SW Sustainment: Nonlinear Regression Comparisons, Residual Fits, and Significance Testing

GALORATH

Christian Smart, Ph.D.
Kim Roye
Galorath Federal

Galorath Presenters

Kim Roye
Senior Analyst

- Former Mathematical Statistician for the U.S. Census Bureau
- Provided cost support for Department of Defense hardware programs for over ten years

Dr. Christian Smart
Chief Scientist

- Former Director for Cost Analytics and Parametric Estimating for the U.S. Missile Defense Agency
- Oversaw development of the NASA/Air Force Cost Model (NAFCOM)
- Provides subject matter expertise to NASA Headquarters, DARPA, and Space Development Agency
- Recognized expert on parametrics and risk analysis

Agenda

- Purpose: Apply Advanced Nonlinear Regression Methods to Develop SW Sustainment CERs
- Four Methods Considered: LOLS, MRLN, ZMPE, ZMAPE
- 11 CERs developed

  Tested Residual Fits

  Tested Significance

METHOD OF LEAST SQUARES

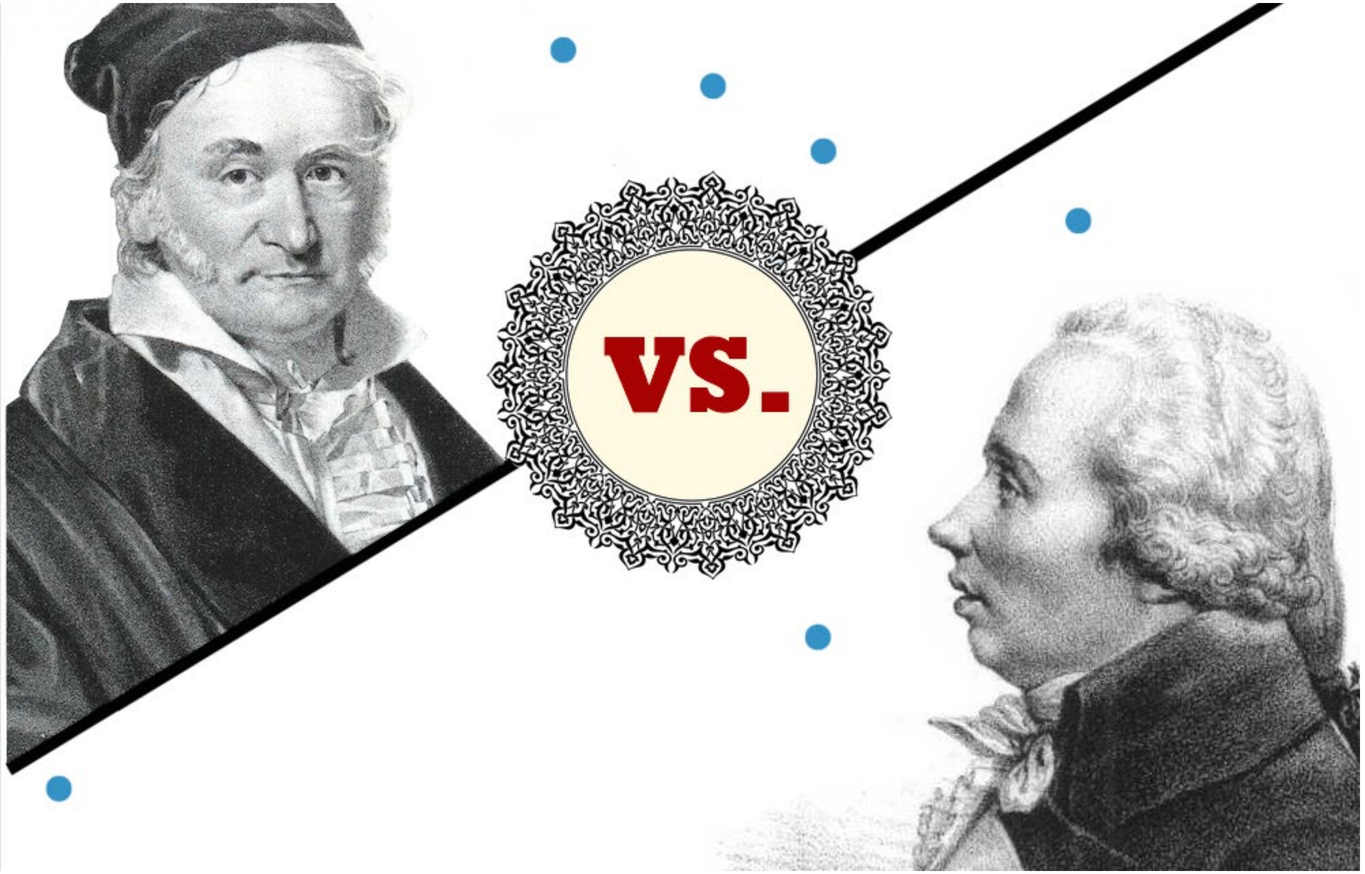
History

The method of least squares was originally used to predict the orbits of heavenly bodies using observed data. Francis Galton applied the technique to find linear predictive relationships between various phenomena, such as relationships between the heights of fathers and sons.

Given the linear equation of the form $Y = a + bX$ and a set of data (X_1, Y_1), (X_2, Y_2),…, (X_n, Y_n), the residuals are defined as $\varepsilon_i = Y_i - (a + bX_i) = Actual - Estimated$

The estimated cost linear regression finds the "best fit" by finding the parameters, a and b, that minimize the sum of the squares of the residuals

$$\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (Y_i - (a + bX_i))^2 = \sum_{i=1}^{n} (Actual_i - Estimated_i)^2$$



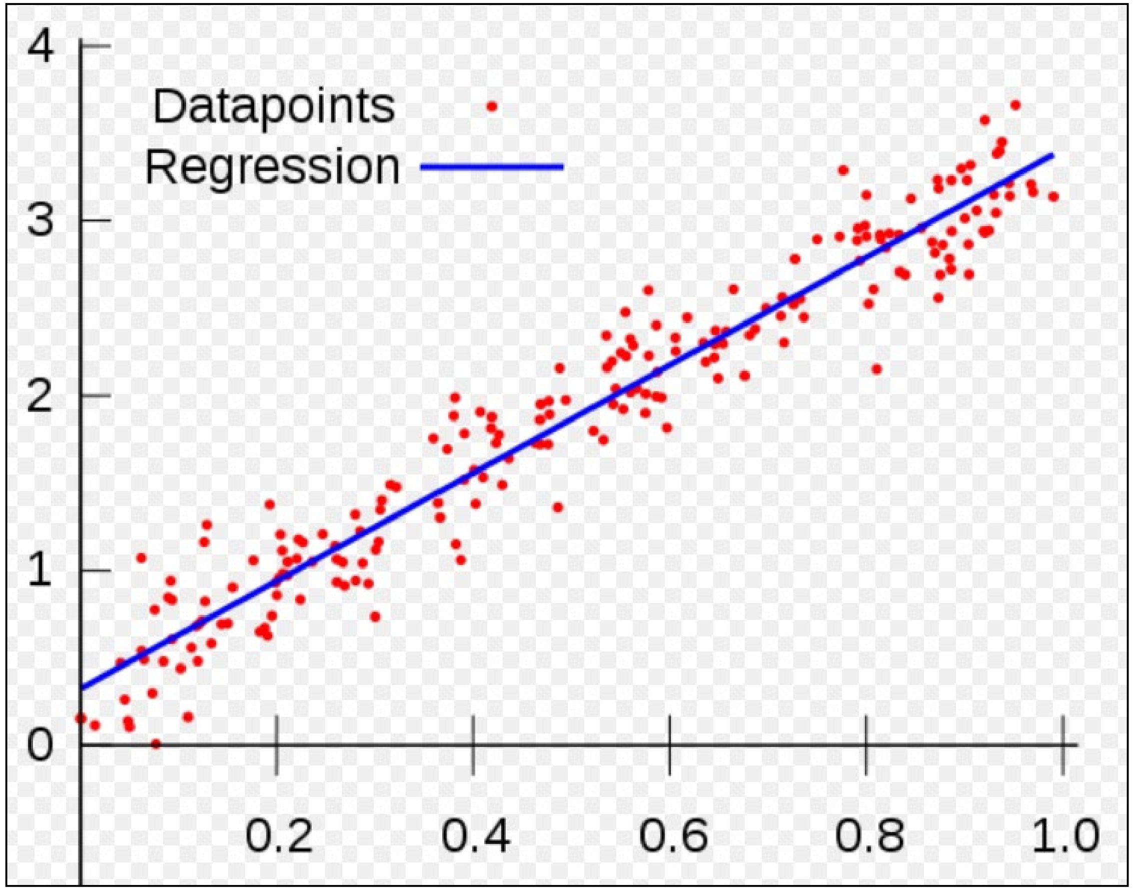
Least Squares method was first developed by mathematicians Legendre and Gauss in the early 19th century

REGRESSION ANALYSIS

History

The method of least squares in commonly called regression because Francis Galton applied the technique to find linear predictive relationships between various phenomena, such as relationships between the heights of fathers and sons.

Galton found a positive correlation between these heights but found a tendency to return or "regress" toward the average height, hence the term "regression analysis."



Fun fact: Francis Galton and Charles Darwin were first cousins.

