 4 drivers, **6** DOF

Scoring Method for Power Function Form

- For a linear regression equation in the CAPS model, the score was calculated as a weighted average of the normalized known drivers:

$$S_{linear} = \frac{\sum w_n \bar{x}_n}{\sum_{x_n\ known} w_n}$$

- For a power function equation, the desired form of the score is a weighted geometric mean of the normalized known drivers:

$$S_{power} = e^{\frac{\sum w_n \overline{\ln x_n}}{\sum_{x_n\ known} w_n}}$$

where

$$\overline{\ln x_n} = \begin{cases} \dfrac{(\ln x_n) - (\ln x_n)_{min}}{(\ln x_n)_{max} - (\ln x_n)_{min}}, & x_n \text{ continuous} \\ \dfrac{x_n - (x_n)_{min}}{(x_n)_{max} - (x_n)_{min}}, & x_n \text{ binary} \end{cases}$$

- Where x_n is not known, the n^{th} term drops out of the numerator and denominator in the score

Power Function Form

- It can be shown that the form for the score reduces the regression equation

$$y = A + B \cdot S_{power}{}^{C}$$

to the desired power function form:

$$y = A + Q \cdot \prod_{x_n \text{ cont.}} x_n{}^{P_n} \cdot \prod_{x_n \text{ bin.}} P_n{}^{x_n}$$

with constants *P* and *Q* functions of *B*, *C*, weightings, and max/min values of drivers (all constants)

- This form can be used on sparse or full datasets

Dataset for Testing

- Created a dataset representative of typical cost estimating problem
- 100 datapoints, 4 independent variable drivers
  - Driver 1 continuous, lognormally distributed, mean value 500, coefficient of variation 0.65, minimum value 100
  - Driver 2 continuous, lognormally distributed, mean value 15, coefficient of variation 5.0, minimum value 0
  - Driver 3 binary, 33% of data has value of 1
  - Driver 4 binary, 50% of data has value of 1
- Dependent variable values set by the equation

$$y = 100 + 20 \cdot {x_1}^{0.6} \cdot {x_2}^{0.3} \cdot 3^{x_3} \cdot 1.2^{x_4} \cdot \varepsilon$$

A Q P_1 P_2 P_3 P_4

with error term ε lognormally distributed, mean value 1, coefficient of variation 0.4, minimum value 0

- Underlying behavior of the data is known
  - Regression results can be compared against the expected result

Validation – 100% Populated

- For the test dataset, regression of the form

$$y = (A + Q \cdot {x_1}^{P_1} \cdot {x_2}^{P_2} \cdot {P_3}^{x_3} \cdot {P_4}^{x_4}) \cdot \varepsilon$$

with objective to minimize the value

$$f_{obj} = \sum_{n=1}^{100} {\varepsilon_n}^2$$

ZMPE

- Optimize *A*, *Q*, P_1, P_2, P_3, P_4
- Solution:

A	0.0
Q	18.9
P_1	0.63
P_2	0.34
P_3	2.89
P_4	1.14

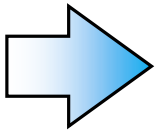
Score-Based

- Convert regression equation to

$$y = (A + Be^{C \cdot \frac{w_1 \overline{\ln x_1} + w_2 \overline{\ln x_2} + w_3 \overline{\ln x_3} + w_4 \overline{\ln x_4}}{w_1 + w_2 + w_3 + w_4}}) \cdot \varepsilon$$

- Optimize *A*, *B*, *C*, w_1, w_2, w_3, w_4
- Solution and re-conversion:

A	0.0
B	32.1
C	7.10
w_1	17%
w_2	66%
w_3	15%
w_4	2%



A	0.0
Q	18.9
P_1	0.63
P_2	0.34
P_3	2.89
P_4	1.14

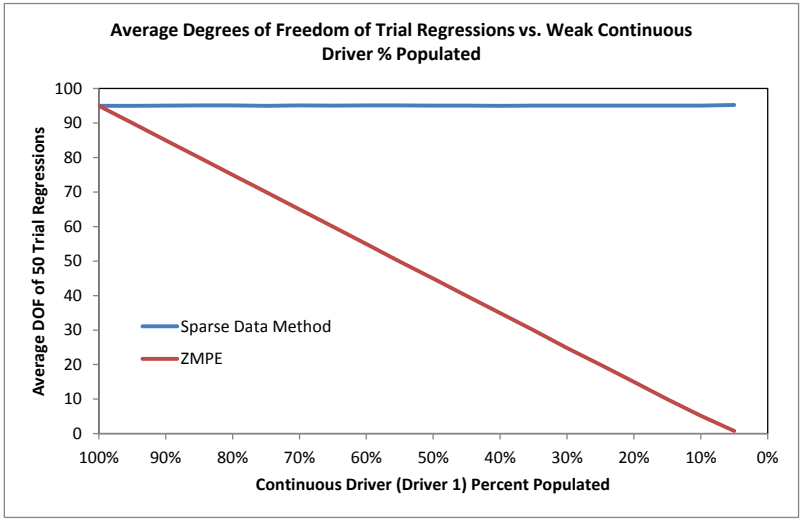
Score-Based method reproduces ZMPE solution on fully-populated dataset

Score-Based Regression vs. ZMPE

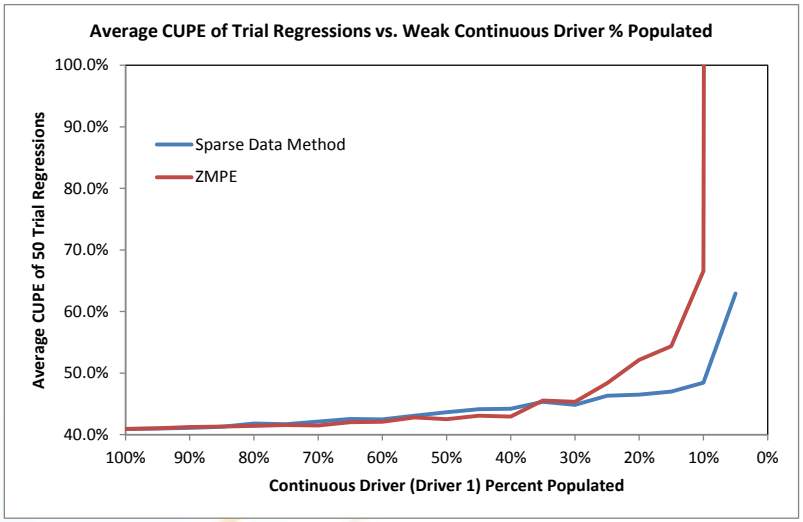
- Validated Score-based method is equivalent to ZMPE on a fully-populated dataset
- Next step: sparsely-populated test cases
  - Individual drivers sparsely populated
    - 50 regressions, each with randomly-selected values removed, at every 5% interval of population percent between 100% and 5% populated
    - Other 3 drivers fully populated
  - All drivers sparsely populated
    - 200 regressions, each with randomly-selected values removed, at every 5% interval of overall population percent between 100% and 5%
- Comparison metric: Characteristic Underlying Percent Error
  - CUPE is measured across the entire dataset, including values that were removed to simulate sparseness of data
  - Defined as $CUPE = \sqrt{\frac{\sum_{n=1}^{100} \epsilon_{\%n}{}^2}{DOF}}$ where $\epsilon_{\%n}$ is the percent error between the actual *y* and the regression equation's predicted *y* for the n^{th} datapoint
  - Measures how well the regression (with incomplete data) captures the underlying relationship (if the data were complete)

