Don’t Let the Financial Crisis Happen to You:

Why estimates using power CERs are likely to experience cost growth

Eric Druker, Richard Coleman, Peter Braxton
2009 ISPA/SCEA Professional Development and Training Workshop
St. Louis

Contents

Introduction

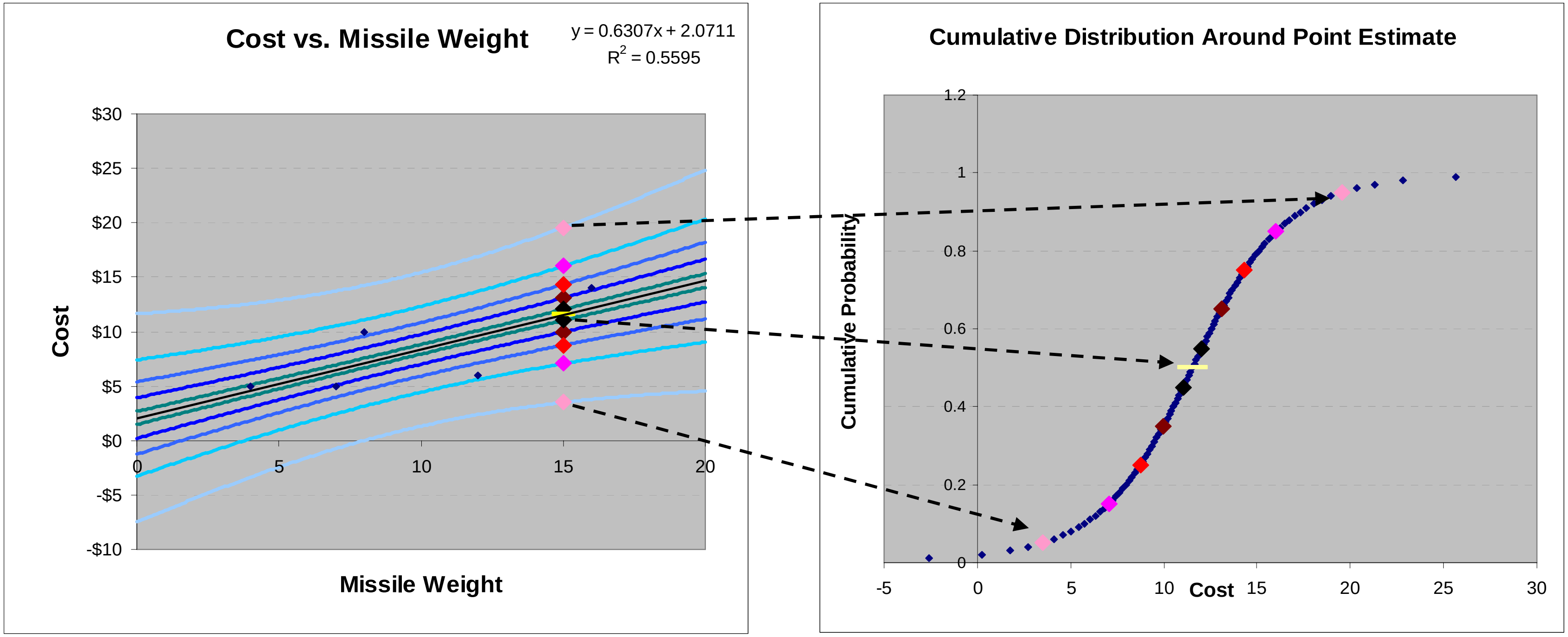
- In similar fashion to the finance industry, most of the methods used in cost estimating and risk analysis are rooted in mathematics
- Over the past 20 years, the finance industry has been grappling with the fact, first described by the French mathematician Benoit Mandelbrot[1], that stock prices are not normally distributed
  - Of particular importance, the probability of significant price drops is higher than described by the tails of the normal distribution
- Underestimating the probability of significant deviations is known as "Kurtosis Risk"
  - Distributions that address this issue are known as fat-tailed distributions
- Although a well-studied phenomenon, kurtosis risk is a recurring theme whenever blame for financial crises is placed on sophisticated mathematical models
  - The failure of Long Term Capital Management in the late 1990's
  - In the latest crises, the financial models used to package mortgages into complex securities underestimated the probability that housing prices would fall
- Is there a lesson that the cost estimating community can learn from finance?
  - Furthermore, is there guidance the cost estimating community can use to mitigate kurtosis risk?