LINEAR REGRESSION

- Widely used technique
- Given an equation of the form:

$$Y = a + bX$$

- And a set of data:

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

- The residuals are defined as:

$$\varepsilon_i = Y_i - (a + bX_i) = Actual - Estimated$$

- This is also referred to as the "error" term since it is the difference between the actual cost and the estimated cost linear regression finds the "best fit" by finding the parameters a and b that minimize the sum of the squares of the residuals

$$\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (Y_i - (a + bX_i))^2 = \sum_{i=1}^{n} (Actual_i - Estimated_i)^2$$

NONLINEAR REGRESSION

- In the spacecraft and defense industry it is more common to see nonlinear relationships between cost and cost drivers
- The power equation is ubiquitous

$$Y = aX^b$$

- In this case Y typically represents cost in $, but can also represent effort (hours, full-time equivalents)
- X typically represents weight or some other performance parameter
- The equation can also be modified to accommodate multiple cost drivers
- The value of the b parameter in the power equation is usually less than 1, indicating economies of scale in design and production
- Linear regression is simple - the calculations can be done by hand, but nonlinear regression requires more sophisticated methods, often the use of a computer

RESIDUALS

- The residuals of the power equation can either be additive or multiplicative
- Additive residuals have the form

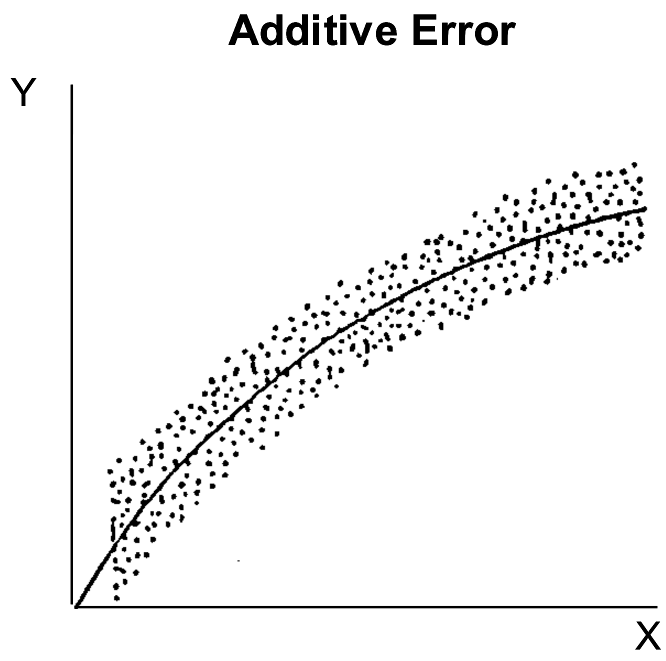
$$Y = aX^b + \varepsilon$$

- Multiplicative residuals have the form

$$\boxed{Y = aX^b \varepsilon}$$

- Multiplicative residuals are more appropriate for the spacecraft and defense industry in most applications because of wide variations in size, scope, and scale of the systems that are estimated
  - As a result we are primarily interested in the percentage difference between actual and estimated costs, not the absolute difference
- For example, if historical data ranges from $50 million to $1 billion, better to analyze percentage differences

RESIDUAL COMPARISON



For data with wide ranges we are more interested in residuals as percentages (multiplicative) than as dollars (additive)

MULTIPLICATIVE RESIDUALS

- For the Power Equation with Multiplicative Residuals, i.e.,

$$Y = aX^b \varepsilon$$

- The Regression Estimates Vary Based on the Variation of the Residual

$$\varepsilon = \frac{Y}{aX^b}$$

- Also Common to Adjust This to Treat ε as a Percentage, i.e., Set

$$Y = aX^b(1+\varepsilon)$$

$$\varepsilon = \frac{aX^b - Y}{aX^b} = \frac{Estimate - Actual}{Estimate}$$

- Actual Cost = Estimate +/- Percentage of Estimate

RESIDUALS ARE RANDOM VARIABLES

VARIATION NOT DUE TO INDEPENDENT VARIABLES

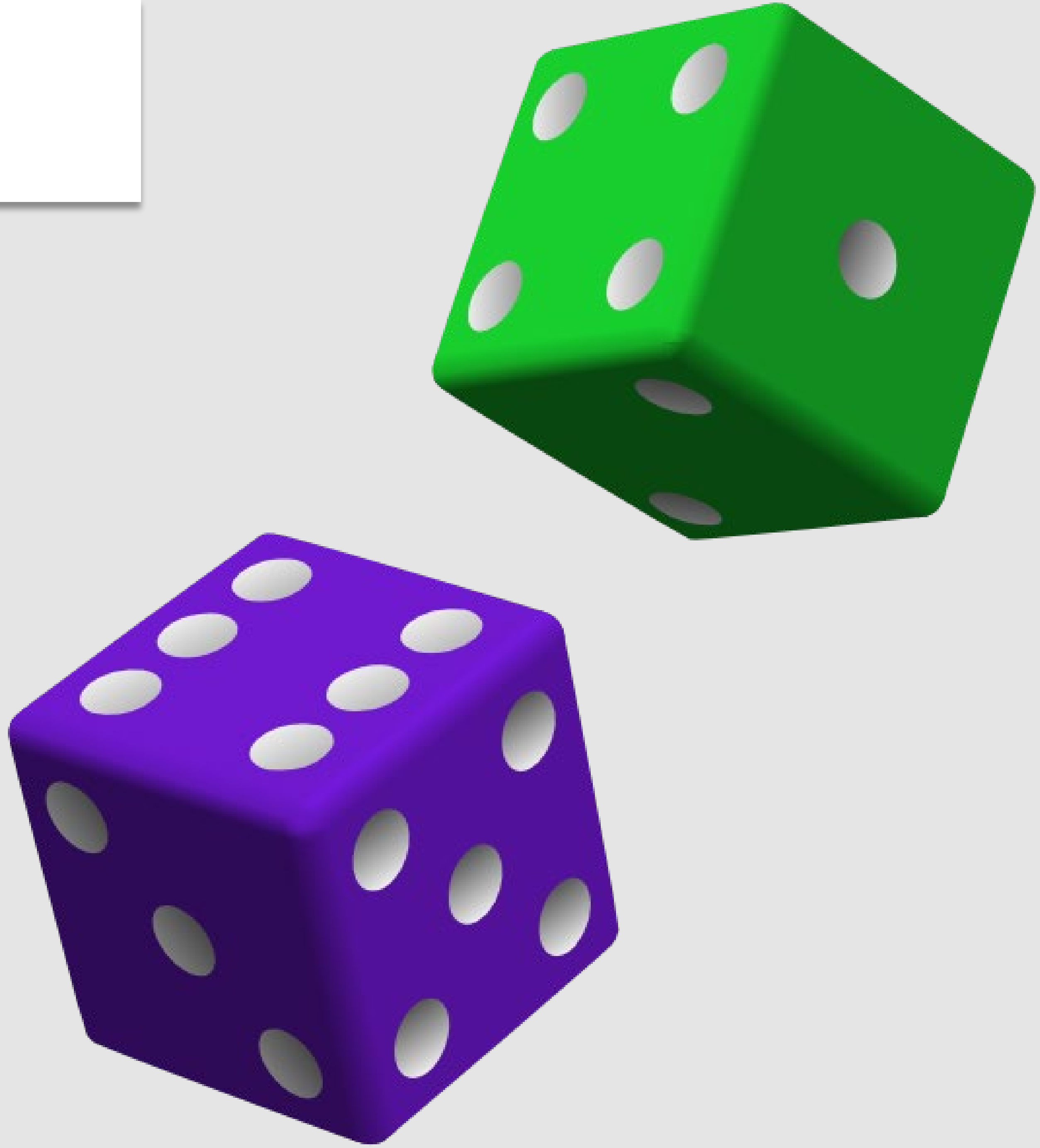
Residual variation is that which is due to the unexplained variation in your model

LINEAR MODELS

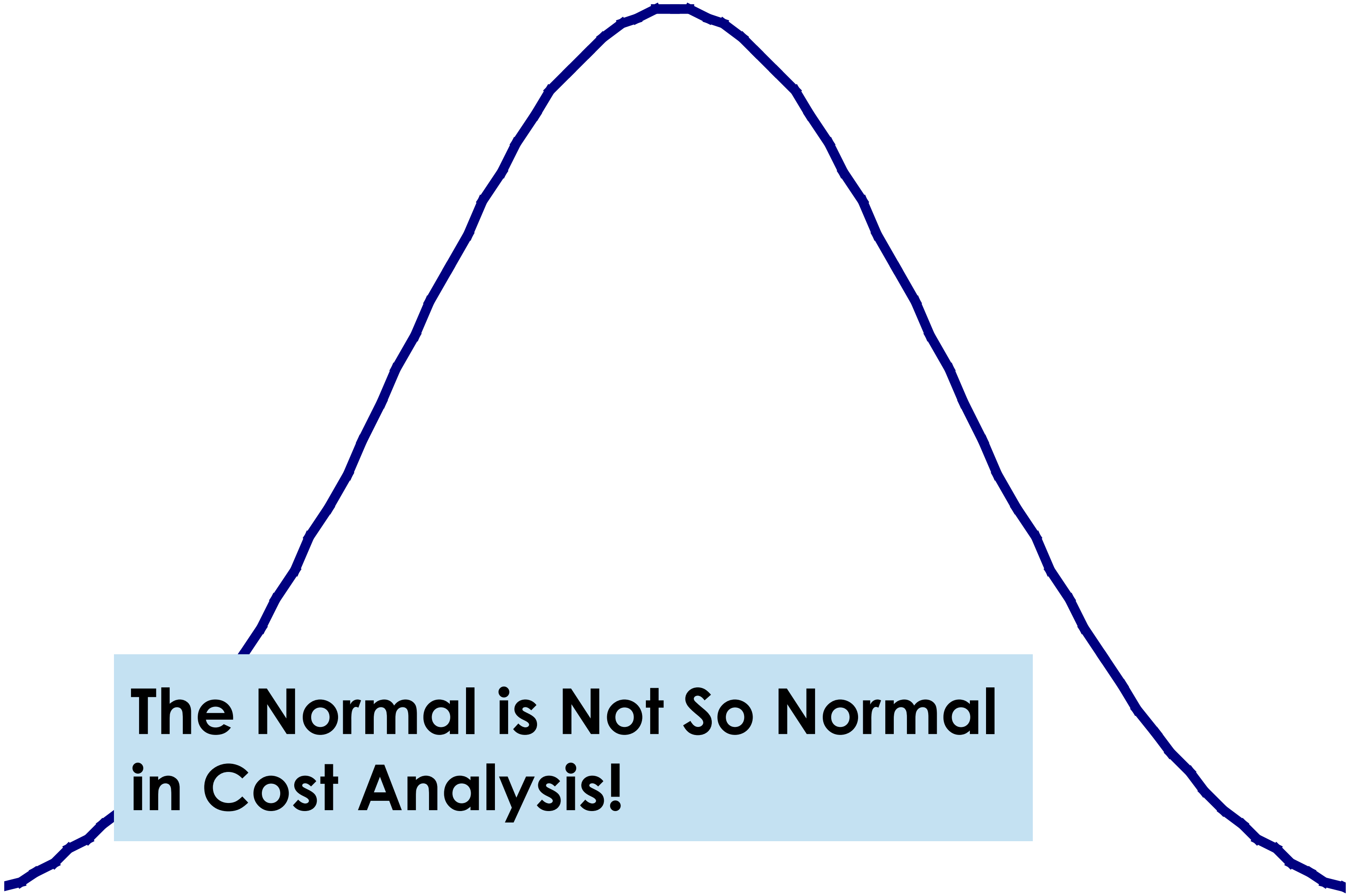
Assumption is that the variation is normally distributed

NONLINEAR MODELS

Varation is typically lognormal; can also be modeled without a specific assumption of a probability distribution (non-parametric)



NORMAL DISTRIBUTION



The Normal is Not So Normal in Cost Analysis!

