NORTHROP GRUMMAN

DEFINING THE FUTURE

# Taking the Next Step:
## Turning OLS CER-Based Estimates into Risk Distributions

SCEA National Conference
June 2008

Chrissy M. Kanick
Eric R. Druker
Richard L. Coleman
Matthew M. Cain
Peter J. Braxton

Northrop Grumman Corporation

Presented at the 2008 SCEA-ISPA Joint Annual Conference and Training Workshop - www.iceaaonline.com


# Topics

- Introduction
  - Cost Estimating Relationships (CERs)
  - Confidence and Prediction Intervals
  - S-Curves
- Methodology
  - Bivariate Linear Regression
  - Bivariate Non-Linear Regression
  - Multivariate Linear Regression
- Simulating Prediction Distributions
- Other Issues
- Conclusion


E.R. Druker, eric.druker@ngc.com, 703 408-0589

# Introduction

- With customers now routinely desiring a range of potential costs, rather than a point estimate, there has been an increased focus on risk and uncertainty in cost estimates
- In particular, more and more customers are making budgeting decisions based on probabilistic cost estimates, also known as S-Curves
- One of the hurdles in incorporating risk and uncertainty into estimates is that risk analysis requires a slightly different skill set than cost estimating
  - Rather than statistics, there is a greater focus on probability
    - "Risk analysis puts the prob in prob/stat"
  - Exposure to modeling and simulation (M&S) – a computer science discipline – is also desirable
- Although full-scale risk analysis may not be necessary or feasible for every organization, there are some relatively simple steps that can be used to produce the uncertainty around traditional cost estimating methods


E.R. Druker, eric.druker@ngc.com, 703 408-0589

# Introduction (cont.)

- For Ordinary-Least-Squares-based CER estimates, the uncertainty distribution around the point estimate can be determined using little more than the ANOVA statistics
  - These ANOVA statistics should already exist as part of the regression analysis performed to develop the CER
- This paper 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
- It will then briefly touch on the modeling techniques needed to implement them into a Monte Carlo simulation


E.R. Druker, eric.druker@ngc.com, 703 408-0589

# Cost Estimating Relationships (CERs)

- A mathematical relationship that defines cost as a function of one or more parameters such as performance, operating characteristics, physical characteristics, etc.[1]
- A CER suggests that there is one or multiple independent variables (cost drivers) that can be utilized to generate a best estimate for the cost of a program or system
- Types of CERs:
  - Traditional
    - Costs estimated as functions of cost driver(s)
  - Rates/Factors/Ratios
    - When these are derived as CERs, the equation has been forced through the origin (zero y-intercept)
- Great care needs to be taken to:
  - Verify that the data is the most recent available data
  - Verify that the data is consistent and robust
  - Verify that the data has been appropriately normalized

[1] NASA Cost Estimating Handbook 2004



E.R. Druker, eric.druker@ngc.com, 703 408-0589

Presented at the 2008 SCEA-ISPA Joint Annual Conference and Training Workshop - www.iceaaonline.com

# Confidence and Prediction Intervals

NORTHROP GRUMMAN

- A Confidence Interval suggests that the true mean value for a parameter is contained within a calculated range
  - It is a measure of the uncertainty in the regression line
- The confidence level is determined by choosing an alpha (α) between 0 and 1. By varying the α, the user can vary the level of confidence that is required for the interval
  - Example: An α of 0.10 ascertains a confidence level of 90%
  - One is 90% certain that the true value of the mean lies within the interval
- The prediction interval differs from the confidence interval in that it is a measure of the uncertainty around the estimate developed using a CER rather than the mean
  - The prediction interval represents the bands around the final cost a system at a fitted X rather than bands around the mean of the distribution of the final cost of a system
- Note that the width of a prediction interval will always be greater than the width of a confidence interval since the prediction interval includes both the error in the regression coefficients and the error in the prediction

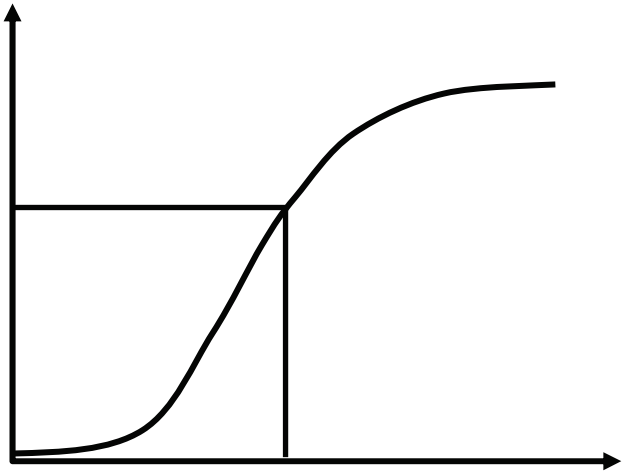


E.R. Druker, eric.druker@ngc.com, 703 408-0589

# A Very-Brief Introduction to S-Curves

NORTHROP GRUMMAN

- S-Curves are the cumulative distribution function for the cost of a system
  - Also known as probabilistic cost estimates
- S-Curves are generally driven by two main factors
  - Cost Estimating Variance
    - Labor estimates
      - Data-Driven
      - SME-Driven
    - Escalation/Inflation Rates
    - Material Costs
    - Productivity (e.g., hrs/SLOC, hrs/$ft^2$)
  - Schedule/Technical Risks and Opportunities
    - Discrete Events
    - Continuous Events



