

***Cost Estimating Relationship Risk
And S-Curves
based on the
Cost Estimating Body of Knowledge***

***Christian Smart, Ph.D., CCEA
Director, Cost Estimating and Analysis
Missile Defense Agency***

San Diego -- June 2015

Assessing Risk for a CER

- This training module focuses on how to assess risk and uncertainty for CERs that are derived from regression analysis
- Sources of uncertainty are addressed
 - Input uncertainty
 - Model uncertainty
 - External Factors
- The importance of addressing model uncertainty is discussed in detail
 - Standard errors and prediction intervals are compared

Motivation

- Cost estimates project years into the future
 - Cost estimates are inherently uncertain, regardless of whether risk is incorporated (i.e., explicitly modeled)
- Numerous cost models and cost modeling platforms enable cost risk estimating and cost risk has become an integral part of the cost estimating process
 - Cost risk is included in NASA policy directives, and in the Weapon Systems Acquisition Reform Act of 2009 for Dept. of Defense programs
- Even though cost risk analyses are commonly conducted for government projects, these attempts often result in a vast understatement of risk

Terminology

- Cost Growth:
 - Increase in cost of a system from inception to completion
- Cost Risk:
 - Predicted Cost Growth
- Uncertainty and Risk
 - Range of possible estimates vs. calibration
- Risks and Opportunities
 - Bad vs. good outcomes for events which may happen

Cost Risk Analysis

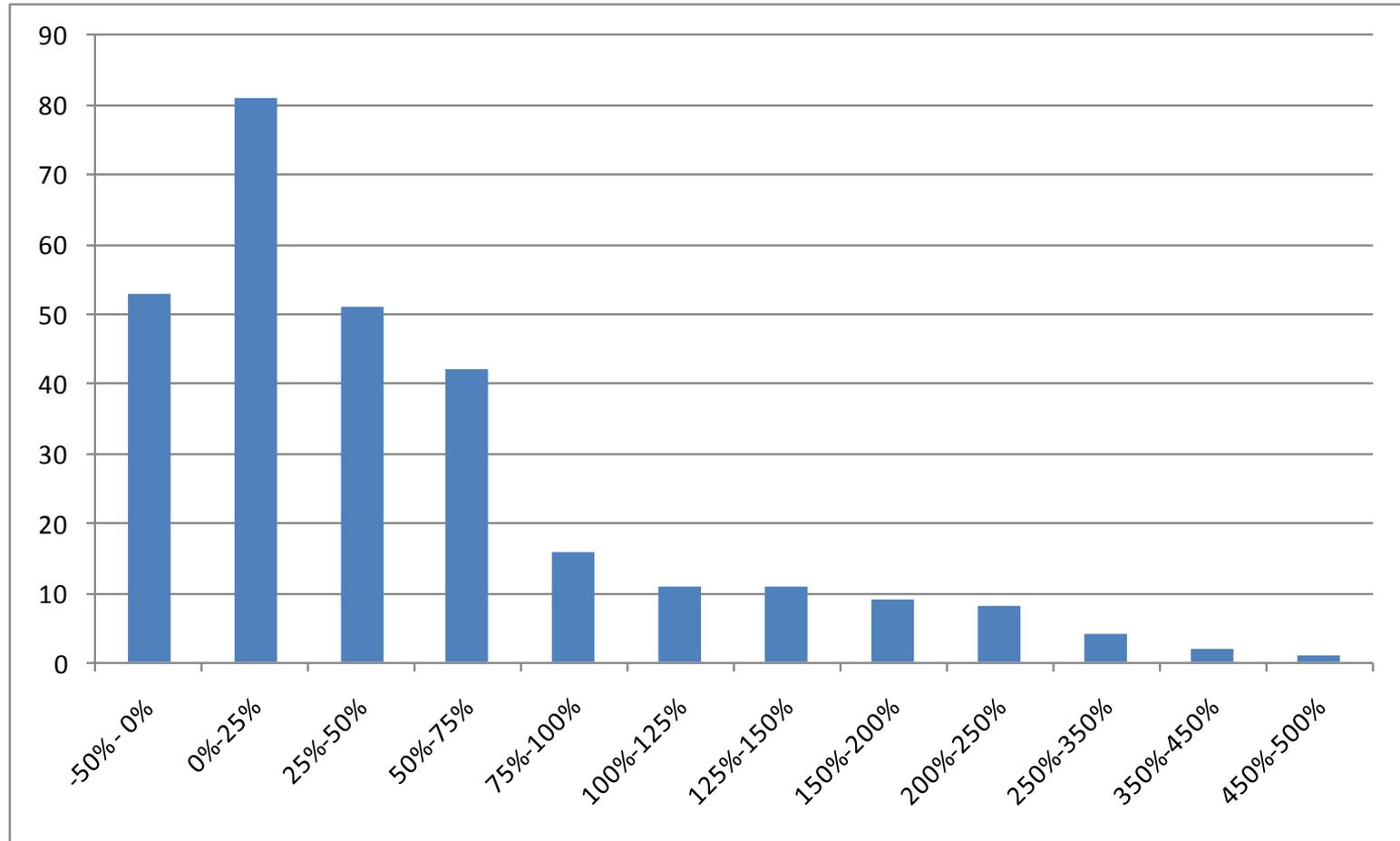
- The cost analysis profession recognizes the importance of risk and has incorporated risk analysis as an **integral** part of the estimating process
- Sources of uncertainty include (but are not limited to)
 - Scope and requirements (what is being estimated)
 - Technology readiness
 - Manufacturing readiness
 - Inflation
 - Model uncertainty

**“The only certainty is uncertainty”
Pliny the Elder AD 23-79
Roman Senator
Died at the Mt. Vesuvius Eruption**

Cost Growth Data

- If our cost estimates are realistic, then our point estimates should be high enough that cost growth is not a big problem
- However this is not the case in practice – for a database of 289 NASA and DoD missions, Smart (2011) found that:
 - Over 80% of development projects experience cost growth
 - Average (mean) cost growth is over 50%
- Thus point estimates should be expected to have an 80% chance of cost growth
- Thus cost risk analysis is a must, not just “nice to have”

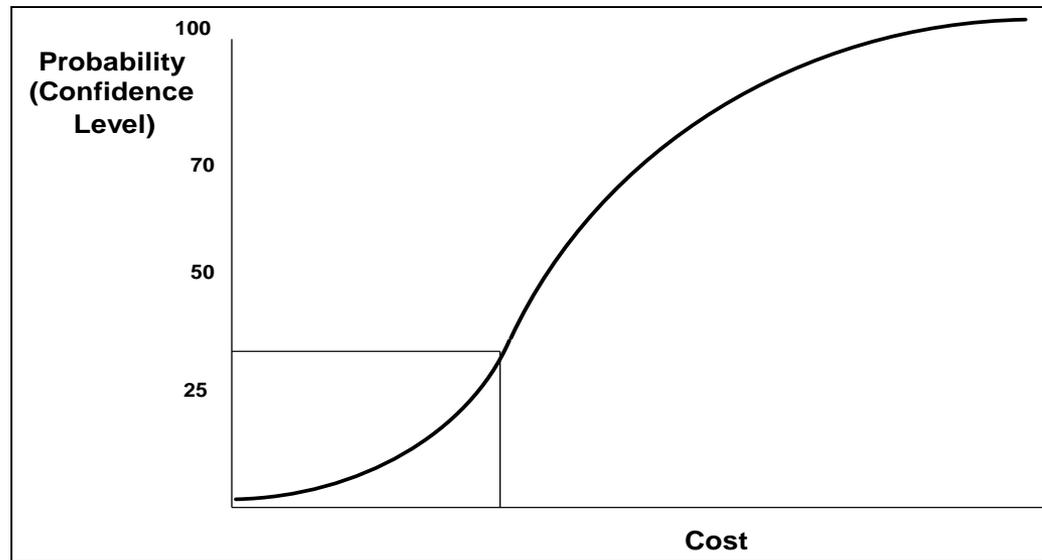
Cost Growth Histogram



Historical cost growth for 289 NASA and DoD missions

Cost Risk and S-Curves

- S-curves, or cumulative distribution functions, are a common way to display the results of risk analysis



- Note that the term “confidence level” is used interchangeably with “percentile of the cumulative probability distribution”

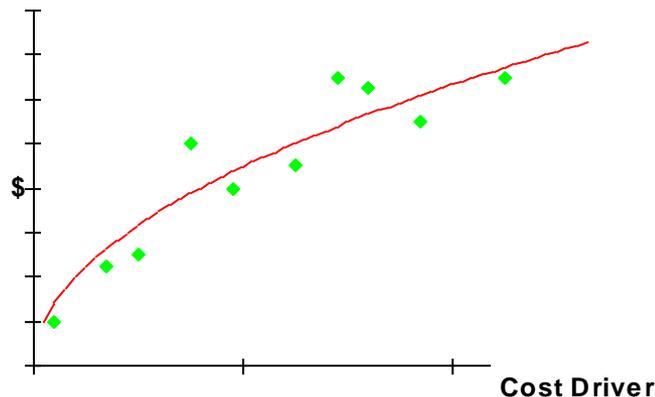
Sources of Risk Understatement

Area	Source	Mean & 50 th	Standard Deviation	80 th
Cost	Errors Which Seem "Always To Understate"	Understate	-	Understate
	Lack Of Basis In Historical Data	Understate	-	Understate
	Omissions of Elements	Understate	-	Understate
	Systematic Understatement In Non-linear CERs	Understate	-	Understate
Risk	Omission Of Risks And Elements Of Bias	Understate	Understate	Understate
	Omission Of Elements Of Variability	-	Understate	Understate
	Inadequate Determination Of Cost Relationships	-	Overstate	Overstate
	Failure To Include Functional Correlation	-	Understate	Understate
	Errors Which Seem "Always To Understate"	-	Understate	Understate
	Omission Of Correlation Of Any Type	-	Understate	Understate
	Insufficient Data Causing Unrecognized Wide(r) Prediction Intervals	-	Understate	Understate
	Systematic Understatement In Non-linear CERs	-	Understate	Understate

What Percentile Are We At Now (And Where Are We Going?), R. Coleman, E. Druker, P. Braxton, B. Cullis, C. Kanick, SCEA 2009, DoDCAS 2010.

