

Uncertainty Analysis and the Project Cost Estimating Capability

NASA Cost Symposium
August 2014

Andy Prince – MSFC/Engineering Cost Office

Brian Alford – Victory Solutions MIPSS Team/Booz Allen Hamilton

Blake Boswell – Victory Solutions MIPSS Team/Booz Allen Hamilton

Matt Pitlyk – Victory Solutions MIPSS Team/Booz Allen Hamilton



**Engineering
Cost
Office**



Booz | Allen | Hamilton



Abstract



Engineering
Cost
Office

- Uncertainty and Risk Analysis is an increasingly important part of cost estimating. In addition to the increased demand from decision makers to see this type of analysis, NASA policy often requires certain programs to report a confidence level along with the cost estimate or a range of costs "with a confidence level established by a probabilistic analysis". This presentation will cover how to incorporate uncertainty in an estimating model. This includes calculations of prediction intervals around a regression based Cost Estimating Relationship. It will compare the effects of using just input uncertainty, just model uncertainty, and both input and model uncertainty on the final results of the model. Finally, there will be a demonstration of how this math has been implemented and automated in the Project Cost Estimating Capability (PCEC) and how to add uncertainty to a PCEC model.





Contents



**Engineering
Cost
Office**

- Input Uncertainty
- Model Uncertainty
 - What is Model Uncertainty
 - Regression Overview
 - Confidence vs Prediction Intervals
 - Calculating Intervals
- Examples of Including Uncertainty
- Uncertainty in PCEC Demo





Input Uncertainty



**Engineering
Cost
Office**

Input Uncertainty is variation around the inputs to a model

- Type that most people are familiar with and many include in their models
- Important because the final values for many parameters are not known at the time of estimating
- Sources include:
 - Uncertainty around the final design
 - Variation in historical data
 - Changing requirements
 - Uncertainty of who will perform the work
 - Variation in price of materials
 - Weather



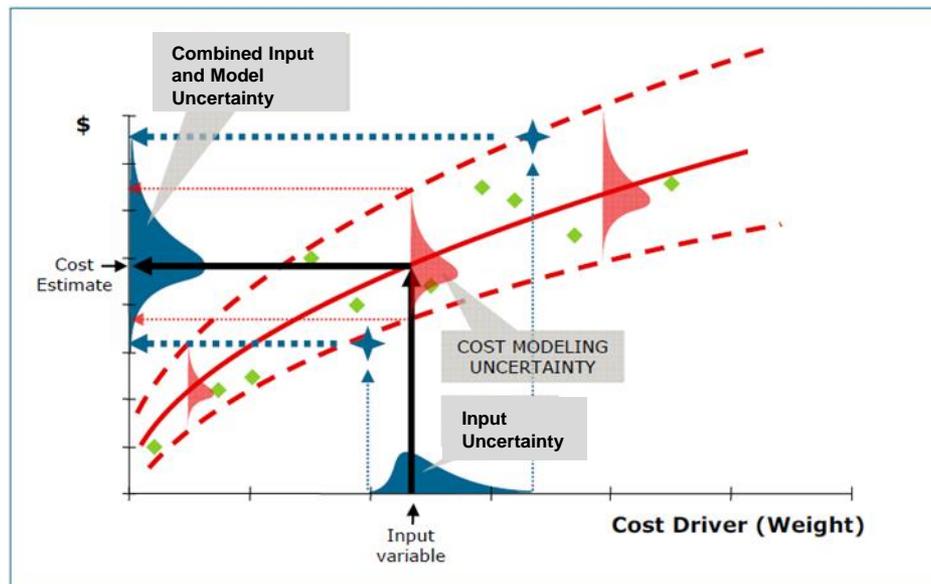
Input Uncertainty is Not Enough for Regression Equations



Engineering
Cost
Office

Model Uncertainty is the uncertainty around the estimate produced by a regression equation

- Holding the input value steady will still produce an output distribution because of the modeling error
- The total uncertainty of the model is a combination of two types of variation: input and model



