Modeling Prediction Intervals using Monte Carlo Simulation Software

2016 ICEAA Professional Development
& Training Workshop

James R. Black

Qing Qing Wu

June 2016

Bios

- Presenter: James “Jay” Black has 12 years of cost estimating experience and currently works as senior operations research analyst for the Administration for Children and Families within the U.S. Department of Health and Human Services. In this role, he supports the Grants Center of Excellence software suite used to administer 1200 grant programs in eight Federal departments. Jay has a Masters in Systems Engineering from Johns Hopkins University and holds a current CCE/A certification.
- Co-author: Qing Qing “Q” Wu is a cost analyst for the Cost Effectiveness Branch at the Naval Surface Warfare Center Carderock Division. She supports the Naval Sea Systems Command 05C Cost Engineering & Industrial Analysis Division in their Weapon Systems Division. She has a Bachelor’s degree in Mathematics from the Macaulay Honors College at The City College of New York.

Presentation Summary

- References/Acknowledgements:
  - CEBok Module 9 - Cost and Schedule Risk Analysis
  - 2014 ICEAA Workshop presentation prepared by Dr. Christian Smart (MDA) and Marc Greenberg (NASA)
  - Joint Agency Cost Schedule Risk and Uncertainty Handbook (CSRUH, Feb 2014)
- Presentation abstract:
  - The use of a prediction interval (PI) is a simple method of quantifying risk and uncertainty for a Cost Estimating Relationship (CER) derived from an Ordinary Least Squares (OLS) regression
  - Yet, few cost estimators implement PIs in their estimates despite their frequent use of CERs

This presentation will provide a step-by-step tutorial for modeling a PI for an example CER using Monte Carlo Simulation software and will identify the beneficial impact on the coefficient of variation (CV)

Cost Estimating Relationships (CERs)

- Definition: A Cost Estimating Relationship (CER) is a mathematical expression of cost as a function of one or more independent variables
- CERs are often developed using regression analysis to fit an equation to a data set
- Examples of equations used for CERs include:

Linear CER: $y = a + bx$

Nonlinear CERs: $y = ax^b$

$y = ab^x$

$y = a + bx^c$

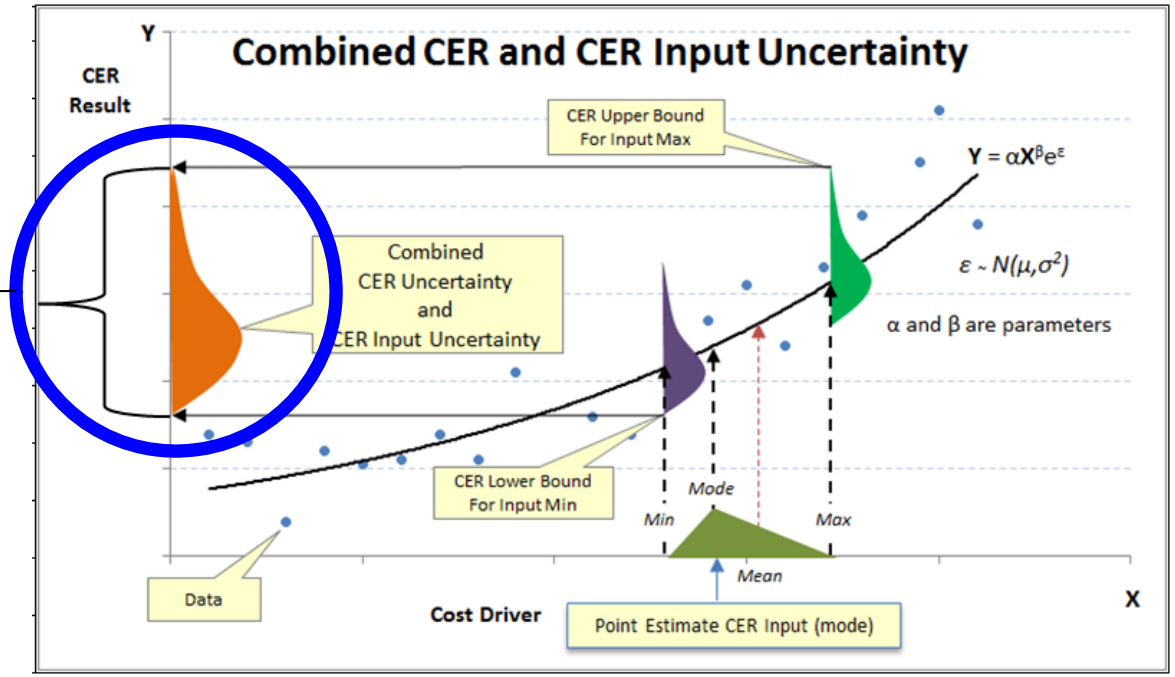
where y = Cost

x = Technical Parameter

Modeling Uncertainty

- CERs do not perfectly fit historical data upon which they are based
- This results in an underlying uncertainty distribution about an estimate
  - The outcome of a CER represents only one point on an uncertainty distribution (typically mean or median)

This brief will model this uncertainty



Modeling Uncertainty (cont.)

Model uncertainty is variation about the dependent variable, i.e., cost

For a linear CER: $Y = a + bX + \varepsilon$

Often used to create weight based estimates

For a nonlinear CER: $Y = aX^{b}\varepsilon$

Often used to model learning curve

where ε represents the error between the estimated cost and the actual cost Y; the estimate uncertainty is captured by the Prediction Interval

Example: Modeling Uncertainty for a Linear CER

- For example, consider a linear CER: $Y = a + bX + \varepsilon$
- Using Monte Carlo simulation software (e.g. Palisade @Risk or Oracle Crystal Ball), define a distribution for ε:
  - ε = normal(mean = 0, std dev = prediction error)
  - OR
  - ε = student-t(midpoint = 0, scale = prediction error, degrees of freedom)

Ok, so how do you define prediction error?

Prediction Interval Equation

$$\hat{Y} \pm t_{\alpha/2,\,df} \times SEE \sqrt{\frac{n+1}{n} + \frac{(X - \overline{X})^2}{\sum X_i^2 - n\overline{X}^2}}$$

Prediction Error

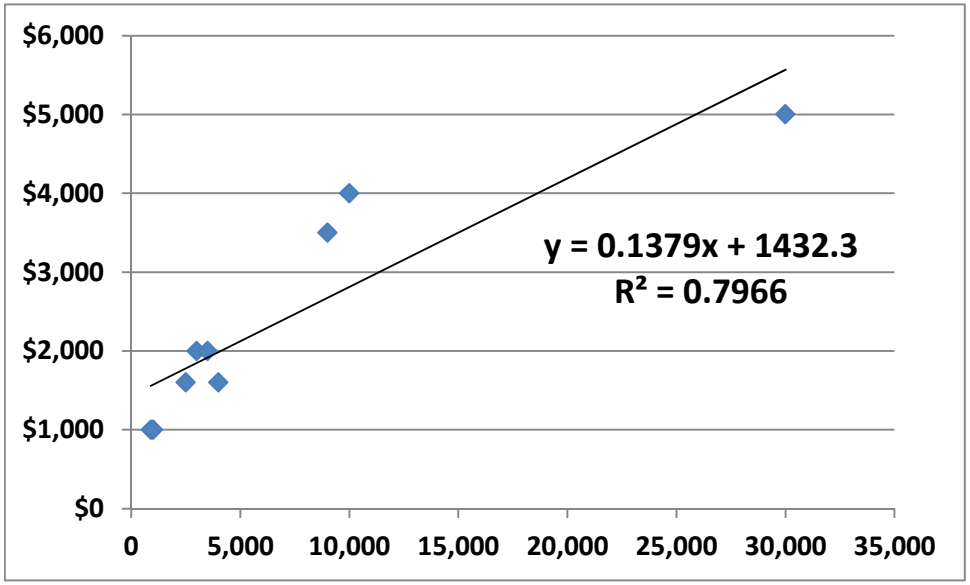
- $\hat{Y}$ = Calculated Value from Regression Line
- $t_{\alpha/2,\,df}$ = t Critical Value (T.INV.2T function in Excel)
- SEE = Standard Error of the Estimate (STEYX function in Excel)
- n = number of observations
- $\overline{X}$ = average of X
- $\sum X_i^2 - n\overline{X}^2$ = sum of squared deviations of X from its mean (DEVSQ function in Excel)