[1] *The (Mis)Behavior of Markets: A Fractal View of Risk, Ruin, and Reward*, by Benoît Mandelbrot and Richard L. Hudson; Basic Books, 2004

Fat-Tailed Distributions

- It turns out that the most common form of regression used in parametric estimates uses a fat-tailed distribution to describe risk around the point estimate
- For Ordinary Least Squares regression techniques, uncertainty around the point estimate is distributed as a t-distribution
  - OLS assumes the population is distributed normally; the fat-tailed distribution arises from a lack of knowledge about the population
  - The t-distribution applies in whatever space in which the function is linear, thus for a power or exponential Cost Estimating Relationship (CER), this is in log-space
- Despite the t-distribution being mathematically correct, most cost estimators and risk analysts model cost risk using a lognormal distribution
- This paper will demonstrate how modeling cost risk using a lognormal distribution will lead to cost growth when power or exponential CERs are used to estimate costs
  - It will then present guidance for cost estimates using CERs using fat-tailed distributions to mitigate kurtosis risk and prevent future cost growth

Cost Risk Around CER-Based Estimates



- For CER-Based Estimates, the prediction intervals fully describe the cost risk distribution
- For log-OLS regressions, risk, which is distributed as a t-distribution in log space, becomes distributed as a log-t distribution
  - Similar to the lognormal distribution where: if X ~ normal, then Y = e^X is distributed as lognormal

Log t distribution description: www.soa.org/files/pdf/edu-exam-c-table07.pdf

[1] *Taking the Next Step: Turning CER-Based Estimates into Risk Distributions*, Eric R. Druker, Richard L. Coleman, Christina M. Kanick, Matthew M. Cain, Peter J. Braxton, SCEA 2008

Booz | Allen | Hamilton

Implications of the log-t distribution

- There are two primary implications of the log-t distribution that need to be accounted for in the cost estimate
  1. For both the lognormal and log-t distributions, the mean cost estimate is a function of both the mean ***and standard deviation*** of the distributions in log space
     - Thus when multiple point estimates developed using log-OLS CERs are added together, the resulting estimate will be **below the 50th percent confidence level** – this is known as the portfolio effect
     - This is a well known and well studied phenomenon: factors such as the PING factor[1] exist to adjust the mean estimate to account for this
  2. The high kurtosis of the log-t distribution causes the mean cost estimate to be *even higher* than that of a lognormal distribution
     - Thus trying to substitute a lognormal distribution for a log-t distribution, even when adjustments are performed, still can cause an understatement of the mean estimate
- The conclusion is inescapable: Modeling cost risk around log-OLS CERs using a lognormal distribution, although better than ignoring risk, can result in cost growth on those programs

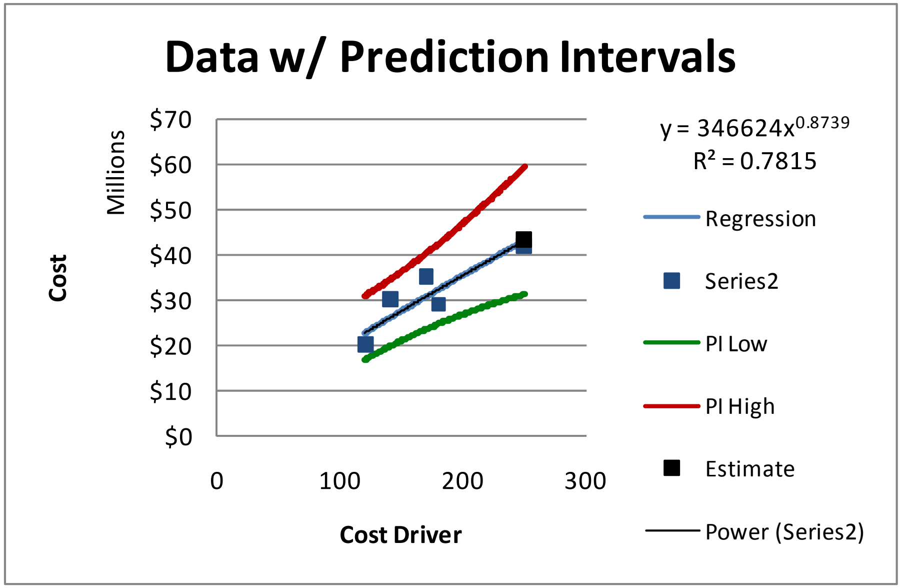
[1]Hu, S., "The Impact of Using Log-Error CERs Outside the Data Range and PING Factor," 5th Joint Annual ISPA/SCEA Conference, Broomfield, CO, 14-17 June 2005