1

NORMAL DISTRIBUTION

Most commonly encountered probability distribution – many random phenomena follow this

2

ALSO KNOWN AS

Bell Curve for its symmetric shape

Gaussian Distribution for one of its discoverers

3

CENTRAL LIMIT THEOREM

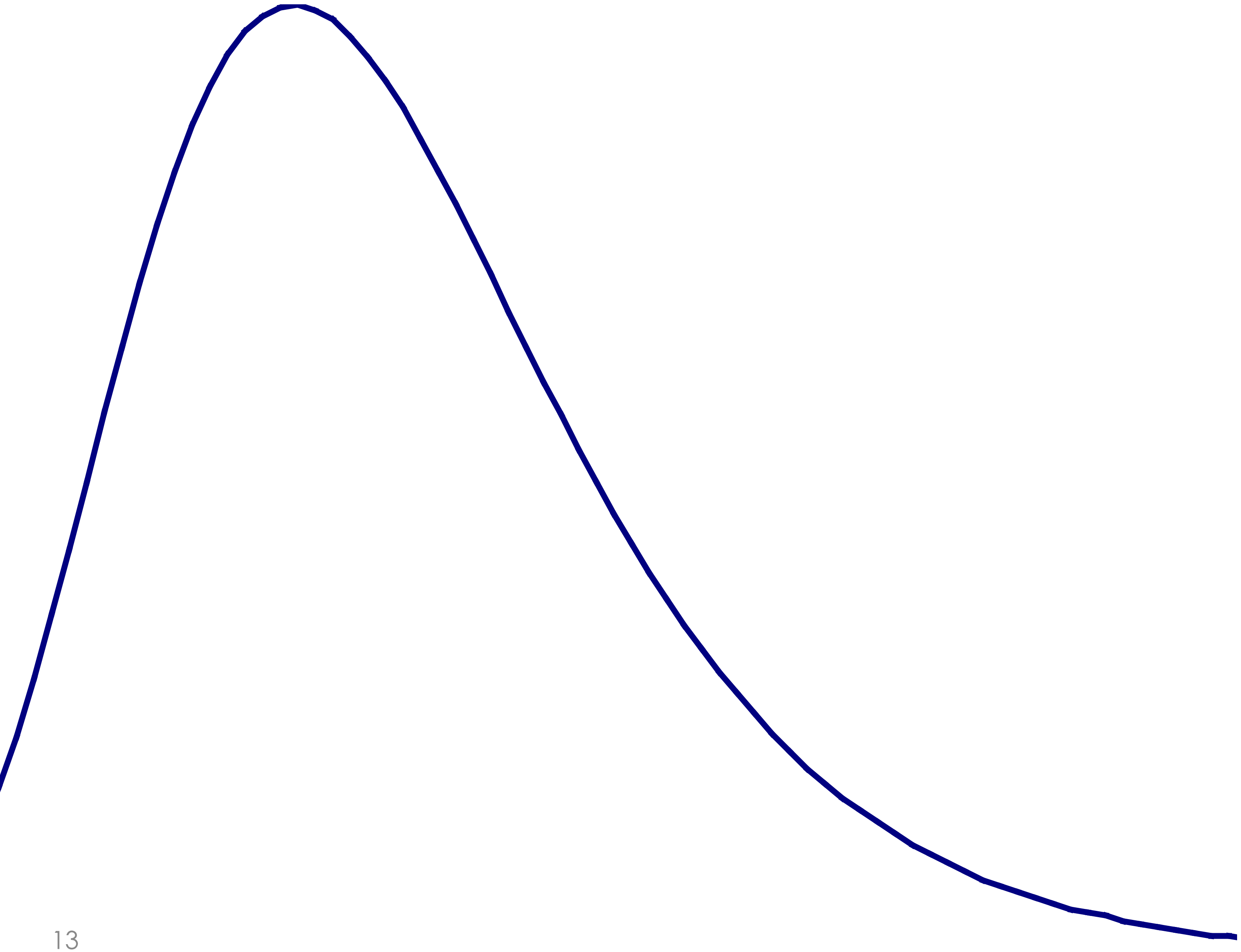
Often the case that the sum of many random phenomena is normally distributed

4

COST ANALYTICS

Rarely the case that cost is normally distributed, issues of skew and large amounts of variation

LOGNORMAL DISTRIBUTION



1

LOGNORMAL DISTRIBUTION

Skewed Distribution

Bounded Below by Zero

2

CONNECTION TO NORMAL

If *x* is Lognormally Distributed, *y* = *ln*(*x*) is Normally Distributed

3

THE DEVIL IS IN THE (DE)TAILS

Lognormal – Heavier Right Tail than the Normal Distribution

4

COST ANALYTICS

Better Alternative for Cost Modeling than the Normal

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

- Maximum Likelihood Estimation is a widely used statistical technique that serves as a framework for the CER methods we discuss
- Let $A_1, \ldots, A_n$ represent the observed data and $X_1, \ldots, X_n$ represent random variables where A_i results from observing the random variable X_i
- The likelihood function, which represents the likelihood of obtaining the sample results, is

$$L(\theta) = \prod_{i=1}^{n} Pr(X_i = A_i \mid \theta)$$

- The Maximum Likelihood Estimate of q is the vector that maximizes the likelihood function
- Maximum Likelihood Estimation is an established popular statistical technique
  - Major advantage – likelihood function is almost always available

APPLICATION OF MLE

- Applying MLE to lognormal residuals yields Log-transformed Ordinary Least Squares (LOLS), which minimizes:

$$\sum_{i=1}^{n}\left(\ln y_i - \ln \beta_0 - \beta_1 \ln X_{i1} - \ldots - \beta_p \ln X_{ip}\right)^2$$

- LOLS is easy to calculate
- Popular technique
- Log-scale trendline in Excel plots
- However it has issues:
  - Estimates the median vice the mean
  - Results in estimates that are biased low
  - This is an issue since estimates are typically added to other estimates in budget formulation

LOLS ALTERNATIVES

1. ZMPE ("Zimpy")

OVERCOMING BIAS

Dr. Steve Book Developed the Zero-Bias Minimum Percent Error (ZMPE) Method as an Alternative to LOLS

OBJECTIVE

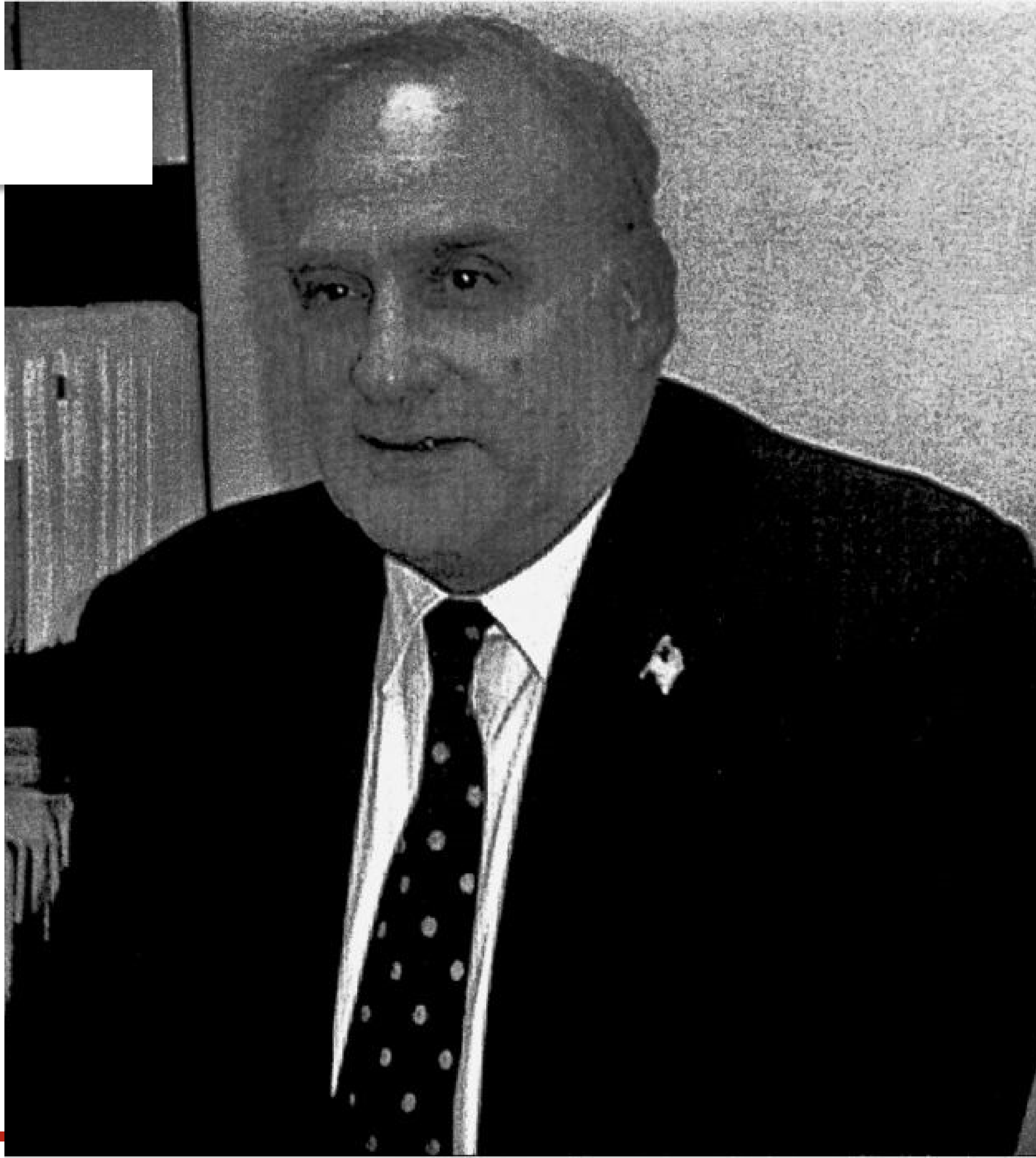
Minimize Squared Percent Error

$$\sum_{i=1}^{n}\left(\frac{y_i - f(x_i,\beta)}{f(x_i,\beta)}\right)^2$$

CONSTRAINT

Objective is minimized subject to a bias constraint:

$$\sum_{i=1}^{n}\left(\frac{y_i - f(x_i,\beta)}{f(x_i,\beta)}\right) = 0$$



LOLS ALTERNATIVES

2. MRLN ("Merlin")

PARAMETRIC EVIDENCE

In one of the presenters' experience, the lognormal distribution fits residuals well for spacecraft and defense cost

METHOD

Apply MLE to Estimate the Mean of the Lognormal – Maximum likelihood Regression of Log Normal error (MRLN)

OBJECTIVE

Minimize

$$\frac{n}{2}ln\theta + \frac{1}{2\theta}\sum_{i=1}^{n}\left(lny_i - ln\beta_0 - \sum_{j=1}^{p}\beta_j lnX_{ij} + \frac{\theta}{2}\right)^2$$

Where n = number of data points, θ = the log-space sample variance

LOLS ALTERNATIVES

3. ZMAPE

HEAVY TAILS

Many financial phenomena have heavy tails

Mixed evidence of this for spacecraft

If so, no reason to minimize variance – it could be infinite!

