Air Force Life Cycle Management Center

Integrity - Service - Excellence

Data Driven Confidence Regions for Cost Estimating Relationships

Outline

- Introduction
- Review of Confidence Intervals
- Example Problems
  - Learning Curve
  - Weight Curve
- Practical Implementation and Summary

Introduction

- **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** μ **is**

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

where:
α **is the specified significance level**
z^* **is the (two tail) critical value from** $N(0,1)$
σ **is the population standard deviation**

- **If** σ **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 its 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 β_j can be treated independently using the regression results b_j , s_{b_j}

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

- Result is a confidence interval “box” within the parameter space

Multivariate Models

- **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 $\widetilde{\beta}$ by computing**

$$\left(S(x_i,\widetilde{\beta}) - S(x_i,b)\right) \le p\, s^2 F(p, N-p, 1-\alpha)$$

where

s^2 **is the mean squared error of the estimate**

$F(p, N-p, 1-\alpha)$ **is the Fisher distribution**

Multivariate Linear Models

- **When the model is linear the Jacobian is constant and the can be evaluated for any different parameter as**

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

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**

Learning Curve Model

- After an approximation for the sum we have

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

- 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

$$\tilde{\tilde{x}} = \frac{F + L + 2\sqrt{F\,L}}{4}$$

Learning Curve Model

- With the Lot midpoints we are now to

$$\overline{\boldsymbol{y}} = \boldsymbol{T_1}\, \widetilde{\boldsymbol{x}}^{\log_2 \boldsymbol{LC}}$$

  - Still nonlinear, but only 1 variable

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

$$\ln \overline{\boldsymbol{y}} = \ln \boldsymbol{T_1} + \boldsymbol{b}\, \ln \widetilde{\boldsymbol{x}}$$

