SCEA

The Society of Cost Estimating and Analysis

ispa

An Analytical Framework for CER Development: A Good Fit is Not Good Enough!

Presented by:
Christian Smart, Ph.D.
George Culver
Science Applications International Corporation

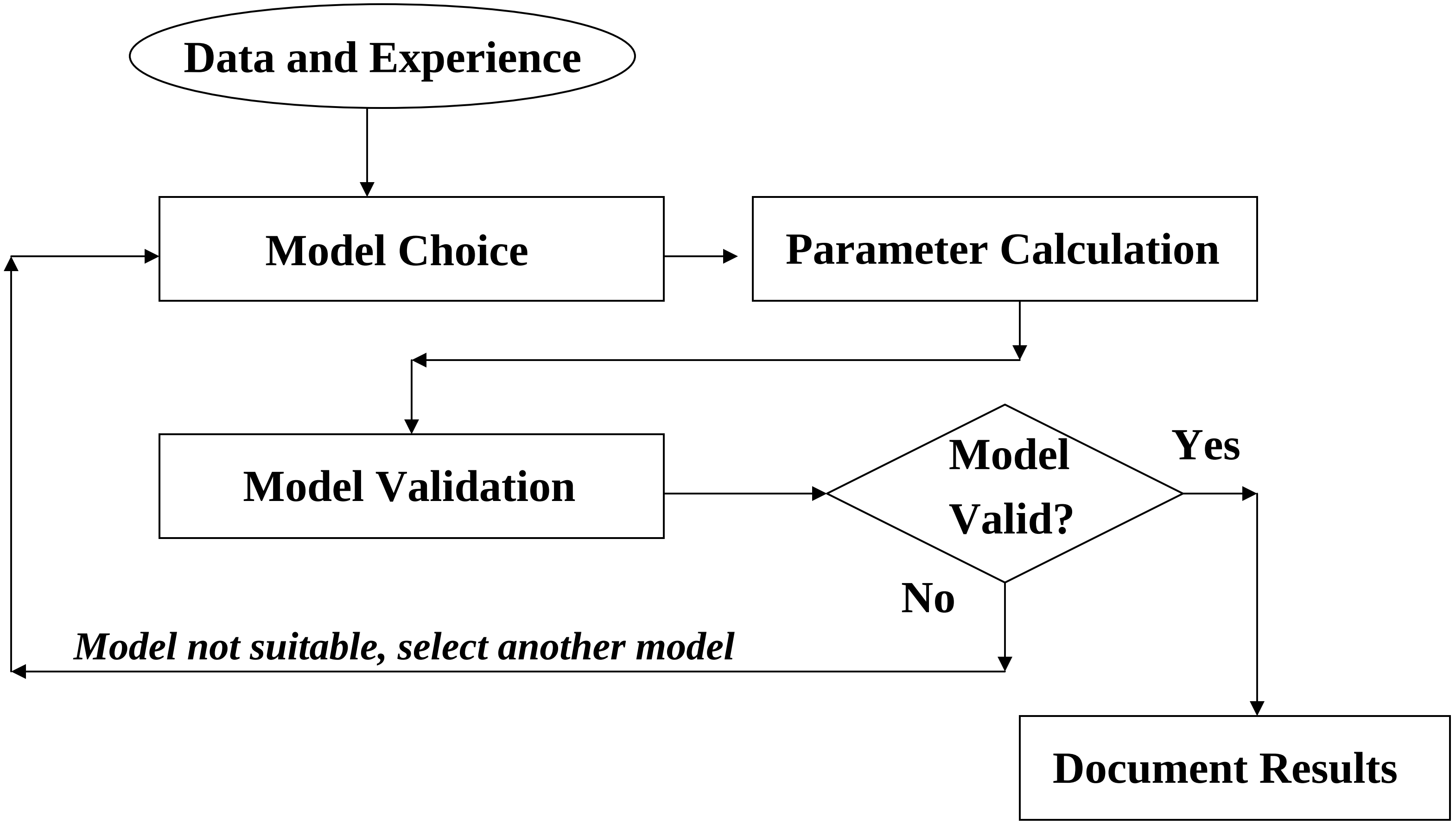
Acknowledgements

- **The Authors are Indebted to the Work of Numerous Other Cost Analysts who Have Pioneered Cost Estimating Relationship Theory and Practice, Including:**
  - **Dr. Steve Book, MCR**
  - **Mr. Pierre Foussier, 3f**
  - **Dr. Shu-Ping Hu, Tecolote Research**
  - **Mr. Don Mackenzie, Wyle Labs**
  - **Dr. Matt Goldberg, Institute for Defense Analysis (IDA)**

Agenda

- Cost Estimating Relationships (CERs)
  - Regression Analysis
    - Linear Regression
    - Power Equation
    - Nonlinear Regression
      - Additive Vs. Multiplicative Residuals
  - Maximum Likelihood Estimation
  - Three Popular CER Methods
    - Log-Transformed Ordinary Least Squares
    - Minimum Unbiased Percent Error (MUPE)
    - Minimum Percent Error/Zero Percent Bias Constraint
  - Examples Comparing The Three Approaches
  - CER Development in the Context of a Parametric Model Development Framework

Model Development Framework



Data and Experience

Model Choice

Parameter Calculation

Model Validation

Model Valid?

Yes

No

Model not suitable, select another model

Document Results

This Framework Applies to CER Development in Particular and Mathematical Modeling in General.

Adapted from *Loss Models*, 3rd. Edition

Model Choice

- ***Goal: Use Historical Data to Accurately Predict Cost of Planned Programs.***
- **It Is Important When Developing Models to Limit Our Choices, Since Given Enough Models To Choose From, There Will be at Least One Model That Appears to Fit the Data Well, but Will Not Help us Effectively Predict Future Cost.**
  - **Experience is a Useful Guide in Limiting the Universe of Choices.**
- **In This Section, We Limit Our Choices to Statistical Methods ("Regression Analysis"), Nonlinear Regression, and Multiplicative Residuals.**
  - **Explanations are Given for Reasons why These Choices are Made in the Following Charts.**

Cost Estimating Relationships

- Cost Estimating Relationships (CERs) are One Way to Discern Trends from Historical Data in Order to Predict the Cost of Future Programs.
- CER Are Developed From Historical Data Using Statistical Techniques Such As Regression Analysis.
  - Regression Analysis Relates One or More Cost Drivers ("Independent" Variables) to Cost ("Dependent" Variables).
- Example:

$$\mathit{Estimated\ Cost} = 1.5 \cdot \mathit{Weight}^{0.5}$$

Where Weight is in Pounds (lbs.) and Cost is in Millions of US$.

  - When Weight = 30,000 lbs.:

$$\mathit{Estimated\ Cost} = 1.5 \cdot 30000^{0.5} \approx \$260\,\mathit{Million}$$

Linear Regression

- 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:

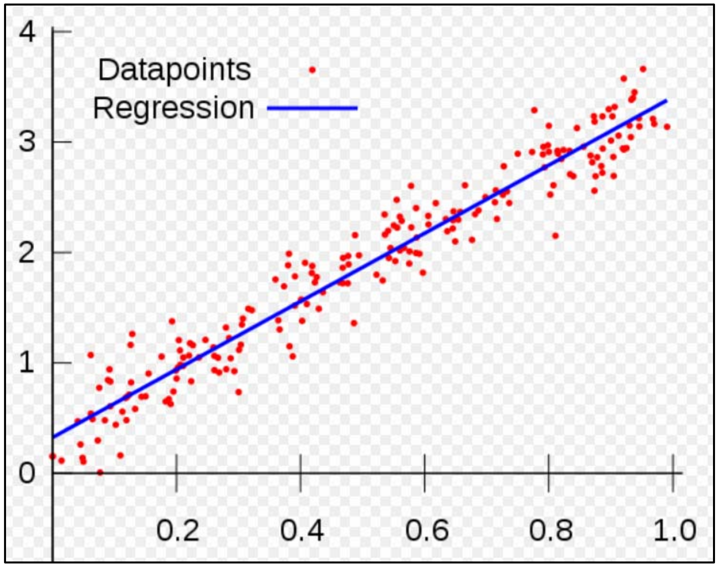
$$\varepsilon_i = Y_i - (a + bX_i) = Actual - Estimated$$

- 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.

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

Least Squares and Regression Analysis

- The Method of Least Squares was First Developed by the Mathematicians Legendre and Gauss in the Early 19th Century, Who Used it to Predict the Orbits of Heavenly Bodies Using Observed Data.
- Francis Galton Later Applied This Technique to Find Linear Predictive Relationships Between Various Phenomena, Such as the Relationship 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."