Non-linear regression example adapted from 2008 NASA Cost Estimating Handbook



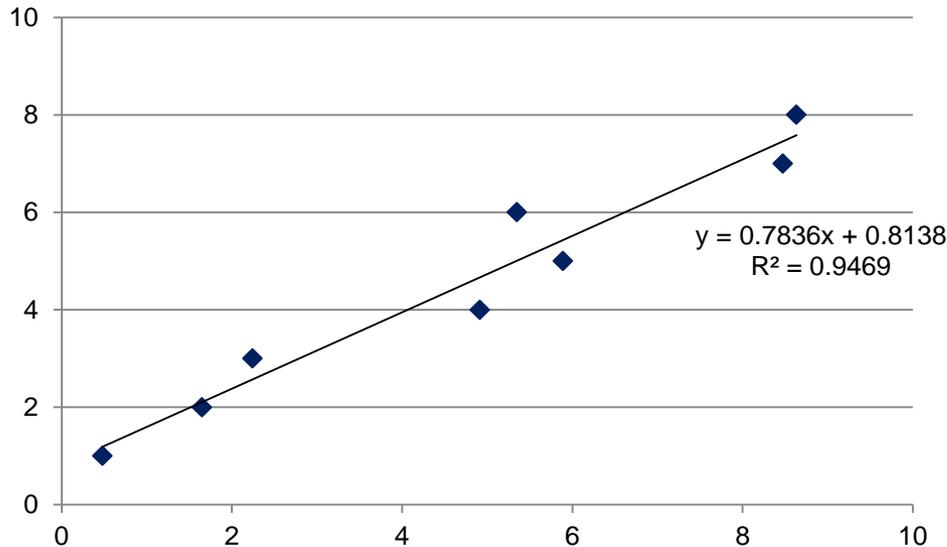


Regression Overview



Engineering
Cost
Office

Ordinary Least Squares regression finds the line through the data that minimizes the sum squared error. This produces a *mean* estimator.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.973075
R Square	0.946876
Adjusted R Square	0.938021
Standard Error	0.609812
Observations	8

ANOVA

	df	SS	MS	F
Regression	1	39.76877	39.76877	106.9424
Residual	6	2.231226	0.371871	
Total	7	42		

	Coefficients	Standard Err	t Stat	P-value
Intercept	0.813835	0.416583	1.953596	0.098566
X Variable 1	0.783633	0.075777	10.3413	4.78E-05

We use information from the regression analysis to produce confidence and prediction intervals which help us estimate.



Confidence vs Prediction intervals

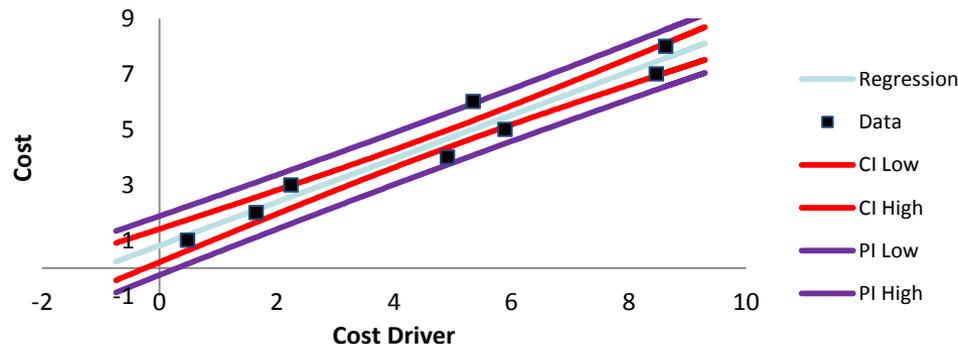


Engineering
Cost
Office

Confidence and Prediction intervals are related but provide bounds for different types of estimation.

- **Confidence intervals** give bounds for estimating the true value of the regression line (the true mean of Y) at a particular x value (or vector if there are multiple independent variables).
- **Prediction intervals** give bounds for estimating another observation (particular value of Y) at a particular x value (or vector if there are multiple independent variables).

Confidence & Prediction Intervals



Prediction intervals include two types of variation: the error around estimating the true mean and the natural variation around the true mean.

Because the estimate of the mean necessarily has less variation than the estimate of observations that produce that mean, confidence intervals are smaller than prediction intervals.



Why are Prediction Intervals Bigger?