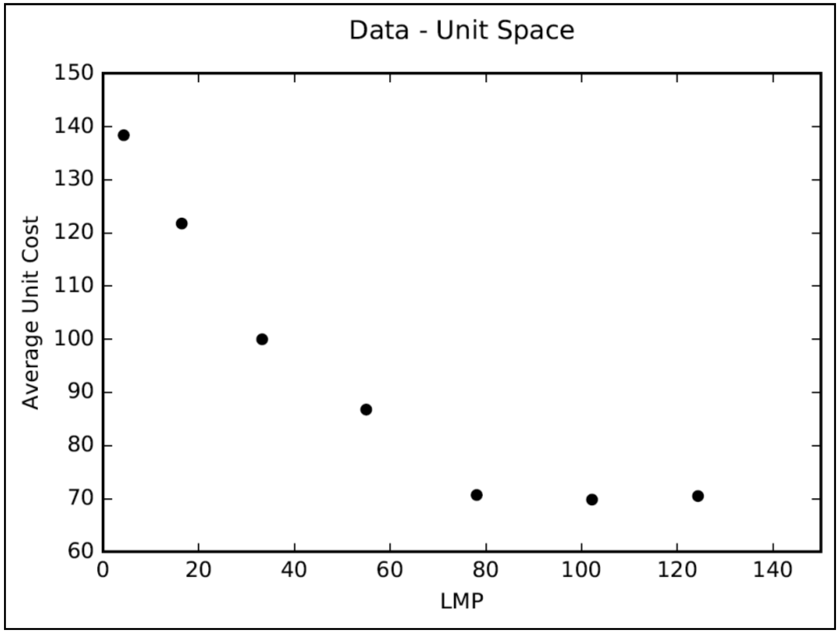
where

$\boldsymbol{b} = \log_2 \boldsymbol{LC}$ is the learning curve exponent

Learning Data Set

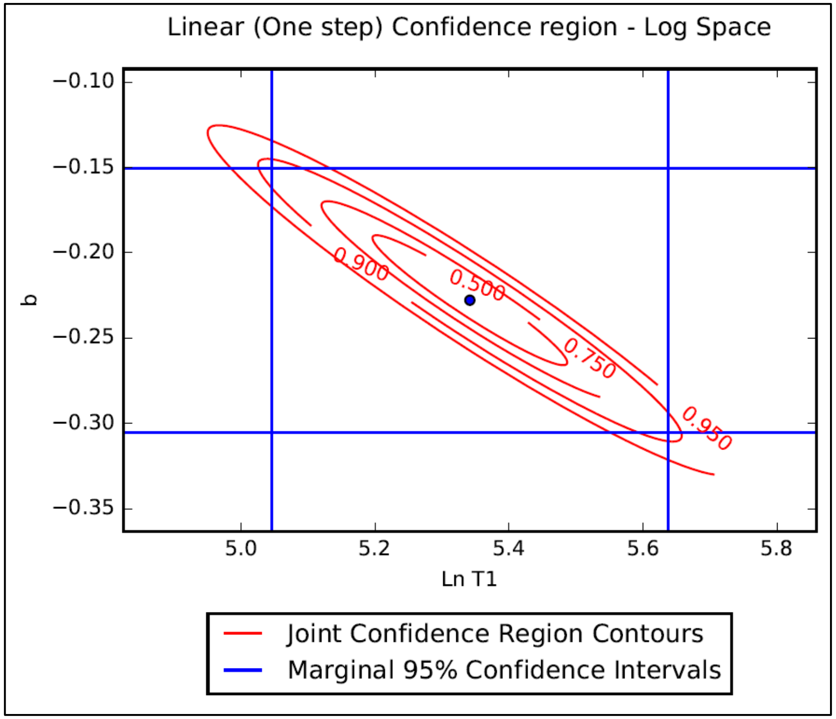
- **Representative data set used for all tests**

Lot	1	2	3	4	5	6	7
First Unit	1	11	24	45	67	91	115
Last Unit	10	23	44	66	90	114	134
Average Unit Cost	138.39	121.78	100.00	86.78	70.71	69.84	70.51