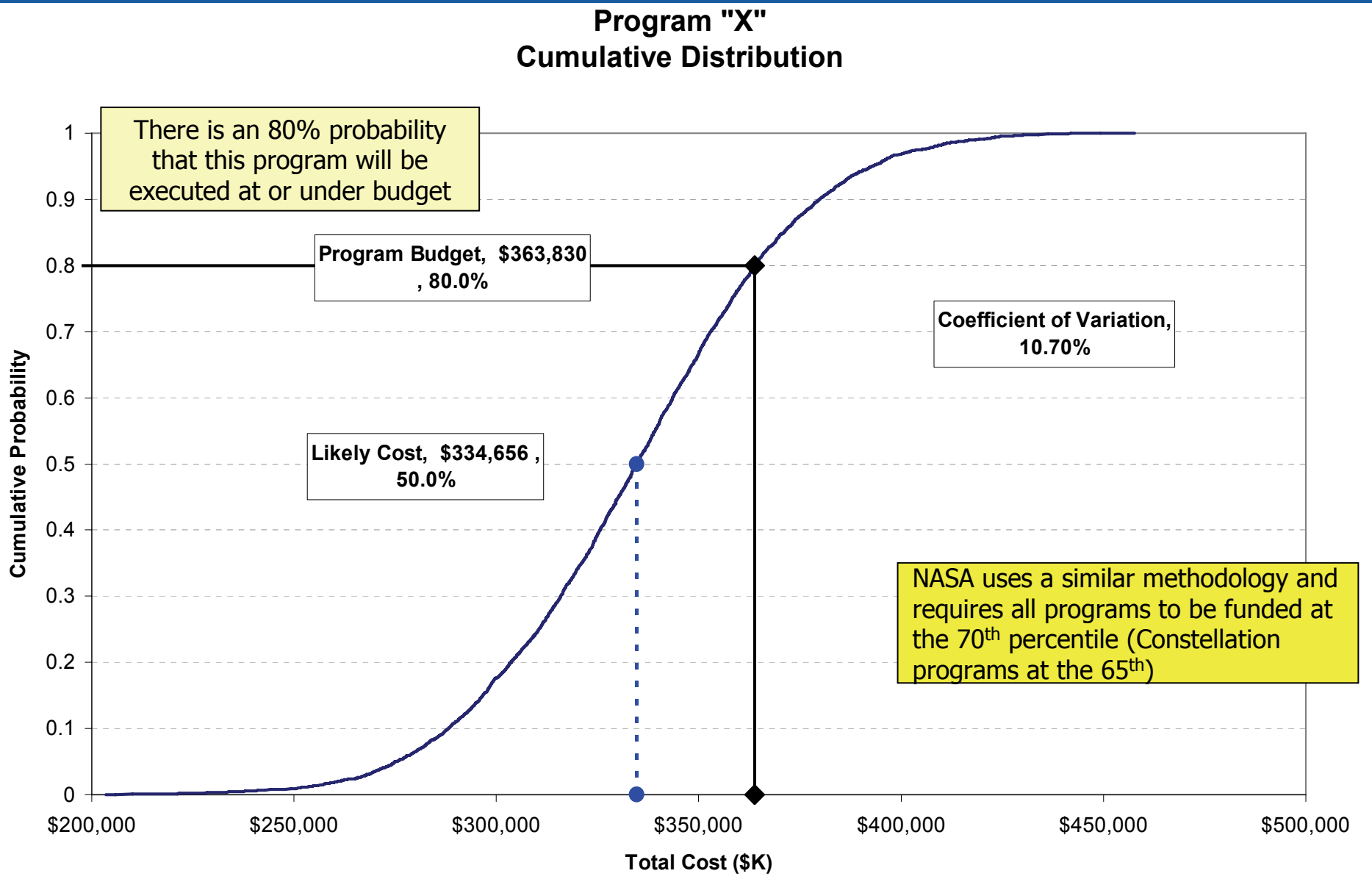
- Two key measures are derived from these S-Curves
  - Confidence level of the estimate
    - What is the probability that the program will finish at or under budget?
  - Uncertainty in the estimate
    - What is the range of possibilities for the final cost of this program?


E.R. Druker, eric.druker@ngc.com, 703 408-0589

Presented at the 2008 SCEA-ISPA Joint Annual Conference and Training Workshop - www.iceaaonline.com

# Sample Program S-Curve

NORTHROP GRUMMAN






E.R. Druker, eric.druker@ngc.com, 703 408-0589

# Why S-Curves?

- Studies have shown[1,2] that 75-85% of DoD programs experience cost overruns
  - This suggests that as an industry, our estimates are not at the 50th percentile, but rather at about the 20th percentile
- Recognizing this, agencies are taking the initiative to budget at higher percentiles of cost
  - This is true in particular for high-risk space and software development programs
- In order to determine the appropriate funding level for programs, it is imperative that the risk and uncertainty around estimates be assessed
  - Thus S-Curves must be developed

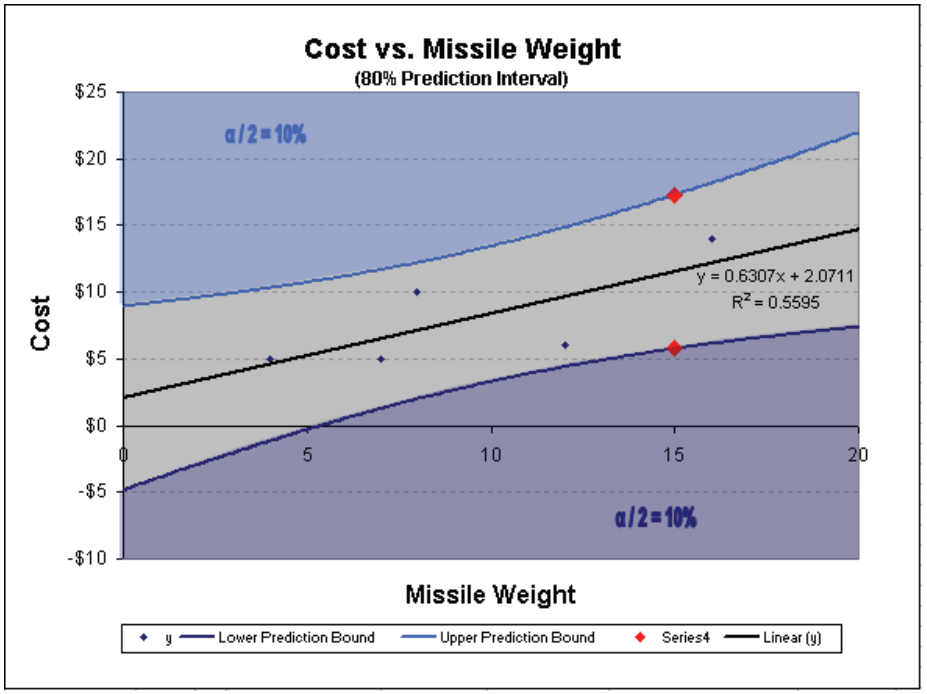
1 Schaffer 2004 study, referenced from *Cost Estimating Requirements to Support New Congressional Reporting Requirements*. Coonce et. Al. NASA PM Challenge, February 2008