Score-Based Regression vs. ZMPE

Weaker Continuous Driver Sparse



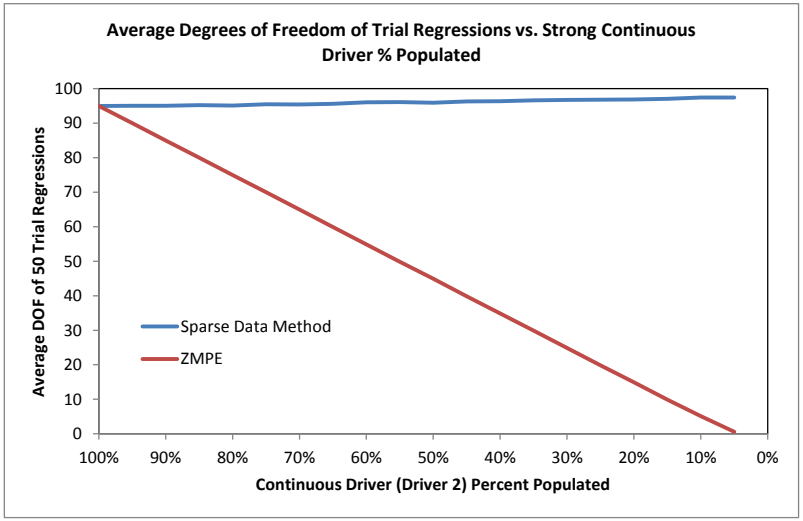
- Degrees of Freedom
  - ZMPE regression DOF decreases linearly with % Populated
  - Score-based regression retains all DOF from the full dataset



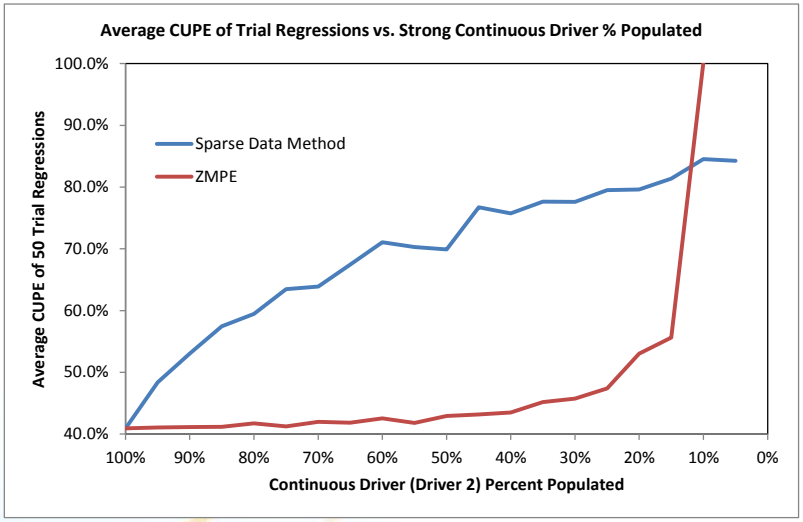
- CUPE of resultant estimating relationship against full dataset
  - Models show similar performance down to 30% populated
  - Below 30% populated, score-based method is better able to model the underlying relationship

Score-Based Regression vs. ZMPE

Stronger Continuous Driver Sparse



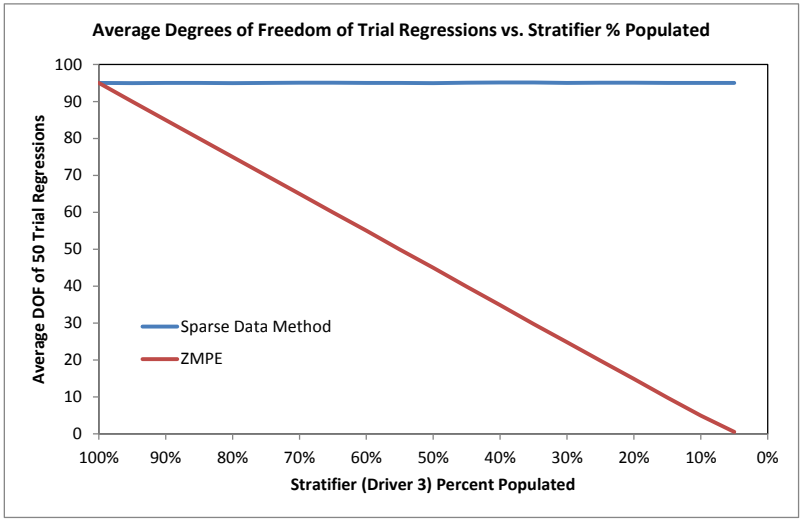
- Degrees of Freedom
  - ZMPE regression DOF decreases linearly with % Populated
  - Score-based regression retains all DOF from the full dataset