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
- Cost Estimating Relationships are often developed using regression analysis to fit an equation to a data set



Equation Forms

- Examples of equations used for CERs include:

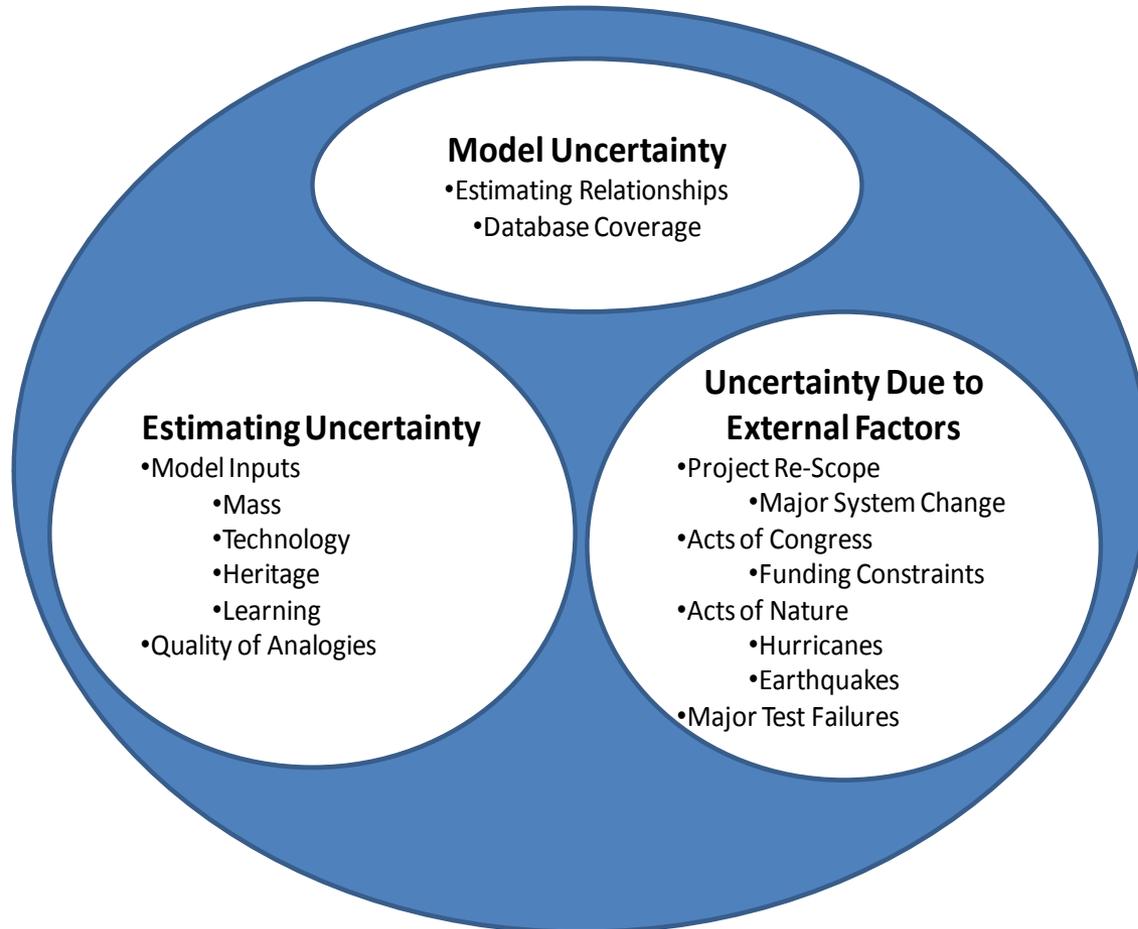
- Linear CER: $y = a + bx$

- Nonlinear CERs:
 $y = ax^b$
 $y = ab^x$
 $y = a + bx^c$

where $y = \text{Cost}$
 $x = \text{Technical Parameter}$

- For more on this subject, see the material presented in PAR02: Cost Estimating Relationships

Sources of Uncertainty



Source: “Cost Risk for CLV,” Charles Hunt, 2006

CER Uncertainty

- Sources of uncertainty:
 - Estimating uncertainty
 - Accounted for by modeling uncertainty on the CER independent variables (aka variates)
 - Model uncertainty
 - Accounted for by modeling uncertainty on cost, the CER's dependent variable (aka co-variate)
 - Accounted for in the standard error, confidence intervals, and prediction intervals
 - External factors
 - Partially accounted for by the standard error and the prediction interval, to the extent to which these factors are in the historical data
 - The process of removing outliers from the data set (“cherry-picking”) will remove much of this effect on a CER

Process for Modeling CER Uncertainty

- Assess estimating uncertainty
- Assess model uncertainty
 - Standard Errors and Prediction Intervals
- Assess external factor uncertainty
- Combine these sources of uncertainty into an S-curve for the CER

Example Data

- This notional data set for missile costs, weights, and new design percentages will be used throughout the training

Development \$ Millions (BY12)	Weight Lbs.	New Design %
\$1,000	1,000	70%
\$2,000	3,000	100%
\$1,600	2,500	30%
\$1,000	900	90%
\$2,000	3,500	50%
\$3,500	9,000	50%
\$5,000	30,000	70%
\$4,000	10,000	100%
\$1,600	4,000	20%

- Also, for the example missile system being estimated, the planned weight is equal to 12,000 lbs., with new design equal to 100%

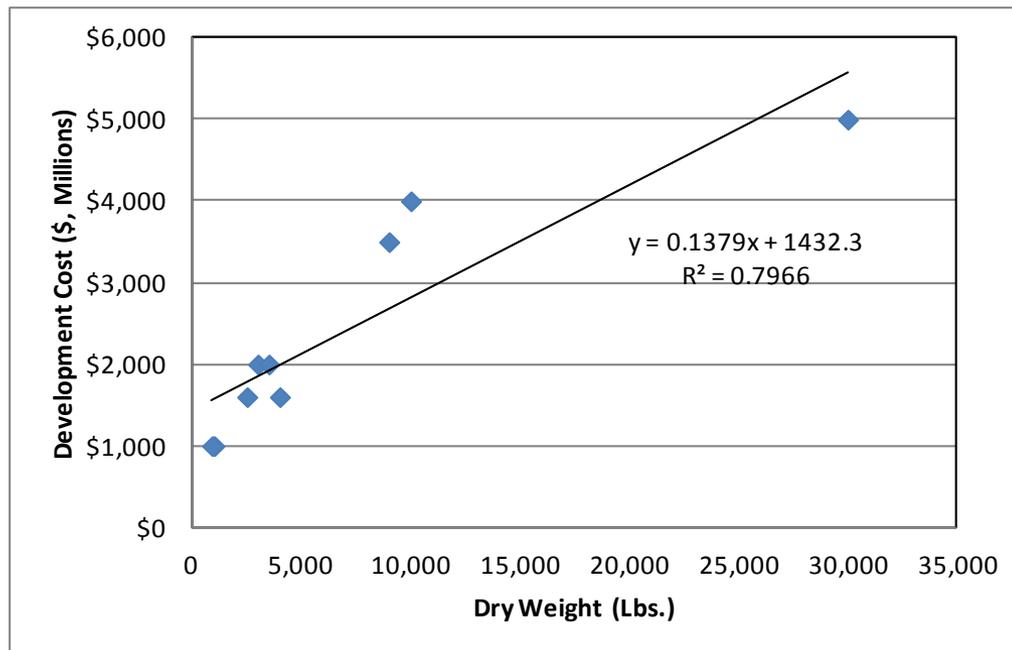
Example CER: OLS

- For the example data set, the best ordinary least squares fit for cost and weight is given by the equation

$$\text{Cost} = \$1,432 \text{ Million} + 0.1379 * \text{Weight}$$

where Weight = mass, in pounds

Cost = \$ Millions, in Base Year FY12

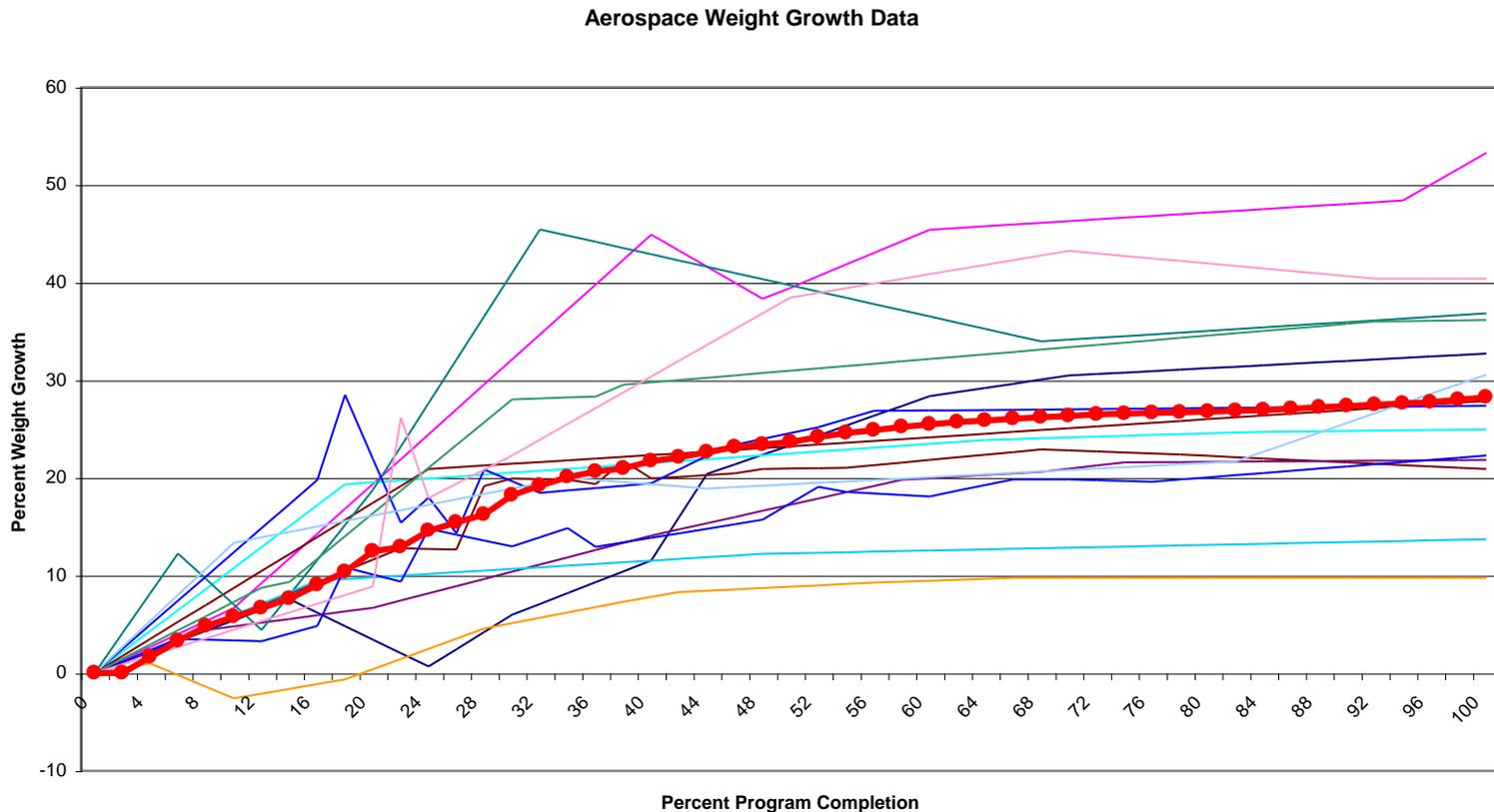


Measuring Estimating Uncertainty

- Also called technical risk, this involves the assessment of uncertainty about the CER's independent variables
- For the example, this involves assigning probability distributions for weight and new design
- Two approaches:
 - Use data (preferred)
 - Possible for quantitative variables for which you have historical growth data, such as weight
 - Use judgment
 - Have to rely on this in many instances, if you don't have historical growth data, especially for qualitative data, such as new design