Linear CER Example: Modeling Uncertainty with X = 5000

Example Dataset

Development $M (BY12$)	Weight Lbs.
$1,000	900
$1,000	1,000
$1,600	2,500
$2,000	3,000
$2,000	3,500
$1,600	4,000
$3,500	9,000
$4,000	10,000
$5,000	30,000

OLS Regression*



$$Y = 0.1379x + 1432.3 + \varepsilon$$

- Define distributions using Monte Carlo Simulation software:
  - x = triangular(low = 4000, most likely = 5000, high = 7000)
  - ε = student-t(midpoint = 0, scale = prediction error, degrees of freedom = 9 – 1 – 1 = 7)

** Note: the use of an Excel trendline and the focus on R^2 is for presentation brevity, make sure you consider T & F-Stat, R^2 adj, and other fit measures when running and evaluating a regression on your own*

Linear CER Example: Modeling Uncertainty with X = 5000

$$\text{Prediction Error} = SEE\sqrt{\frac{n+1}{n}+\frac{(X-\bar{X})^2}{\sum X_i^2 - n\bar{X}^2}}$$

Example Dataset

Development $M (BY12$)	Weight Lbs.
$1,000	900
$1,000	1,000
$1,600	2,500
$2,000	3,000
$2,000	3,500
$1,600	4,000
$3,500	9,000
$4,000	10,000
$5,000	30,000

Set Up the Inputs

n=count(Development $)
SEE=STEYX(Development $, Weight)
Avg=AVERAGE(Weight)
Devsq=DEVSQ(Weight)