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.**

Additive and Multiplicative Residuals

- **The Residuals of the Power Equation Can Either Be Additive or Multiplicative.**
- **Additive Residuals Have the Form**

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

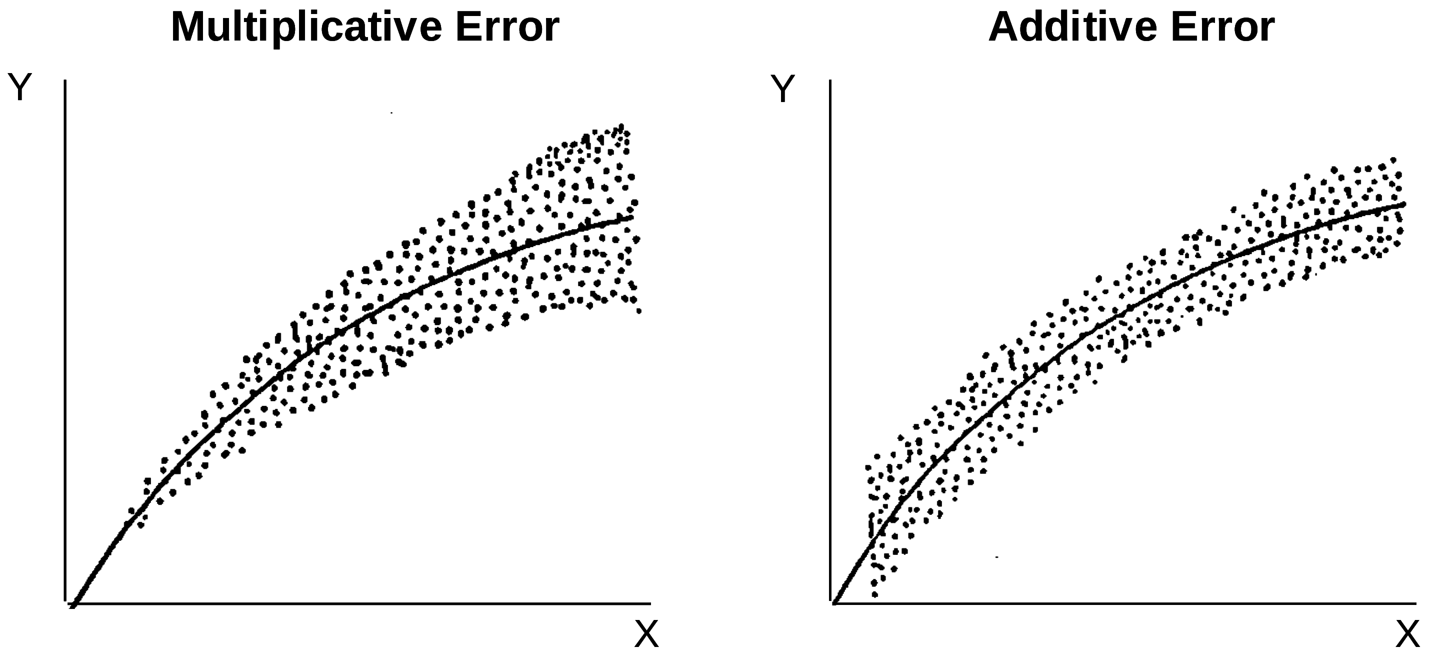
- **Multiplicative Residuals Have the Form**

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

- **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.**

Residuals Comparison

- **The Commonly-Used Regression Techniques Considered in This Presentation are all Based on the Multiplicative Error Assumption.**



The Focus of This Section is on Nonlinear Regression Methods for Equations of the Form Y =aX^b, With Multiplicative Residuals.

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**

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

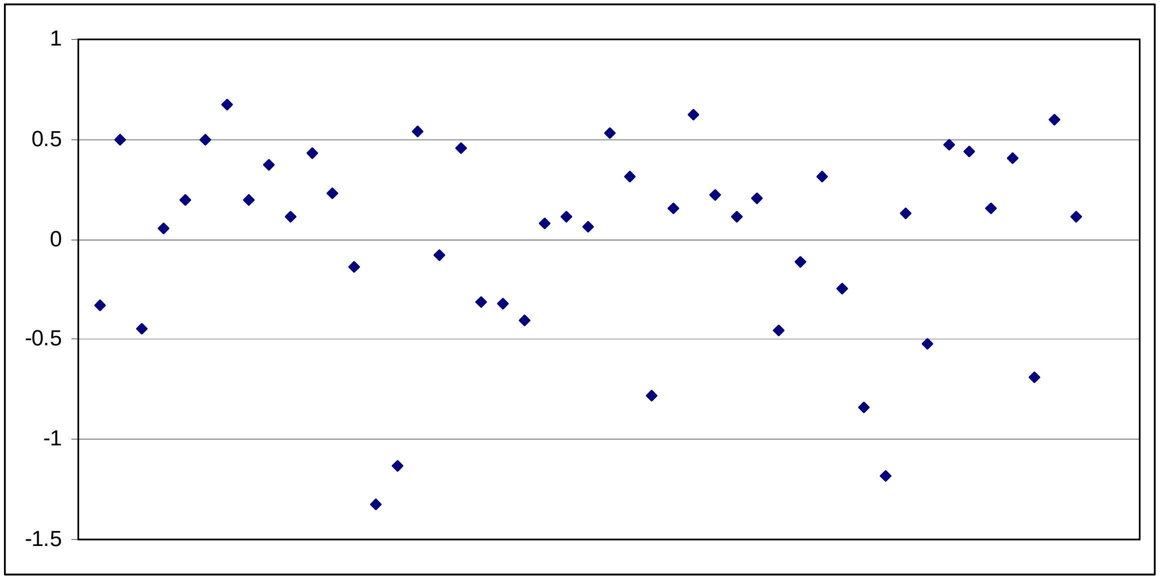
$$\varepsilon = \frac{aX^{b} - Y}{aX^{b}} = \frac{Estimate - Actual}{Estimate}$$

- **Actual Cost = Estimate +/- Percentage of Estimate**

Multiplicative Residual Example

- If the Estimate is Greater Than the Actual the Residual is Greater Than Zero.
- If the Estimate is Less Than the Actual the Residual is Less Than Zero.
- Note the Lack of Symmetry.
  - For Estimates Above the Actual, the Maximum Value of the Residual is *1*.
  - For Estimates Below the Actual, the Minimum Value Has No Bound!

Residuals for a Subsystem CER in the NASA/Air Force Cost Model (NAFCOM)

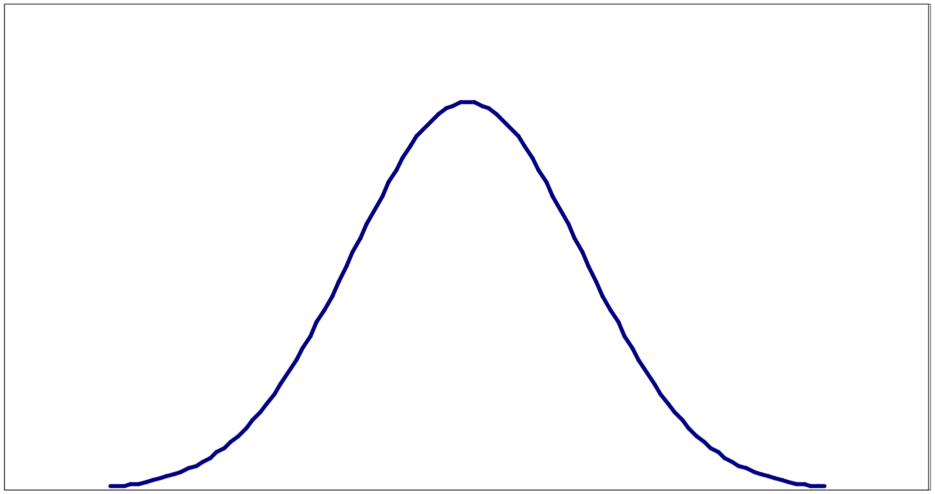


Residuals Are Random Variables

- **For a "Good" Regression Model, The Cost Drivers Explain All (or Most) of the Variation in the Historical Data That Can Be Explained.**
  - **It is Typically Assumed That any Remaining Variation is Random.**
    - **Either Due to Non-Repeatable Random Phenomena (e.g., Test Failures) That Are Truly Random Phenomena and Will Not Help Predict Future Cost, or Due to Our Ignorance.**
      - **Statistics Has Been Called "The Science of Ignorance."**
- **The Multiplicative Residuals That Represent This Unexplained Variation are Thus Treated as Random Variables.**
- **For Linear Regression, it is Assumed that the Additive Residuals are Normally Distributed.**
- **For Nonlinear Regression for CER Development, Residuals Assumed to Follow Normal, Lognormal, Gamma, or Treated without Making Such an Assumption (Non-Parametric).**

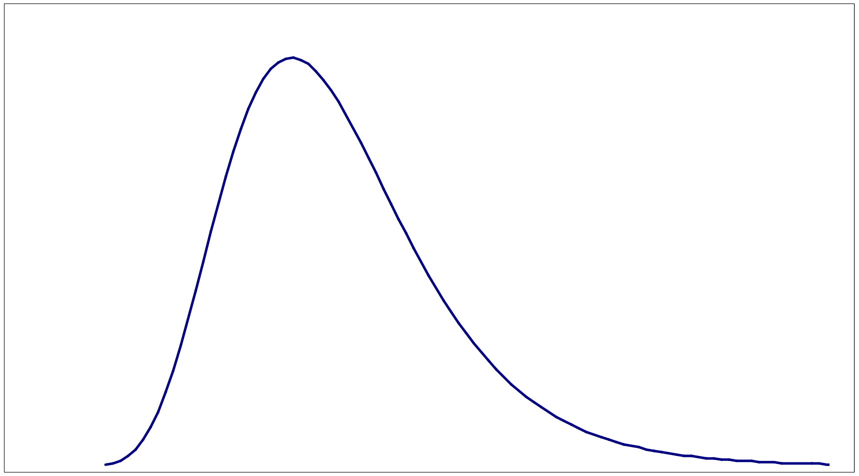
Normal Distribution

- **The Most Common Probability Distribution.**
- **Many Random Phenomena Follow This Distribution.**
- **Also Called the "Bell Curve," Noted for Its Symmetry and Thin Tails.**
- **If Cost is a Sum of Many Random Independent Phenomena, the Central Limit Theorem Indicates This May be the Appropriate Distribution.**