Weight Growth

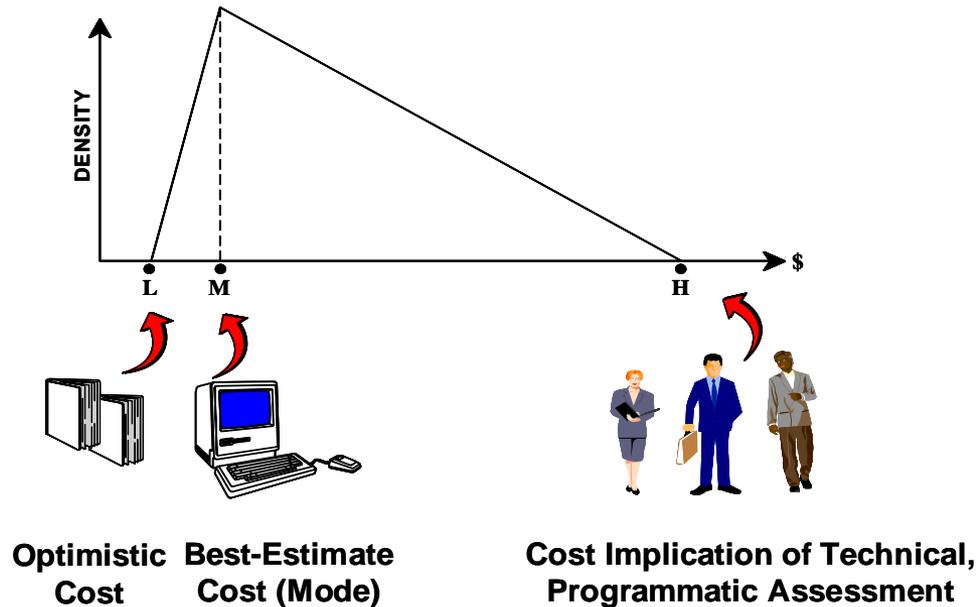
- For a database of satellite weights, from inception to completion of development, dry weight grew on average 28%*



* Source: Tim Anderson, "Satellite Remaining Weight Growth"

Using Judgment

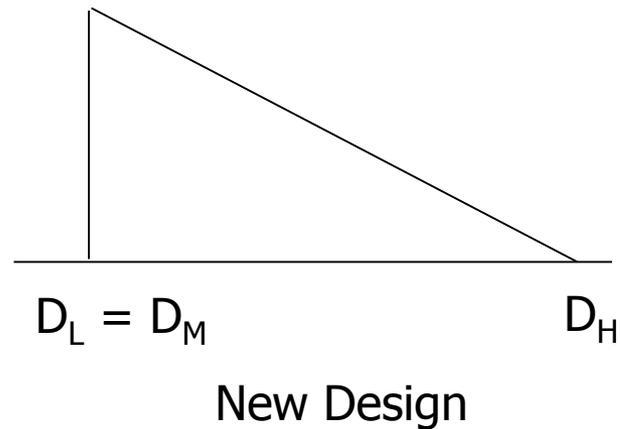
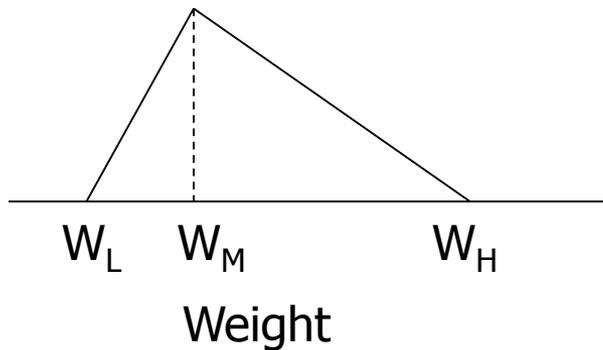
- Involves assigning low, most likely, and high values



- Care must be taken to avoid over-optimism
 - Actual example
 - A system's solid rocket motor was described as being "just like" an existing design, but "twice as large"
 - Taken at face value, this would indicate little new design, but the final cost exhibited significant levels of new design

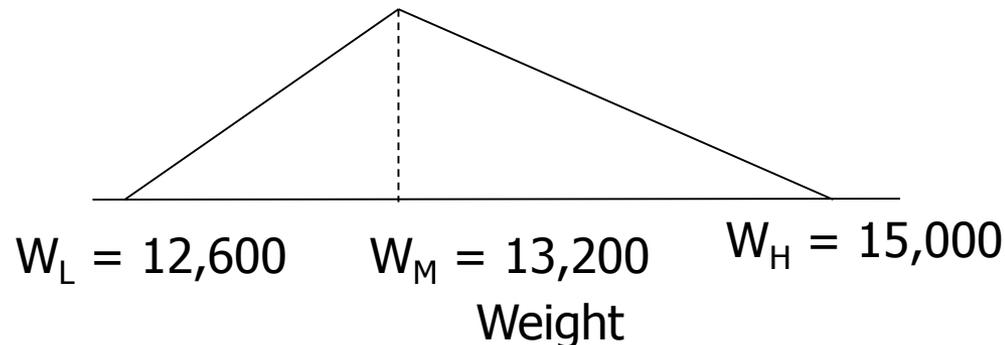
Using Judgment

- Solicit input from technical personnel
- Be careful – risk solicited from experts tends to have a tight range



Estimating Uncertainty for Example

- The data from Anderson is for satellites – missiles have constraints on weight growth
- Solution – use weight growth data for launch stages
- Historically, launch stages have exhibited weight growth in the range of 5-25%, with a median around 10%
- One solution
 - Low = $1.05 \times \text{Planned Weight} = 1.05 \times 12,000 = 12,600$ lbs.
 - Most Likely = $1.10 \times \text{Planned Weight} = 1.10 \times 12,000 = 13,200$ lbs.
 - High = $1.25 \times \text{Planned Weight} = 1.25 \times 12,000 = 15,000$ lbs.

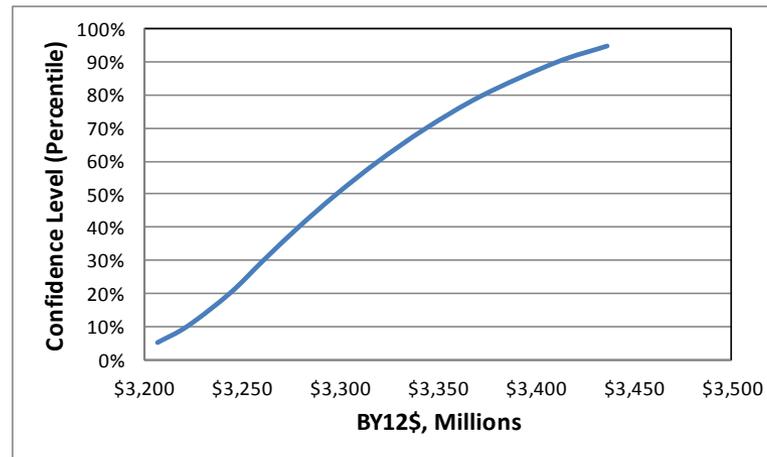


Estimating Uncertainty for Example (2)

- Note that the planned weight is not on the distribution
- This means that the point estimate is likely not achievable
- Refinements that will improve the simple triangle used:
 - Set the low and high weights to be percentiles of the triangle, not absolute bounds
 - Assign a lognormal instead of a triangle – better representation of risk seen in practice, since in reality we do not typically see absolute limits placed on uncertainty, other than that we know neither weight nor cost can be less than zero

Estimating Uncertainty S-Curve

- ❑ Monte Carlo simulation using weight uncertainty provides a range of \$3,200 to \$3,450 (5th-95th percentiles)
- ❑ Mean = \$3,300 and standard deviation is only \$70, which amount to a coefficient of variation equal to only 2%
- ❑ Not realistic – several studies (Braxton 2011, Smart 2011) show that cost growth data indicates coefficients of variation should be much higher, anywhere from 30-70%
- ❑ This is where many analysts stop, but there is much more to a full, credible cost risk analysis



Cost Risk Realism

- ❑ Monte Carlo simulation using weight uncertainty provides a range of \$3,200 to \$3,450 (5th-95th percentiles)
- ❑ Mean = \$3,300 and standard deviation is only \$70, which amount to a coefficient of variation equal to only 2%
- ❑ Not realistic – several studies (Braxton 2011, Smart 2011) show that cost growth data indicates coefficients of variation should be much higher, anywhere from 30-70%
- ❑ This is where many analysts stop, but there is much more to a full, credible cost risk analysis

Model Uncertainty

- CERs do not perfectly fit historical data upon which they are based
 - There are non-repeatable random effects that cannot be predicted
 - For example
 - Parts breaking during testing
 - Strikes, which are difficult to predict
- 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, depending upon the CER methodology

Modeling Model Uncertainty

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

For a linear CER:

$$Y = a + bX + \varepsilon$$

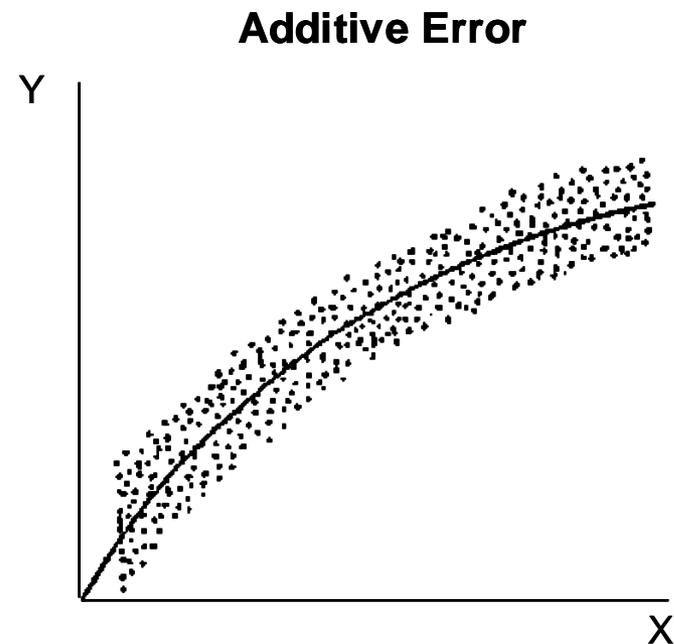
For a nonlinear CER:

$$Y = aX^b \varepsilon$$

where ε represents the error between the estimated cost and the actual cost Y

Residual Distributions

- Ordinary least squares CERs are based on the assumption of additive error
- Log-transformed CERs are based on the assumption of multiplicative error