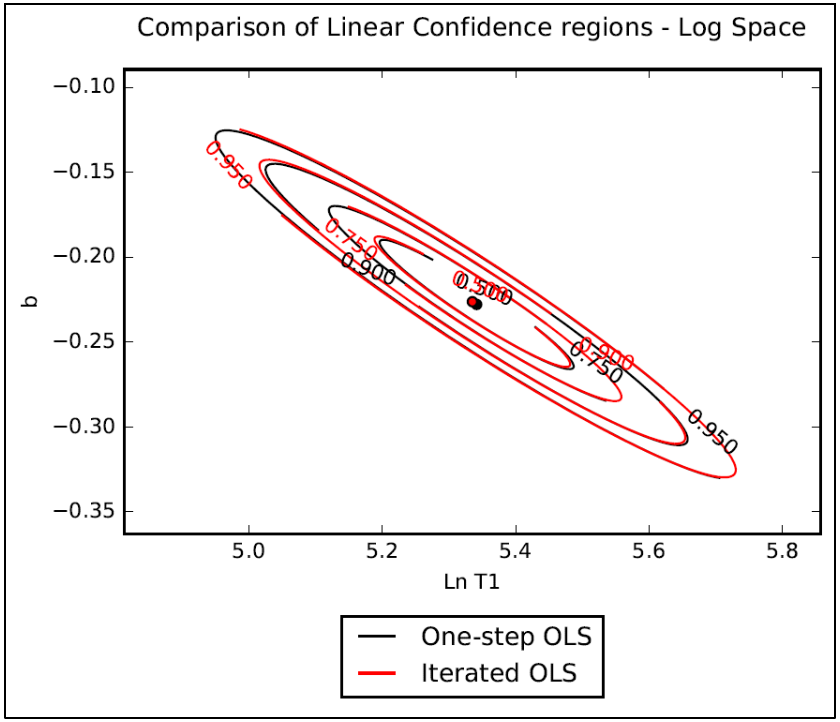
Transformed Model Results

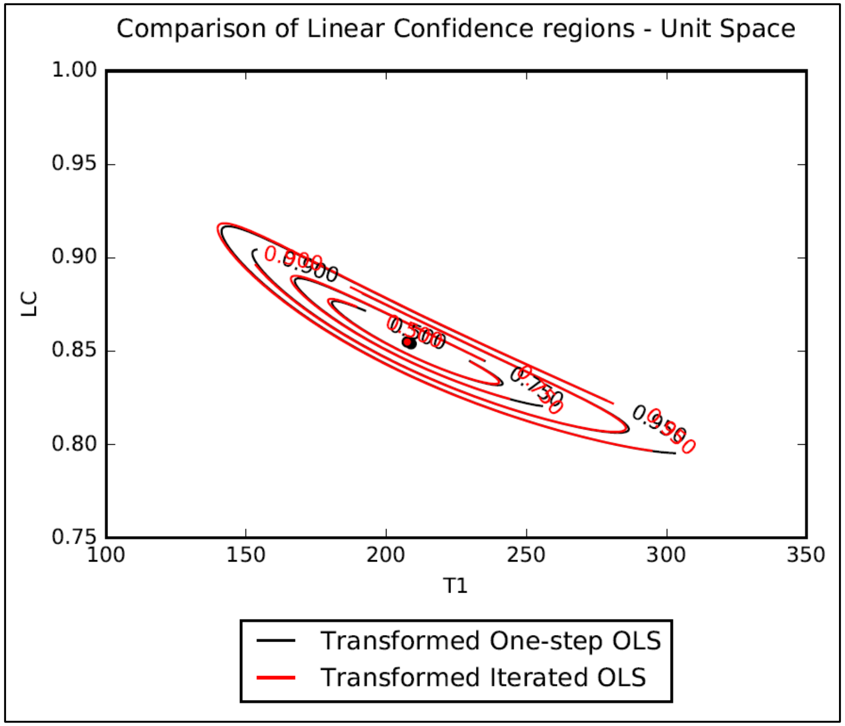
- **The product of marginal distributions is not the same as the joint confidence regions, even for the linear problem**
  - **Ellipse captures the covariance of the parameters**
  - **Points outside the ellipse shouldn't be considered reasonable pairs at the specified confidence level**



Transformed Model Results

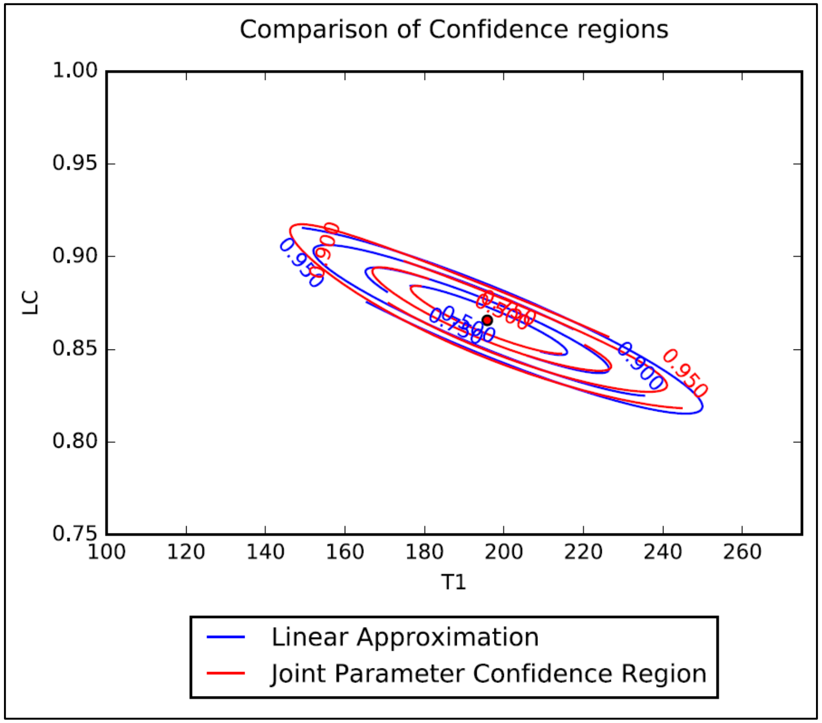
- **The ellipsoidal confidence regions in the transformed space are non-ellipsoidal in unit space**
  - **The result of nonlinear transformations**