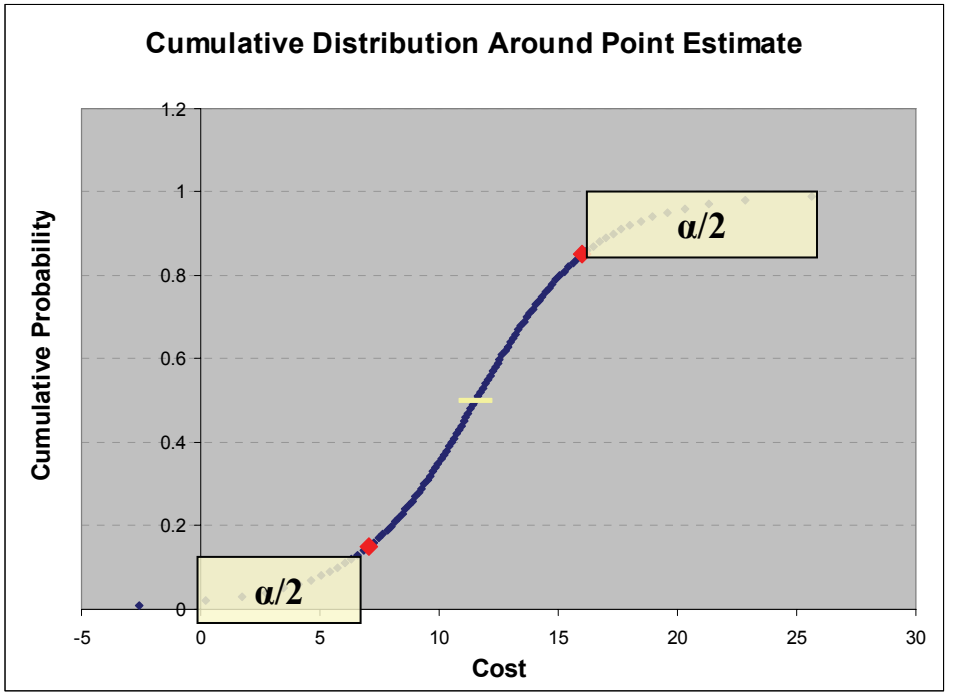
2 NAVAIR Cost Growth Study, R. L. Coleman, M.E. Dameron, C.L. Pullen, J.R. Summerville, D.M. Snead, 34th DoDCAS and ISPA/SCEA 2001


E.R. Druker, eric.druker@ngc.com, 703 408-0589

# Basic Theory - OLS

- OLS methods assume that error around the (true) regression line is distributed normally and therefore symmetrically
  - This implies that the Prediction Interval about the (fitted) line is a t (also symmetric)
- Prediction intervals give a range for the estimate for which there is a set probability of the final cost being outside
  - Because 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 using a confidence level ($0<\alpha<1$), we can generate the distribution implied by the prediction intervals