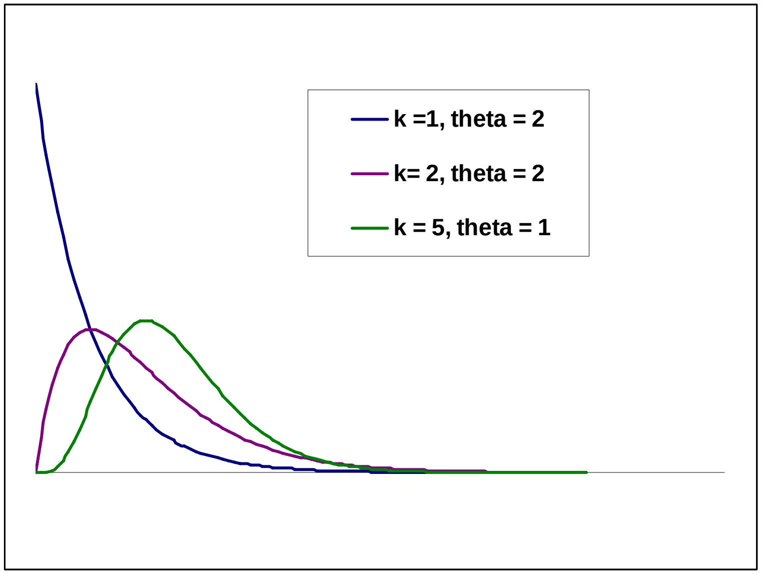
Lognormal Distribution

- **Lognormal Distribution is a Skewed Distribution.**
- **If *X* is Lognormally Distributed, *Y* = *ln(X)* is Normally Distributed.**
- **Has Fatter Tails Than the Normal Distribution.**
- **Bounded Below by Zero, Unbounded Above.**
- **If Cost is a Function is Multiplicative Factors (e.g., Test Failues Cause a Percentage Increase in Cost Rather Than a Fixed Amount Increase), then Complex Projects are Likely to be Lognormally Distributed (Multiplicative Analog to Central Limit Theorem).**
  - **These Aspects Make the Lognormal Appealing for Cost Modeling.**



Gamma Distribution

- The Gamma Distribution is a Flexible Distribution.
- Can Resemble a Lognormal, Can Also Resemble an Exponential Distribution.
  - Indeed the Gamma Distribution is the Sum of Independent Exponential Distributions.
- PDF is Given by: $f(x) = x^{k-1} \frac{e^{-\frac{x}{\theta}}}{\Gamma(k) \cdot \theta^k}$



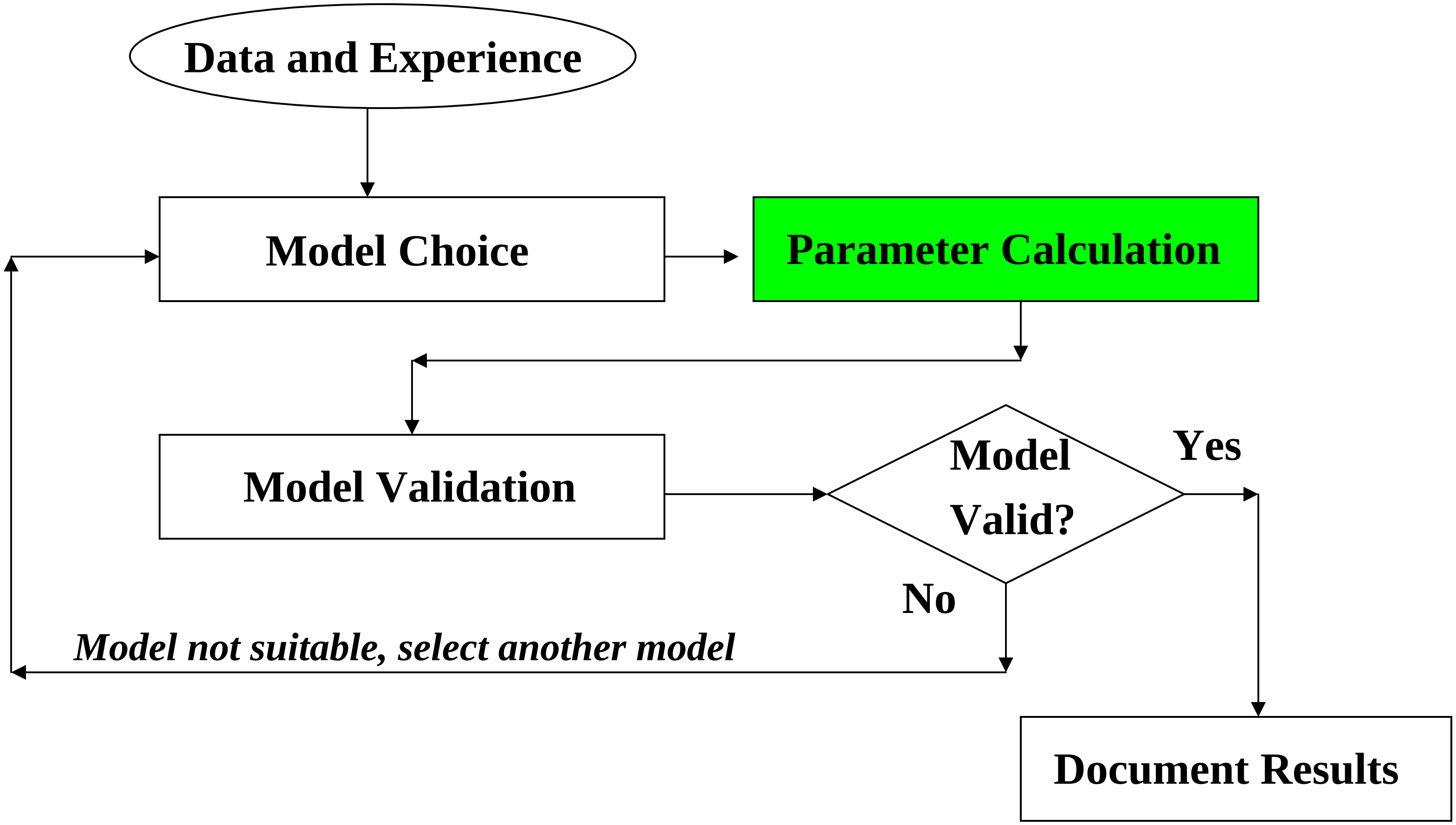
Parametric Vs. Non-Parametric

- When the Data Follow an Observable Pattern, Based Either on Preliminary Data Analysis or Through Experience, Parametric Analysis is Preferred.
  - Assume Residuals Follow Lognormal, Gamma, or Normal, for Example.
  - For Example, NASA Cost Data are Skewed, Which Makes Intuitive Sense, Because Cost Cannot be Less Than Zero, but There is No Upper Limit.
    - Leads to Assumption of Lognormal or Gamma Residuals.
- When the Data Do Not Follow an Observable Pattern, or There is No Reason to Assume an Underlying Pattern in the Data, Non-Parametric Analysis May Be Suitable.
  - Data Sets are Small.
  - No Reason to Assume Similarity with Other Data.
- However, if Non-Parametric Techniques Are Used, Must Be Careful to Ensure Models are Valid, Since Techniques May Be Similar Enough to a Parametric Technique that the Non-Parametric Version Inherits Some Features of the Parametric Version.

Parametric Vs. Non-Parametric

- Another Issue with Non-Parametric Techniques is the Lack of Rich Techniques for Developing Confidence Intervals, Prediction Intervals, Covariance Matrices, and Other Useful Metrics and Methods Available for Parametric Models.
- Indeed, Some Statistical Techniques Do Not Exist for Nonparametric Problems.
  - As Shown by Bahadur and Savage in Their 1956 Paper “The Nonexistence of Certain Statistical Procedures in Nonparametric Problems”, There Is
    - No Effective Hypothesis Test for the Population Mean
    - No Effective Confidence Interval for the Population Mean
    - No Effective Point Estimate for the Population Mean
    - No Confidence Interval Will Fit the Data Well.
- Makes Model Validation Problematic for Non-Parametric Methods.
- Note: Parametric Techniques Do Not Necessarily Involve Assuming the Residuals Follow a Particular Probability Distribution.
  - Can Be Much Weaker, Such as Assuming a Constant Coefficient of Variation.

