Data Driven Confidence Regions for Cost Estimating Relationships
Outline

- Introduction
- Review of Confidence Intervals
- Example Problems
  - Learning Curve
  - Weight Curve
- Practical Implementation and Summary
Cost Estimates are models that contain uncertain parameters
- Parameters drive the model output uncertainty

Most analysts use regression analysis to estimate parameters
- Computed values are based on the sample data

Monte Carlo Methods are useful to assess the model uncertainty
- Some characterization of the parameter uncertainty is required
Goal

- Review parameter confidence regions for linear and nonlinear regression models

- Compute the regions for some common (nonlinear) Cost Estimating Relationships (CERs)
  - Determine just how “non-linear” the parameter confidence regions are

- Discuss current limitations and recommendations
REVIEW OF CONFIDENCE INTERVALS
One Variable

- **Given data** \( \{x_i\}_{i=1}^N \)

- The confidence interval for the population mean \( \mu \) is

\[
P \left( \bar{x} - z^* \frac{\sigma}{\sqrt{N}} < \mu < \bar{x} - z^* \frac{\sigma}{\sqrt{N}} \right) > (1 - \alpha)
\]

where:
- \( \alpha \) is the specified significance level
- \( z^* \) is the (two tail) critical value from \( N(0,1) \)
- \( \sigma \) is the population standard deviation

- If \( \sigma \) is unknown then \( s \) and the Student’s t distribution \( t^* \) critical value can be used

\[
P \left( \bar{x} - t^* \frac{s}{\sqrt{N}} < \mu < \bar{x} - t^* \frac{s}{\sqrt{N}} \right) > (1 - \alpha)
\]
Confidence Interval Interpretations

- For any significance level for any parameter, based on the data set, the Confidence Region for the mean value of that parameter is a fixed interval.

- The mean value is either in the interval or it's not.
  - The probability that the population mean is in the range is 0 or 1.

- The interpretation must be about the confidence interval computation process providing the intervals that contain the true population \((1 - \alpha)\%\) of the time.
Marginal Confidence Intervals

- When solving for models involving more than one parameter typically regression is used

- Each parameter $\beta_j$ can be treated independently using the regression results $b_j, s_{bj}$

$$ P\left( b_j - t^* \frac{s_{bj}}{\sqrt{N}} < \beta_j < b_j + t^* \frac{s_{bj}}{\sqrt{N}} \right) > (1 - \alpha) $$

- Result is a confidence interval “box” within the parameter space
For a given confidence level \((1 - \alpha)\) we can check to see if expectation model produces a significantly different answer at a new parameter value \(\tilde{\beta}\) by computing

\[
\left( S(x_i, \tilde{\beta}) - S(x_i, b) \right) \leq p s^2 F(p, N - p, 1 - \alpha)
\]

where

- \(s^2\) is the mean squared error of the estimate
- \(F(p, N - p, 1 - \alpha)\) is the Fisher distribution
When the model is linear the Jacobian is constant and the can be evaluated for any different parameter as

\[(\tilde{\beta} - b)^T D^T D (\tilde{\beta} - b) \leq p s^2 F(p, N - p, 1 - \alpha)\]

where

\[D\]

is the system Jacobian

With this formulation we can very quickly check lots of points using just matrix vector multiplication

All confidence regions are also ellipses whose shape is determined by \(D\)

- Ratio of ellipse axes is related to Pearson Correlation coefficient
Multivariate Nonlinear Models

- For Nonlinear models the Jacobian \((D)\) is not constant

- We could do one of the following to compute confidence regions of the parameters
  - Evaluate the model a lots of different points to find the true confidence region
  - Assume the Jacobian is constant or that it doesn’t change much that it is and use the same \((D)\) to compute a linear approximation to the true confidence regions

- Model evaluation probably not be as fast as matrix vector multiplication
  - But it is embarrassingly parallel

- Sometimes a (nonlinear) transformation on the variables or the data (or both) can yield a linear model
  - Not guaranteed to exist
LEARNING CURVE MODEL
Learning Curve Model

- Basic learning curve model form is

\[ y = T_1 x^{(\log_2 LC)} \]

where
\( T_1 \) is the cost of the theoretical first unit
\( LC \) is the learning curve slope percent

- Nonlinear model with 2 parameters and 1 variable
  - Can be made linear by applying logarithm

- For production lot average cost we have

\[ \frac{1}{L - F + 1} \sum_{k=F}^{L} y_k = \frac{T_1}{L - F + 1} \sum_{k=F}^{L} x_k^{(\log_2 LC)} \]

where
\( L \) is the last unit in the lot
\( F \) is the first unit in the lot

**Integrity - Service - Excellence**
After an approximation for the sum we have

\[ \bar{y} = T_1 \left( \frac{(L + 0.5)^{(\log_2 LC)} - (F - 0.5)^{(\log_2 LC+1)}}{(L - F + 1)(\log_2 LC + 1)} \right) \]

Nonlinear model with 2 parameters and 2 variables
- L and F are not really independent variables
- Using a Lot midpoint can simplify to one variable

The equation above has a lot midpoint of

\[ \tilde{x} = \left( \frac{(L + 0.5)^{(\log_2 LC+1)} - (F - 0.5)^{(\log_2 LC+1)}}{(L - F + 1)(\log_2 LC + 1)} \right)^{\frac{1}{\log_2 LC}} \]