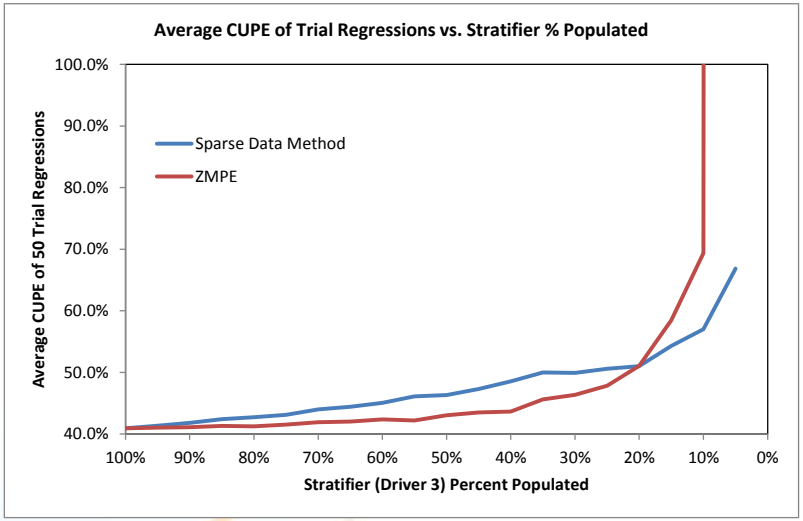
- CUPE of resultant estimating relationship against full dataset
  - ZMPE performs much better above very low population percentages
  - Score-based method only proves better able to capture underlying relationship once ZMPE DOF becomes very small

Score-Based Regression vs. ZMPE

Binary Driver Sparse



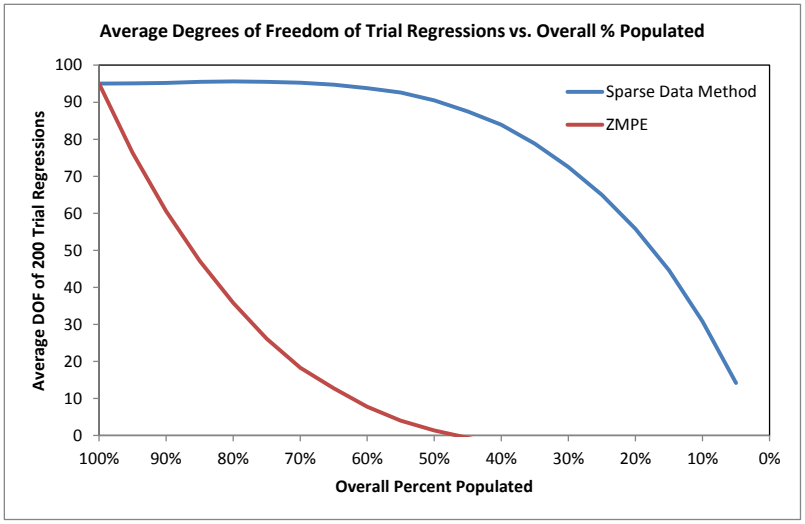
- Degrees of Freedom
  - ZMPE regression DOF decreases linearly with % Populated
  - Score-based regression retains all DOF from the full dataset