n=9
SEE=682.93
Avg=7100
Devsq=672620000

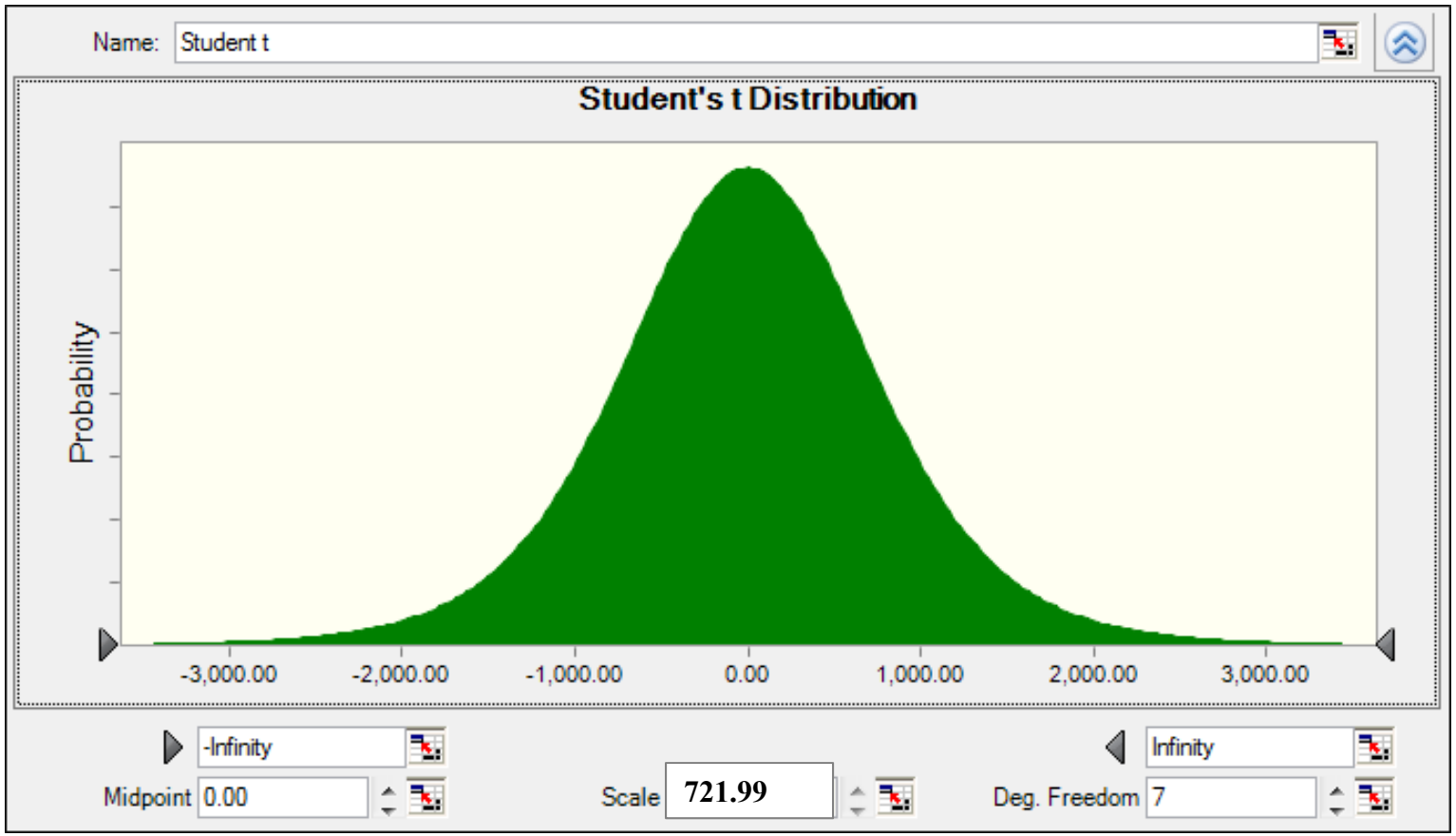
X= 5000

Evaluate Prediction Error where X = 5000

=SEE*(SQRT(((n+1)/n)+(((X-Avg)^2)/Devsq)))

=721.99

Linear CER Example: Enter Inputs in Student-t Distribution



Inputs to the Student-t distribution:

- Midpoint: 0
- Scale: Prediction Error = 721.99
- Deg. Freedom: n-k-1 = 9 – 1 – 1 = 7

Linear CER Example: Define Distributions and Set Forecast

- Define distributions using Monte Carlo Simulation software:
  - x = triangular(low = 4000, most likely = 5000, high = 7000)
  - ε = student-t(midpoint = 0, **scale = prediction error = 721.99,** degrees of freedom = 9 – 1 – 1 = 7)

$$Y = 0.1379x + 1432.3 + \varepsilon$$

- Also using Monte Carlo Simulation software, set the forecast on the dependent variable

Run the simulation* and capture the results

** Note: prudent estimators will also assign correlation between each distribution before running the simulation, for more details on correlation see CEBok Module 9 - Cost and Schedule Risk Analysis*

Linear CER Example: Run the Simulation Using Two Scenarios

- The following slides identify two scenarios and their results when using the linear CER example:

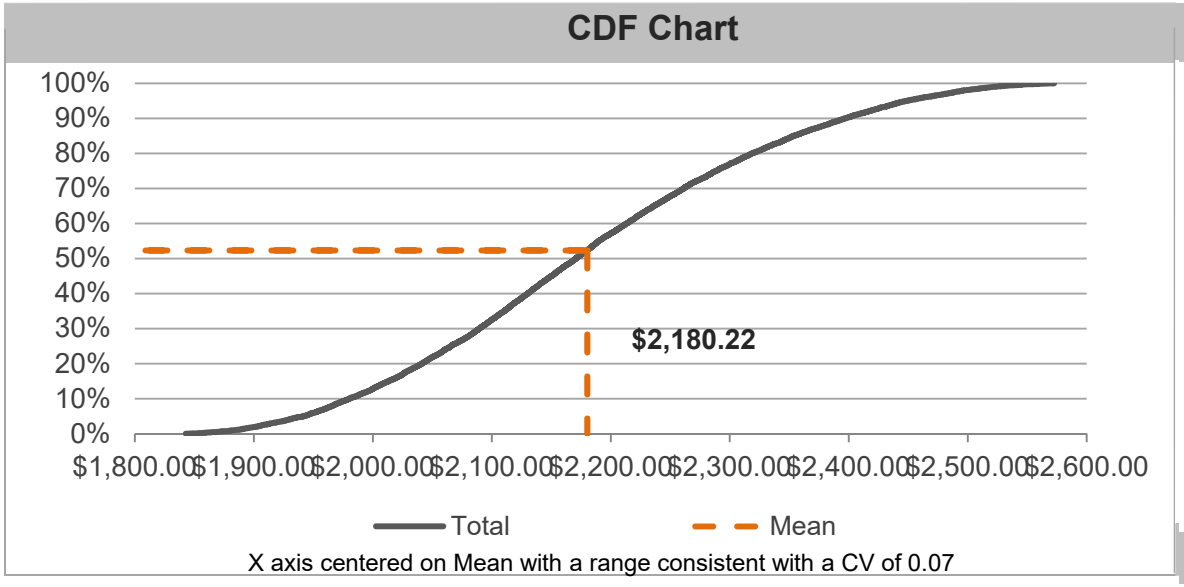
$$Y = 0.1379x + 1432.3 + \varepsilon$$

  - Scenario #1: Constraining the simulation to only consider the triangular distribution on the independent variable X
  - Scenario #2: Allowing the simulation to consider both the triangular distribution on the independent variable X as well as the student-t distribution on ε

The results of Scenario #2 identify the beneficial impact on the CV when modeling the PI

Scenario #1 Results

Risk Only on Independent Variable



Deciles	
90%	$2,396.70
80%	$2,318.78
70%	$2,261.32
60%	$2,213.26
50%	$2,170.60
40%	$2,130.13
30%	$2,089.70
20%	$2,041.35
10%	$1,980.35
CV	7.0%

+14% more $

Basis and Values of Risk Parameters

Risk Parameter	Min	Most Likely	Max
Weight Dist	Weight Low (10%) 4000	Weight Most Likely 5000	Weight High (90%) 7000

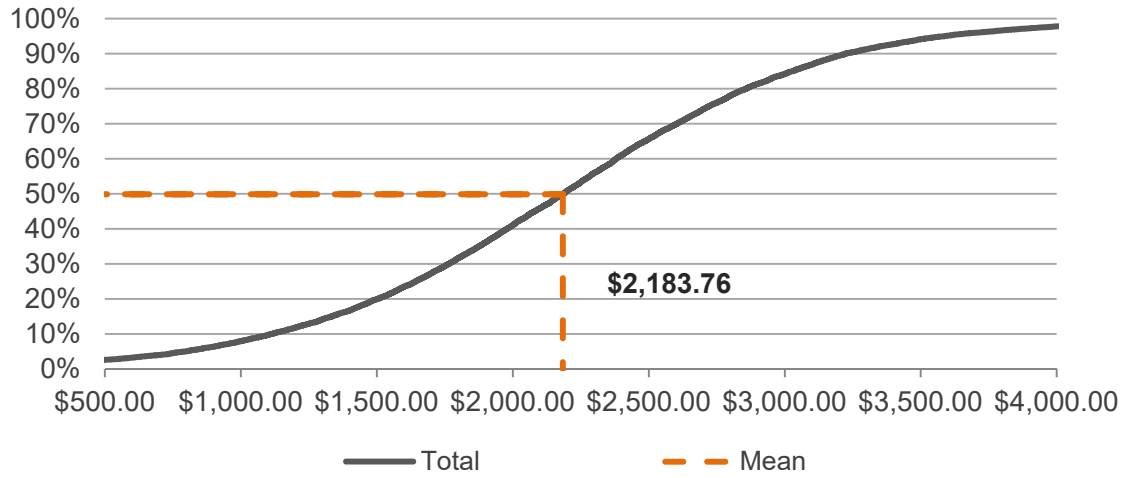
Note: Regression of the original dataset had a $R^2 = 0.7966$