METHOD

Minimize Absolute Percent Error

$$\sum_{i=1}^{n}\left|\frac{y_i - f(x_i,\beta)}{f(x_i,\beta)}\right|$$

CONSTRAINT

Objective is minimized subject to a bias constraint:

$$\sum_{i=1}^{n}\left(\frac{y_i - f(x_i,\beta)}{f(x_i,\beta)}\right) = 0$$

GOODNESS OF FIT METRICS

Changes in Metric Calculations

R^2

Traditional linear regression R^2 has a cross-term that vanishes when the regression equation is linear; it does not vanish in the nonlinear case and can lead to negative R^2 values when the equation is nonlinear. Instead we use “Pearson’s R^2,” which is the square of the correlation coefficient between the actual and estimated costs

Standard Error

Linear regression calculates the standard deviation of the differences between the actual and estimated costs. We use “standard percent error” which is the standard deviation of the difference between the actual and estimated costs as a percentage of the estimated costs

Bias

Bias is always zero for a linear regression, the bias is negative for LOLS. We calculate bias as the average percentage error rather than the absolute error

APPLICATION TO SW SUSTAINMENT DATA

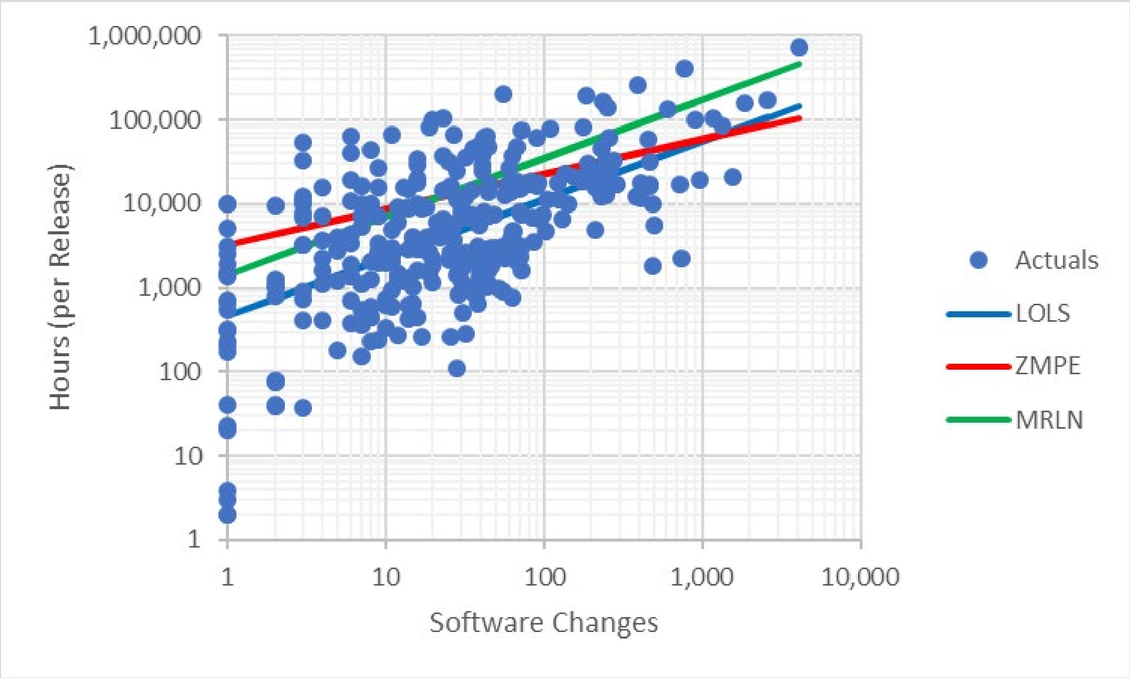
OVERVIEW

- Bias is a significant issue for all LOLS CERs – ranged from -33% to -200%
- MRLN, ZMPE, and ZMAPE all eliminate bias
- MRLN, ZMPE, and ZMAPE also significantly reduce the standard percent errors
- Recommended CER is highlighted in yellow
- For MRLN and LOLS, checked residuals to see if they are lognormal
- For MRLN, if residuals did not fit a lognormal, calculated Zero-percent bias Minimum Absolute Percent Error as well (4/11 CERs)
- The log-space R^2s differ from the recommended Pearson's R^2 in all cases – in one case, it calls into question the significance of the CER (SW Baseline size 61% vs. 35%)
- Looking at actuals vs. estimates plot, there is a tendency across most CERs to overestimate smaller effort (< 1,000 hours) – recommend segmenting the data

VARIABLES

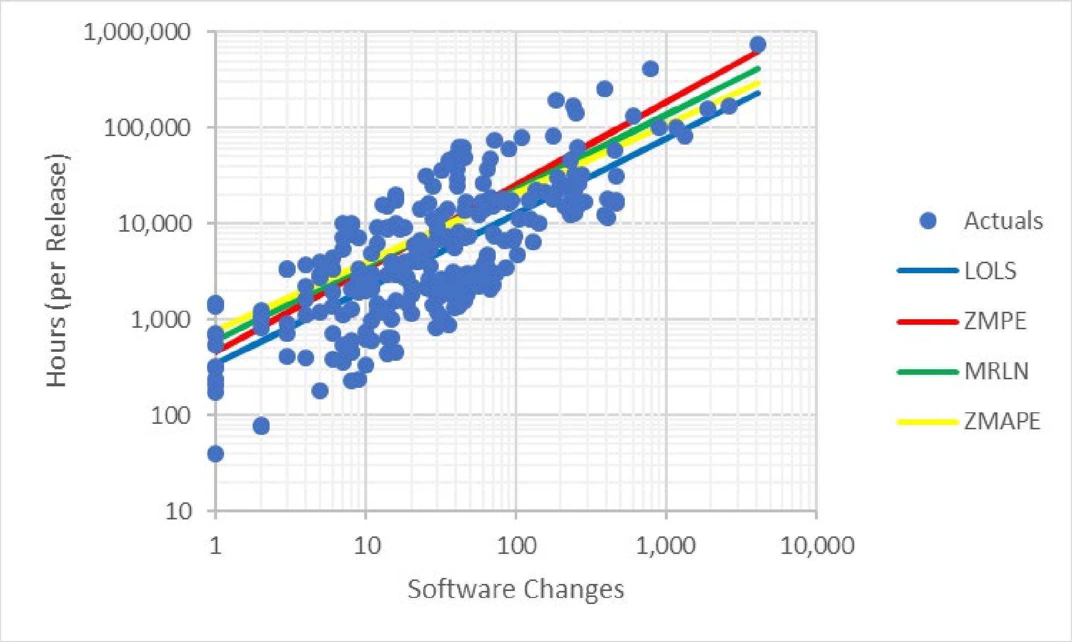
- Dependent variable is effort in hours per software release
- Data were trimmed – lowest and highest 10% od data points in terms of hours/software change were trimmed
- Primary independent variable is software changes
- Other independent variables considered include:
  - Super Domain
  - Commodity
  - Total System Requirements
  - Total Requirements Implemented
  - External Interfaces Modified
  - SW Baseline Size
  - Backlog
  - Change Type
  - % Change Type

ALL DATA



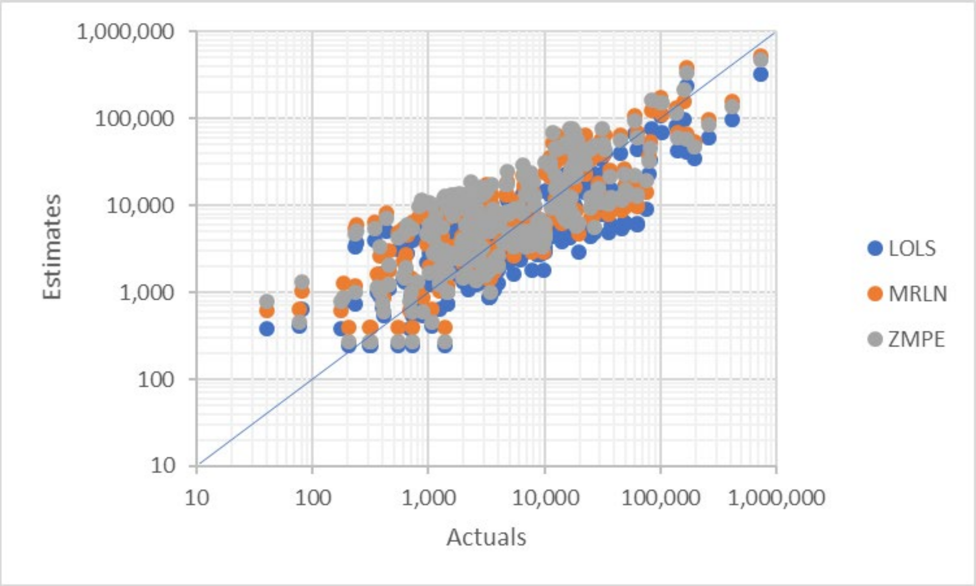
Method	Int	SC	R^2	SPE	Bias
LOLS	462.5	0.6929	44.27%	638.81%	-209.66%
MRLN	1432.3	0.6929	44.27%	194.79%	0.00%
ZMPE	3298	0.4177	33.66%	171.95%	0.00%