Source: Eskew and Lawler (1994)

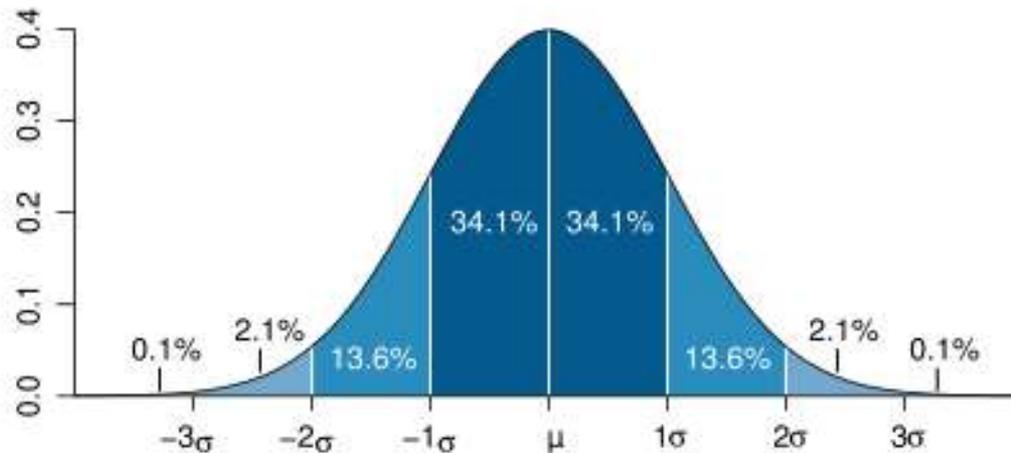
Measuring Model Uncertainty

- Given a value of the technical parameter x , say x_i , the actual cost corresponding to it is y_i
- What would our estimate be of the cost associated with x ?
- The estimate would have to be $\hat{Y}_i = A + Bx_i$, because A and B are proxies for the true coefficients a and b
- We can measure the quality of a CER by calculating the standard error of the estimate (SEE) according to the squared-error formula

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{Y}_i)^2$$

Normal Distribution and Standard Deviation

- 68.2% of the area under a Normal distribution is contained within *one* standard deviation about the mean
- 95% of the area under a Normal distribution is contained within *two* standard deviations about the mean
- 99.7% of the area under a Normal distribution is contained within *three* standard deviations about the mean



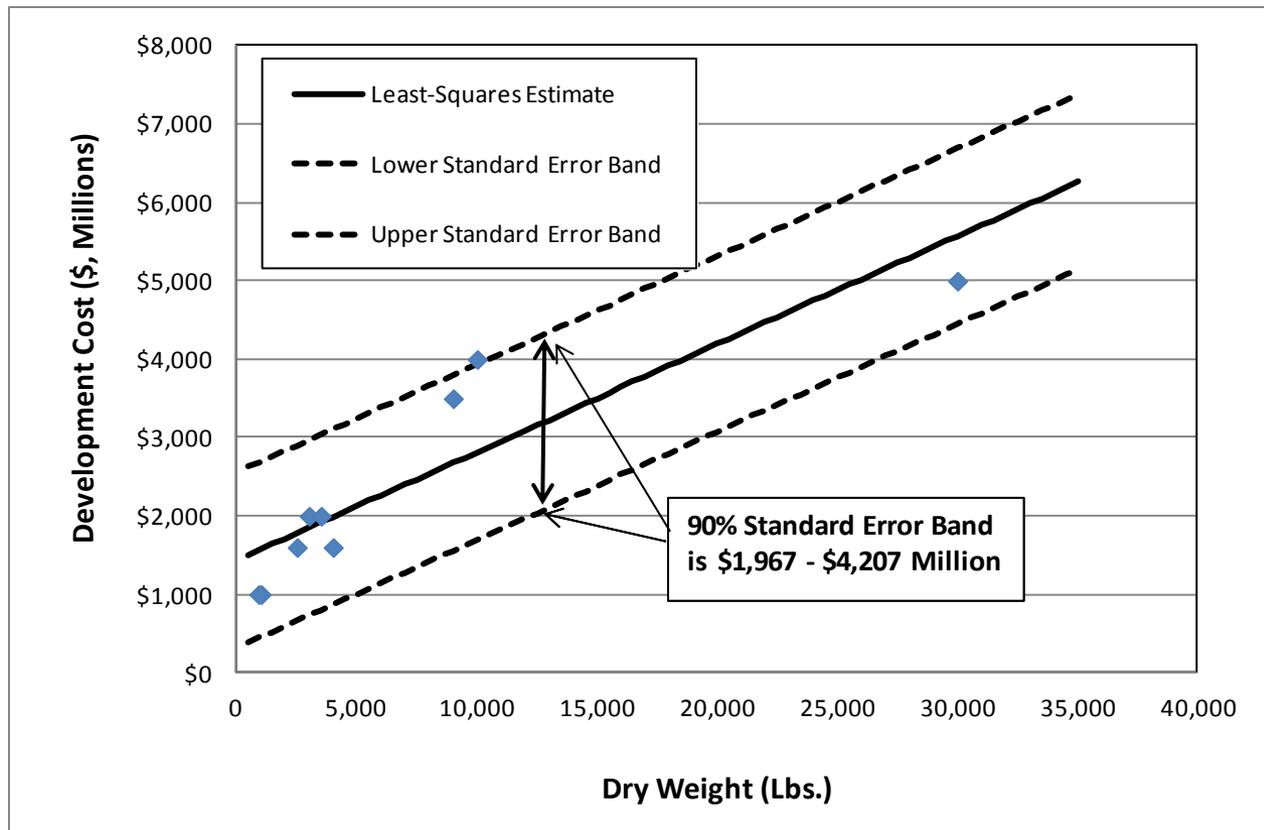
Standard Error Bands

- A two standard deviation error band can be calculated around the estimate using the standard deviation
- For the example linear CER

Development \$ Millions (BY12)	OLS Estimate	Estimate - Actual	(Estimate - Actual)^2
\$1,000	\$1,570	\$570	\$325,128
\$2,000	\$1,846	-\$154	\$23,716
\$1,600	\$1,777	\$177	\$31,347
\$1,000	\$1,556	\$556	\$309,592
\$2,000	\$1,915	-\$85	\$7,234
\$3,500	\$2,673	-\$827	\$683,268
\$5,000	\$5,569	\$569	\$324,102
\$4,000	\$2,811	-\$1,189	\$1,413,008
\$1,600	\$1,984	\$384	\$147,379
		Variance =	\$466,396
		SEE =	\$683

Standard Error Bands - Graph

- For the example CER, standard errors are often graphed around the estimate using multiples of the SEE



Standard Errors Vs. Confidence and Prediction Intervals

- Many of the estimates use the standard error of the estimate (SEE) as a proxy for the uncertainty around the estimate
 - The SEE is the standard error of the points around the regression line
 - It is not the standard error of an estimate made using the regression
- To find the error around an estimate made using a regression, we must use prediction intervals
- For Ordinary-Least-Squares-based CER estimates, the uncertainty distribution around the point estimate can be determined using little more than the analysis of variance (ANOVA) statistics
 - These ANOVA statistics should already exist as part of the regression analysis performed to develop the CER
- This section will provide an easy-to-follow guide for producing these uncertainty distributions for various types of CERs including:
 - Bivariate ordinary least squares (OLS)
 - Linear and Linear Transformed
 - Multivariate OLS

Calculating the Variance

$$\text{Var}(\hat{Y}) = \text{Var}(A + Bx) = \text{Var}((\bar{Y} - B\bar{x}) + Bx) = \text{Var}(\bar{Y} + B(x - \bar{x}))$$

$$= \text{Var}(\bar{Y}) + \text{Var}(B(x - \bar{x}))$$

$$= \text{Var}(\bar{Y}) + (x - \bar{x})^2 \text{Var}(B)$$

$$= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) + (x - \bar{x})^2 \text{Var}(B)$$

$$= \frac{1}{n} \sigma^2 + (x - \bar{x})^2 \left(\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) = \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

Bounding the Mean Cost at Cost-Driver Value x

- A confidence interval based on the variance of the estimated cost \hat{Y} bounds the mean cost of all elements (to which the CER applies) that have the value x for the technical parameter that drives the CER-based cost
- The degree of confidence associated with this interval is $(1-\alpha)100\%$, enforced by the choice of the appropriate percentage point of the t distribution, namely $t_{\alpha/2, n-2}$

$$\hat{Y} \pm t_{\alpha/2, n-2} SEE \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

Calculating the Variance of the Difference

- What we want is not a bound on the mean cost – we want to estimate the actual cost
- The width of a prediction interval for an estimate of the element's actual cost, corresponding to the technical parameter value x , is proportional to the variance of the difference between the estimate and the actual value:

$$\begin{aligned} \text{Var}(\hat{Y} - Y) &= \text{Var}(\hat{Y}) + \text{Var}(Y) = \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) + \sigma^2 \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \end{aligned}$$

Prediction Interval Equation

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

- \hat{Y} = Calculated Value from Regression Line (ANOVA)
- $t_{\alpha/2, df}$ = t Critical Value (tinv($\alpha/2, df$) function in Excel)
- SEE = Standard Error of the Estimate (ANOVA)
- n = number of observations (ANOVA)
- \bar{X} = average of X (calculated)
- $\sum X^2$ = Sum of X^2 (calculated)

Bounds on Estimates in the OLS Model

- First step in derivation is to calculate the variance (i.e., the square of the standard deviation) of the estimated cost \hat{Y} corresponding to the technical parameter value x
- It is important to note that the estimated cost \hat{Y} is an estimate of the mean cost corresponding to a technical parameter value of x
- We will see that $\text{Var}(\hat{Y})$, a measure of probable error of the estimate, depends explicitly on the technical parameter value x , in particular on how far away it is located from the central X location, namely
- It follows that using a CER to estimate far out of the range of the historical data base is risky