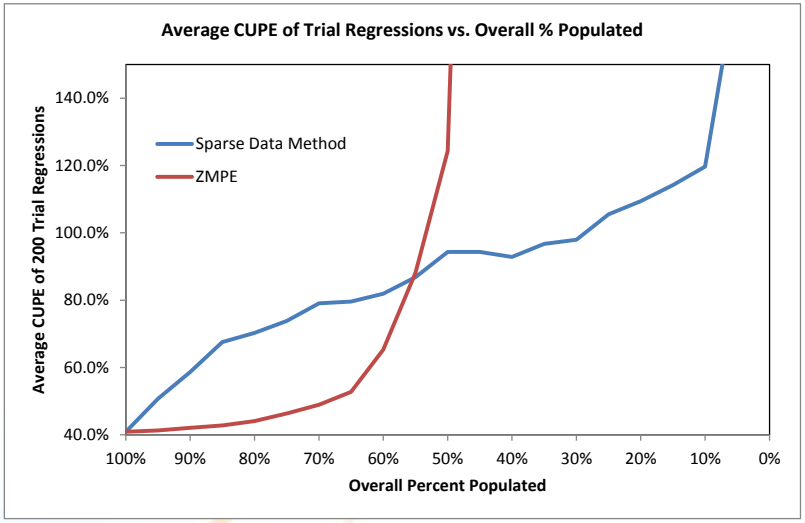
- CUPE of resultant estimating relationship against full dataset
  - ZMPE performs slightly better above 20% populated
  - Score-based method proves better able to capture underlying relationship once ZMPE DOF becomes small

Score-Based Regression vs. ZMPE

All Drivers Sparse



- Degrees of Freedom
  - ZMPE regression DOF drops significantly as overall population % drops – ZMPE unusable most of the time below ~50% populated
  - Score-based regression DOF drops much more slowly – only when no drivers are known is a DOF lost



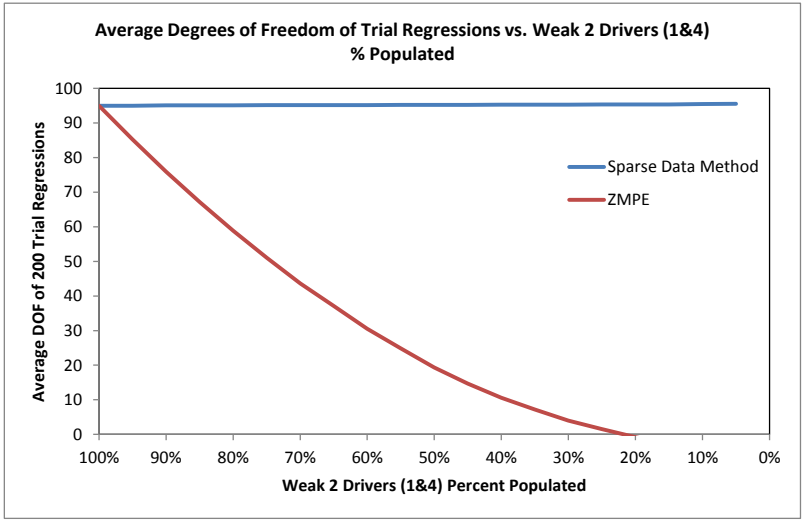
- CUPE of resultant estimating relationship against full dataset
  - ZMPE performs better until DOF becomes very low (around 50% populated)
  - Score-based method can regress a relationship at significantly lower population percentages