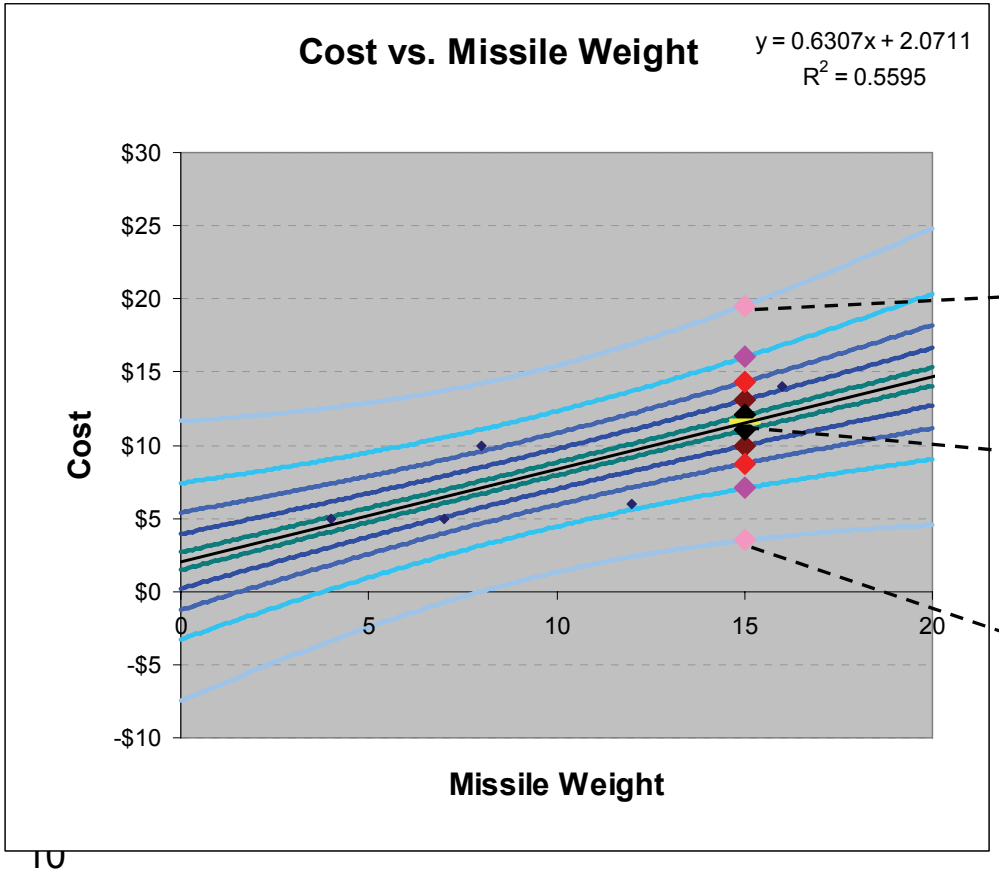

E.R. Druker, eric.druker@ngc.com, 703 408-0589

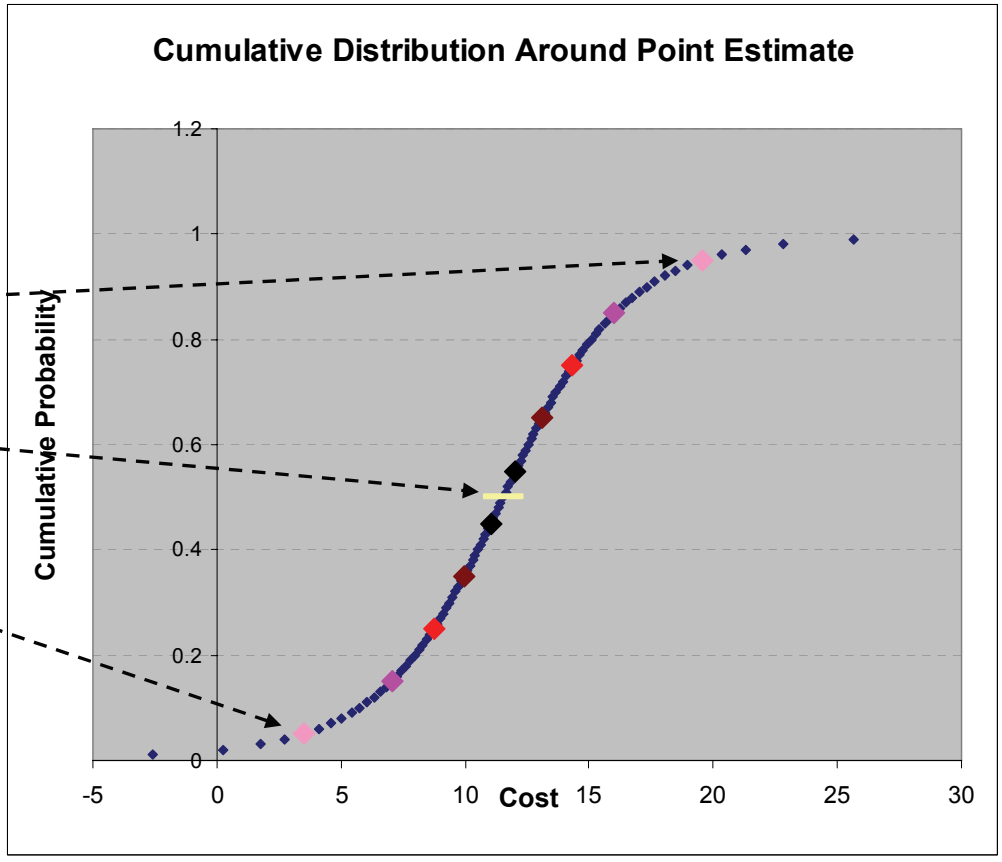
Presented at the 2008 SCEA-ISPA Joint Annual Conference and Training Workshop - www.iceaaonline.com

# Risk Distribution around the Estimate - Example

NORTHROP GRUMMAN

- The below graphs demonstrate further how prediction intervals translate into uncertainty distributions
- For any prediction interval defined by an α
  - The upper prediction bound is at the (1-α/2)-th percentile on the cumulative distribution
  - The lower prediction bound is at the (α/2)-th percentile on the cumulative distribution





E.R. Druker, eric.druker@ngc.com, 703 408-0589

# Prediction Interval Equation

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

**This is the data set from SCEA CostProf Module 8 - Regression Analysis.*



E.R. Druker, eric.druker@ngc.com, 703 408-0589

Presented at the 2008 SCEA-ISPA Joint Annual Conference and Training Workshop - www.iceaaonline.com

# Prediction Interval Uncertainty Distribution Needed Statistics

- This example shows how to produce prediction interval distributions around a bivariate OLS CER
- Highlighted boxes will be utilized to calculate the prediction interval
  - Step 1: Run an ANOVA on the data set
  - Step 2: Compute average of $X$ and well as the $\Sigma X^2$

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.7480121 |
| R Square | 0.559522 |
| Adjusted R Square | 0.4126961 |
| Standard Error | 3.0171528 |
| Observations | 5 |

| Not Computed as Part of Excel ANOVA | |
|---|---|
| Average of x | 9.4 |
| Sum of x^s | 529.0 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 34.69036697 | 34.69036697 | 3.810783573 | 0.145970637 |
| Residual | 3 | 27.30963303 | 9.103211009 | | |
| Total | 4 | 62 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | 2.0711009 | 3.323394495 | 0.623188406 | 0.577325343 | -8.505423614 | 12.64763 | -8.5054236 | 12.6476254 |
| X Variable 1 | 0.6307339 | 0.323101566 | 1.952122837 | 0.145970637 | -0.397519438 | 1.658987 | -0.3975194 | 1.65898733 |

**This is the data set from SCEA CostProf Module 8 - Regression Analysis.*



E.R. Druker, eric.druker@ngc.com, 703 408-0589

# Bivariate Linear Regression

| | |
|---|---|
| Random Number | 0.8043 |
| Prediction Interval | 61% |
| α | 39% |

← =(1-0.8043)*2

- It is important to note the relationship between the α and the random number in the prediction interval formula
- The random number generated will vary uniformly between 0 and 1
  - If the random number is < 0.5 then the corresponding α is the random number * 2
  - If the random number is ≥ 0.5 then the corresponding α is the (1 -random number) * 2
- Using the α value, the percentile of cost for that run of the simulation is determined
  - If the random number is < 0.5 then the corresponding percentile is α/2
  - If the random number is ≥ 0.5 then the corresponding percentile is 1 - α/2
- For example:
  - A random number of .8 will be equivalent to the 80th percentile of cost
  - This is equivalent to the upper bound of the 60% confidence interval

**This is the data set from SCEA CostProf Module 8 - Regression Analysis.*



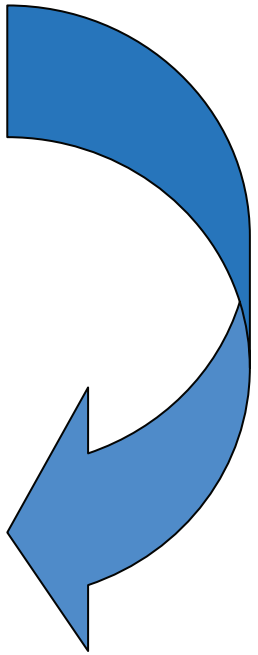
E.R. Druker, eric.druker@ngc.com, 703 408-0589

# Bivariate Linear Regression

- To use the prediction interval for risk analysis, the prediction interval equation along with the data from the ANOVA, the manually calculated data, and a random number draw are used to choose a point on the CDF which is the value of the final cost for that run of the Monte Carlo simulation

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

| | |
|---|---|
| Random Number | 0.8043 |
| Prediction Interval | 61% |
| alpha | 39% |
| X | 15 |
| Value of Regression Line | 11.53 |
| Value of Final Cost for THIS Run | 15.30 |