Model Development Framework
Parameter Calculation



Parameter Calculation

- **There are Numerous Ways to Calculate the Parameters of a Cost-Estimating Relationship, but in This Presentation, We Consider One Method.**
- **The Method Presented, Maximum Likelihood Estimation, is a Widely Used Statistical Technique that Serves as a Unifying Framework for the Three CER Methods Presented.**

Maximum Likelihood Estimation

- **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**

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

- **The Maximum Likelihood Estimate of θ is the Vector That Maximizes the Likelihood Function.**
- **Maximum Likelihood Estimation is a Popular Statistical Technique.**
  - **Major Advantage – Likelihood Function is Almost Always Available.**
- **The Three CER Methods Considered in This Section All Have a Connection to Maximum Likelihood Estimation.**
- ***Parameter Calculation for Each of the Three CER Methods Considered Can be Viewed in the Context of Maximum Likelihood.***

Maximum Likelihood Lognormal Residuals

- **For $Y_i = f(X_i, \beta) \cdot u_i$, where**

  β = vector of coefficients of the CER

  Y_i = actual cost of the ith data point

  X_i = vector of cost drivers for the ith data point

  u_i = residual of the ith data point
- **Likelihood Function for Lognormal Distribution**

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

- **If we set $\mu = 0$, We are Estimating the Median**
  - **The Lognormal is Used to Model the Distribution of the Residual.**
  - **When $u = 1$, the Actual Matches the Estimate.**
  - **For a Lognormal, Median $= e^{\mu} = e^{0} = 1$.**
- **Why the Median?**

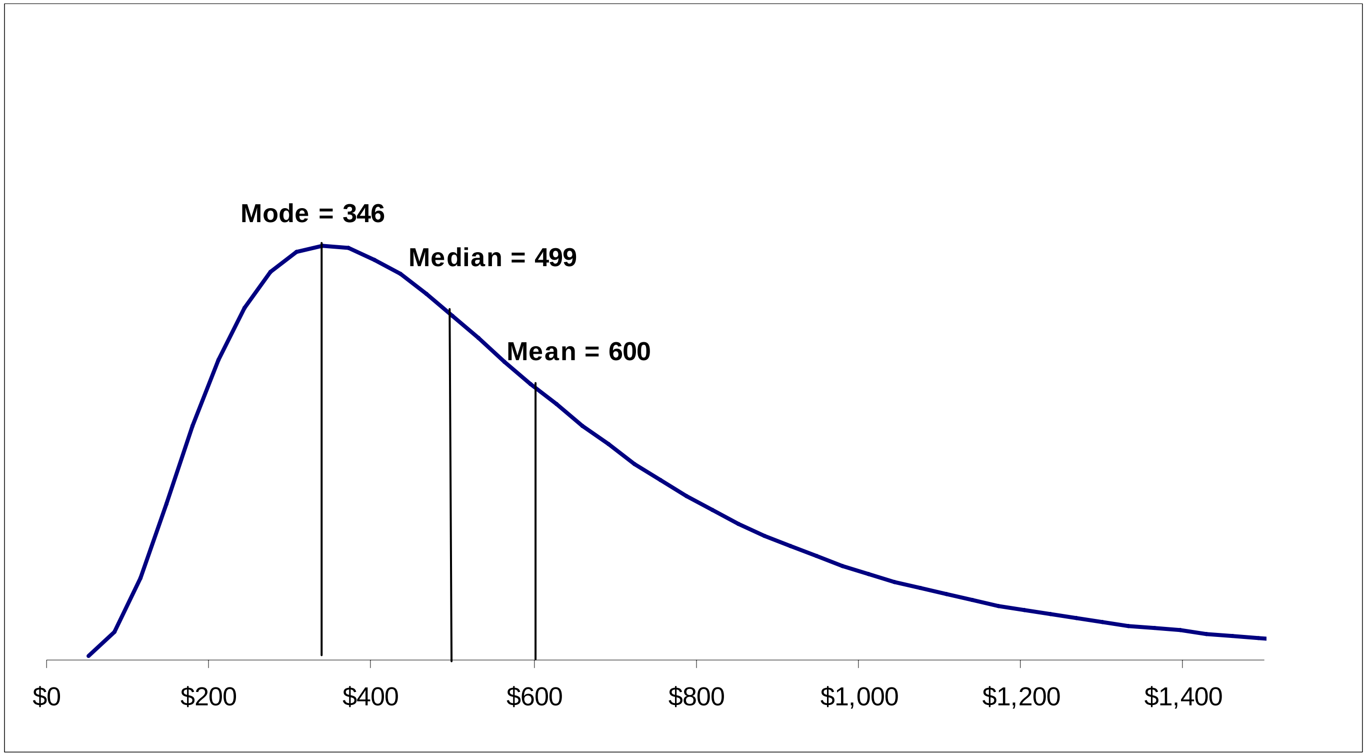
Lognormal – The Median is the Message

- **The Three Most Commonly Encountered Measures of Centrality are the Mean, Median, and Mode.**
- **Mean = “Expected Value,” For a Sample of *n* Data Points This is**

$$\sum_{i=1}^{n} \frac{X_i}{n}$$

- **Median = 50th Percentile, the Point at Which Half the Population is Less Than This Value, and Half is Greater.**
- **Mode = “Most Likely,” The Peak of the Distribution.**
- **For a Normal Distribution, Mean = Median = Mode.**
- **For a Lognormal, Mode < Median < Mean.**
- **For a Lognormal, The Mean is Always Greater Than the 50th Percentile, and Can Be Any Percentile Greater Than the 50th: 90th, 95th, etc.**
- **For This Reason, a Better Metric for the Center of a Lognormal is the Median.**
  - **Common to Report the Median Rather Than the Mean as the “Average” of Skewed Data (Income, House Prices, etc.).**

Lognormal
Mean Vs. Median Example



Lognormal Maximum Likelihood Estimation of the Median

- **From the Lognormal Likelihood Function**

$$Pr(U \le u) = \int_{-\infty}^{u} \frac{1}{U\sqrt{2\pi\theta}} e^{-\frac{(\ln u - 0)^2}{2\theta}} dU$$

- **We Want to Analyze This in Terms of *Y=f(X,β)·u***
  - **Since *u=Y/f(X,β)*,**

$$Pr(Y \le y) = Pr(u \cdot f \le y) = Pr\left(u \le \frac{y}{f}\right)$$

$$= \int_{-\infty}^{y} \frac{1}{y/f\sqrt{2\pi\theta}} \frac{1}{f} e^{-\frac{(\ln(y/f))^2}{2\theta}} dy$$

$$= \int_{-\infty}^{y} \frac{1}{y\sqrt{2\pi\theta}} e^{-\frac{(\ln(y/f))^2}{2\theta}} dy$$

Lognormal Maximum Likelihood Estimation of the Median (cont'd.)

- **Therefore, the Likelihood Function is**

$$L(\beta,\theta)=\prod_{i=1}^{n}\left(\frac{1}{y_i(2\pi\theta)^{\frac{n}{2}}}\right)\cdot e^{-\frac{1}{2\theta}\sum_{i=1}^{n}ln\left(\frac{y_i}{f_i}\right)^{2}}$$

- **Since the Logarithm Function is Monotonically Increasing, Can Take the Log of the Likelihood Function and Maximize That Instead (Usually Easier to Do):**

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

  - **Note: We Ignore Constants as They Do Not Affect the Maximization.**

- **Note That This is the Same as Minimizing the Negative of this Result:**

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

Minimizing the Log Likelihood Function

- **In Order to Minimize the Likelihood Function, First Minimize with Respect to θ.**
  - **To Minimize, We Take the Partial Derivative with Respect to θ, Set Equal to Zero, and Solve for θ.**
    - **Taking the Derivative Yields**

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

- **Setting Equal to Zero and Solving Yields**

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

Minimizing the Log Likelihood Function (cont'd.)

- Plugging in the Value for θ into the Log Likelihood Function Yields

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

- Ignoring Constants, This Becomes

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

- This is Equivalent to Minimizing

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

- This is the Least Squares of the Log of the Differences Between the Actuals and the Estimates.
- Notice the Similarity to Linear Regression.

Log-Transformed Ordinary Least Squares (LTOLS)

- **What We Have Derived is a Generalization of Log-Transformed Ordinary Least Squares in the Context of Maximum Likelihood.**
- **In Log-Transformed Ordinary Least Squares, Apply a Logarithmic Transform to Both the Actual and the Estimated Costs.**
- **For the Power Equation $Y=aX^b$ This Transforms the Equation From a Nonlinear Equation to a Linear One:**

$$\ln Y = \ln\left(aX^{b}\right) = \ln a + b \ln X$$

- **The Parameters Can Be Easily Calculated in a Spreadsheet.**
- **Must Remember to Transform the *a* Parameter.**
- **The Maximum Likelihood Median Estimator is More General.**
  - **Any Equation Form May Be Used, but Unless the Log Transformed Equation is Linear, May Need Computer to Solve (e.g., Excel‘s Solver Capability).**
  - **Nothing in the MLE Derivation Forces any Particular Functional Form.**

Maximum Likelihood Normal Residuals

- **For the Equation $Y_i = f(X_i, \beta) \cdot u_i$, When the Residuals are *Normally* Distributed, with Mean = *1* and Variance θ, the Likelihood Function, as Demonstrated by Lee (1997), is**

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

- **The Log-Likelihood Function is Thus**

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

Maximum Likelihood
Normal Residuals

- **Maximizing This Expression for θ and Then Substituting back into $l(\beta,\theta)$ Yields the Concentrated Log-Likelihood Function:**

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

- **Note this is the Same as Minimizing**

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

- **Goldberg and Tuow (1997) Note That This Method Is Very Sensitive to Departures From the Normally Distributed Residuals.**
- **If Residuals are Not Normally Distributed or Close to it, Estimates May Not be Robust.**

Minimum Percent Error and Maximum Likelihood

- **As Noted in Goldberg and Tuow (2003), This Method is Very Similar to the Minimum Percent Error Method Developed by Book and Young (1995, 1997), Who Ignore the Final Term and Instead Minimize the Sum of Squared Percentage Errors.**
- **Minimum Percent Error Method Minimizes**

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

- **Thus the Minimum Percent Error Method is a Pseudo-Likelihood Estimator in the Case of Normally Distributed Residuals.**

Minimum Percent Error Bias Constraints

- The Minimum Percent Error (MPE) Method is Biased.
  - Instead of Bias Below the Mean, the MPE Method is Biased High.
    - One Way to Make the Error Term Small is to Make the Estimates Large.
- To Correct for This Book and Lao (1996) Introduced a Bias Constraint.
  - Same Objective Function, But Now Sample Bias is Constrained to be Zero, That Is:

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

- This Method is Referred to as MPE-ZPB or ZMPE (“Zimpy”).
- Not a Parametric Method, But Similar to the Normal MLE.