Standard Error Bands and Confidence and Prediction Intervals

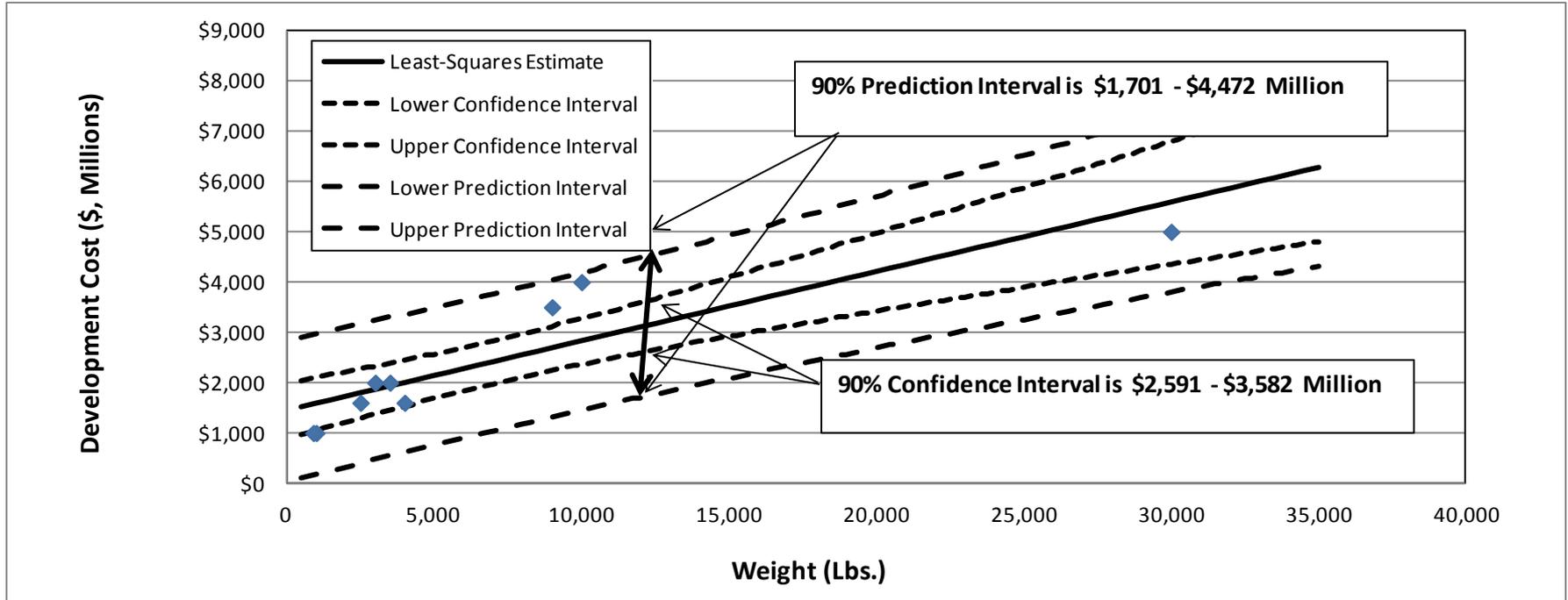
- **A Confidence Interval (CI) is a range within which there is a set probability that the population mean of a parameter is captured**
 - When viewed in terms of an estimate, it is the range around the true *mean*
- **Confidence levels are defined by significance levels (α), which are always between 0 and 1**
 - This significance level is the probability the population mean for a parameter lies outside of the interval
 - It is the probability of committing a “Type I” (or false positive) error
- **By varying α the user can vary the confidence level of the interval**
 - Example: An α of 0.10 represents a confidence level of 90%
 - One is 90% certain that the true value of the mean lies within the interval
- **The prediction interval (PI) differs from the CI in that it is a measure of the uncertainty around the estimate developed using a CER rather than just the average of the estimate**
- **The width of the PI will always be greater than the width of the CI since the PI includes both the error in the regression coefficients *and* the error in the prediction**

Methods for Prediction Intervals

- The method used to calculate the prediction depends on the shape of the model uncertainty distribution
- Three commonly used distributions to assess this uncertainty include:
 - Normal, for Ordinary Least Squares
 - Lognormal, for Log-transformed Ordinary Least Squares
 - Non-parametric, for methods such as Minimum Percent Error
- Note that each distribution is tied to a CER method, since the uncertainty assumption is intrinsically linked to each method

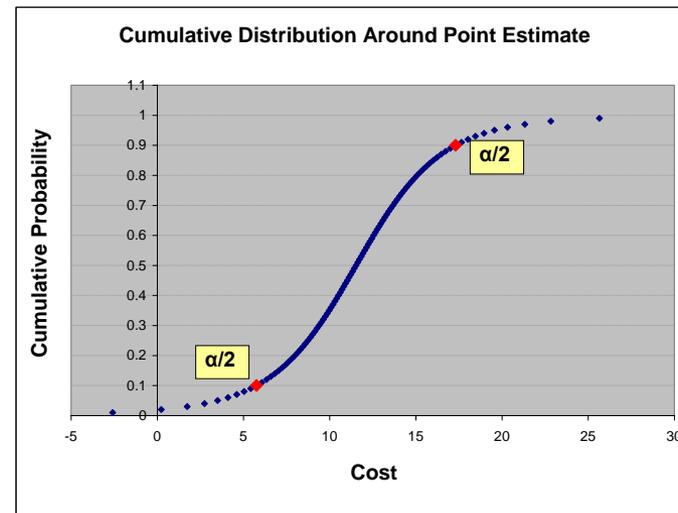
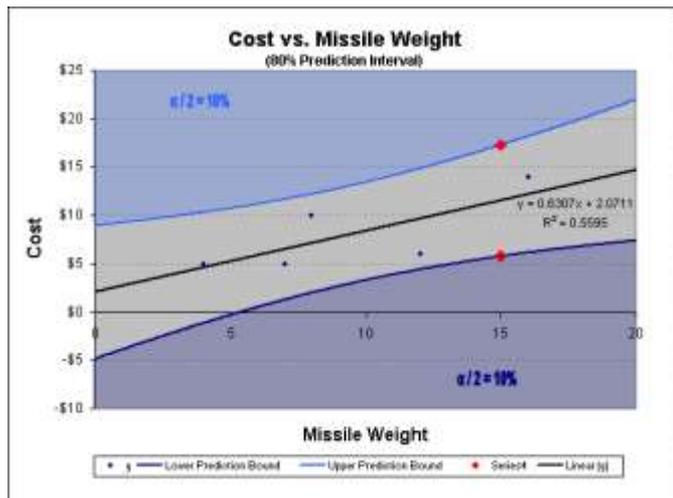
Prediction Interval for Linear CER

- The prediction interval is wider than the confidence interval



Basic Theory - OLS

- OLS methods assume that error around the regression line is distributed normally and therefore is symmetric
 - This implies that the Prediction Interval about the line is also symmetric
- Prediction intervals give a range for the estimate where the probability of the costs being outside the range is known
 - Because the error is symmetric, there is an equal chance of the final costs being outside/above the range as there is the final costs being outside/below the range
- By finding all possible prediction interval lines generated by varying α from 0 to 1 we can generate the true uncertainty around the estimate



Generating an S-Curve

Random Number	0.8
Prediction Interval	0.6
α	0.4

$$=(1 - 0.8) * 2$$



- For fixed input values, an S-curve corresponding to those inputs can be generated by mapping the points on the cumulative distribution of cost back to significance levels (and thus prediction bands)
- Holding the CER inputs constant the prediction interval equation

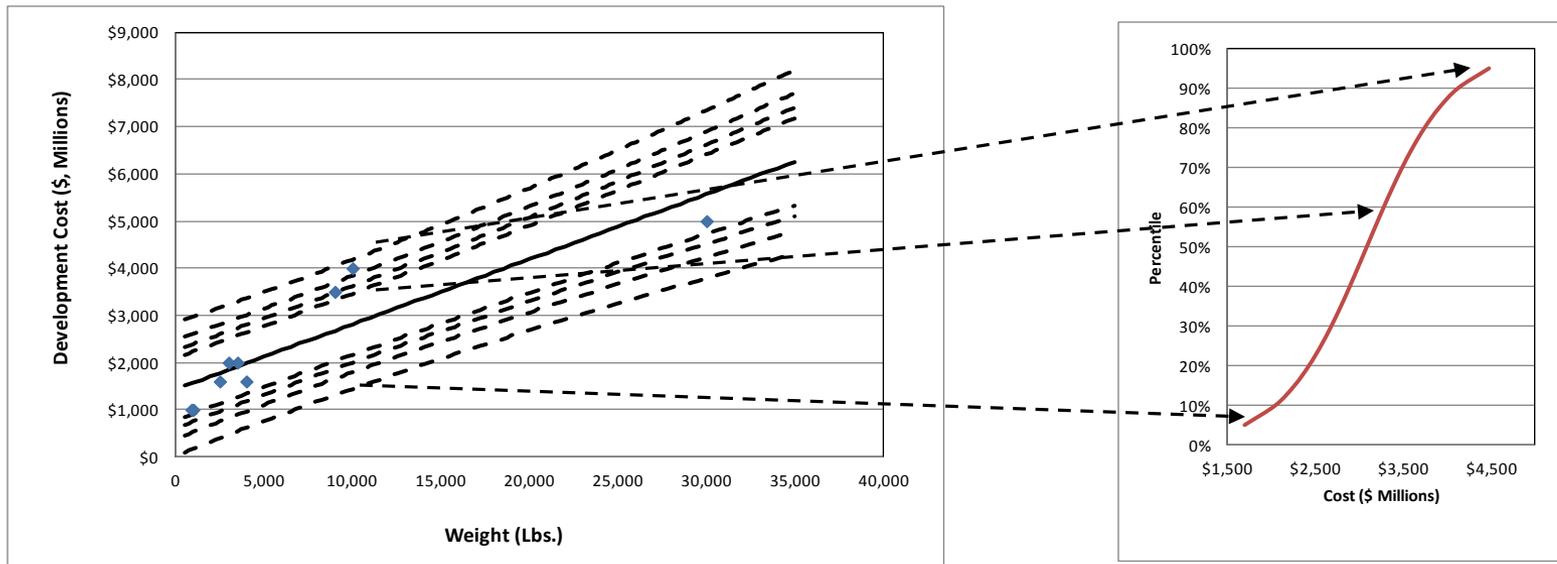
$$\hat{Y} \pm t_{\alpha/2, df} \times SEE \sqrt{\frac{n+1}{n} + \frac{(X - \bar{X})^2}{\sum X_i^2 - n\bar{X}^2}}$$

can be used to generate the percentiles

- For percentiles $< .5$ the corresponding α is the random number * 2
- For percentiles $\geq .5$ the corresponding α is the $(1 - \text{random number}) * 2$