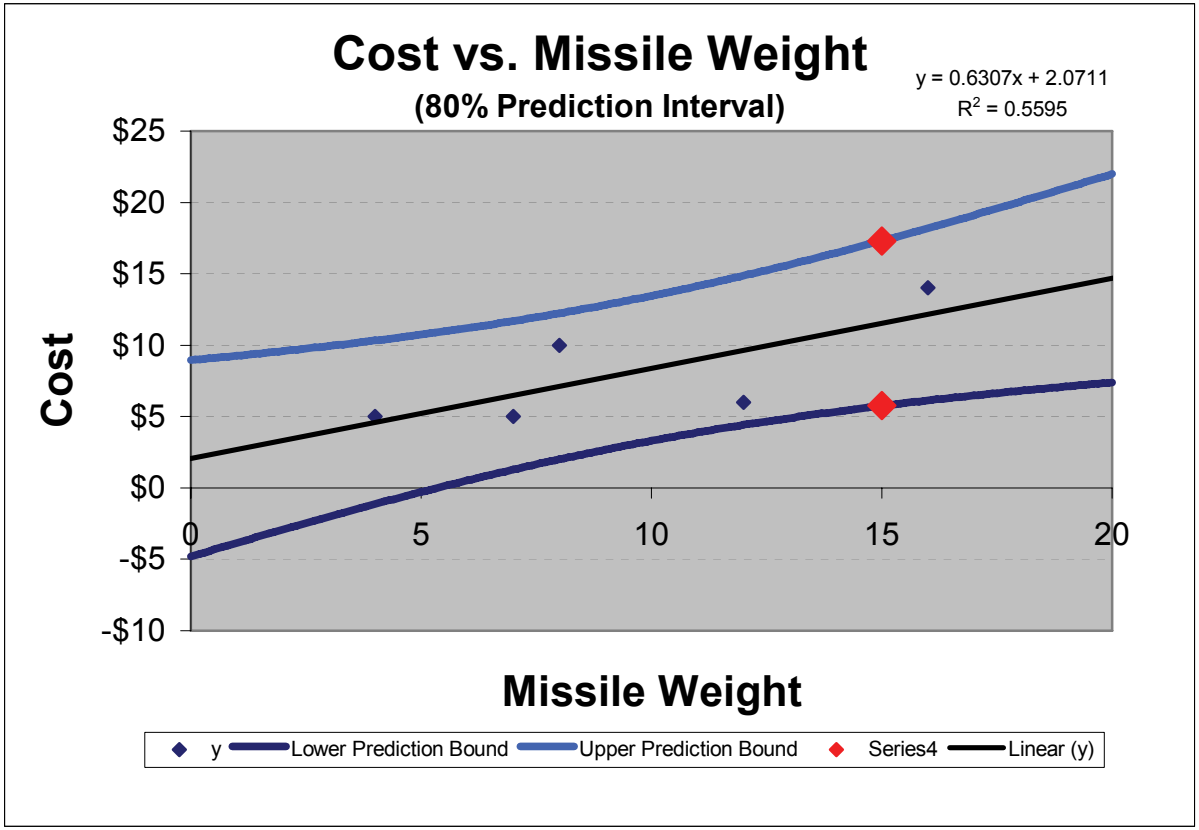
**This is the data set from SCEA CostProf Module 8 - Regression Analysis.*



E.R. Druker, eric.druker@ngc.com, 703 408-0589

# Bivariate Linear Regression

- In the simulation, the α value is a function of the random number drawn for that run. Assuming the traditional Inverse CDF technique is being used to generate instances from a distribution given a random number, the random number generated will vary uniformly between 0 and 1




**This is the data set from SCEA CostProf Module 8 - Regression Analysis.*



E.R. Druker, eric.druker@ngc.com, 703 408-0589

# Bivariate Non-linear Regression

- This 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
  - Exponential CER: $y = ae^{bx}$
  - We evaluate cost given a missile weight of 15 units (x=15)
- 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



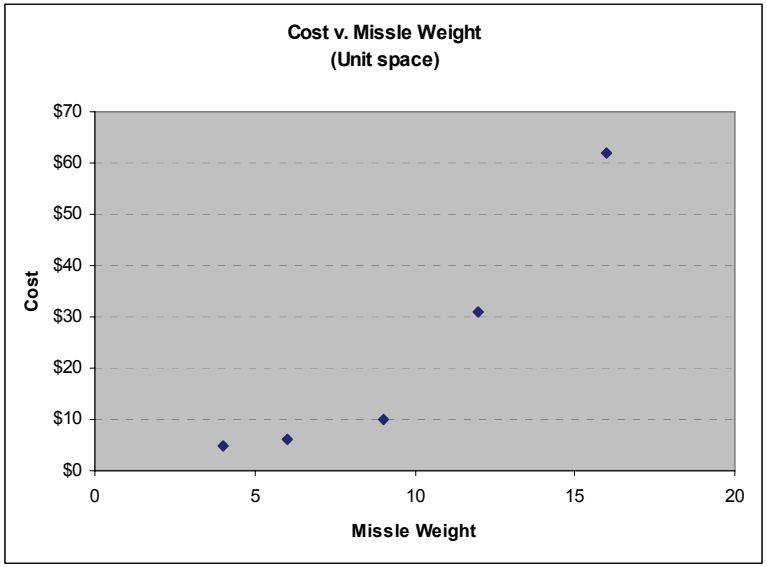
E.R. Druker, eric.druker@ngc.com, 703 408-0589

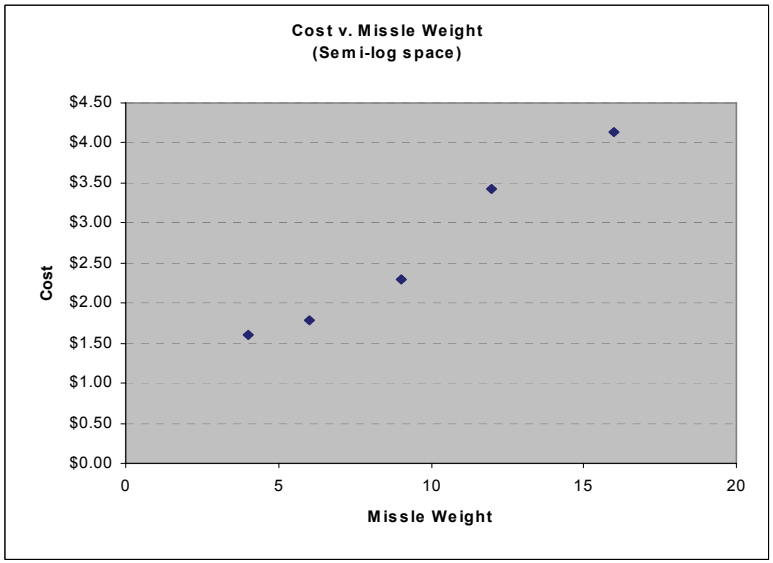
# Bivariate Non-linear Regression Linear Transformation

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

| Weight | Cost |
|---|---|
| 4 | $ 5 |
| 6 | $ 6 |
| 9 | $ 10 |
| 12 | $ 31 |
| 16 | $ 62 |

| Weight | Ln (Cost) |
|---|---|
| 4 | $ 1.61 |
| 6 | $ 1.79 |
| 9 | $ 2.30 |
| 12 | $ 3.43 |
| 16 | $ 4.13 |





E.R. Druker, eric.druker@ngc.com, 703 408-0589

Presented at the 2008 SCEA-ISPA Joint Annual Conference and Training Workshop - www.iceaaonline.com