MPE-ZPB and Normal MLE

- **MPE-ZPB is an Approximation of the Normal Maximum Likelihood Estimator.**
- **MPE-ZPB Objective is to Minimize:**
  (Subject to Zero Bias Constraint)

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

- **Normal MLE Objective is to Minimize**

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

- **As Has Been Noted, Dominant Term is**

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

- **Minimizing this Term is Same as Minimizing MPE-ZPB Objective Function**
  - **Second Term in MLE Assigns a Penalty for Over-Estimating, Assures Solution is Asymptotically Unbiased, So it is Similar to MPE-ZPB Bias Constraint.**

MPE-ZPB and Normal MLE Example

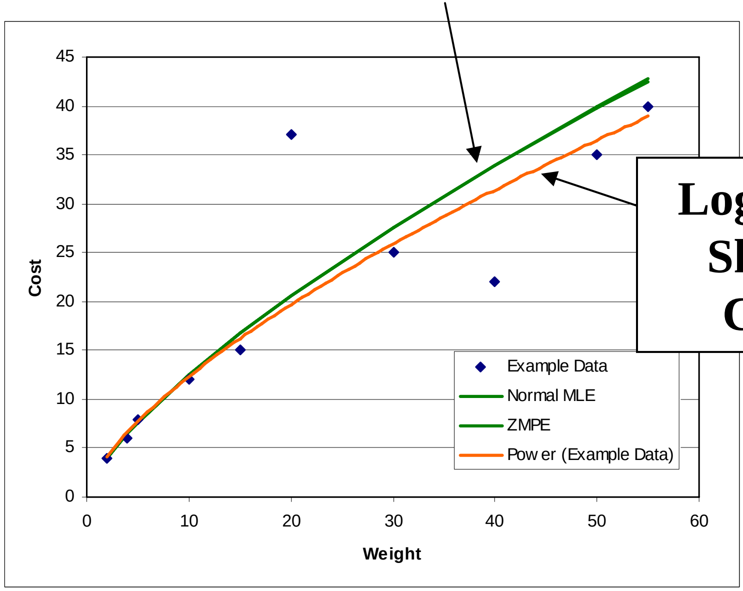
- For the Data Displayed in the Table and Graphically Displayed in the Charts:
  - Normal MLE Fit is

$$Est.Cost = 2.41Wt.^{0.716}$$

  - MPE-ZPB Fit is

$$Est.Cost = 2.39Wt.^{0.719}$$

Weight	Cost
2	4
4	6
5	8
10	12
15	15
20	37
30	25
40	22
50	35
55	40



MPE-ZPB and Normal MLE Fits are Coincident

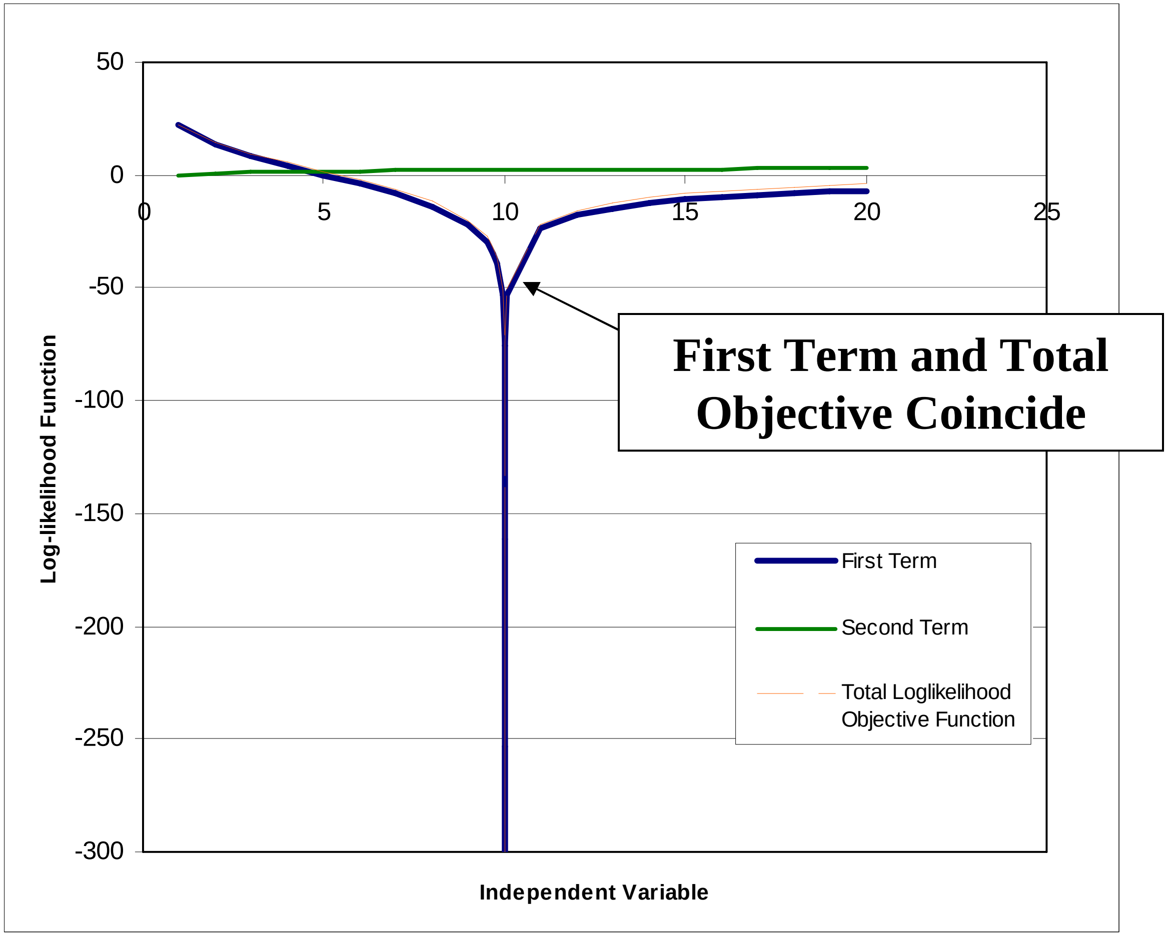
Loglinear Fit Shown for Contrast

MPE-ZPB and Normal MLE

- **Normal MLE and ZMPE Solutions are Very Similar Since they are Minimizing the Same Dominant Term and are Both "Unbiased."**
  - **MLE Solution is Asymptotically Unbiased (Unbiased for "Large" Samples).**
  - **MPE-ZPB Solution is Unbiased Regardless of Sample Size.**
- **One Advantage that MPE-ZPB has is Lack of Bias Regardless of Sample Size.**
  - **Cost Estimates are Often Based on Small Samples, so MLE Solution may be Biased.**
- **On the Other Hand, MPE-ZPB is Tied to the Assumptions of the Normal MLE.**
  - **Need Normally Distributed (Multiplicative) Residuals to Ensure Consistent Solutions in Many Cases.**

Normal MLE Log Likelihood Example

- For one Data Point with *Actual Cost = 10*, it is easy to see that the objective Function is Dominated by the First Term.



First Term and Total Objective Coincide

Maximum Likelihood Gamma Residuals

- **When the Residuals Follow a Gamma Distribution, The Negative Log-likelihood Function is**

$$l(\beta)=\sum_{i=1}^{n}\left(\frac{y}{f(x_i,\beta)}+ln\ f(x_i,\beta)\right)$$

- **This Can Be Minimized by Iteratively Minimizing The Sum of Percent Squared Errors Until the Estimates Converge:**

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

- **Note *k* is the Iteration Number.**
- **This Method Was First Developed by Nelder (1968) and Wedderburn (1974), Who Called the Method Iteratively Re-Weighted Least Squares (IRLS) and Re-Discovered by Hu in the 1990s, Who Called it Miminum Unbiased Percentage Error (MUPE).**