Risk Distribution around the Estimate - Example

- For any prediction interval defined by an α
 - The upper prediction bound is at the $(1-\alpha/2)^{\text{th}}$ percentile on the cumulative distribution
 - The lower prediction bound is at the $(\alpha/2)^{\text{th}}$ percentile on the cumulative distribution

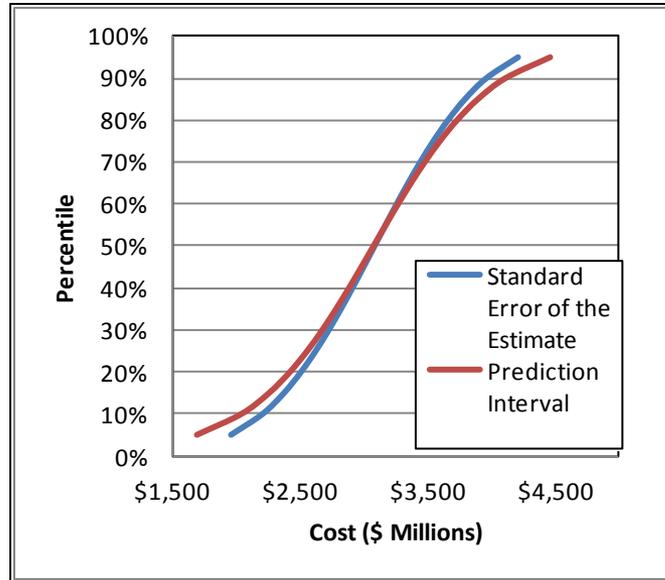


Generating the S-Curve from the Prediction Intervals

- The S-curve can be generated by varying the critical value of the t distribution for the equation, holding the CER input(s) constant:

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

Prediction Interval Distribution vs. SEE



Percentile	SEE	PI	Difference
5%	\$1,964	\$1,702	-13%
10%	\$2,212	\$2,052	-7%
15%	\$2,379	\$2,269	-5%
20%	\$2,512	\$2,432	-3%
25%	\$2,626	\$2,567	-2%
30%	\$2,729	\$2,686	-2%
35%	\$2,824	\$2,793	-1%
40%	\$2,914	\$2,895	-1%
45%	\$3,001	\$2,992	0%
50%	\$3,087	\$3,087	0%
55%	\$3,173	\$3,182	0%
60%	\$3,260	\$3,280	1%
65%	\$3,350	\$3,381	1%
70%	\$3,445	\$3,489	1%
75%	\$3,548	\$3,607	2%
80%	\$3,662	\$3,742	2%
85%	\$3,795	\$3,906	3%
90%	\$3,962	\$4,122	4%
95%	\$4,210	\$4,473	6%

- ❑ The prediction interval results in more uncertainty than using the standard error of the estimate
- ❑ Above the median, the higher the percentile, the more the SEE underestimates
 - Below the median, the SEE overestimates

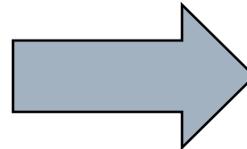
Non-linear Regression

- **The following example shows how to produce prediction interval distributions using OLS around a non-linear CER**
 - Similar to the linear CER example, we assume that missile weight is a driver of cost but now, the relationship is non-linear
 - Power Equation CER: $y = ax^b$
 - We evaluate development cost given a missile weight, using the same example data as used for the linear CER
- **Non-linear CERs, first, must be converted into a linear relationship before performing OLS regression.**
 - Commonly referred to as transforming to log or semi-log space
- **Once the data has been transformed, the remaining steps are no different than producing prediction interval distributions from a bivariate linear CER**

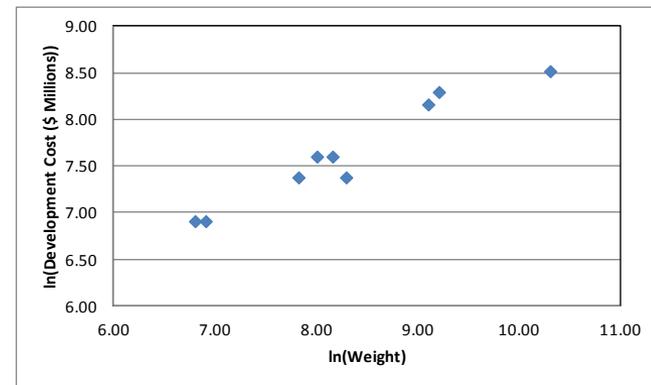
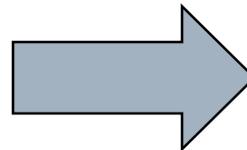
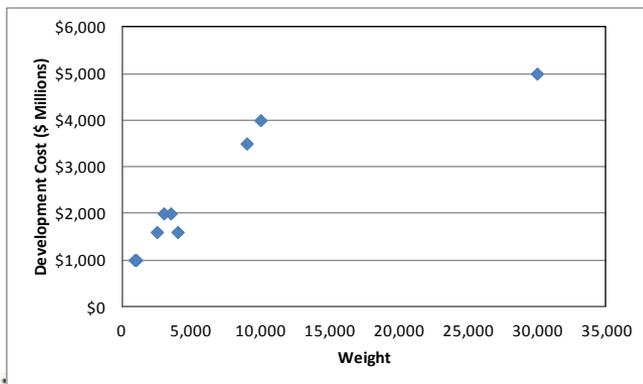
Non-linear Regression Linear Transformation

- Transform the CER into log space by taking the natural log (\ln) of both sides such that $\ln y = \ln a + b \ln x$
- Scatter plot reveals linear relationship in semi-log space

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



Ln(\$)	Ln(Weight)
6.91	6.91
7.60	8.01
7.38	7.82
6.91	6.80
7.60	8.16
8.16	9.10
8.52	10.31
8.29	9.21
7.38	8.29



Regression Residuals in Linear Space

- Standard error of the estimate is calculated in log space

Development \$ Millions (BY12)	LOLS Estimate	$\ln(\text{Estimate}) - \ln(\text{Actual})$	$(\ln(\text{Estimate}) - \ln(\text{Actual}))^2$
\$1,000	\$1,036	0.0355	0.0013
\$2,000	\$1,799	-0.1058	0.0112
\$1,600	\$1,642	0.0258	0.0007
\$1,000	\$983	-0.0174	0.0003
\$2,000	\$1,944	-0.0284	0.0008
\$3,500	\$3,124	-0.1136	0.0129
\$5,000	\$5,720	0.1345	0.0181
\$4,000	\$3,294	-0.1942	0.0377
\$1,600	\$2,079	0.2619	0.0686
		Variance =	0.0216
		SEE =	0.1471

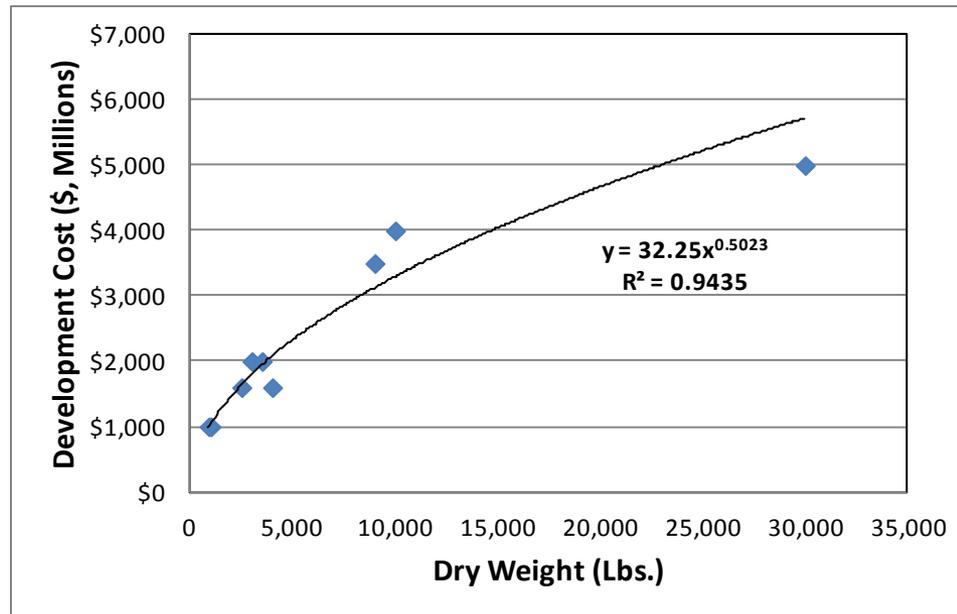
Log-Transformed Linear Least Squares

- For the example data set, the best ordinary least squares fit for cost and weight is given by the equation