Why Estimates Made From Power CERs Overrun

Cause	Description	Effect
Log-linear regression	For log-linear OLS regressions, error is symmetric in log space, this means it is skewed in unit space	Increases the mean cost estimate
Regression error	No regression is perfect, the error between the best-fit regression line and the historical data increases the uncertainty surrounding the cost estimate	Increases the standard deviation around the cost estimate
Distance of cost driver from mass of data	The error in the estimate is minimized if the cost driver is as the center of mass of the cost drivers from the historical data. Estimating away from this center increases uncertainty surrounding the cost estimate	Increases the standard deviation around the cost estimate
Number of data points	The fewer data points used in the regression, the more uncertainty there is surrounding the cost estimate; also increases the kurtosis of the probabilistic cost estimate	Increases the standard deviation around the cost estimate

For power CERs, the **mean** cost estimate rises when the standard deviation increases, thus all three of these effects lead directly to an increase in the mean cost estimate

Toy Problem Introduction



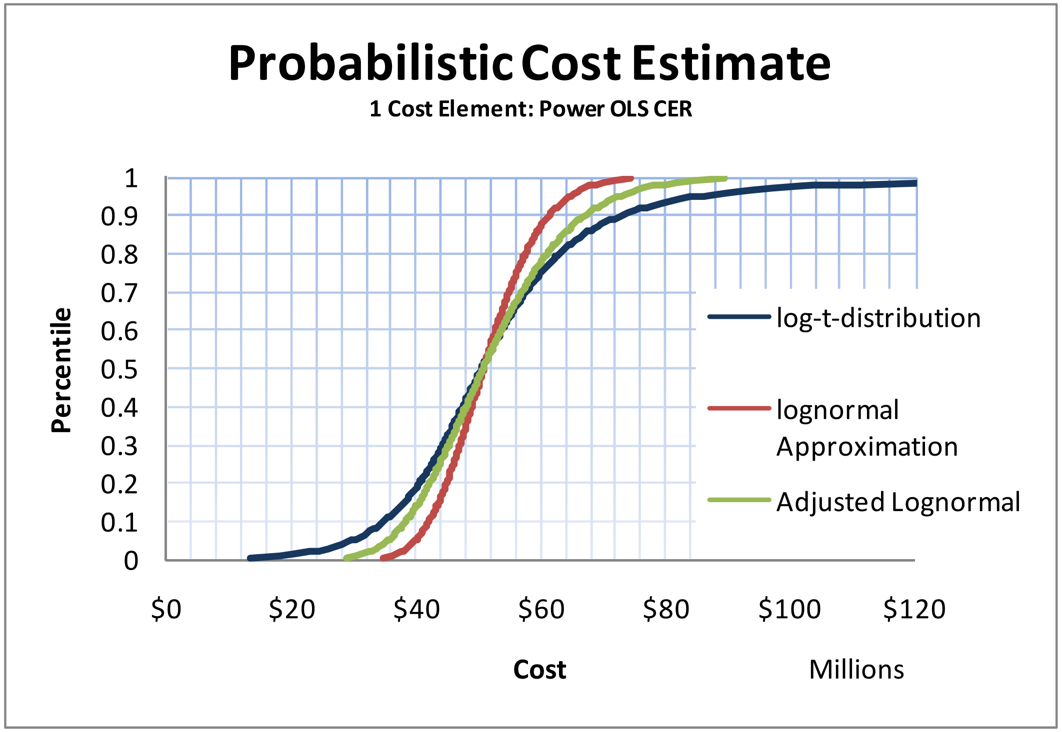
Regression Statistics	
Multiple R	0.88405226
R Square	0.7815484
Adjusted R Square	0.70873119
Standard Error	0.14829338
Observations	5