Engineering
Cost
Office

Confidence intervals:

Here we are trying to predict the mean y value and so are only concerned about the variation around the regression line estimate.

$$\begin{aligned}\text{Mean estimate} &= \hat{y}_0 | x_0 \\ \text{var}(\hat{y}_0 | x_0) &= \text{var}(x_0 \hat{\beta} | x_0) \\ &= x_0 \text{var}(\hat{\beta}) x_0' \\ &= x_0 \hat{\sigma}^2 (X'X)^{-1} x_0' \\ &= \hat{\sigma}^2 x_0 (X'X)^{-1} x_0'\end{aligned}$$

Where $\hat{\beta}$ is the estimated regression coefficients and X is the design matrix.

Prediction intervals:

Here we are trying to predict a new y value (not the mean of the y's) and so are also concerned with the variation of the y values themselves, not just around the mean. To account for this, we add another term.

$$\begin{aligned}\text{New observation estimate} &= \tilde{y}_0 = \hat{y}_0 | x_0 + \varepsilon_0 \\ \tilde{y}_0 &= \hat{y}_0 | x_0 + \varepsilon_0 \\ \text{var}(\tilde{y}_0 | x_0) &= \text{var}(\hat{y}_0 | x_0 + \varepsilon_0) \\ &= x_0 \text{var}(\hat{\beta}) x_0' + \text{var}(\varepsilon_0 | x_0) + 2\text{cov}(x_0 \hat{\beta}, \varepsilon_0 | x_0) \\ &\quad \text{(no covariance because of OLS assumptions)} \\ &= x_0 \hat{\sigma}^2 (X'X)^{-1} x_0' + \hat{\sigma}^2 \\ &= \hat{\sigma}^2 [x_0 (X'X)^{-1} x_0' + \mathbf{1}]\end{aligned}$$

The inclusion of the additional term results in a “+1” in the final step which makes the prediction intervals larger.





Calculating Intervals



Form of an interval:

Output from regression equation \pm critical value from t distribution \times Standard Error

Ordinary Least Squared dictates that we use a t-distribution because

- We assume normally distributed data around means
- Do not know the actual variance of the residuals (σ^2) and so must use an estimate (SEE)

The difference between confidence and prediction intervals is in the Standard Error term which we have already seen. So

$$\text{C.I.} = \hat{y}_0 | x_0 \pm t_{(\text{confidence level, residual degrees of freedom})} * \text{SEE} * \sqrt{x_0 (X'X)^{-1} x_0'}$$

$$\text{P.I.} = \hat{y}_0 | x_0 \pm t_{(\text{confidence level, residual degrees of freedom})} * \text{SEE} * \sqrt{1 + x_0 (X'X)^{-1} x_0'}$$

SEE (Standard Error of the Estimate) comes from our regression analysis and is our estimate for sigma.

Note: These formulas are more common written as:

$$\text{C.I.} = \hat{y} | x_0 \pm t_{\frac{\alpha}{2}, df} \times \text{SEE} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum X^2 - n\bar{X}^2}}$$

$$\text{P.I.} = \hat{y} | x_0 \pm t_{\frac{\alpha}{2}, df} \times \text{SEE} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum X^2 - n\bar{X}^2}}$$

However, these formulas do not extend to the case with multiple independent variables. Thus matrix algebra (and notation) is used.



Producing and Using Prediction Intervals



Engineering
Cost
Office

Now that we know what predictions intervals tell us and how they are different than confidence intervals, how do we use them?

The use we are concerned with is producing estimated y values for a Monte Carlo simulation. Because Monte Carlo simulations rely on the range of future observations rather than the mean of the observations, prediction intervals are used to randomly sample from the distribution of future observations.

1. Calculate the output of the regression equation for the x value(s). This is the mean estimate.
2. Randomly sample from a t distribution with degrees of freedom equal to the residual degrees of freedom from the regression analysis.
3. Calculate the Standard Error for the regression equation at the given input vector.

$$se(\tilde{y}|x_0) = SEE\sqrt{1 + x_0(X'X)^{-1}x_0'}$$

where SEE = the Standard Error of the Estimate for the regression analysis and X is the design matrix. The design matrix is of size [(number of data points) x (number of variables + 1)] where the first column is all 1's, the second column is the data points for the first variable in the regression, the third column is the data points for the second variable in the regression, etc.

Now we have all the part for the expression:

Output from regression equation + critical value from t distribution \times Standard Error

This will produce a single potential future observation of y either above or below the estimated mean (because the sample from the t distribution may be positive or negative) which can then be used in one trial of a Monte Carlo simulation.