Initial Results

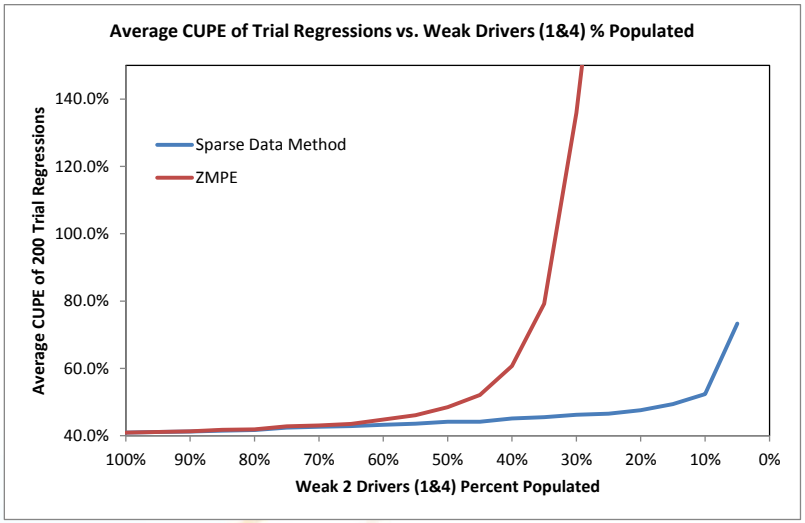
- Sparse data regression method is able to come to solutions for sparser datasets than ZMPE
- Quality of fit depends on how well strong drivers are known
  - Underlying trends are modeled well when a weak driver is sparse
  - Underlying trends are lost when a strong driver is sparse; better to exclude points for which strong drivers are not known
- One more test case: both strong drivers fully populated, both weak drivers sparsely populated

Score-Based Regression vs. ZMPE

Only Weak Drivers Sparse



- Degrees of Freedom
  - ZMPE gradually loses degrees of freedom as population % drops
  - With two strong drivers 100% populated, score-based regression never omits any data from the regression



- CUPE of resultant estimating relationship against full dataset
  - Score-based method finds the underlying relationship as well as or better than ZMPE regardless of population %
  - Regressions only use datapoints with known values for strong drivers

Score-Based method improves regression when secondary drivers are sparsely populated

Application – Satellite System Test Schedule (1)

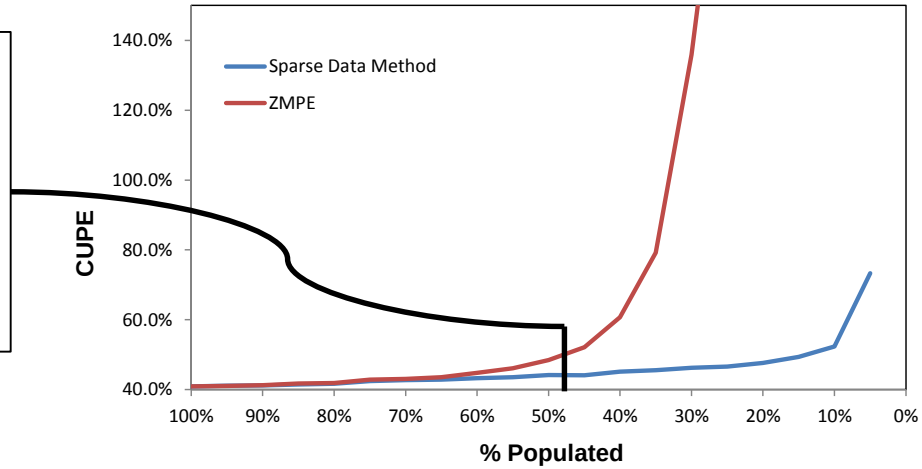
- Satellite system test schedule database*
  - 96 programs
  - 37 potential schedule drivers
  - 20 drivers are very sparsely populated (15% populated or less)
  - Overall Percent Populated: 48%

Driver	Pop %
Demonstration	100%
New/Block Change	100%
Incumbent	100%
New	100%
Option	99%
IMINT / Remote Sensor	99%
Comsat	99%
GFE P/L	69%
Design Life	100%
Qual Vehicle	45%
Dry Weight	100%
BOL Power	84%
GEO Orbit	96%
HEO/MEO Orbit	96%
LEO Orbit	96%
Mission Types	100%
Number of Payloads	98%