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.236029043	0.236029	10.73302	0.046561406
Residual	3	0.065972783	0.021991		
Total	4	0.302001826			

- This toy problem will be used to demonstrate how substituting the lognormal distribution for the log-t distribution can lead directly to cost growth
- For the toy problem, a significant power CER was used to estimate costs
  - The estimate is at the high end of the historical data
  - Very few data points (5) were used in the regression to demonstrate the effect of kurtosis risk

S-Curve Comparison



Assumptions

# Data Points	5
Distance of X from Mass of Data	2.1 StDev's
CV of Regression (Log Space)	0.86%
CER Adjustment Factor	1.49

% Error

Percentile	Logormal	Adj Lognormal	Log-t
10%	20%	9%	
20%	10%	4%	
30%	6%	2%	
40%	3%	1%	
50%	0%	0%	
60%	-2%	-1%	
70%	-5%	-2%	
80%	-9%	-3%	
90%	-16%	-8%	
Mean	**-12%**	**-11%**	
Mean %-ile	**52%**	**55%**	**72%**

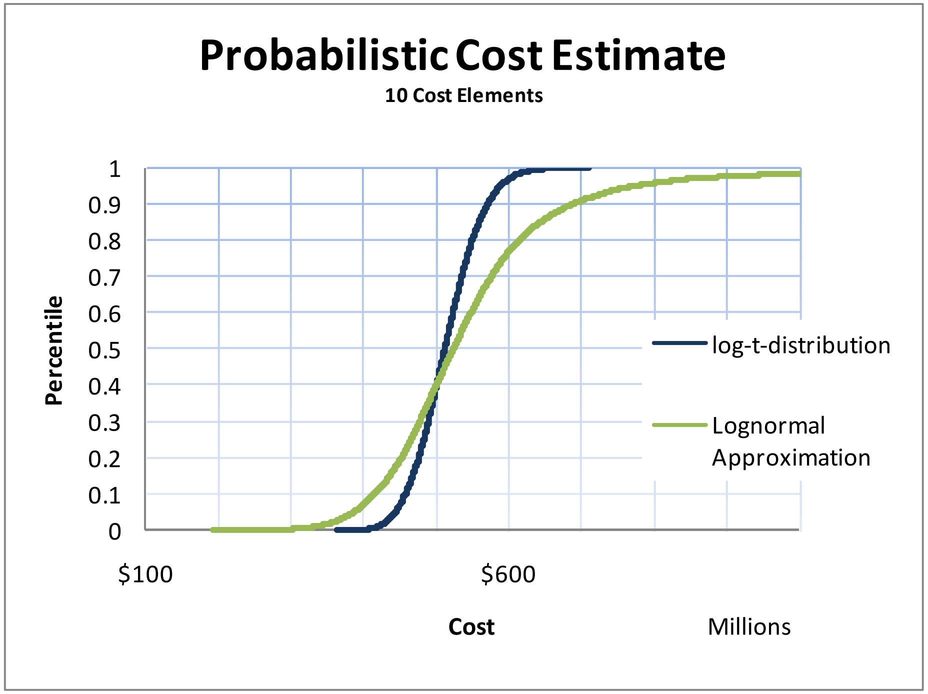
	CV	Excess Kurtosis
t-Distribution	608.0%	9,500.00
Lognormal	15.0%	0.43
Adjusted Lognormal	22.5%	0.82