TRIMMED DATA



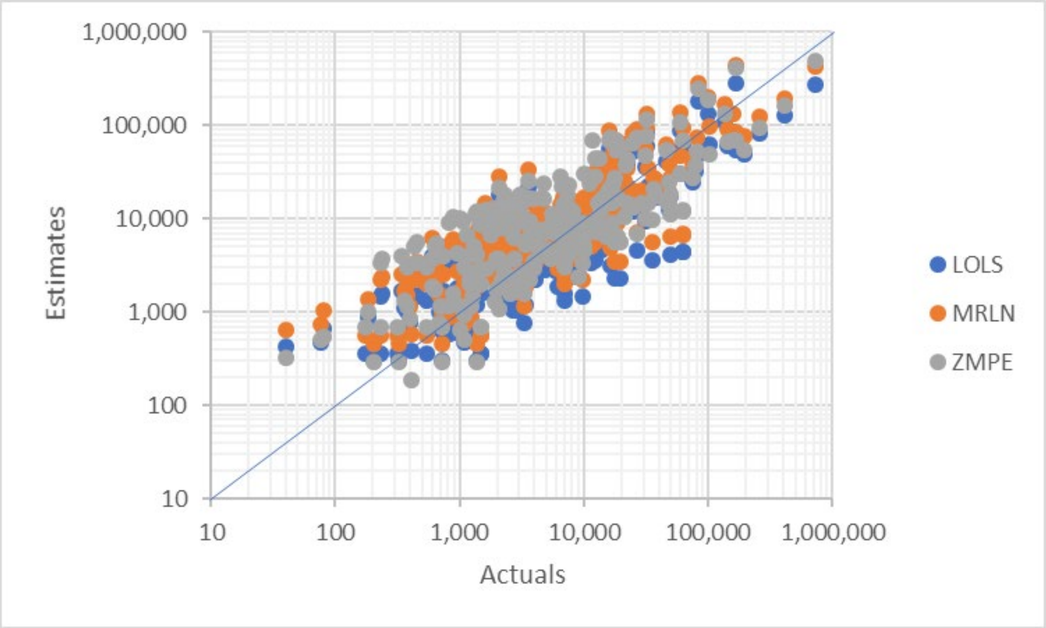
Method	Int	SC	R^2	SPE	Bias
LOLS	340.7	0.7858	61.75%	215.63%	-75.82%
MRLN	599	0.7858	61.75%	114.75%	0.00%
ZMPE	453.7	0.8713	63.13%	113.46%	0.00%
ZMAPE	753.1	0.7192	60.24%	117.53%	0.00%

TRIMMED DATA – SW Changes and Super Domain



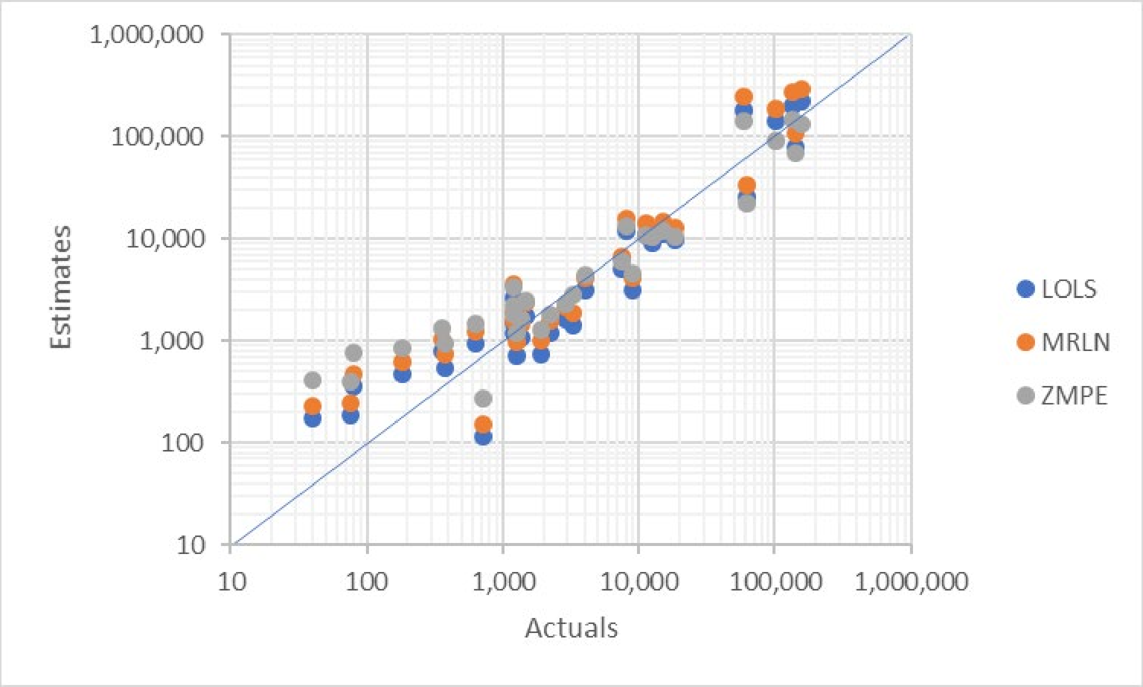
Method	AIS	ENG	RT	SUP	SC	R^2	SPE	Bias
LOLS	242.1	386	735.7	698.7	0.7341	71.33%	187.20%	-61.51%
MRLN	391.1	623.5	1188.2	1128.4	0.7341	71.33%	109.34%	0.00%
ZMPE	268.2	798.3	994.2	731.2	0.741	66.64%	101.91%	0.00%

TRIMMED DATA – SOFTWARE CHANGES AND COMMODITY TYPE



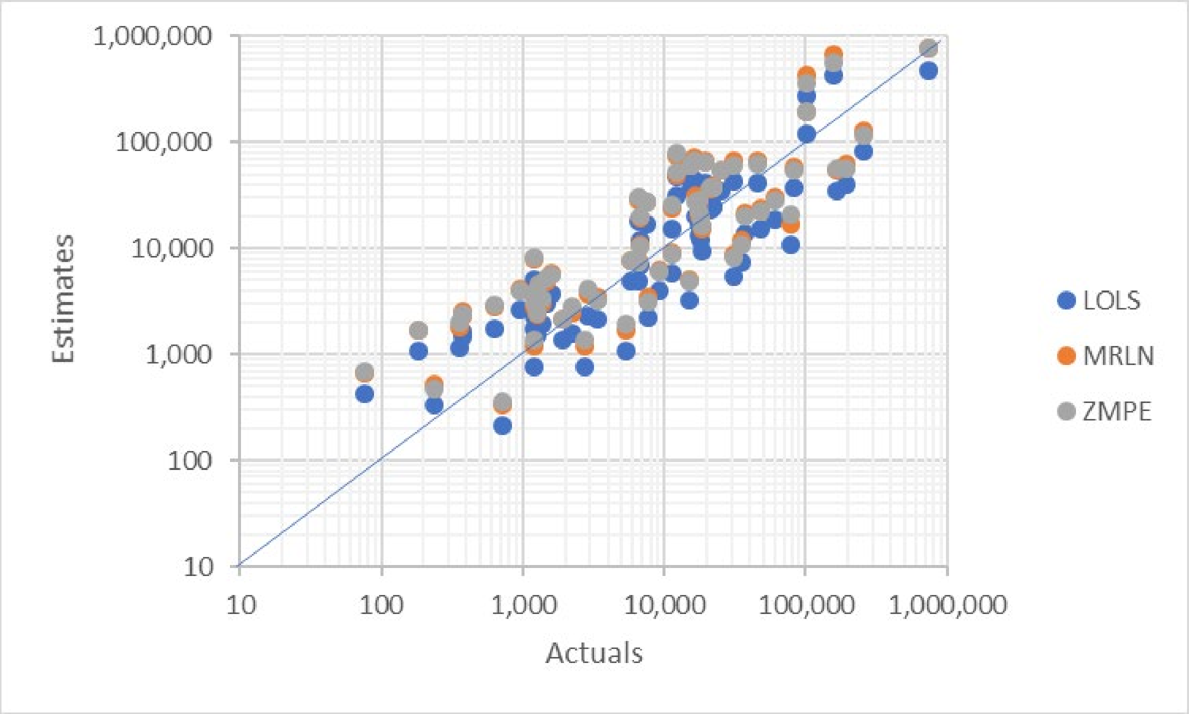
Method	Aviation	Business	C4ISR	ChemBio	Fire	Missiles	Simulation	Space	Test	Vehicles	SC	R^2	SPE	Bias
LOLS	1450.5	301.4	363.5	182.4	1530.9	1114.1	576.4	1005.3	1740.8	424.6	0.6645	60.83%	197.87%	-53.81%
MRLN	2229.1	463.6	559.2	280.7	2353.6	1717.9	886.4	1545.9	2682.2	653.0	0.6645	60.92%	123.58%	0.00%
ZMPE	1027.8	294.3	680.1	79.8	976.9	849.0	216.9	760.3	774.0	319.8	0.7671	63.30%	99.83%	0.00%

TRIMMED DATA – SOFTWARE CHANGES AND TOTAL SYSTEM REQUIREMENTS



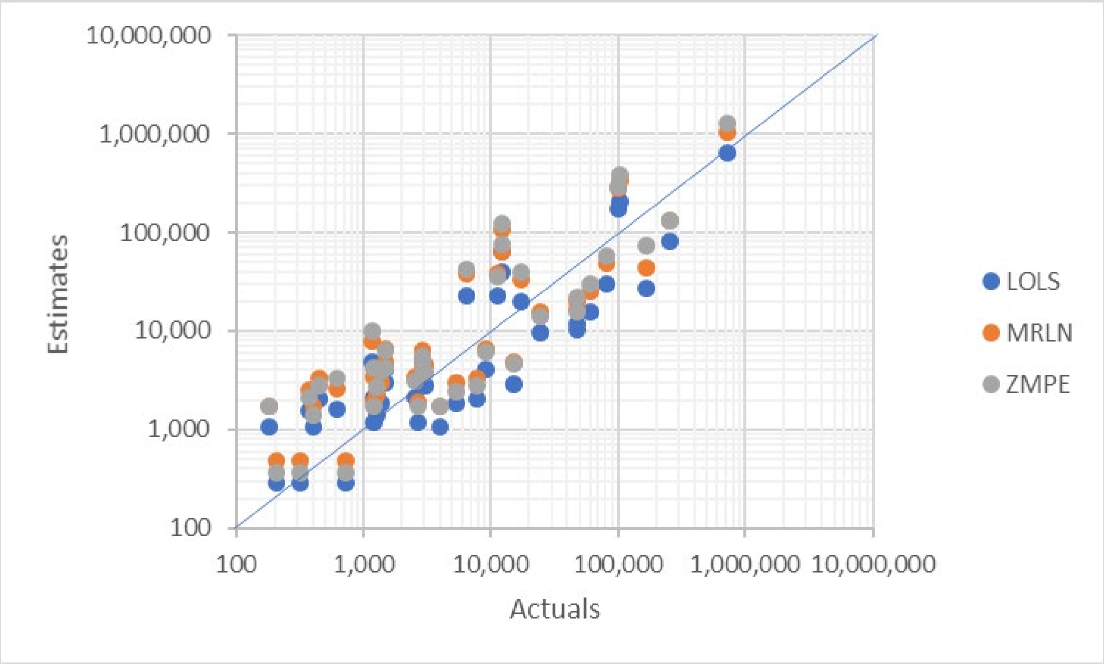
Method	Int	SC	Tot Req	R^2	SPE	Bias
LOLS	608.7	0.9807	-0.2111	78.73%	125.62%	-33.55%
MRLN	813	0.9807	-0.2111	78.73%	90.29%	0.00%
ZMPE	1466.0	0.7971	-0.2120	77.50%	71.21%	0.00%