Simple heuristic

\[ \tilde{x} = \frac{F + L + 2\sqrt{FL}}{4} \]
With the Lot midpoints we are now to

$$\bar{y} = T_1 \tilde{x}^{\log_2 LC}$$

Still nonlinear, but only 1 variable

Using the lot midpoint and applying a logarithm again we get a linear model

$$\ln \bar{y} = \ln T_1 + b \ln \tilde{x}$$

where

$$b = \log_2 LC$$ is the learning curve exponent
Learning Data Set

- Representative data set used for all tests

<table>
<thead>
<tr>
<th>Lot</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Unit</td>
<td>1</td>
<td>11</td>
<td>24</td>
<td>45</td>
<td>67</td>
<td>91</td>
<td>115</td>
</tr>
<tr>
<td>Last Unit</td>
<td>10</td>
<td>23</td>
<td>44</td>
<td>66</td>
<td>90</td>
<td>114</td>
<td>134</td>
</tr>
<tr>
<td>Average Unit Cost</td>
<td>138.39</td>
<td>121.78</td>
<td>100.00</td>
<td>86.78</td>
<td>70.71</td>
<td>69.84</td>
<td>70.51</td>
</tr>
</tbody>
</table>

Data - Unit Space

Average Unit Cost vs. LMP
- The product of marginal distributions is not the same as the joint confidence regions, even for the linear problem
  - Ellipse captures the covariance of the parameters
  - Points outside the ellipse shouldn’t be considered reasonable pairs at the specified confidence level
Transformed Model Results

- The ellipsoidal confidence regions in the transformed space are non-ellipsoidal in unit space
  - The result of nonlinear transformations
Nonlinear Regression Results

- The linear approximation method provides a good match to the model evaluation confidence region
  - For this model and data set, the problem is only “slightly nonlinear”
  - The true confidence region is not ellipsoidal

![Comparison of Confidence regions](image)

- Linear Approximation
- Joint Parameter Confidence Region
Nonlinear Regression Results

- The Transformed OLS confidence regions actually overstate the true confidence region obtained from model evaluation.

- If used, this could overstate the model outcomes.

- The error assumptions drive the differences.
  - Transformed OLS has lognormal errors in unit space.

![Comparison of Transformed Linear and Confidence regions](image-url)
WEIGHT CURVE MODEL
Comparing Two “Similar” Models

- In this example, the parameter confidence regions for two weight CERs are compared.

- The two models have “similar” fit statistics, but the model form yields drastically different confidence regions.

- Model 1 – \( y = \theta_1 x^{\theta_2} \)

- Model 2 – \( y = \theta_1 (1 - e^{-\theta_2 x}) \)
Representative data set used for all tests

<table>
<thead>
<tr>
<th>Data point</th>
<th>Cost $K</th>
<th>Weight (lbs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs 1</td>
<td>3,106.64</td>
<td>77.05</td>
</tr>
<tr>
<td>Obs 2</td>
<td>29,166.32</td>
<td>1,236.77</td>
</tr>
<tr>
<td>Obs 3</td>
<td>4,820.48</td>
<td>232.14</td>
</tr>
<tr>
<td>Obs 4</td>
<td>34,111.22</td>
<td>863.36</td>
</tr>
<tr>
<td>Obs 5</td>
<td>6,387.04</td>
<td>224.40</td>
</tr>
<tr>
<td>Obs 6</td>
<td>20,871.60</td>
<td>720.44</td>
</tr>
<tr>
<td>Obs 7</td>
<td>28,621.92</td>
<td>959.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data point</th>
<th>Cost $K</th>
<th>Weight (lbs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs 8</td>
<td>19,796.80</td>
<td>332.50</td>
</tr>
<tr>
<td>Obs 9</td>
<td>7,526.40</td>
<td>269.42</td>
</tr>
<tr>
<td>Obs 10</td>
<td>6,002.24</td>
<td>123.84</td>
</tr>
<tr>
<td>Obs 11</td>
<td>11,668.48</td>
<td>316.15</td>
</tr>
<tr>
<td>Obs 12</td>
<td>6,329.12</td>
<td>59.77</td>
</tr>
<tr>
<td>Obs 13</td>
<td>4,683.20</td>
<td>59.17</td>
</tr>
<tr>
<td>Obs 14</td>
<td>21,068.72</td>
<td>369.12</td>
</tr>
</tbody>
</table>
Models were designed to have similar fit statistics

- Use $SE$ to measure fit since $R^2$ may not mean much for nonlinear problems

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>233.91</td>
<td>40,021.07</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.6991</td>
<td>0.0013</td>
</tr>
<tr>
<td>$SE$</td>
<td>4,525.4</td>
<td>4,273.5</td>
</tr>
</tbody>
</table>
Parameter Confidence Regions

- The regions are quite different in terms of scale and shape

- For both, the linear approximation is only a good approximation in a close neighborhood of the solution
Practical Implementation and Summary

How should we model parameter uncertainty?

- Excel based Monte Carlo tools treat inputs as independent, then apply a correlation
  - Using Pearson’s Correlation can only yield the linear approximation
  - Using Rank correlations is better but not the complete answer (fails if non-monotonic)
- There are methods that require some additional information
  - Conditional Method
  - Multivariate Inverse Transform sampling
  - Copulas?

The objective of this paper

- Highlight the parameter confidence intervals for nonlinear models
  - Critical input to model uncertainty and quantification
- Provide some simple examples
- Solicit feedback from others
QUESTIONS?