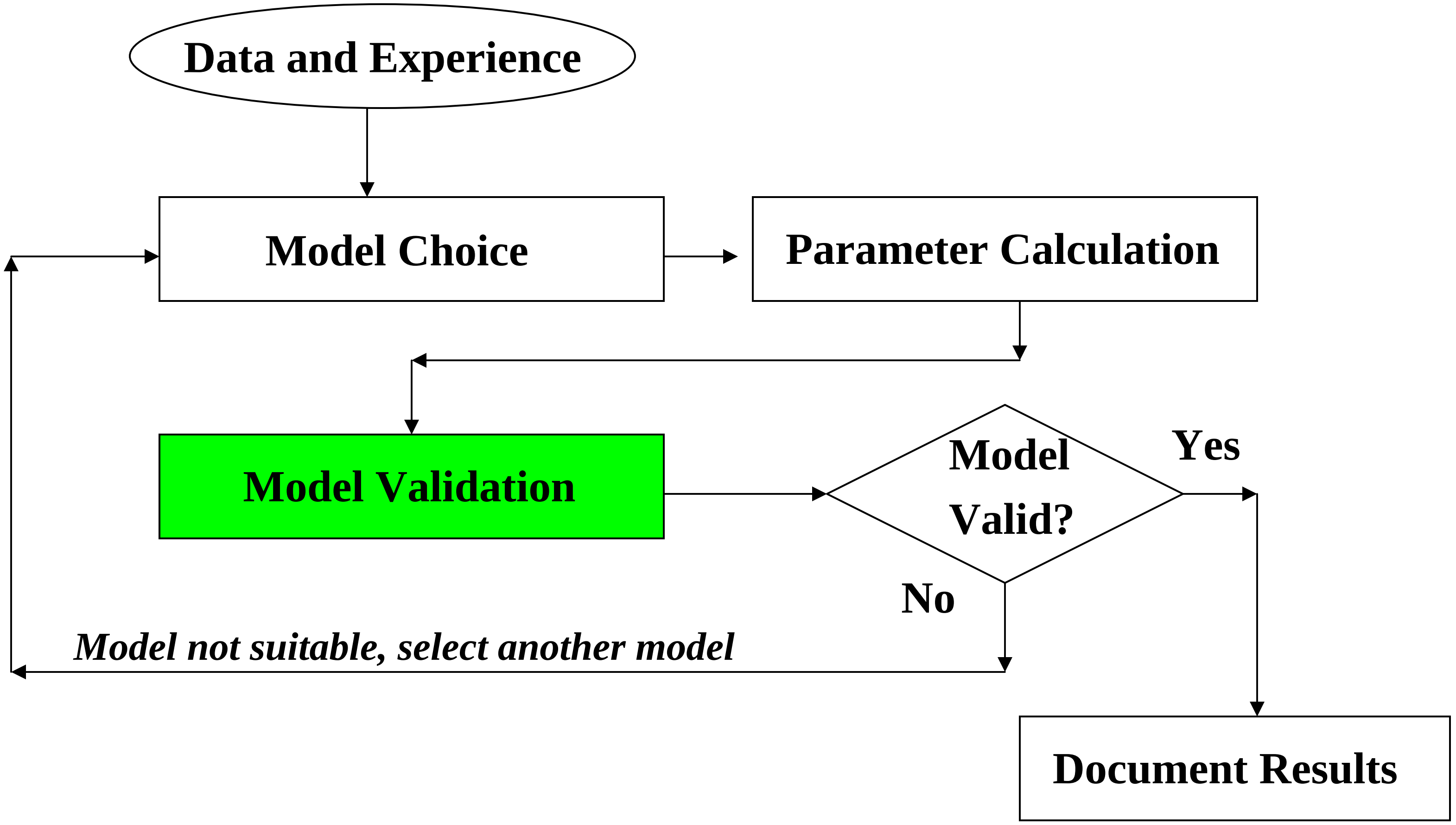
IRLS/MUPE

- **In the Case of Gamma Residuals, IRLS/MUPE is a Maximum Likelihood Estimator.**
  - **Also a Generalized Linear Model (GLM).**
- **However, IRLS/MUPE Does Not Depend Upon the Assumption of Gamma Residuals.**
- **The Likelihood Method Was Generalized by Wedderburn to Consider Quasi-likelihood, Which Has Good Statistical Properties, But Only Requires a Constant Coefficient of Variation.**
  - **Constant Coefficient of Variation Distributions Include Both Gamma and Lognormal Distributions.**

Summary of Three Methods

- **Log-Transformed OLS, MPE-ZPB, and IRLS/MUPE All Share a Common Connection in Maximum Likelihood Estimation.**
- **Log-Transformed OLS is a Maximum Likelihood Estimator of the Median of Lognormally Distributed Multiplicative Residuals.**
  - **Parametric Method**
- **MPE is a Pseudo-Likelihood Estimator of the Mean of Normally Distributed Multiplicative Residuals.**
  - **Bias Constraint Added.**
  - **Not Directly Parametric But May Has Parametric Properties.**
- **IRLS/MUPE is a Maximum Likelihood Estimator of the Mean of Gamma Distributed Residuals.**
  - **Also More General, Quasi-Likelihood.**
  - **Parametric Method.**

Model Validation



Model Validation

- **Parameter Calculation is the End of the Process for Many Cost Analysts.**
- **Once Coefficients Have Been Calculated, Many Analysts Begin Applying the New Equations.**
- **But We are Not Done Yet!**
- ***Still Need to Check Model Validity.***
  - **Do The Estimates Do a Good Job of Replicating Actual Cost?**
  - **Do the Underlying Assumptions Hold True?**

Model Validation – Goodness of Fit

- One Commonly Used Method to Validate Models is to Determine the Goodness of Fit.
  - Do the Estimates "Fit" the Actual Cost?
- Commonly Used Metrics
  - Actual Cost Vs. Estimated Cost
    - Pearson's R^2
    - Standard Percent Error
    - Percent Bias
  - Actual Parameters vs. Calculated Parameters
    - Consistency
    - Efficiency

Goodness of Fit: Pearson's R^2

- **Pearson's *R^2***
  - **Pearson Correlation of Actual Vs. Estimated Cost, in Unit Space.**
  - **Proportion of Variation of the Estimate that can be Attributed, to Variations of the Actual Cost.**
  - **In Excel, "=CORREL(A1:An,B1:Bn)^2" Where The Actual Costs are in the Cell Range *A1:An* and the Estimates are in the Range *B1:Bn.***
  - **Higher *R^2s* are Better than Lower *R^2s*.**

Goodness of Fit Metrics
SEE

- **Standard Percent Error of the Estimate**
  - **%SEE =** $$\sqrt{\frac{1}{n-k}\sum_{i=1}^{n}\left[\frac{y_i - f(x_i)}{f(x_i)}\right]^2} \times 100\%$$

    Where n is the Number of Data Points in the Sample, and k is the Number of Parameters.
  - **All Else Equal, it is Desirable to Have Low Standard Percent Error.**
  - **Will be Lowest for the MPE-ZPB Method (by Design) (Book, 2006).**
  - **Foussier has Noted that This Metric Distorts the True Underlying Error (Foussier, 2008).**
    - **Has Proposed Average Absolute Percent Error as a Better Measure than the Squared Error.**

Goodness of Fit Metrics
Percent Bias

- **Bias**
  - **Percentage Bias =** $\frac{1}{n}\sum\left[\frac{f(x_i)-y_i}{f(x_i)}\right]$
  - **MPE Without the Bias Constraint Produces Estimates that are Biased Upwards.**
  - **Log-Transformed Ordinary Least Squares Produces Estimates that are Biased Low.**
    - **Estimating the Median, which for a Lognormal is Always Less than the Mean.**
    - **Can be Corrected for with a Simple Factor.**
  - **It is Desirable to Have Estimates that Have Zero Bias if you are Interested in Estimating the Mean.**

Goodness of Fit Metrics
Consistency

- **Consistency**
  - **An Estimator is Consistent if for all $\delta > 0$ and any θ,**

$$\lim_{n \to \infty} Pr\left(\left| \hat{\theta}_n - \theta \right| > \delta \right) = 0$$

  - **Why This Matters: It’s Important That the Technique Converges to the True Parameter as the Sample Size Increases.**
    - **Without This we Have No Guarantee Our Estimated Coefficients Resembles the True Underlying Coefficients.**
  - **One of the Most Important Metrics, Often Overlooked.**
  - **Necessary to Have a Reliable Model.**
  - **Maximum Likelihood Methods are Consistent.**

Goodness of Fit Metrics
Mean Square Error and Efficiency

- **The Mean-Squared Error (MSE) of an estimator is**

$$E\left[\left(\hat{\theta}-\theta\right)^{2} \mid \theta\right]$$

- **An Estimator $\hat{\theta}$ is a Uniformly Minimum Variance Unbiased Estimator (UMVUE) if it is Unbiased and for any True Value of θ There is no Other Unbiased Estimator That has a Smaller Variance.**
- **An Estimator with That is UMVUE is Efficient, in That it Achieves the Lower Bound.**
  - **In Practice This Means That the Estimated Coefficient Will Likely Be Closer to the True Estimate Than That Calculated with Another Estimator.**
  - **Maximum Likelihood Estimates are UMVUE.**

Validating Model Hypotheses

- If a Maximum Likelihood Method Has Been Used, Need to Check to See if Residuals Fit the Assumed Shape.
  - Fit Used in the Negative Sense (“Not Reject”).
- Three Commonly Used Tests to Validate
  - Chi-Square
  - Kolmogorov-Smirnov (K-S)
  - Anderson-Darling (A-D)
- Chi-Square and K-S are Both Simple and Easy to Compute.
- A-D is More Powerful and Considered a Good Test for Departure from Normality.
- Chi-Square Gives More Weight to Low Probability Intervals.
- A-D Gives More Weight to the Tails of the Distribution.