$$\text{Cost} = 32.25 \text{ Weight}^{0.5023}$$

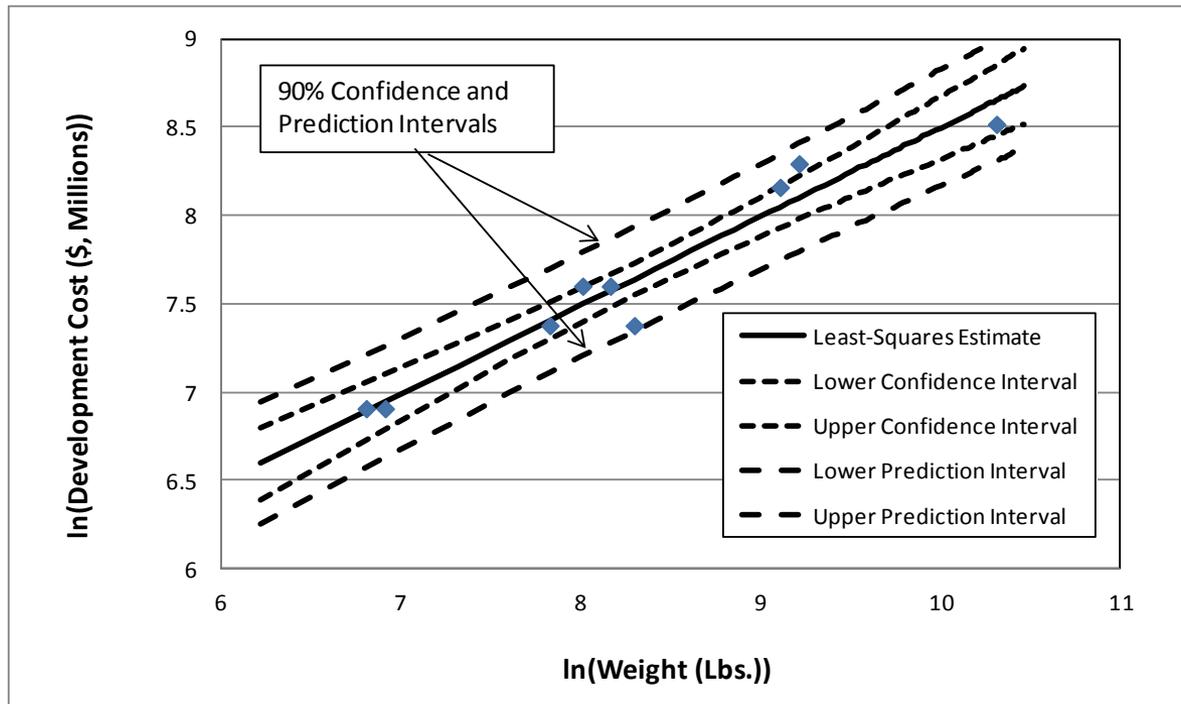
where Weight = mass, in pounds

Cost = \$ Millions, in Base Year FY12



Prediction Interval in Log Space

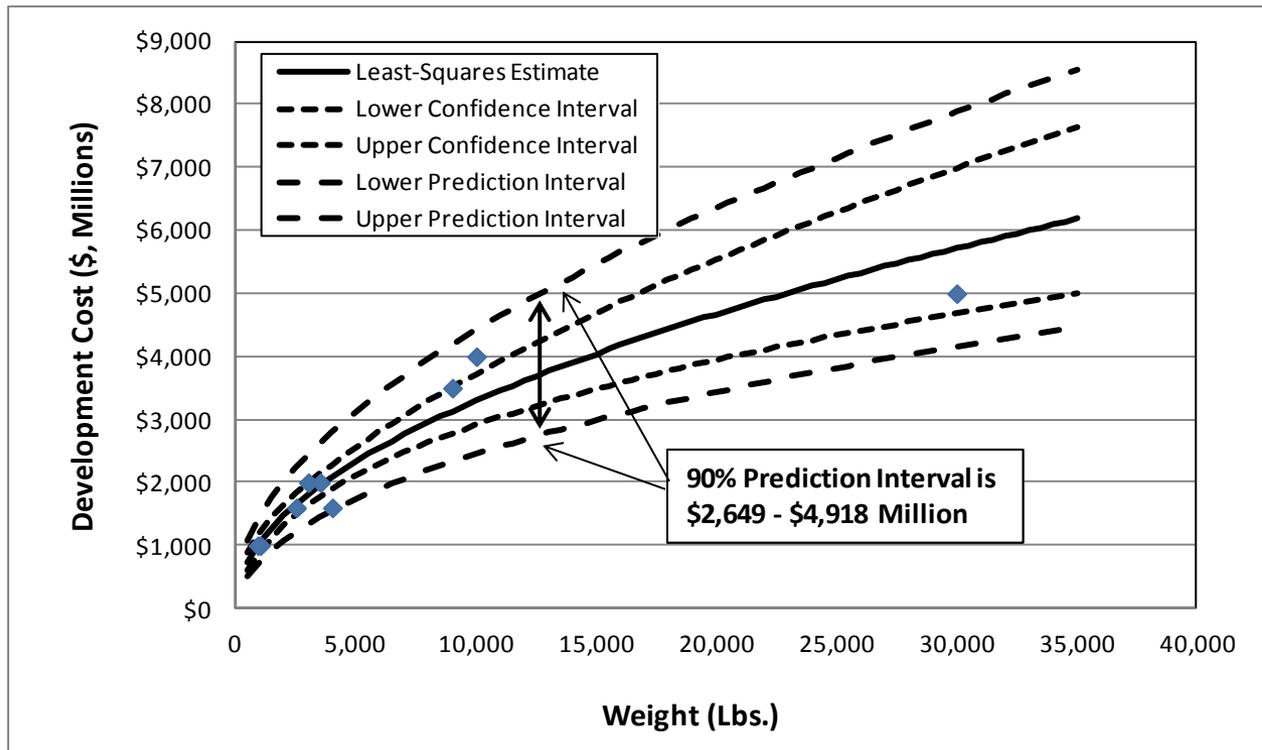
- Apply the same methodology and prediction interval equation to the data while still in log space



Prediction intervals and regression in log space resemble those in the linear example

Non-linear Regression

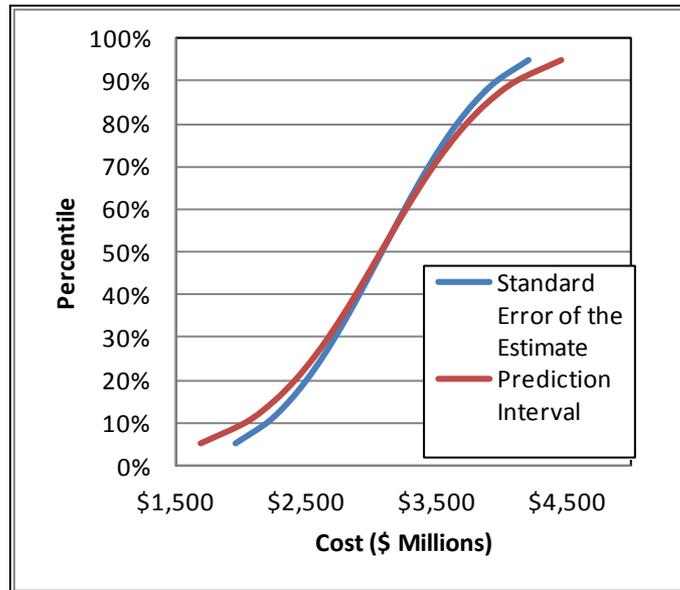
- **The final step is to transform the model back to unit space**
 - Since this is an power equationCER, take the exponent of all cost (y) values in slog space to get back to unit space
 - Notice that error increases along with the cost driver



Generating the S-Curve

Non-Linear

- The S-curve can be generated by varying the critical value of the t distribution for the equation, holding the CER input(s) constant, as with the linear equation, but an extra step is needed to transform the percentiles to unit space from log space



Percentile	SEE	PI	Difference
5%	\$2,424	\$2,290	-5%
10%	\$2,557	\$2,470	-3%
15%	\$2,651	\$2,588	-2%
20%	\$2,728	\$2,681	-2%
25%	\$2,795	\$2,760	-1%
30%	\$2,858	\$2,831	-1%
35%	\$2,917	\$2,898	-1%
40%	\$2,974	\$2,962	0%
45%	\$3,031	\$3,024	0%
50%	\$3,087	\$3,087	0%
55%	\$3,145	\$3,151	0%
60%	\$3,204	\$3,218	0%
65%	\$3,267	\$3,289	1%
70%	\$3,335	\$3,366	1%
75%	\$3,409	\$3,453	1%
80%	\$3,494	\$3,555	2%
85%	\$3,596	\$3,682	2%
90%	\$3,728	\$3,858	3%
95%	\$3,932	\$4,161	6%

Multivariate Linear Regression

- Although it uses matrices, creating prediction intervals using multivariate linear regression is no more difficult than doing so for bivariate linear regressions
- The equation for the $(1-\alpha)$ prediction interval around any estimate is:

$$\mathbf{Z}^T \hat{\boldsymbol{\beta}} \pm t_{\frac{\alpha}{2}, m-n} \sigma \sqrt{1 + \mathbf{Z}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{Z}}$$

Where:

- \mathbf{Z} is the matrix containing the values of the independent variable for this prediction (the final entry being 1, signifying the intercept)
- $\boldsymbol{\beta}$ is the matrix containing the best-fit coefficients (with the final entry being the intercept)
 - It follows directly that $\mathbf{Z}^T \boldsymbol{\beta}$ represents the estimate
- \mathbf{X} is the matrix containing the independent variable data points used to build the regression

MPE-ZPB

- The Minimum Percent Error with Zero Percent Bias (MPE-ZPB) method (Book 2006) pursues the Minimum-Percentage-Error goal directly
 - Computes minimum-percentage-error cer, subject to constraint that percentage bias be exactly zero
 - CER derived using “constrained optimization” – another capability of *Excel Solver*

$$\text{Minimize } F(a, b, c) = \sum_{k=1}^n \left(\frac{y_k - a - bx_k^c}{a + bx_k^c} \right)^2, \text{ subject to}$$

the constraint

$$\%Bias(a, b, c) = \sum_{k=1}^n \left(\frac{a + bx_k^c - y_k}{a + bx_k^c} \right) = 0$$

Non-Parametric Methods

- MPE-ZPB is a non-parametric method, which means there is no underlying assumption about the shape of the uncertainty distribution for the dependent variable
- In 2006 Steve Book proposed using the bootstrap technique as an ad-hoc approach for developing prediction intervals in such cases
- “Bootstrap” statistical sampling appears to be an appropriate technique to consider
 - The bootstrap method of error estimation was introduced by B. Efron in 1977 and has a 28-year history behind it
 - It is a “distribution-free” method, so it does not require the usual (and questionable) distributional assumptions, e.g., normal or lognormal error distributions or even homoscedasticity
 - It works with additive- or multiplicative-error models and all algebraic functional forms
- For more information see Steve Book’s presentation on this subject (Book, 2006)

Other Issues

- **Results may yield negative costs if the prediction interval is wide**
 - This happens commonly when the distribution around a rate or factor is being used
 - A solution to this is to use cost on cost (rather than a fixed percentage) CERs
 - If the prediction interval of cost on cost data is wide enough that there are significantly common instances negative costs, the usefulness of the CER is questionable
 - If this happens uncommonly, then it is harmless

Prediction Interval Conclusions (1)

- One of the benefits of this methodology is that it takes into account several of the common issues estimators have with CERs
 - CERs with high CVs
 - The larger the CV of the regression, the larger the CV of the prediction interval cumulative distribution
 - Estimating outside the range of data
 - Because the prediction interval for an estimate widens as the cost driver moves away from the center of mass of the regression, the prediction interval cumulative distribution becomes wider as estimates are made outside the range of the data
 - Low number of data points
 - The fewer data points, the wider the t distribution will be, and thus the prediction intervals
- This method can also be combined with other risk analysis methods
- Generating uncertainty distributions from CERs is one simple way of accounting for risk in cost estimates

Prediction Interval

Conclusions (2)

- This is remedy for the oft-repeated injunction to “never use a CER outside the range of the data”
 - This may be a perfectly reasonable proscription outside cost and risk analysis, but in cost and cost risk analysis, the analyst must routinely operate outside the range of the data
 - It is the nature of the development that the object being developed is routinely bigger, faster, stealthier (or commonly, smaller) than heretofore, and to forswear CERs outside of their data range is to abandon them almost everywhere
- The prediction interval, of course, affords no immunity against incorrect CERs or against factors that may apply in realms outside the data that is unknown to the analyst
- The prediction interval, however, gives the analyst the ability to use a CER wherever it is needed and to correctly characterize the resultant uncertainty so long as the analyst is aware of the other possibilities just mentioned