Driver	Pop %
Wet Mass	11%
No. of Deployables	11%
Max Data Rate	11%
Active Thermal Control	11%
CC&DH Redundancy	10%
NiH2 Battery	11%
Battery Capacity	10%
Bus Nominal Voltage	3%
Propellant Wt	7%
Deployed Solar Array	11%
Solar Array Area	10%
RCS Isp	7%
Total Thrust	7%
Instrument Dry Mass	11%
Instrument Types	13%
Instrument Cost	11%
Bus New Design	11%
Instruments New Design	11%
No. Customers	11%
No. Organizations	11%

48% Populated

On test data
- Method captures underlying behavior better than ZMPE
- Combined weighting of strong drivers: 50-100%, average 85%



* Burgess, E. "Predicting System Test Schedules." presented to the Space System Cost Analysis Group, July 2005

Application – Satellite System Test Schedule (2)

- Result:

$$[\text{Test to 1st Launch (months)}] = 0.22 + 1.46\,[\text{Score}]^{4.40}$$

- Score is a function of 14 Drivers
  - 5 sparsely-populated drivers are influential
  - Conversion to standard form:

$$[\text{TT1L (months)}] = 0.22 + 0.41\,[\text{DL}]^{0.18}\,[\text{wt}]^{0.15}\,[\text{Inst Types}]^{0.20} \cdot [\text{Miss Types}]^{0.30}\,[\text{Prop Wt}]^{0.08}\,[\#\ \text{Orgs}]^{0.33}\,1.28^{[Act\ TC]} \cdot 1.22^{[IMINT]}\,1.20^{[Opt]}\,1.19^{[New\ Des]}\,1.17^{[GEO]}\,1.10^{[HEO]}\,1.08^{[Incbt]}$$

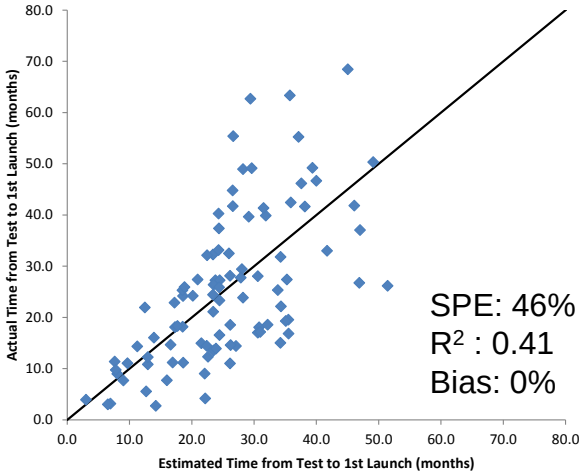
Relative Impact of Drivers

Driver	Weight
Design Life	29%
Dry Weight	14%
Instrument Types	9%
Mission Types	8%
Propellant Wt	7%
Deployed Solar Array	6%
Active Thermal Control	6%
IMINT / Remote Sensor	5%
Option	4%
New	4%
GEO Orbit	4%
No. Organizations	3%
HEO/MEO Orbit	2%
Incumbent	2%

Combined weighting of fully-populated drivers: **72%**

- With this as a starting point, continue typical SER development process
  - Reduce to a reasonable number of drivers
  - Compare against results of other regressions
  - Etc.
- ZMPE regression on this dataset would be forced to discount the 3rd strongest driver