- Above are the results from a lognormal approximation using the standard error of the estimate, a lognormal distribution using an adjusted standard deviation, and the true log-t distribution
- Observations:
  - Both the lognormal distributions underestimate high percentiles and the mean

Toy Problem Findings

- Misuse of normal or lognormal distributions to characterize cost risk around CER-based estimates can lead to error in the analysis
  - At best, all high percentiles will be underestimated
  - At worst, all high percentiles will be significantly underestimated and both the **mean and the percentile it falls on the risk curve will be underestimated**
- This can have even more significant effects when multiple log-t distributions are summed for either a whole program estimate or in portfolio analysis
- For log-linear OLS regressions, the point estimate from the regression line is the median cost and uncertainty is skewed right
  - When these median point estimates are added together, because uncertainty is skewed right, the sum is below the 50^{th} percentile
  - This is similar to how the modes do not sum together for non-symmetrical triangular distributions
- The next slide demonstrates this effect

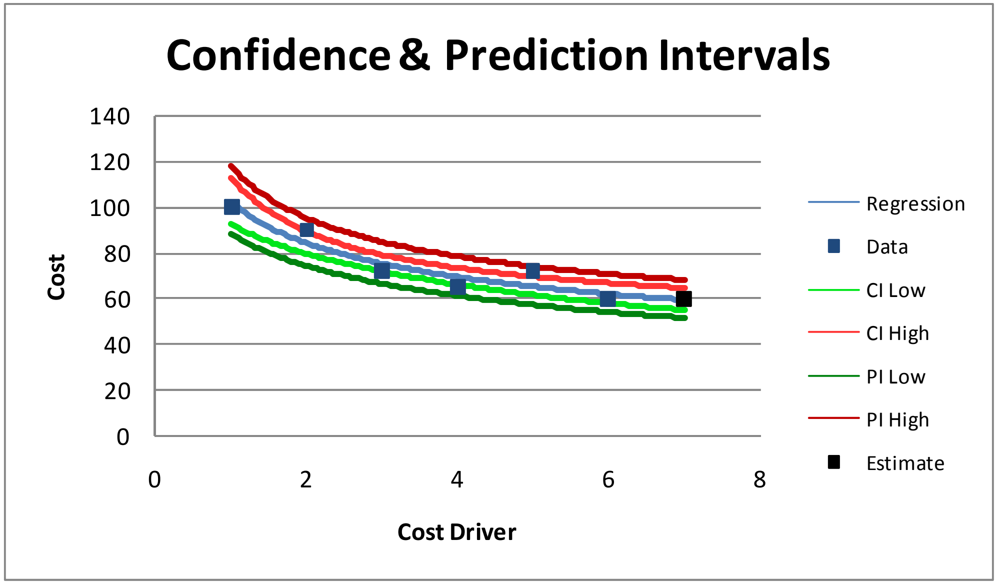
Sum of Multiple Power CER-Based Estimates