TRIMMED DATA – SOFTWARE CHANGES AND TOTAL REQUIREMENTS IMPLEMENTED



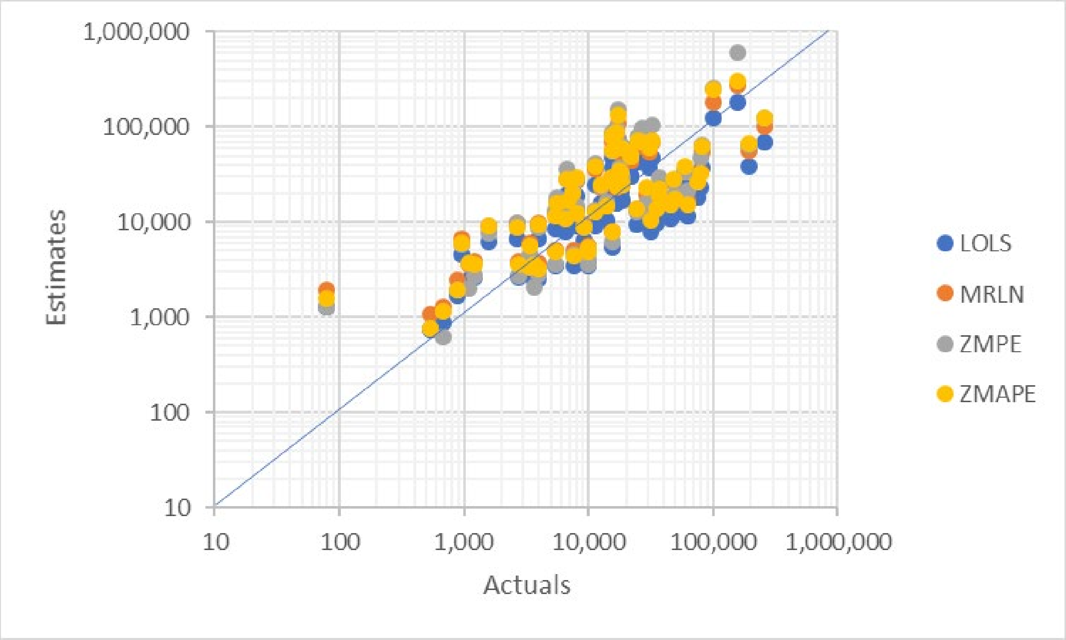
Method	Int	SC	Tot Req	R^2	SPE	Bias
LOLS	330.5	0.9671	-0.1085	61.29%	173.11%	-58.08%
MRLN	522.5	0.9671	-0.1085	61.29%	102.84%	0.00%
ZMPE	474.1	0.9496	-0.0706	68.81%	101.66%	0.00%

TRIMMED DATA – SOFTWARE CHANGES AND EXTERNAL INTERFACE MODIFIED



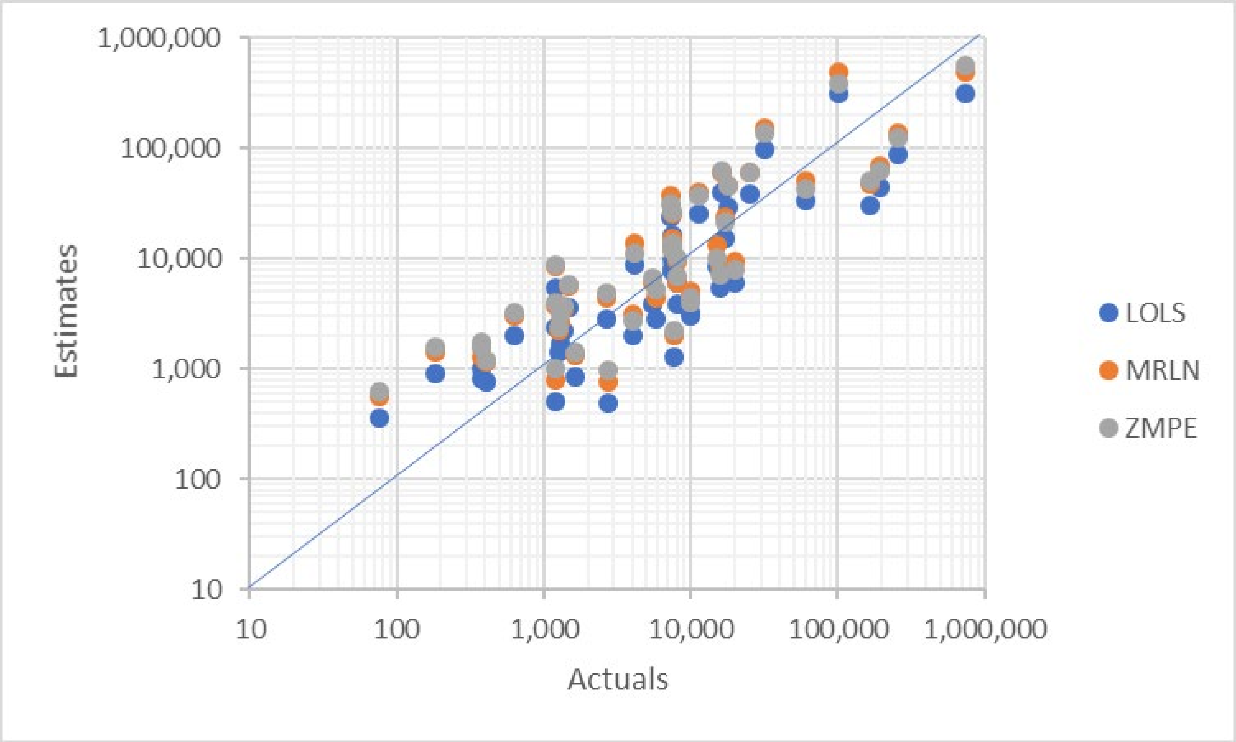
Method	Int	SC	EI Mod	R^2	SPE	Bias
LOLS	292.5	0.9386	-0.1076	85.80%	174.57%	-62.61%
MRLN	475.6	0.9386	-0.1076	85.80%	99.63%	0.00%
ZMPE	360.3	0.9851	-0.01364	86.97%	94.81%	0.00%

TRIMMED DATA – SOFTWARE CHANGES AND SW BASELINE SIZE



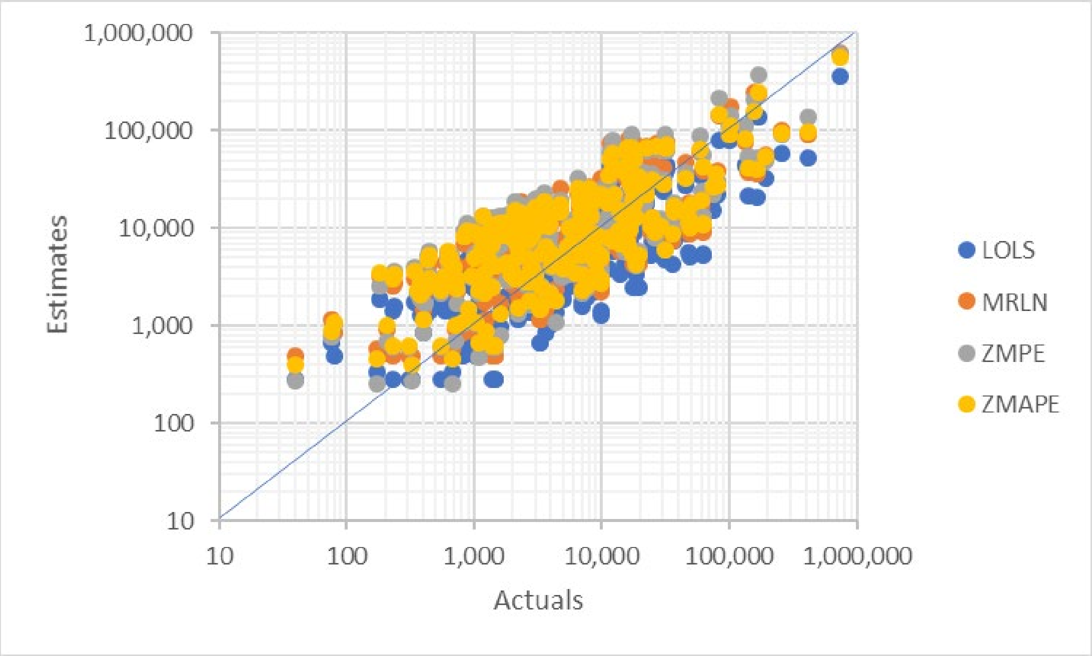
Method	Int	SC	SW BL Size	R^2	SPE	Bias
LOLS	1219.5	0.7465	-0.0368	34.87%	141.10%	-46.55%
MRLN	1787.1	0.7465	-0.0368	34.87%	90.64%	0.00%
ZMPE	400.2	0.8691	0.0460	28.92%	86.98%	0.00%
ZMAPE	2660.4	0.8228	-0.0894	35.77%	91.85%	0.00%

TRIMMED DATA – SOFTWARE CHANGES AND BACKLOG



Method	Int	SC	SW BL Size	R^2	SPE	Bias
LOLS	756.6	1.0178	-0.3631	56.93%	165.07%	-55.67%
MRLN	1177.7	1.0178	-0.3631	56.92%	99.37%	0.00%
ZMPE	909.9	0.9841	-0.2633	71.23%	97.01%	0.00%