Actual vs. Estimated Test Schedule



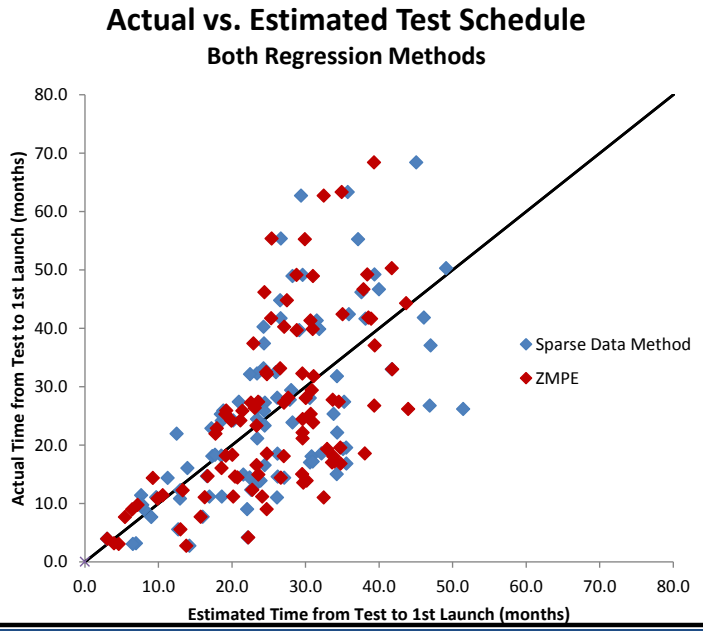
Application – Satellite System Test Schedule (3)

- ZMPE Regression Result:

$$[\text{TT1L (months)}] = -561 + 525\,[\text{DL}]^{0.002}\,[\text{wt}]^{0.01}\,[\text{Miss Types}]^{0.02} \cdot 1.01^{[IMINT]}\ 1.01^{[Opt]}\ 1.01^{[New\ Des]}\ 1.01^{[GEO]}\ 1.004^{[Incbt]}$$

- Only uses fully-populated drivers
- Result is much more dominated by constant terms than score-based regression result

- Score-based Regression Result
  - SPE: 45.9%
  - R^2 : 0.41
- ZMPE Regression Result
  - SPE: 45.4%
  - R^2 : 0.37



Score-Based method improves upon ZMPE for proof-of-concept case

Conclusions & Next Steps (1)

- Extended concept of scoring method for sparse datasets from CAPS study to generic power form regression

- Scoring method captures influence of sparsely-populated independent variables where traditional regression cannot
  - Traditional regression methods would force drivers or data to be omitted
  - Scoring method allows inclusion, informing regression
  - Better captures the influence of sparsely-populated drivers
  - Provided the dominant drivers are fully populated

Conclusions & Next Steps (2)

- Recommended approach to sparse dataset regression
  - Omit only data for which expected strong driver(s) are unknown
  - Employ scoring to allow inclusion of data with sparsely-populated secondary drivers
  - Verify the majority of the explanatory power of the regression comes from the fully-populated drivers (combined weightings in the score outweigh sparse drivers)

- Next Steps
  - Examine sensitivities of method
    - How much explanatory power must be held by fully-populated IVs to ensure CUPE is not worse than traditional regression?
    - How sparse is too sparse – when should an IV not be included?
  - Examine suitability of method in NRO CAAG CER development

BACKUP

Conversion from Scoring Form to Traditional Power Form

Scoring Form reduces to traditional power form:

$$y = A + B \cdot S_{power}{}^{C} \quad \longrightarrow \quad y = A + Q \cdot \prod_{x_n \text{ cont.}} x_n{}^{P_n} \cdot \prod_{x_n \text{ bin.}} P_n{}^{x_n}$$

with the following formulas for constants *P* and *Q*:

$$P_n = \begin{cases} \dfrac{C\, w_n}{(\sum w_n) \cdot ((\ln x_n)_{max} - (\ln x_n)_{min})}, & x_n \text{ continuous} \\ \dfrac{C\, w_n}{\sum w_n}, & x_n \text{ binary} \end{cases}$$

$$Q = \frac{B}{e^{\sum_{x_n \, continuous} \frac{C\, w_n \, (\ln x_n)_{min}}{(\sum w_n) \cdot ((\ln x_n)_{max} - (\ln x_n)_{min})}}}$$

NATIONAL RECONNAISSANCE OFFICE

SUPRA ET ULTRA