# Bivariate Non-linear Regression

NORTHROP GRUMMAN

- Run ANOVA on the data set in semi-log space
  - Along with the ANOVA output, compute the average missile weight ($\overline{X}$) and sum the squared missile weights ($\sum X^2$).
  - The highlighted cells are needed to calculate the prediction intervals

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.984578517 |
| R Square | 0.969394856 |
| Adjusted R Square | 0.959193141 |
| Standard Error | 0.219820479 |
| Observations | 5 |

| Not Computed as Part of Excel ANOVA | |
|---|---|
| Average of x | 9.4 |
| Sum of x^s | 533.0 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 4.591597758 | 4.591597758 | 95.02273701 | 0.002293591 |
| Residual | 3 | 0.144963129 | 0.048321043 | | |
| Total | 4 | 4.736560886 | | | |

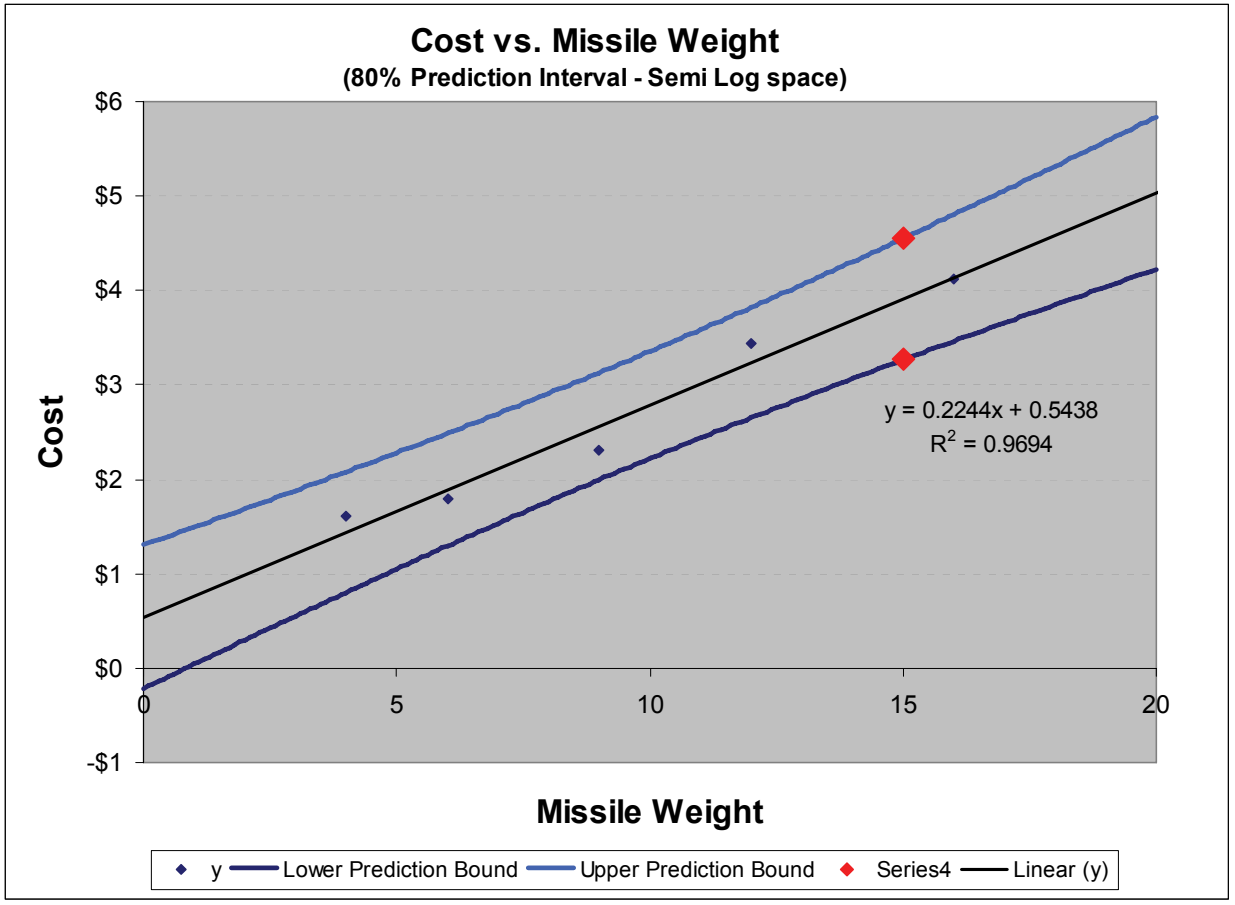
| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | 0.54380709 | 0.237656301 | 2.288208174 | 0.106131002 | -0.212521326 | 1.300135507 | -0.212521326 | 1.300135507 |
| x | 0.224380183 | 0.023018167 | 9.747960659 | 0.002293591 | 0.151126104 | 0.297634263 | 0.151126104 | 0.297634263 |