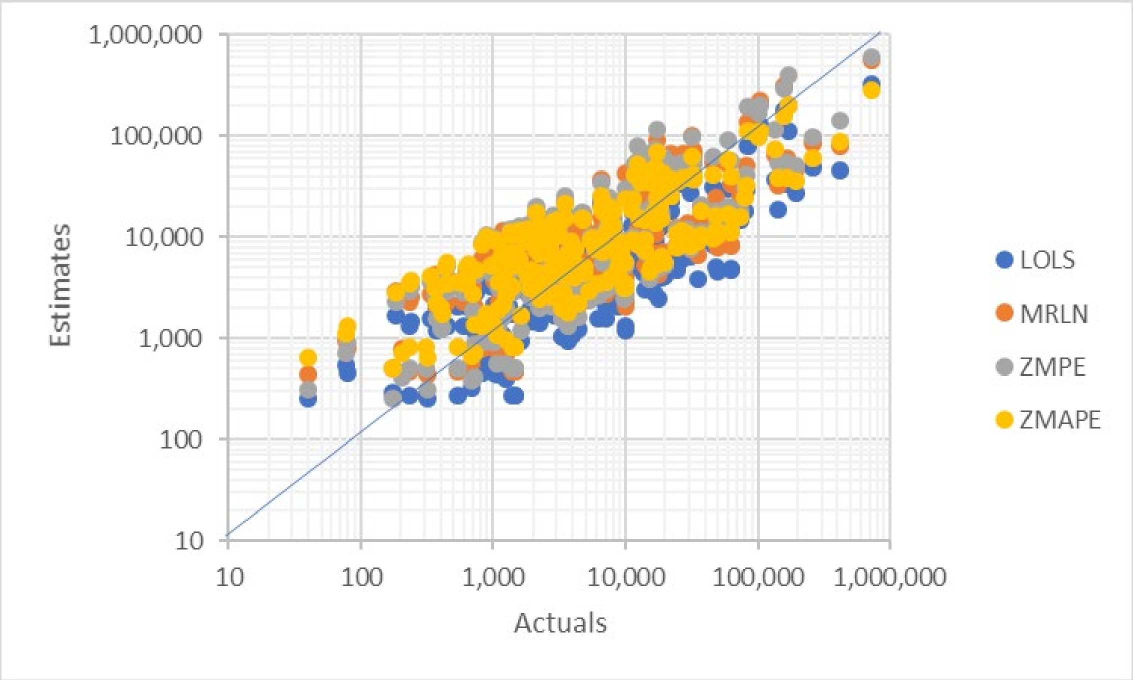
TRIMMED DATA – SOFTWARE CHANGES AND CHANGE TYPE



Method	C	E	H	M	O	SC	R^2	SPE	Bias
LOLS	331.6	530.5	382.2	281.1	284.6	0.7864	70.70%	221.14%	-73.97%
MRLN	573.8	923.9	666.0	489.5	493.7	0.7861	70.69%	119.60%	0.00%
ZMPE	252.2	530.5	382.2	281.1	284.6	0.8187	68.57%	109.12%	0.00%
ZMAPE	459.5	1020.0	525.6	630.7	398.0	0.7589	73.04%	112.34%	0.00%

GALORATH

TRIMMED DATA – SOFTWARE CHANGES AND CHANGE TYPE %



Method	Int	E	M	C	O	SC	R^2	SPE	Bias
LOLS	337.8	0.0945	0.0218	0.0298	0.0136	0.7678	64.22%	227.90%	-73.88%
MRLN	587.4	0.0945	0.0218	0.0298	0.0136	0.7678	64.22%	123.82%	0.00%
ZMPE	375.1	0.0391	0.0675	-0.0308	-0.0021	0.8187	65.30%	112.88%	0.00%
ZMAPE	687.2	0.0271	0.0427	-0.0239	0.0119	0.7020	61.94%	117.22%	0.00%

Residual Fits for MRLN CERs

MRLN CERS

- MRLN is optimal when the residuals are lognormally distribution
- Residuals are defined as:

$$Actual = Estimate * \varepsilon$$

so

$$\varepsilon = \frac{Actual}{Estimate}$$

- We test each CER fit with MRLN to examine its residuals

RESIDUAL STATISTICS

CER	Mean	St. Dev	Skew	Kurtosis	N
All Data	1.0	1.9	4.3	26.4	329
All - Trimmed	1.0	1.1	1.9	5.9	263
Super Domain	1.0	1.1	2.3	9.4	263
Commodity	1.0	1.2	3.9	22.9	263
Total Reqmts	1.0	0.9	2.5	11.4	32
Total Reqmts Imp	1.0	1.0	1.6	4.8	65
EI Mod	1.0	1.0	1.2	3.4	41
SW Baseline Size	1.0	0.9	1.3	3.5	69
Backlog	1.0	1.0	1.4	4.3	45
Change Type	1.0	1.2	2.2	8.1	263
Change Type %	1.0	1.2	2.5	10.3	263

DISTRIBUTIONS CONSIDERED

- Lognormal – key assumption we need to test
- Exponential – cursory visual examination of residuals tends to look exponential
- Gamma – flexible distribution that can resemble a lognormal
- Weibull – flexible distribution with 3 parameters

GOODNESS-OF-FIT

- Anderson-Darling
  - Goodness-of-fit test
  - Emphasis is on detecting goodness-of-fit in the tails (weighted)
  - Good at detecting departure from Gaussian (normal) distribution, and by extension, lognormality
- Comparing Distributions
  - Bayesian Information Criterion
  - Lognormal is Better Than All Others Considered for All CERs
    - Battle for second – Exponential finished second in 6/11 fits

GALORATH

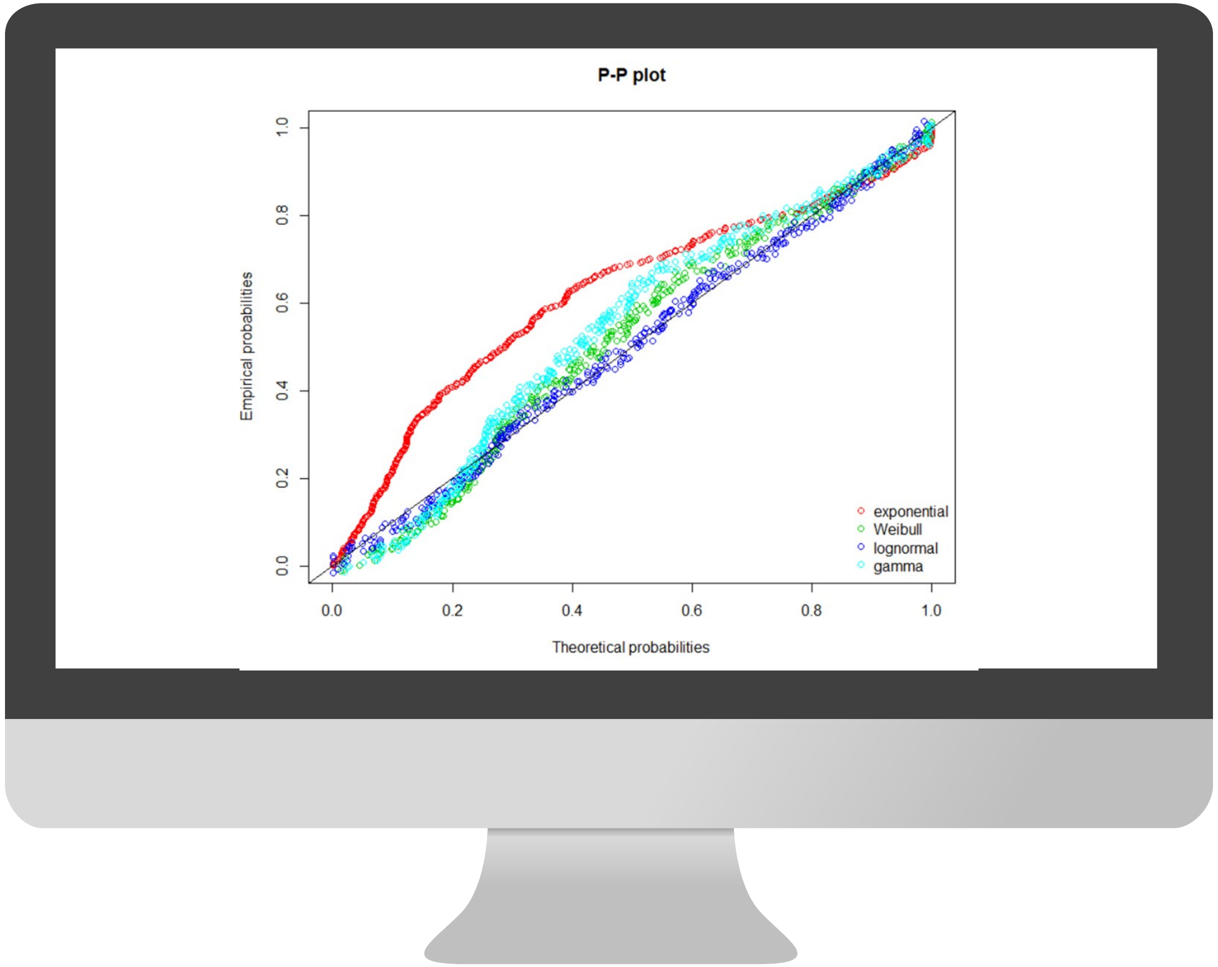
RESIDUAL RESULTS

- Cannot reject assumption of lognormality for 7/11 CERs
- Lognormal better fit than other distributions considered
- If lognormality rejected, recommend non-parametric CER – ZMPE or ZMAPE

			Bayesian Information Criterion Rank			
CER	P-value	Result	Lognormal	Exponential	Gamma	Weibull
All Data	0.08	Do Not Reject	1	4	3	2
All - Trimmed	4.0x10^-4	Reject	1	4	3	2
Super Domain	0.28	Do Not Reject	1	2	3	4
Commodity	0.36	Do Not Reject	1	4	2	3
Total Reqmts	0.79	Do Not Reject	1	4	2	3
Total Reqmts Imp	0.12	Do Not Reject	1	2	3	4
EI Mod	0.11	Do Not Reject	1	2	3	4
SW Baseline Size	0.04	Reject	1	3	2	4
Backlog	0.11	Do Not Reject	1	2	3	4
Change Type	1.8x10^-4	Reject	1	2	4	3
Change Type %	1.2x10^-5	Reject	1	2	4	3

GRAPHICAL COMPARISON EXAMPLE – ALL DATA

WHICH IS THE BEST?