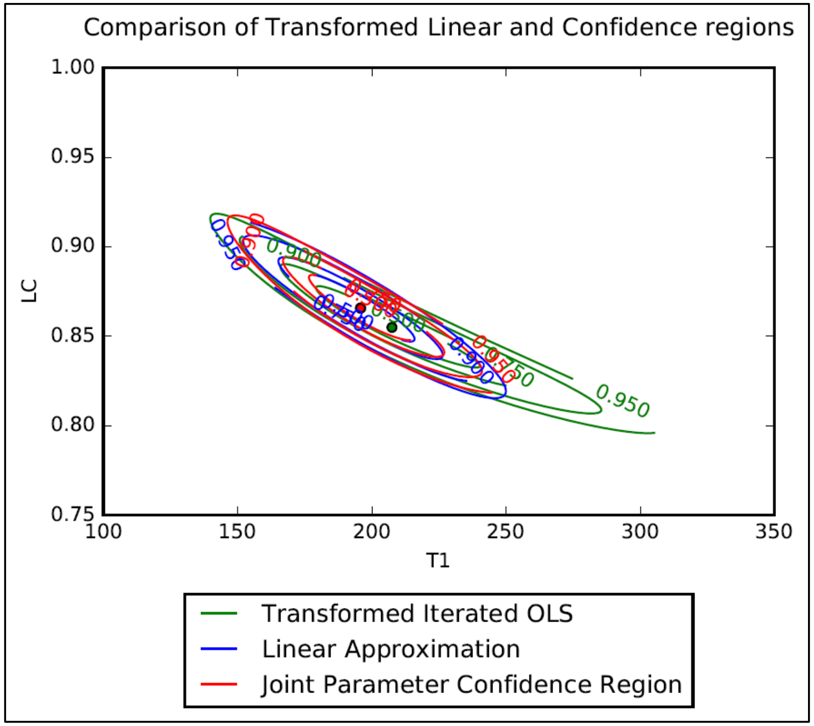
Nonlinear Regression Results

- **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**



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**



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 – $y = \theta_1 \left(1 - e^{-\theta_2 x}\right)$

Weight Data Set

- **Representative data set used for all tests**

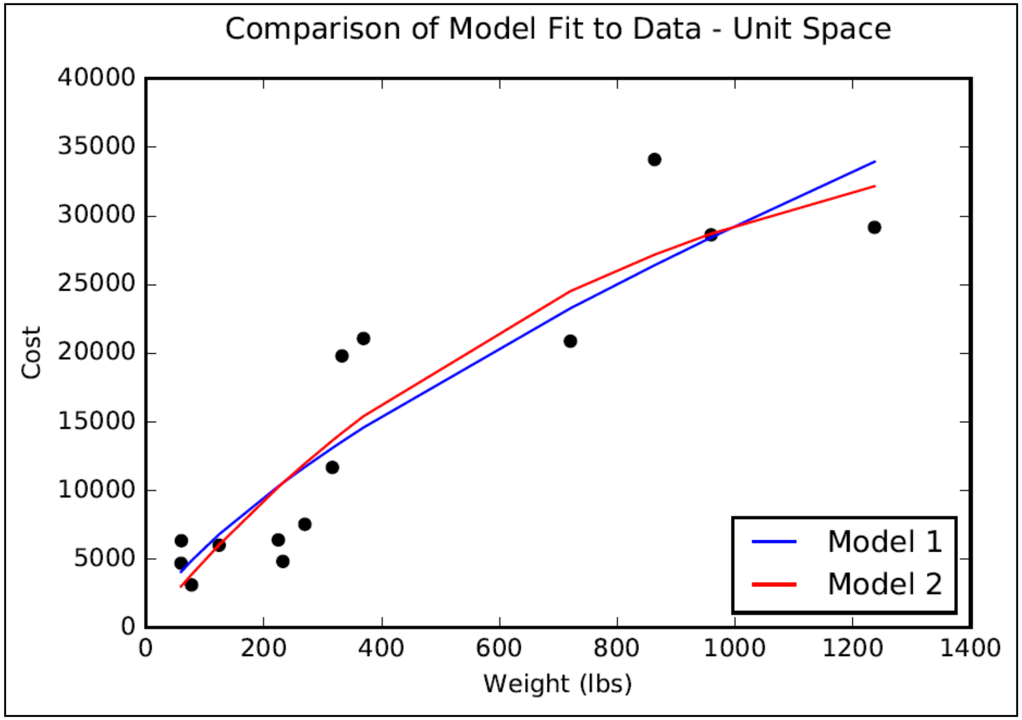
Data point	Cost $K	Weight (lbs)
Obs 1	3,106.64	77.05
Obs 2	29,166.32	1,236.77
Obs 3	4,820.48	232.14
Obs 4	34,111.22	863.36
Obs 5	6,387.04	224.40
Obs 6	20,871.60	720.44
Obs 7	28,621.92	959.33

Data point	Cost $K	Weight (lbs)
Obs 8	19,796.80	332.50
Obs 9	7,526.40	269.42
Obs 10	6,002.24	123.84
Obs 11	11,668.48	316.15
Obs 12	6,329.12	59.77
Obs 13	4,683.20	59.17
Obs 14	21,068.72	369.12