E.R. Druker, eric.druker@ngc.com, 703 408-0589

# Bivariate Non-linear Regression Prediction Interval in Semi-log Space

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



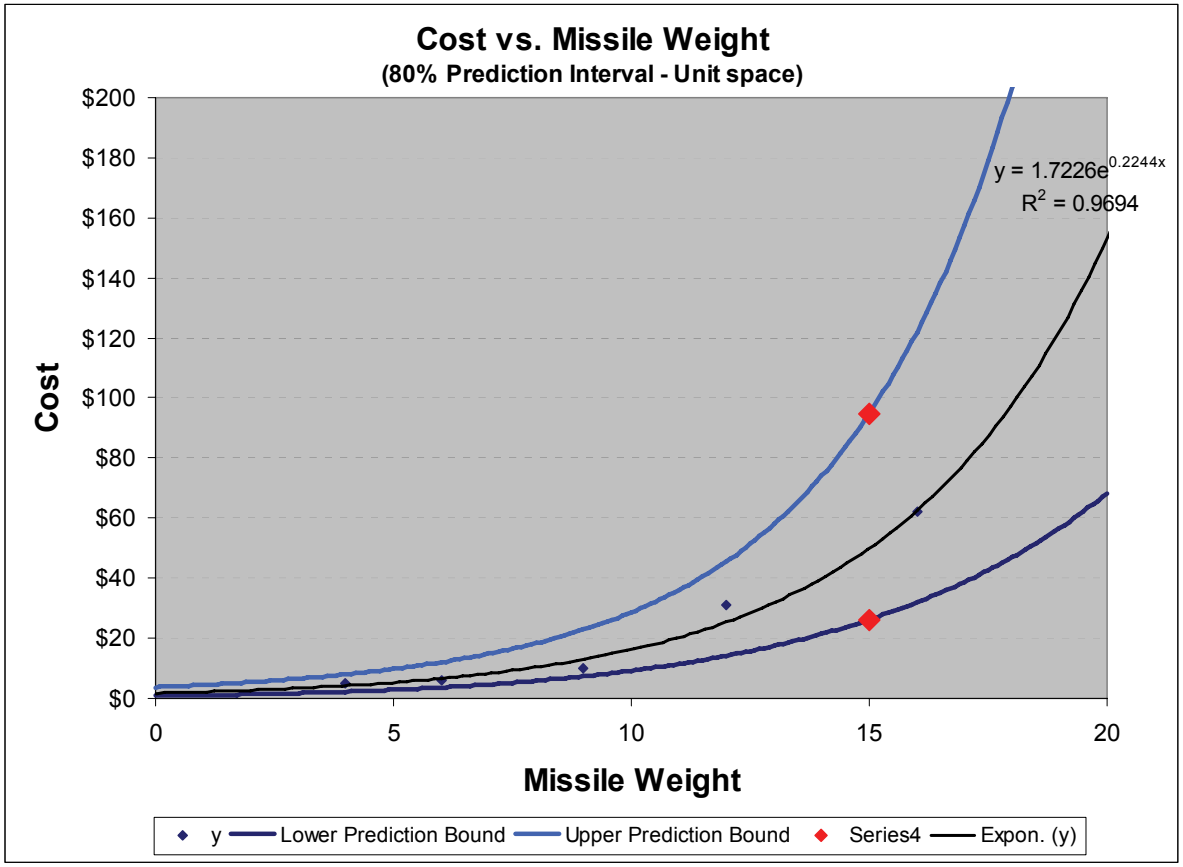

Prediction intervals and regression in semi-log space resemble those in the bivariate linear example

Presented at the 2008 SCEA-ISPA Joint Annual Conference and Training Workshop - www.iceaaonline.com

# Bivariate Non-linear Regression

NORTHROP GRUMMAN

- Final step is to transform back to unit space
  - Since this is an exponential CER, take the exponential of all cost (y) values in semi-log space to get back to unit space

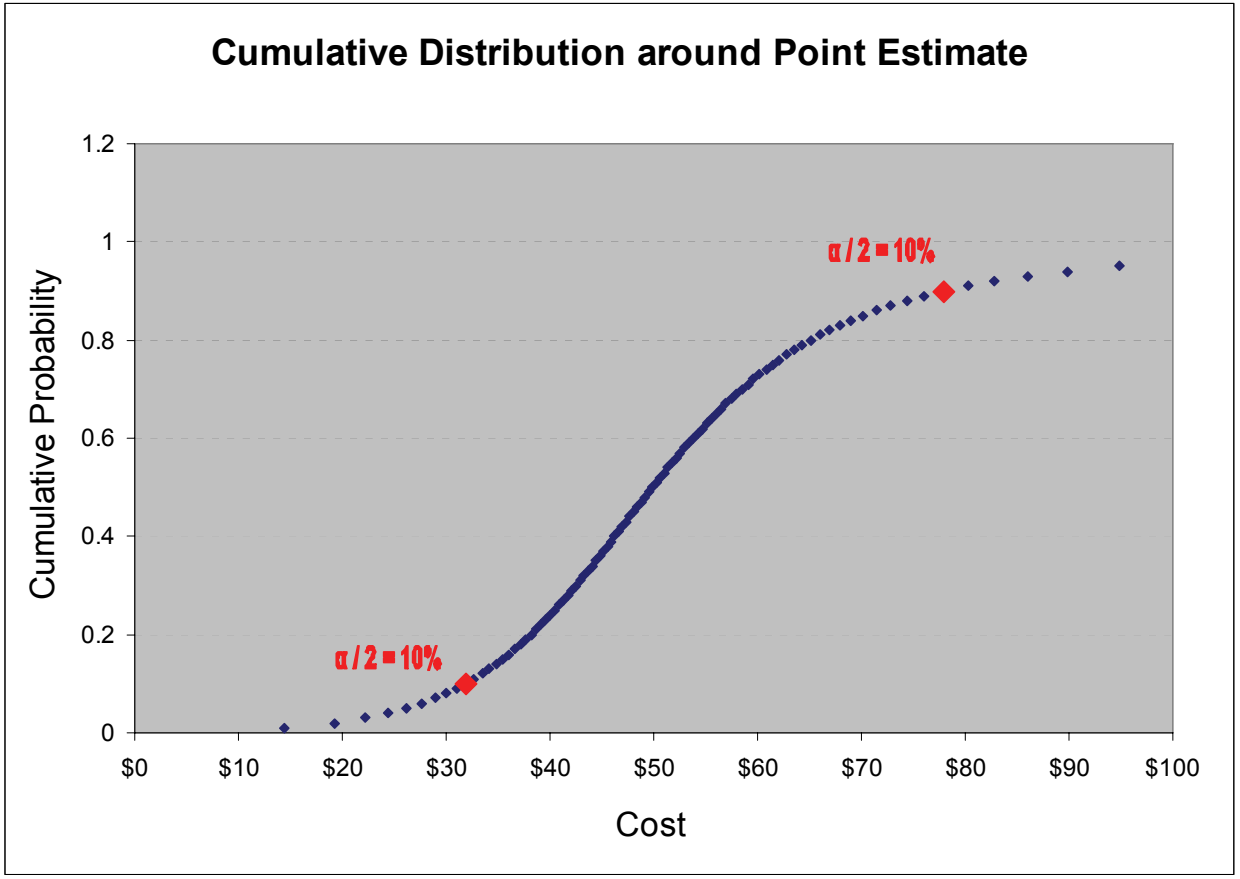





E.R. Druker, eric.druker@ngc.com, 703 408-0589

# Bivariate Non-linear Regression

- Again, the same methodology used in the bivariate linear regression that produced the inverse CDF around the point estimate is applied




| | Semi-log Cost | Unit Cost |
|---|---|---|
| Random Number | 0.671563599 | |
| Prediction Interval | 34% | |
| alpha | 66% | |
| X | 15 | |
| Value of Regression Line | $ 3.91 | $ 49.87 |
| Value of Final Cost for THIS Run | $ 4.04 | $ 57.04 |

For a missile weighing 15 units, the mean of the regression (50th percentile) is $49.87. This snapshot of the random number generated prediction interval estimates cost at $57.04 or the 67th percentile