Prediction Intervals in Excel

- Use the Data Analysis Tool, choose cost data as the “Y-input range” and weight as the “X-input range”
- Data analysis tool provides the number of observations (“n”), degrees of freedom (“residual”), intercept and coefficient of the regression equation, and the standard error of the estimate
- Calculate in Excel the sum of squares of the x-values, the average, and the average squared
- Calculate the inverse of the T distribution using the Excel function “TINV”
 - Take into account the fact that you need two tails, not one

Prediction Intervals in Excel (2)

- Output of Excel data analysis tool for the OLS example

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0.89250958					
R Square	0.79657335					
Adjusted R Square	0.7675124					
Standard Error	682.9318838					
Observations	9					
<i>ANOVA</i>						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	12784117.18	12784117.18	27.41043735	0.001205175	
Residual	7	3264771.705	466395.9579			
Total	8	16048888.89				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	1432.277594	294.5779367	4.862134655	0.001830671	735.7114616	2128.843727
X Variable 1	0.137863876	0.026332525	5.235497813	0.001205175	0.075597349	0.200130402

Assess External Factor Uncertainty

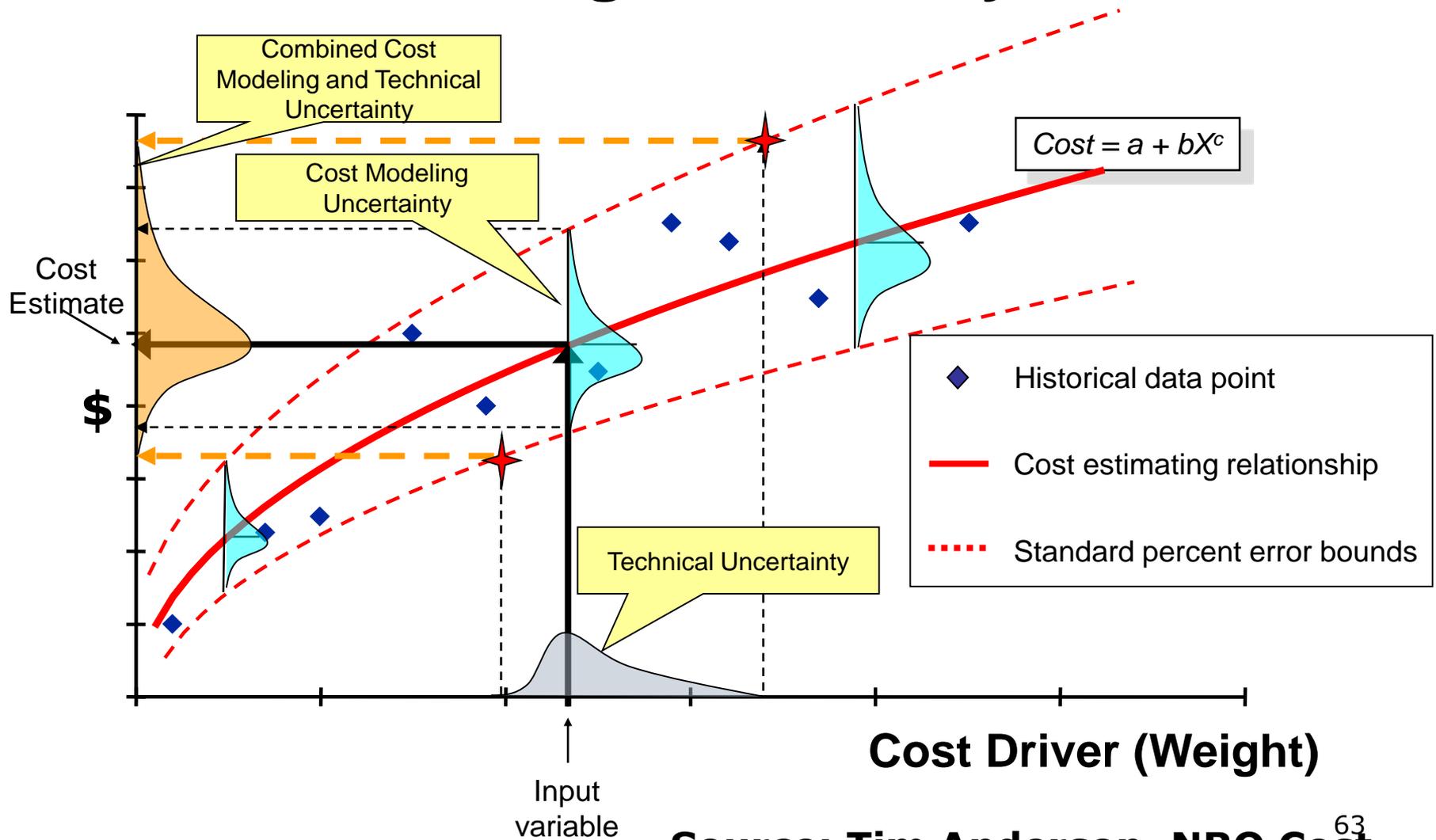
- External factors include labor strikes, acts of God (natural disasters), acts of Congress (national disasters), and other phenomenon not typically explicitly modeled, such as major test failures
- When modeled, typically done as a likelihood, and a consequence
- Likelihood can be modeled via a binomial distribution
 - Happens or doesn't happen
- Consequence can be modeled as a point estimate, or a range of values
- ***Recommendation: to model this, and convince your project manager that he should budget for at least some of these potential mishaps***

Example of External Uncertainty

- The nine data points do not include any test failures; however for the project being estimated there is one highly risky development test
 - If the test fails, the cost to investigate and re-test is \$200 million
 - There is a 30% probability of the test failing
- This can be modeled with a custom distribution:

$$F(x) = \begin{cases} \$0 & \text{for } x < 0.70 \\ \$200 \text{ million} & \text{for } x \geq 0.70 \end{cases}$$

Combining Model and Estimating Uncertainty



Source: Tim Anderson, NRO Cost Group Risk Process

Monte Carlo Simulation

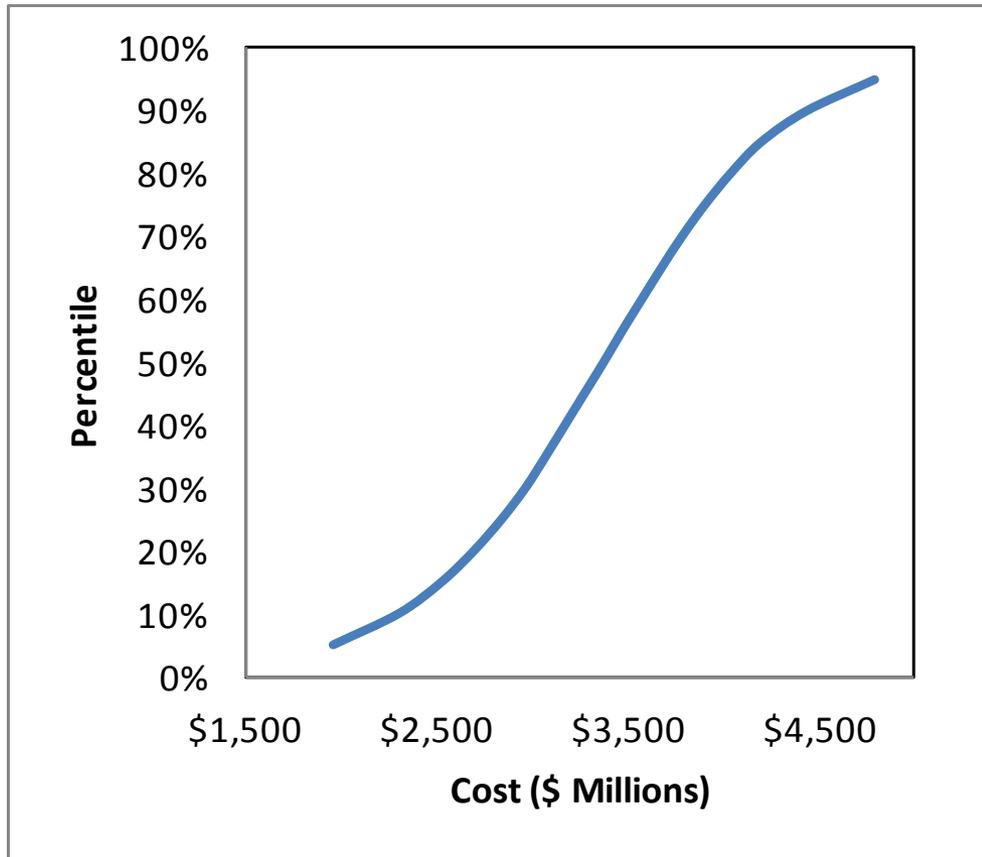
- For the linear CER example, the three sources of risk – estimating, model, and external factors – can be combined via the equation

$$\text{Cost} = \$1,432 \text{ Million} + 0.1379^* \text{ Weight} + \varepsilon + \text{Additional test cost}$$

using Monte Carlo simulation

- Weight is modeled according to a triangular distribution
 - Model error ε is modeled according to the t-distribution (prediction interval)
 - The additional test cost is modeled using a custom distribution
- Note that the value of the error depends upon the weight, since the input affects the prediction interval
- Must take this into account in the simulation

Combined S-Curve for the Linear CER



Mean	\$3,367
Sigma	\$883
Percentile	
5%	\$1,955
10%	\$2,300
15%	\$2,523
20%	\$2,693
25%	\$2,837
30%	\$2,962
35%	\$3,068
40%	\$3,169
45%	\$3,270
50%	\$3,372
55%	\$3,469
60%	\$3,570
65%	\$3,673
70%	\$3,781
75%	\$3,899
80%	\$4,034
85%	\$4,193
90%	\$4,433
95%	\$4,790

References

- ❑ Anderson, T., “NRO Cost Risk Process,” Presentation
- ❑ Anderson, T. “Satellite Remaining Weight Growth,” Presentation
- ❑ Book, S.A., “Prediction Intervals for CER-Based Estimates,” presented at the International Society of Parametric Analysts 26th International Conference, Frascati, Italy, 2004
- ❑ Coleman, R. et al., “What Percentile Are We At Now (And Where Are We Going?),” SCEA 2009, DoDCAS 2010
- ❑ Eskew, H.L. and K.S. Lawler, “Correct and Incorrect Error Specifications in Statistical Cost Models,” *Journal of Cost Analysis*, Spring 1994, page 107
- ❑ Hunt, C., “Cost Risk for CLV,” Presentation
- ❑ Society of Cost Estimating and Analysis, *Cost Estimating Body of Knowledge*, SCEA, 2009