Model Fit Results

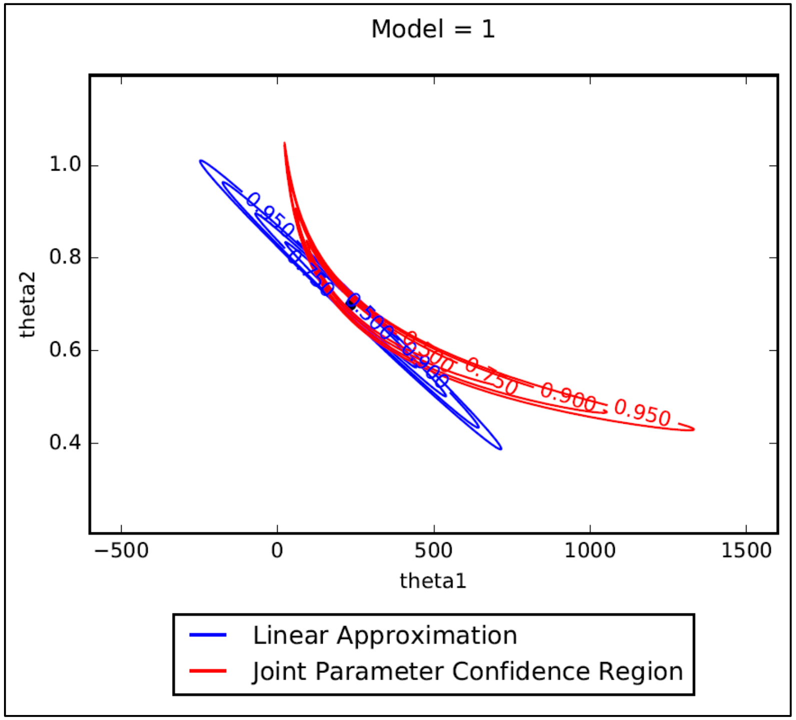
- **Models were designed to have similar fit statistics**
  - Use SE **to measure fit since** R^2 **may not mean much for nonlinear problems**

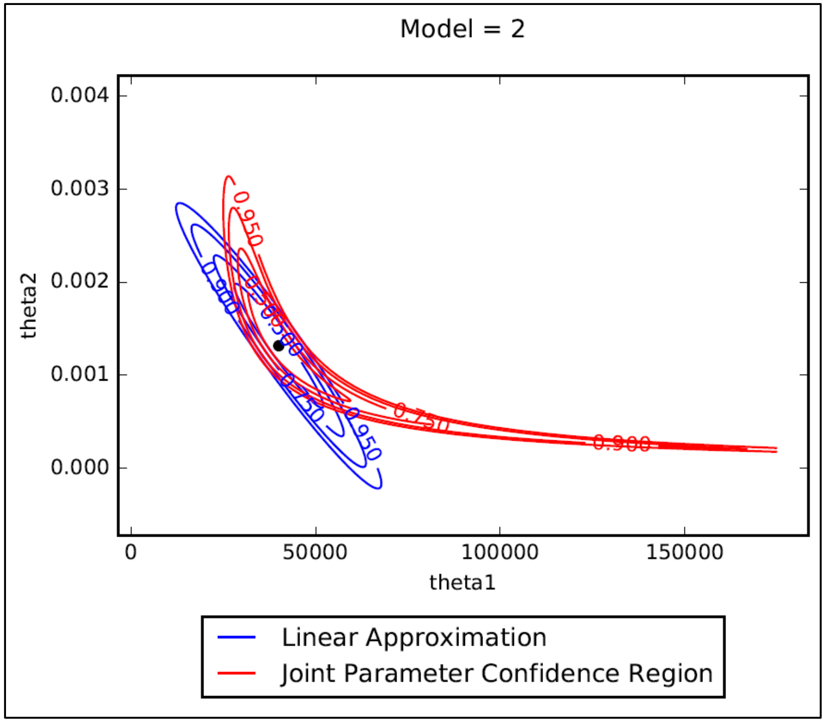
	Model 1	Model 2
θ_1	233.91	40,021.07
θ_2	0.6991	0.0013
SE	4,525.4	4,273.5



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 neighbor hood 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?