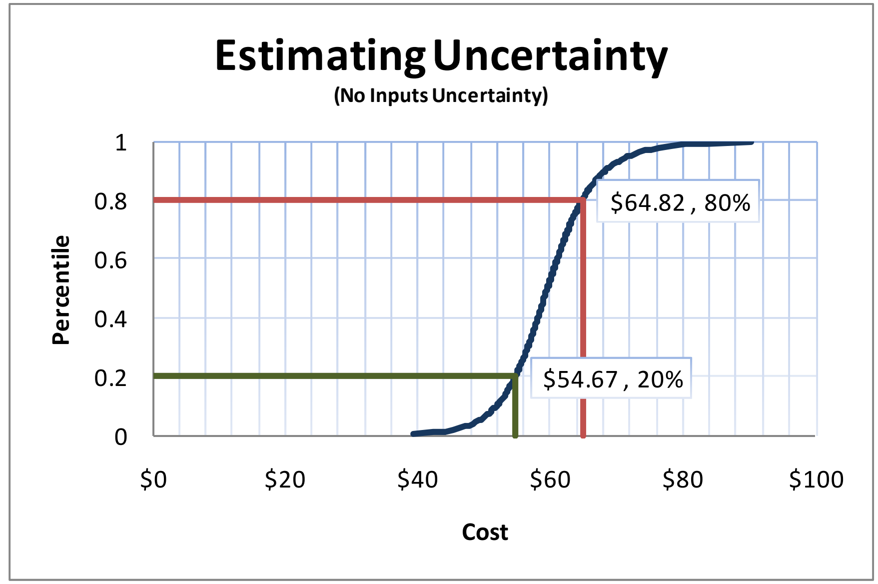


Percentile	Logormal	Log-t	Error
10%	$ 457.48	$ 415.11	10%
20%	$ 475.81	$ 452.34	5%
30%	$ 488.71	$ 477.36	2%
40%	$ 499.96	$ 500.10	0%
50%	$ 511.33	$ 523.14	-2%
60%	$ 522.46	$ 547.59	-5%
70%	$ 534.75	$ 575.79	-7%
80%	$ 549.02	$ 614.15	-11%
90%	$ 570.99	$ 687.28	-17%
Mean	**$513.05**	**$563.66**	-9%
Mean %-ile	**52%**	**72%**	-27%

- This probabilistic cost estimate assumes that 10 cost elements, all estimated using the regression from the toy problem, are summed together to get the total program cost
- When multiple cost elements estimated using power curves are added together, mean and median costs (along with the high percentiles) are understated

Learning Curve Example





- Learning curves are a prime example of when power curves are used to estimate costs
- When a system or pieces of a system is/are estimated using learning curves, the total cost of the system will be understated if kurtosis risk is not accounted for

Recommendations

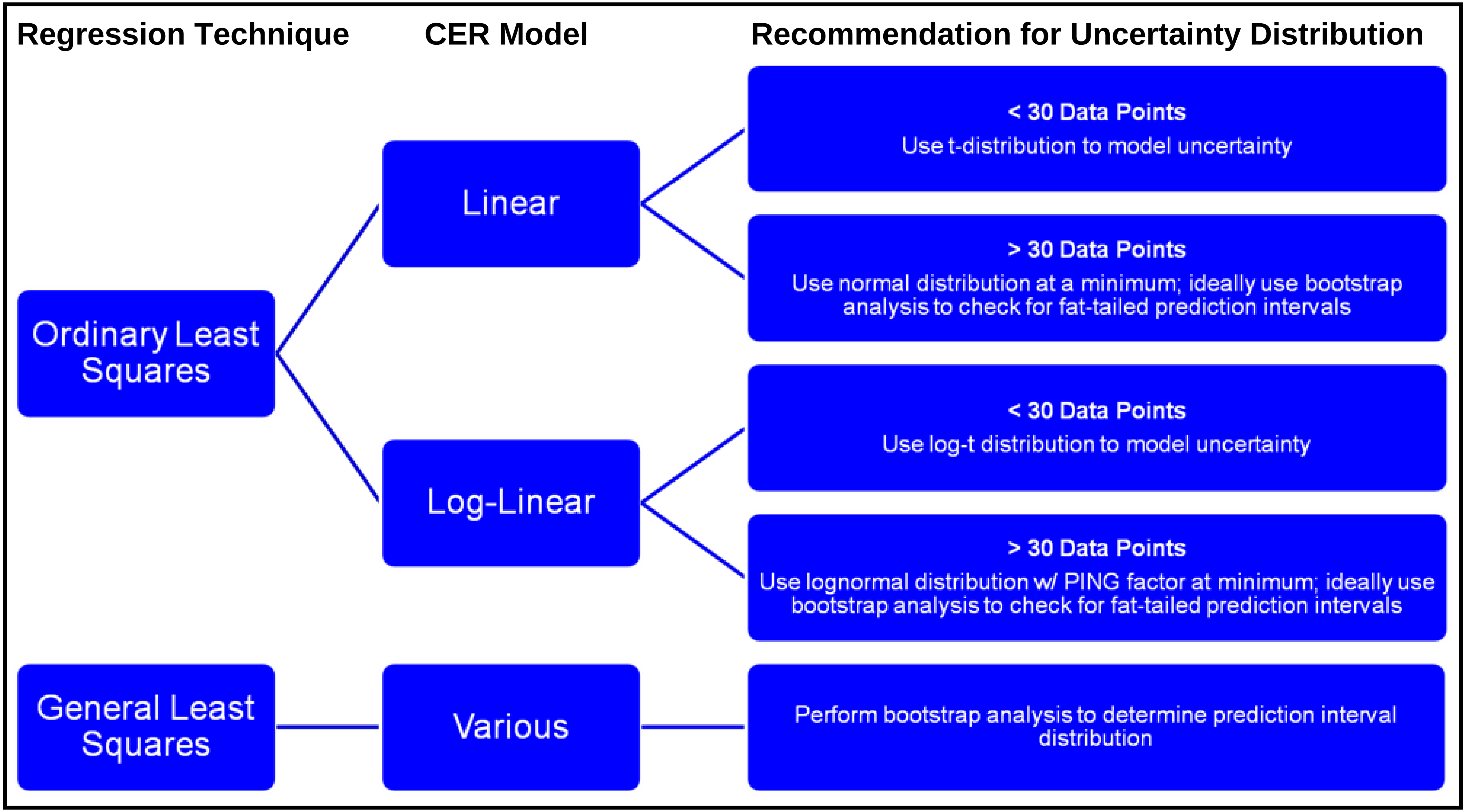
- To prevent the underestimation of cost, it is recommended that **cost risk analysis be performed on all CER-based estimates**
  - At minimum, the PING factor should always be used to adjust log-linear OLS estimates