MRLN Significance Testing

MRLN SIGNIFICANCE - THEORY

MRLN is a Maximum Likelihood Estimation (MLE) method

- **Assumption is that residuals are lognormally distributed**
- **Use Likelihood Ratio Test to compare the improvement in likelihood in going from the null hypothesis to the regression model**

Likelihood Function for Lognormal

$$\prod_{i=1}^{n} p(y_i|x_i;\beta_0,\beta_1) = \prod_{i=1}^{n} \frac{1}{y_i\sqrt{2\pi\theta}} e^{-\frac{(\ln y_i - \ln\beta_0 - \beta_1 \ln x_i)^2}{2\theta}}$$

MRLN SIGNIFICANCE - IMPLEMENTATION

Likelihood Ratio Test

- Calculate the likelihood of the regression model and the likelihood of the null model using the lognormal distributions

  Then calculate the ratio of these two

Chi-Square Test

- -2log(Null Likelihood/Regression Likelihood) follows a Chi-square distribution with one degree of freedom

  Use this to calculate a p-value

  Implemented in Excel

MRLN SIGNIFICANCE - RESULTS

Regression Significance

- All MRLN regressions are statistically significant
- Assumes lognormal residuals – not valid for four CERs

Variable Significance

- Calculated variable significance incrementally (Regression with one variable vs. regression with two, etc.)
- All variables/sets are significant EXCEPT Total Requirements Implemented

ZMPE/ZMAPE Significance Testing

INTRODUCTION

Non-parametric

- Significance testing for ZMPE/ZMAPE is harder than for MRLN as there is no underlying distribution assumed

Theory

- If nothing is known about the distribution, there is no effective hypothesis testing

  - Bahadur and Savage, Annals of Mathematical Statistics, 1956

Practice

- Can provide estimates using bootstrapping

  -Feldman, ISPA-SCEA Annual Conference, 2010

BOOTSTRAPPING

Bootstrapping

Involves resampling the residuals – akin to pulling yourself up by your bootstraps

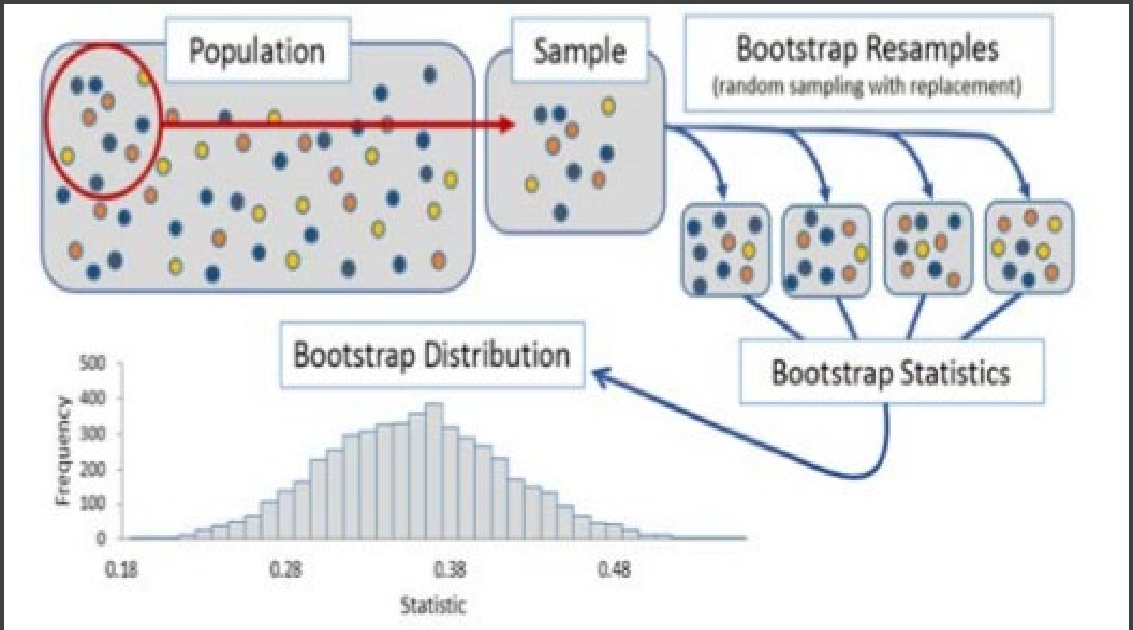
T-like Test

End result will be the calculation of a t-like statistic from the resampled residuals

T-test on the slope coefficient not being equal to zero (equivalent to F-test)

R Implementation

The method is implemented in the R statistical programming language



BOOTSTRAP DETAILS

- Develop ZMPE/ZMAPE solution
- Resample with replacement *N* times from ZMPE/ZMAPE residuals
- For each of the *N* resamples resample *M* times
- Use the outer loop to calculate the standard deviation of the slope coefficient (s_b)
- The inner loop is used to calculate a series of t-like values
- The critical value is b_1 / s_b
- Calculate the number of t-like values that are greater than the critical value, call this #
- The p-value is #/*M*
- If #/*M* <=*0.05*, reject the null hypothesis (regression is significant); otherwise, regression is not significant

SIGNIFICANCE TESTING

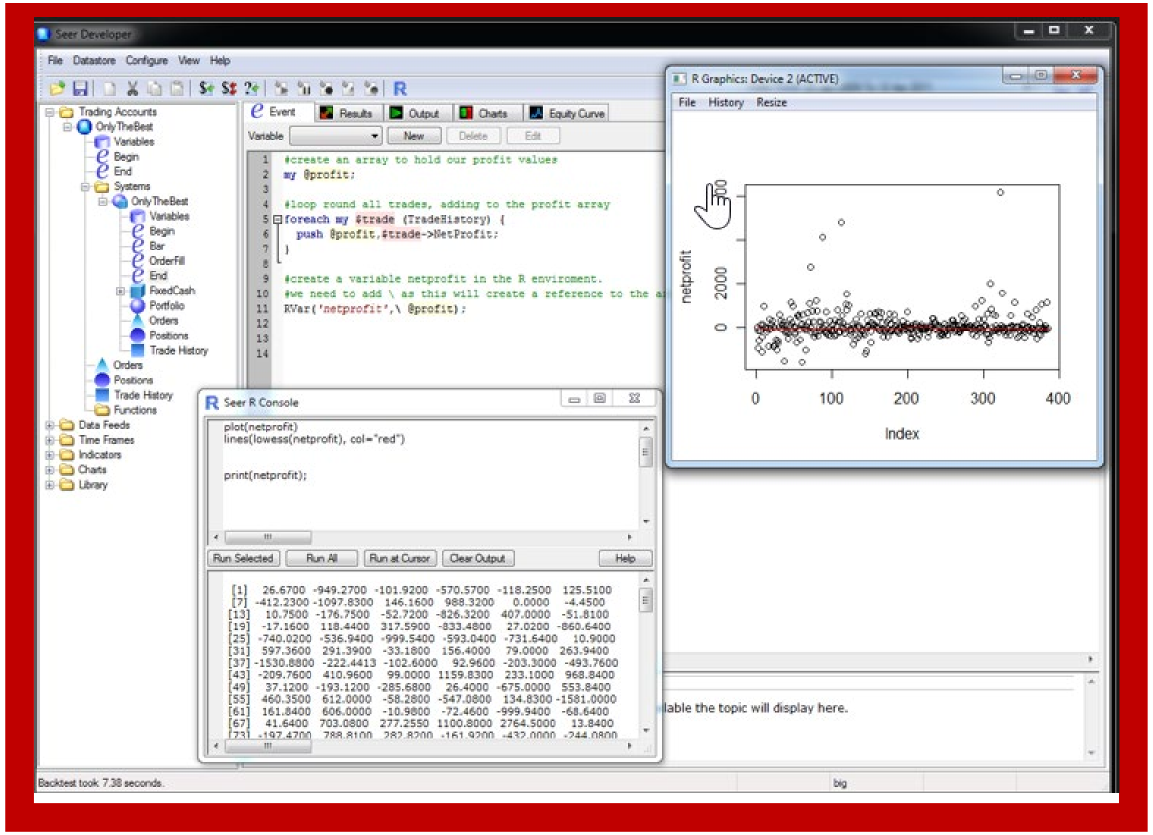
Illustrating the Process

b	$\hat{\beta}_0^*$	σ^*	t_b^*
1	4.257	1.802	0.133
2	2.990	1.459	-0.705
3	5.668	1.421	1.161
4	3.645	1.724	-0.216
5	2.746	0.952	-1.335
6	3.136	1.109	-0.795
7	4.036	1.102	0.016
8	5.699	1.736	0.969
9	4.340	1.514	0.213
10	3.429	1.693	-0.348
⋮	⋮	⋮	⋮
500	2.047	1.524	-1.293

$$t_5^* = \frac{(2.746 - 4.018)}{0.952} = -1.335$$

GALORATH

IMPLEMENTATION IN *R*



R Programming Language

- Freely available platform with a large number of pre-built packages for doing statistical analysis
- One of the leading tools for machine learning and data science

Rsolnp Package

- Leveraged the rsolnp package to calculate ZMPE
- Used built-in resampling capability in Base R for bootstrap

Distribution Fits

- Also added capability to check that the residuals are lognormally distributed using the *fitdistrplus* package.

GALORATH

REFERENCES

- Bahadur, R.R., and L.J. Savage, "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems," The Annals of Mathematical Statistics, 1956
- Book, S.A., and N.Y. Lao, "Deriving Minimum-Percentage-Error CERs Under Zero-Bias Constraints," The Aerospace Corporation, El Segundo, CA, July 1996
- Feldman, D., "Testing for the Significance of Cost Drivers Using Bootstrap Sampling," presented at the 2010 ISPA-SCEA Annual Conference
- Smart, C., "Cutting the Gordian Knot: Maximum Likelihood Estimation for Regression of Log Normal Error," presented at the 2017 ICEAA Annual Conference

BACKUP

MLE – LOGNORMAL RESIDUALS

- For $Y_i = f(X_i, \beta)\varepsilon_i$, where
  β = vector of coefficients of the CER
  Y_i = actual cost of the i^{th} data point
  X_i = vector of cost drivers for the i^{th} data point
  ε_i = residual of the i^{th} data point
- Probability density function for lognormal distribution

$$p(y, \mu, \theta) = \frac{1}{y\sqrt{2\pi\theta}} e^{-\frac{(\ln y - \mu)^2}{2\theta}}$$

- Note that μ is the log-space mean
- If we estimate $\boldsymbol{\mu} = \boldsymbol{ln}(\boldsymbol{Y})$ then in the case of the power equation

$$\boldsymbol{Y} = \boldsymbol{\beta_0} \boldsymbol{X}^{\boldsymbol{\beta_1}}$$

  we are estimating the linear equation $\boldsymbol{\mu} = \boldsymbol{ln\beta_0} + \boldsymbol{\beta_1 ln}(\boldsymbol{X})$
- Note that $\boldsymbol{e^{\mu}} = \boldsymbol{Y} = \boldsymbol{\beta_0 X^{\beta_1}}$ is the median in linear space