Model Validation – Other Aspects

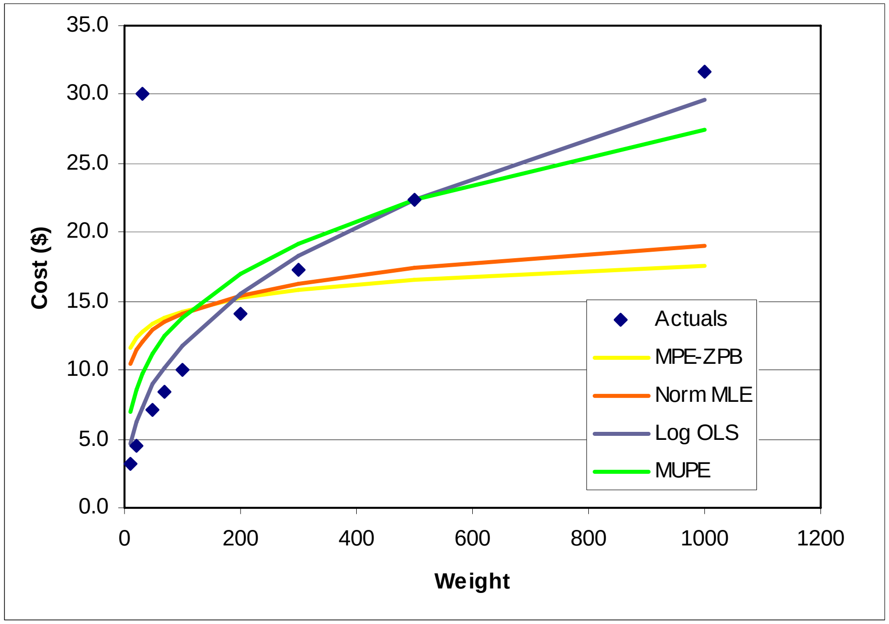
- **If Log-Transformed OLS is Used, Must Check Goodness-of-Fit to Determine That the Residuals Fit a Lognormal.**
  - **Fit is Used in the Negative Sense.**
- **If Using IRLS/MUPE as a Maximum Likelihood Estimator, Must Check to See if Residuals Fit a Gamma Distribution.**
  - **Otherwise the Only Assumption Required is Finite Variance, But Instead of Maximizing Likelihood, Only Maximum Quasi-Likelihood is Guaranteed.**
  - **Quasi-Likelihood Requires Fewer Assumptions, But Has Weaker Optimality Properties as Well (Not “Efficient”).**
- **MPE-ZPB is Posited as a Non-Parametric Method, but as a Good Approximation of the Normal MLE method, Should Check if the Residuals fit a Normal Distribution.**

Implementing the Methods

- **Log-Transformed OLS is the Easiest to Calculate.**
  - **Can Be Implemented in a Spreadsheet Using Native Excel Functions.**
- **MPE-ZPB Requires the Use of a Numerical Routine.**
  - **Can Be Implemented in a Spreadsheet Using an Excel Add-In, Excel Solver.**
- **IRLS/MUPE Also Requires the Use of a Solver-Like Routine.**
  - **Requires Solver to be Applied Iteratively.**
  - **Experience Indicates That IRLS/MUPE Typically Converges in Less Than 10 Iterations (Book, 2006).**

Comparing the Three Methods

Weight	Cost
10	3.2
20	4.5
30	30
40	7.1
70	8.4
100	10
200	14.1
300	17.3
500	22.4
1000	31.6



- **Although MPE-ZPB Has the Lowest Standard Percent Error, The Overall Trend Does Not Match the Actual Data.**
- **MUPE and Log-OLS Have Similar Fit.**
  - **Similar to Results Reported by Mackenzie(2003).**

Log-Transformed Ordinary Least Squares Summary

- **Long-Standing Method Because of Ease of Computation.**
  - **Can Calculate Coefficient By Hand.**
- **Pros**
  - **Computationally Simple.**
  - **Works Well on Skewed Data.**
  - **Has Optimal Properties (Maximum Likelihood Estimator of the Median).**
  - **Parametric Method, So Have Access to Covariance Matrices and Confidence Intervals.**
  - **Can Use Any Equation Form in the Generalized Maximum Likelihood Estimator Form.**
- **Cons**
  - **Underestimates the Mean (Biased Estimator of the Mean, Since the Median is Less than the Mean).**
    - **Can Be Corrected for by Applying an Adjustment Factor.**

MPE-ZPB Summary

- **Recent Innovation – Requires the Use of a Computer.**
- **Pros**
  - **Minimizes the Standard (Percent) Error of the Estimate.**
    - **This Goodness-of-Fit Metric is Best Among All Three Methods.**
  - **Estimator is Unbiased.**
  - **Requires No Parametric Assumptions.**
- **Cons**
  - **Estimator is Not Consistent.**
  - **Model is Not Parametric.**
    - **Confidence Intervals for Population Mean Not Available.**
  - **Method is Not Robust (Sensitive to Departure from Normality).**
    - **Particularly Troublesome for Estimating Skewed Data.**
    - **Model is Similar Enough to Normal MLE to Retain Some of Its Properties.**

IRLS/MUPE Summary

- **Recent Innovation – Requires the Use of a Computer.**
- **Pros**
  - **Requires Weak Assumptions (Finite Variance) But Still Has Confidence Intervals and Covariance Matrices Available to Fully Parametric Methods.**
  - **Method is Maximum Likelihood if Residuals Are Gamma Distributed.**
  - **Asymptotically Unbiased.**
  - **Consistent**
- **Cons**
  - **If Not MLE, Weak Optimality Properties ("Quasi-Likelihood").**
    - **May Not be Efficient.**
  - **Can Be Biased for Small Samples.**

Comparison with Other Industries - Insurance

- The Analogy with Cost Estimating in Insurance is “Loss Modeling.” In Insurance Parlance, a “Loss” is the Amount of a Loss Experienced by a Policyholder.
  - In Insurance, Parametric Models are Used to Estimate not Only Loss Size, but Also Claim Frequency.
  - Log-transformed OLS and MUPE/IRLS Frequently Used to Model Claim Frequency.
    - MUPE/IRLS Referred to as “Gamma Regression” by Casualty Insurance Modelers (Fu and Moncher, 2004).

Comparison with Other Industries – Insurance (2)

- **Fu and Moncher (2004) Report that the Gamma and Lognormal are the Most Widely Used Distributions in Loss Modeling.**
  - **Mention 31 Recent Papers that use Lognormal and 37 that use Gamma distributions.**
  - **Lognormal Also Used in Ratemaking and Reserve Setting (akin to Cost Risk Analysis).**
  - **Also Study Normal but Find Lognormal and Gamma much Better for Modeling Skewed, Positive Data.**
    - **Like "Loss" and "Cost"**
  - **Recommend Against Use of Normal Distribution for Modeling Skewed Data.**
    - **Normal is Symmetric.**

Comparison with Other Industries – Health Care and Labor Economics

- **Costs are Modeled Parametrically in Health Care Economics.**
  - **Log-Transformed OLS and IRLS/MUPE Widely Used.**
  - **Recent Papers Include**
    - **“Estimating Log Models: To Transform or Not to Transform?”, Journal of Health Economics, 2001**
    - **“Comparing Alternative Models: Log Vs. Cox Proportional Hazard,” Health Economics, 2004.**
    - **“Generalized Modeling Approaches to Risk Adjustment of Skewed Outcomes Data,” Journal of Health Economics, 2005.**
    - **“Net Migration and State Labor Market Dynamics,” Journal of Labor Economics, 2004.**

Comparison with Other Industries – Summary

- **All Models Used Either Log-Transformed OLS or IRLS/MUPE (aka "Gamma Regression").**
- **Contrary to Recent Claims, Log-Transformed OLS is a "Modern" Method.**
- **Conclusion**
  - **Log-Transformed OLS and IRLS/MUPE are the Two Leading Methods Used in Other Industries.**

The Framework in Action
NAFCOM CER Development

- **The Authors Have Successfully Applied this Framework to the Development of CERs Included in the Latest Version of the NASA/Air Force Cost Model (NAFCOM).**
  - **Released Spring 2009.**
  - **Also Performed Significant Cost Driver Research and Added Twenty New Data Points.**
  - **Model Choice is Statistically-Derived CERs Using Historical Data.**
  - **Parameter Calculation Method is Log-Transformed OLS.**
    - **Also Investigated the Use of MPE-ZPB.**
  - **All CERs Validated by Calculating Goodness-of-Fit Metrics and by Verifying that the Residuals for Each CER Fit a Lognormal Distribution.**

NAFCOM CER Development
Log-Transformed OLS Vs. MPE-ZPB

- **Although We Decided to Use Log-Transformed OLS, We Also Investigated the MPE-ZPB Method.**
  - **Found Log-Transformed OLS to be a Better Choice**
    - **CER Residuals are Lognormally Distributed, so Log-Transformed OLS is a Valid Method.**
      - **Empirical Data Leads us to Believe Log-Transformed OLS CERs Provide Statistically Consistent and Efficient Estimates for our Data.**
    - **MPE-ZPB CERs Provide Similar Results to Normal MLE, but Residuals are Not Normally Distributed.**
      - **Empirical Data Gives us No Confidence that MPE-ZPB CERs will Provide Statistically Consistent or Efficient Estimates.**
        - » **To the Contrary we Found the Method is Not Robust for Skewed Data Since it is Overly Sensitive to Individual Outlying Data Points.**
        - » **In Line with Intuition and What Professionals in Other Industries Have Found When Working with Skewed Data.**