E.R. Druker, eric.druker@ngc.com, 703 408-0589

Presented at the 2008 SCEA-ISPA Joint Annual Conference and Training Workshop - www.iceaaonline.com

# 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-α) prediction interval around any estimate is:

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

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


E.R. Druker, eric.druker@ngc.com, 703 408-0589

Presented at the 2008 SCEA-ISPA Joint Annual Conference and Training Workshop - www.iceaaonline.com

# Simulating Prediction Distributions

- Once the prediction interval has been generated, the next step is to add it into a risk model
- This allows the distribution around the CER-based estimate to be simulated along with other risk, opportunity and uncertainty distributions
- The general method for achieving this is through the use of a Monte Carlo Simulation
  - For situations where the entire estimate is produced using one CER, a Monte Carlo simulation is not needed
    - The S-Curve can be generated simply by using the formula
  - Situations where Monte Carlo analysis is requred are:
    - When there are multiple CERs used to develop an estimate
    - When uncertainty around the cost driver(s) is being used
- The following slide will quickly outline the most common method for generating draws from a risk/uncertainty distribution
  - This is known as the Inverse CDF Technique
- A method will then be shown for using the Inverse CDF Technique to simulate prediction distributions


E.R. Druker, eric.druker@ngc.com, 703 408-0589

# The Inverse CDF Technique



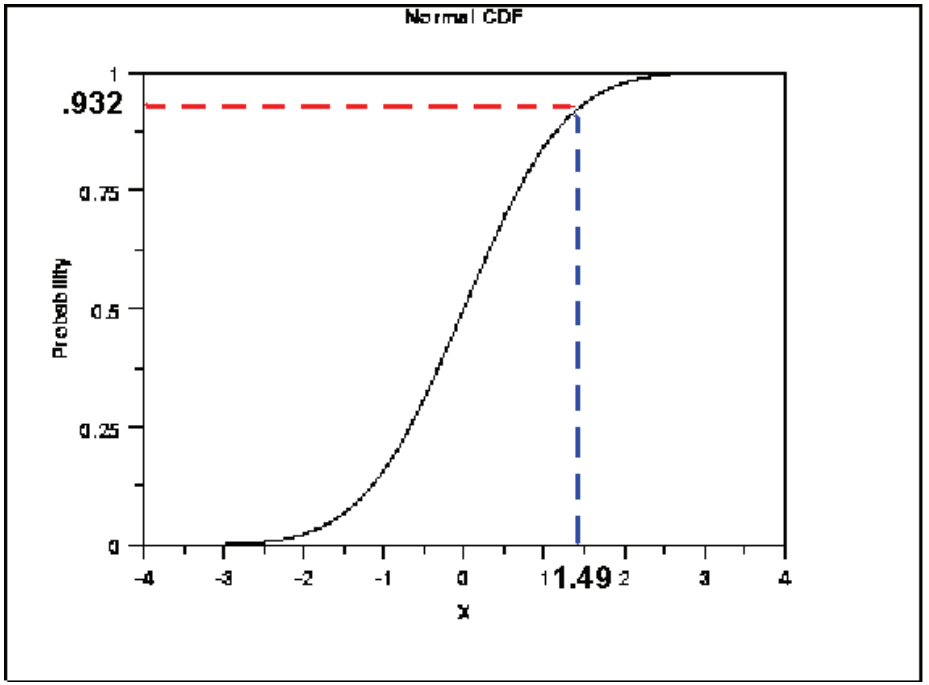
- The basis of most Monte Carlo simulations is the Inverse CDF Technique
- The Inverse CDF Technique is the standard method of generating results from any given cumulative density function
  - It uses a Uniform (0,1) Random Variable
    - Every simulation must contain some sort of Uniform (0,1) random number generator
    - In Excel this function is "=RAND()"
    - There is an entire area of computer science dedicated to the production of "pseudo random numbers" with the goal of producing "the most apparently random" output
- The CDF of a distribution maps a value (x) to the probability that the random variables takes on a value less an or equal to x
  - Therefore, the inverse of the CDF maps a probability (between 0 and 1) to a value from the distribution
  - By generating a random uniform(0,1) random number, we can produce a value from any distribution with an invertible CDF



E.R. Druker, eric.druker@ngc.com, 703 408-0589

# Standard Normal Example

- In Excel, random uniform(0,1) draws can be generated using the =RAND() function
- Using the inverse CDF technique, this random number draw can then be mapped to a value from any invertible distribution
- The example to the right uses the standard normal distribution
  - The random number draw .932 maps to the value 1.49 of the distribution






E.R. Druker, eric.druker@ngc.com, 703 408-0589

Presented at the 2008 SCEA-ISPA Joint Annual Conference and Training Workshop - www.iceaaonline.com

# Simulating Prediction Intervals

NORTHROP GRUMMAN

- Unfortunately, simulating prediction intervals is not quite as simple as simulating the standard normal distribution
- The steps below will allow simulation of prediction intervals
  - Gather all the information needed for the prediction interval equation onto one worksheet
  - Generate a uniform(0,1) random number draw using the RAND() function
    - This random number will update each time a cell is change in the model
    - Using the Lurie-Goldberg Method prediction intervals can have Pearson's correlation applied between them by correlating the random number draws1
  - Enter the Inverse CDF equation and link it up to the random number draw and the information for the prediction interval
  - Write a macro that loops 5000 times and for each loop stores the prediction interval value
  - Sort the results to generate a cumulative distribution function
  - Gather the desired information from the CDF of the prediction interval
    - Means, Standard Deviations, Percentiles
- The prediction interval distribution can be derived with any other risk/uncertainty distributions to develop a risk model


E.R. Druker, eric.druker@ngc.com, 703 408-0589

# Other Issues

- There are two main issues with this method that the user should be aware of before beginning implementation
  - The first is that utilizing CERs that are not statistically based is not advisable
  - This method only applies for OLS regression techniques, where the residuals are distributed normally
  - 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



E.R. Druker, eric.druker@ngc.com, 703 408-0589

# Conclusions

- One of the benefits of a methodology like this is that it takes into account two of the common issues estimators have with CERs
  - The quality of the CER:
    - The larger the CV of the regression, the larger the CV of the prediction interval cumulative distribution
  - "Estimating outside the range of the 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
- Generating uncertainty distributions from CERs is one simple way of accounting for risk in cost estimates
- The S-Curves developed using this and similar methodologies are critical for decision makers as they determine funding levels for programs



E.R. Druker, eric.druker@ngc.com, 703 408-0589

# Conclusions

- 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 development that the object being developed is routinely bigger, faster, stealthier (or commonly, smaller) than heretofore**
    - **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