Going from 20% to 80% confidence requires +14% more $; i.e. the CV of 7% is low

Scenario #2 Results

Risk on Independent Var. & Error Term

CDF Chart



X axis centered on Mean with a range consistent with a CV of 0.4

Deciles	
90%	$3,223.41
80%	$2,858.32
70%	$2,603.07
60%	$2,381.09
50%	$2,185.85
40%	$1,978.97
30%	$1,764.76
20%	$1,502.36
10%	$1,114.86
CV	39.9%

+90% more $

Basis and Values of Risk Parameters

Risk Parameter	Min	Most Likely	Max
PI Dist	**Student-t Distribution Parameters**: Midpoint = 0, Scale = 721.99 (Prediction Error) Degrees of Freedom = 7 (n-2)		
Weight Dist	Weight Low (10%) 4000	Weight Most Likely 5000	Weight High (90%) 7000

Note: Regression of the original dataset had a $R^2 = 0.7966$

Going from 20% to 80% confidence requires +90% more $; the CV improves when the simulation also considers the PI

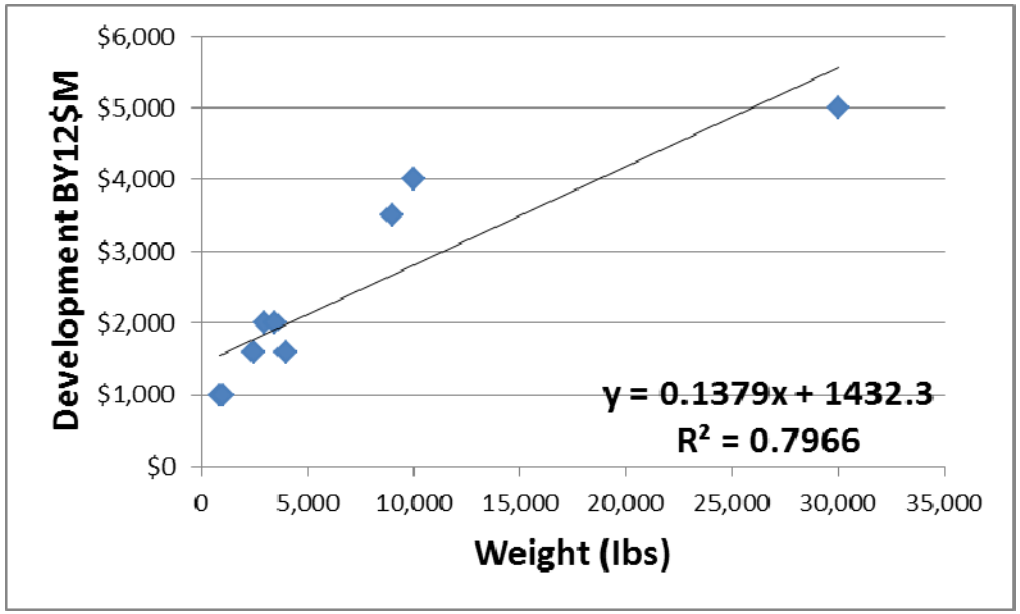
Summary

- Implementing risk on the error term using the prediction interval is not difficult
- Even for regressions with reasonable fit statistics, implementing risk on the error term can produce desirable CVs

BACKUP

Linear Regression Example
Fit Statistics

Development $ Millions (BY12)	Weight Lbs.
$1,000	900
$1,000	1,000
$1,600	2,500
$2,000	3,000
$2,000	3,500
$1,600	4,000
$3,500	9,000
$4,000	10,000
$5,000	30,000



Variable	Coefficient	Std Dev of Coef	Beta Value	T-Statistic (Coef/SD)	P-Value	Prob Not Zero
Intercept	1432.2776	294.5779		4.8621	0.0018	0.9982
Lbs	0.1379	0.0263	0.8925	5.2355	0.0012	0.9988

Goodness-of-Fit Statistics

Std Error (SE)	R-Squared	R-Squared (Adj)	Pearson's Corr Coef
682.9319	79.66%	76.75%	0.8925

Analysis of Variance

Due To	DF	Sum of Sqr (SS)	Mean SQ = SS/DF	F-Stat	P-Value	Prob Not Zero
Regression	1	12784117.1836	12784117.1836	27.4104	0.0012	0.9988