- The following slide presents CER Risk Analysis Guidance regarding how uncertainty around the cost estimate should be assessed for various CER types

- For OLS regression, the recommendation was split into situations where there are > 30 data points vs. < 30 data points
  - Convention wisdom holds that the t-distribution converges to the normal distribution at approximately 30 degrees of freedom
    - Degrees of freedom are subtracted from the # of data points in regression analysis to account for the independent variables, but 30 was chosen because it is close and simpler to remember

- The recommendation is made that when there are >30 data points, bootstrap analysis be performed to determine if the prediction interval distributions exhibit fat-tailed tendencies
  - Because this is not always feasible, minimum recommendations regarding the normal/lognormal distributions, as well as the use of the PING factor, are provided

CER-Risk Analysis Guidance



Next Steps

- Although factors such as the PING factor do a sufficient job of modeling cost risk when the number of data points is high, would it not be easier to just use actual distributions?
  - These factors, although close, still provide only an approximation
  - If the actual distributions are always applied, there is no need to rely on the factors or have guidance depending on the number of data points
- With very little work, risk models can be adapted to model these distributions
  - Although not contained in all COTS risk tools, they can be easily adapted to allow the log-t distribution to be used
  - Ideally the distributions could be included in the COTS risk tools in the future
- An example of this is Booz Allen's Regression & Risk Analysis Methodology Streamliner (RAMS) tool[1]
  - This tool automates both regression and the creation of cost risk distributions around CER-based estimates
  - The tool also includes interface to work with the COTS risk tools

[1] Making Statistical Analysis Accessible: The RAMS Tool for Regression & Risk Analysis, Lytton, Matthew, Druker, Eric, Hogan, Greg. SCEA 2009

Conclusion

- This paper is written for two audiences:
  1. We hope that those estimators already adjusting their log-linear CER based estimates using the common factors understand how kurtosis risk can cause costs to be understated under certain conditions
  2. We hope that those estimators who are not currently adjusting their log-linear CER based estimates are now aware that this causes an underestimation of both cost, cost risk, and cost uncertainty

- It is hoped that the guidance presented here will lead to more awareness surrounding this issue and less underestimation of cost risk

- As a next step, the authors would like to expand this paper to cover the generalized least squares regression methods
  - Dr. Book presented a paper on using the bootstrap method to develop prediction intervals around generalized least squares estimates[1]
  - The authors would like to study how the risk distributions arising from these prediction intervals behave in regards to kurtosis

[1]Prediction Bounds for General Error Regression CERs, Stephen A. Book, DoDCAS 2006