NAFCOM CER Development
Attitude Determination and Control

- CERs Updated for New Version of NAFCOM.
- Cost Drivers for Attitude Determination and Control Include:

Weight
Mission Class
Management Rating
Heritage
Technology Maturity Index
Year of Technology
Stabilization Method
Sensors Rating

- For this CER we Applied Three Methods
  - Log-Transformed OLS
  - MPE-ZPB Method
  - Normal MLE

Attitude Determination and Control CER Comparison

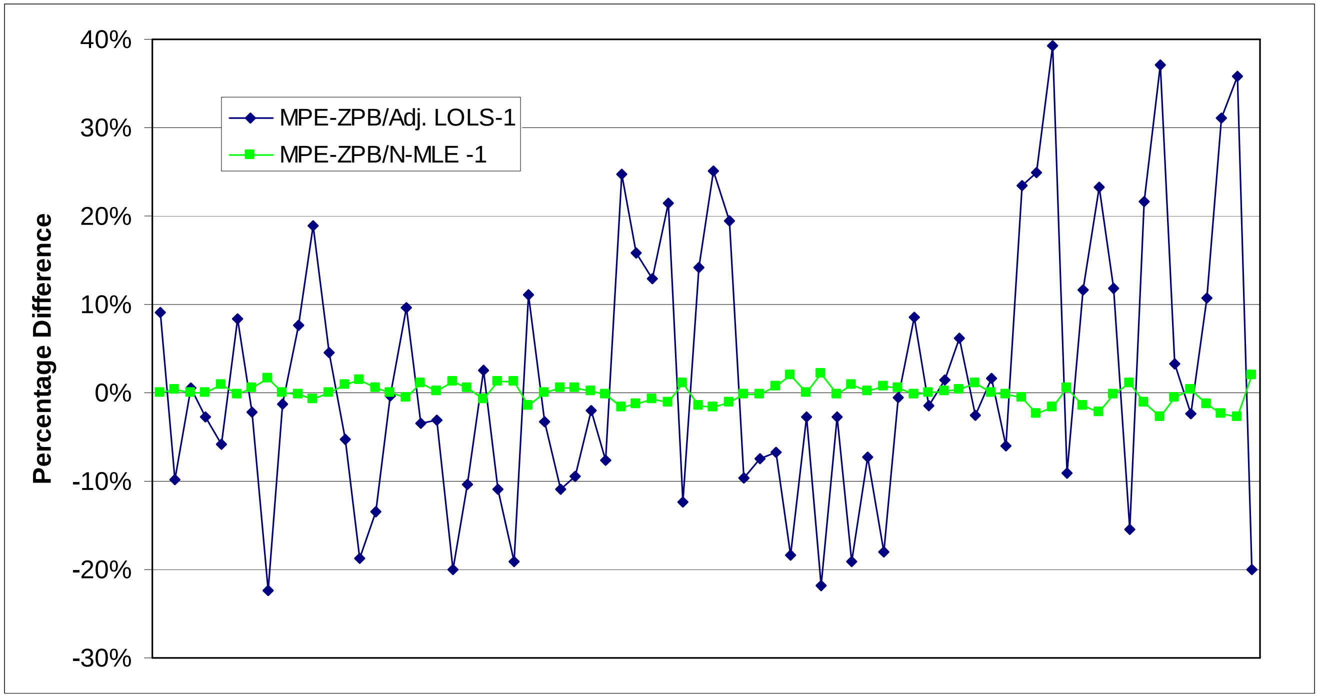
- **The Methods Provide Similar Goodness-of-Fit Metrics.**

	Log-Transformed OLS	Adjusted Log-Transformed OLS	MPE-ZPB	Normal MLE
Pearson R^2	99.2%	99.2%	99.3%	99.3%
Std. % Error	50.2%	46.4%	41.4%	41.5%
Absolute % Error	36.5%	35.0%	36.3%	36.2%
Bias	-7.2%	-0.4%	0.0%	0.0%

- **Note “Absolute % Error” is the Average Absolute Percent Error of the Estimate, Using Degrees of Freedom as the Denominator.**

Attitude Determination and Control CER Validation

- **Still Need to Check Model Hypotheses Are Valid.**
- **In this instance, Normal MLE and the MPE-ZPB CER Provide Almost Exactly the Same Estimates, so the Normal MLE and MPE-ZPB Validity Both Depend Upon Normally Distributed Residuals.**



Attitude Determination and Control CER Validation

- The Anderson-Darling Test Statistic for the Log-Transformed OLS Residuals is 0.3351, Much Less Than Critical Value (0.752) at 5% Significance
  - Cannot Reject Hypothesis That Residuals are Lognormally Distributed.
- The Anderson-Darling Statistic for Normal MLE Residuals is 1.0251, Above Critical Value (0.752) at 5% Significance, so we Reject the Hypothesis that Residuals are Normally Distributed.

Consequences of Model Validity (or Lack Thereof)

- **We Have Empirical Evidence That Log-Transformed OLS Assumptions Are Valid.**
  - **Provides Confidence That the Method is Consistent and Efficient.**
- **We Have Empirical Evidence That Normal MLE Assumptions Are NOT Valid for This Application.**
  - **Method Likely to Not be Consistent or Efficient.**
  - **Have No Confidence That Coefficients Resemble the True Parameters.**
    - **Both Normal MLE and MPE-ZPB are Very Sensitive to Departures from Normality, Indicating a Lack of Robustness.**
- **Found This Pattern Holds for All NAFCOM Subsystem CERs.**

Summary

- **Introduced Rigorous Framework for Developing Parametric Cost Models.**
  - **Emphasized Importance of Thoroughly Validating Models.**
- **Through MLE Framework Found:**
  - **Log-Transformed OLS is a General Method with Optimal Properties.**
  - **Similarities Between Normal MLE and MPE-ZPB.**
    - **MPE-ZPB is an Approximation of the Normal MLE.**
- **MPE-ZPB Not Explicitly Parametric, but Tied to Normal MLE Assumption in Many Instances.**
  - **Both Normal MLE and MPE-ZPB Very Sensitive to Departures of Residuals from Normality.**
- **Log-Transformed OLS and MUPE Widely Used in Other Industries.**
  - **Log-Transformed OLS is a Modern, Relevant Method.**

References

- **Bahadur, R.R., and L.J. Savage, "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems," Annals of Mathematical Statistics 27, 1956, pp 1115-1122.**
- **Book, S.A., and P.H. Young, "General-Error Regression for USCM-7 CER Development," The Aerospace Corporation, El Segundo, CA, 1995**
- **Book, S.A., and P.H. Young, "General-Error Regression for Deriving Cost-Estimating Relationships," Journal of Cost Analysis, Fall 1997, pp. 1-28.**
- **Book, S.A., and N.Y. Lao, "Deriving Minimum-Percentage-Error CERs Under Zero-Bias Constraints," The Aerospace Corporation, El Segundo, CA, July 1996.**
- **Book, S.A., "IRLS/MUPE CERs Are Not MPE-ZPB CERs," Presented at the International Society for Parametric Analysts Annual Conference, Seattle, WA, May 23-26, 2006.**
- **Book, S.A., and P.H. Young, "The Trouble with R^2," *The Journal of Parametrics*, Vol. 26, No. 1, Summer 2006, pp. 87-112.**
- **Eskew, H.L. and K.S. Lawler, "Correct and Incorrect Error Specifications in Statistical Cost Models," *Journal of Cost Analysis*, Spring 1994, page 107.**
- **Foussier, P. M., *From Product Description to Cost: A Practical Approach*, Vols. 1 and 2, Springer-Verlag, London, 2006.**
- **Foussier, P. M. and P. Foussier, "Should We Use the Median Instead of the OLS?," Parametric World, Fall 2008, pp. 8-11.**
- **Goldberg, M.S., and A.E. Tuow, *Statistical Methods Learning Curves and Cost Analysis*, Institute for Operations Research and Management Sciences, Linthicum, MD, 2003.**
- **Hu, S., "The Impact of Using Log-Error CERs Outside the Data Range and Ping Factor," Presented at the Annual Joint ISPA-SCEA Conference, Denver, CO, June, 2005.**

References (2)

- **Ismail, N., and A.A. Jemain, "Comparison of Minimum Bias and Maximum Likelihood Methods for Claim Severity," Casualty Actuarial Society E-Forum, Winter 2009.**
- **Fu, L., and R. Moncher, "Severity Distributions for GLMs: Gamma or Lognormal?", 2004 CAS Spring Meeting, Colorado Springs, CO, 2004.**
- **Klugman, S.A., et al., *Loss Models*, 3rd Ed., John Wiley & Sons, Hoboken, 2008.**
- **Lee, D.A., *The Cost Analyst's Companion*, Logistics Management Institute, McLean, VA, 1997.**
- **Mackenzie, D., "Cost Estimating Relationship Variance Study," AIAA Space Conference, 2003, Long Beach, CA.**
- **Mackenzie, D., et al., "Top Level Spacecraft Cost Distribution Study," Joint Annual ISPA-SCEA Conference, Noordwijk, May, 2008.**
- **Nelder, J. A., "Weighted Regression, Quantal Response Data, and Inverse Polynomials," *Biometrics*, Vol. 24 (1968), pages 979-985.**
- **Wedderburn, R.W.M., "Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method," *Biometrika*, Vol. 61, Number 3 (1974), pages 439-447.**