Notes: The PCEC refers to $(X'X)$ as the “squared design matrix” and includes this with CERs for users to use for including model uncertainty in estimates.

The PCEC includes Prediction Interval calculations for appropriate CERs. Users can use any Monte Carlo addin with the PCEC to add uncertainty to their models.





Why Are Prediction Intervals curved?



Prediction (and confidence) intervals are narrowest near the point (\bar{x}, \bar{y}) and get wider as the x value(s) move away. To understand why, it is easier to refer again to the PI equation for a univariate regression:

$$\text{P.I.} = \hat{y}|x_0 \pm t_{\frac{\alpha}{2}, df} \times SEE \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum X^2 - n\bar{X}^2}}$$

Looking at this expression we can make a few observations:

- When the input value equals the mean, $x_0 = \bar{X}$, the third term under the radical is 0 and the standard error is at its smallest.
- When $x_0 \neq \bar{X}$, the third term under the radical is positive. Moreover, the standard error grows larger as x_0 gets farther from \bar{X} . This happens in either direction because the term is squared and thus always positive.
- The regression equation is most precise near the center of the data set. Estimating outside the range of the data carries large uncertainty because the standard error grows large.
- The standard error must be calculated for every input value. During simulations where x_0 is varied (input uncertainty), the standard error must be recalculated for each trial.
- Increasing the number of data points (n) adds precision to the C.I.'s more than the P.I.'s. This is because increasing n makes the second and third terms under the radical vanish to 0. These are the only terms under the radical for C.I.'s, but P.I.'s have the 1 as well.



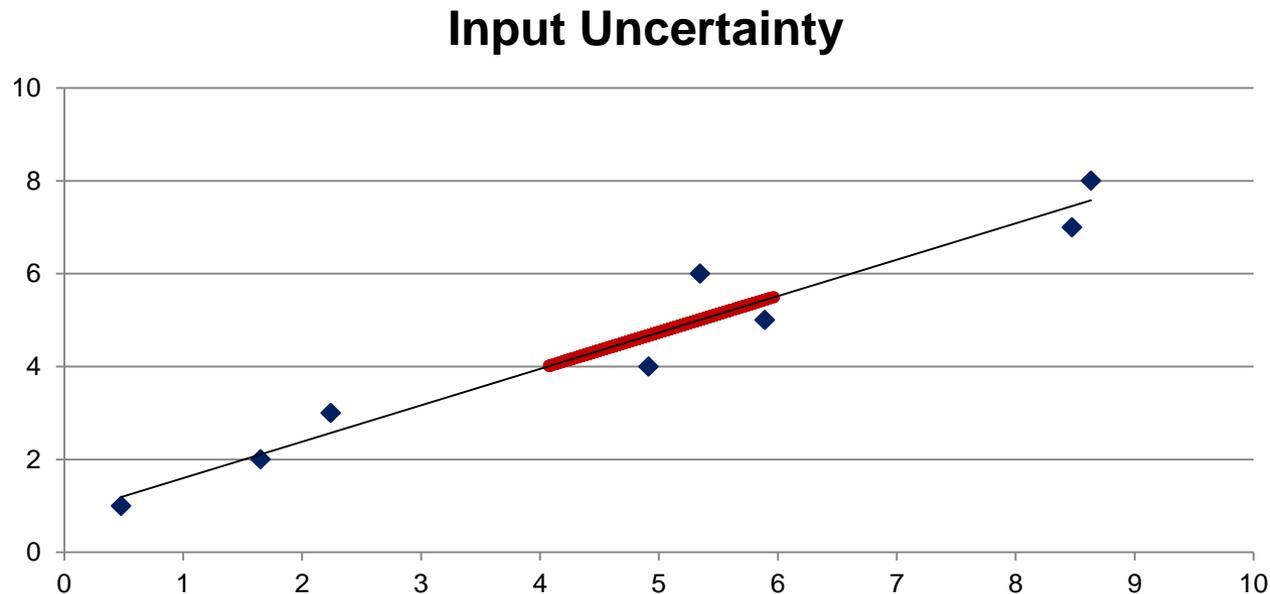


Output Comparisons: Input Uncertainty



Engineering
Cost
Office

This graph illustrates the results of including only input uncertainty in a simulation.



It assumes that the output from the regression equation is the desired predicted value, which is rarely the case when incorporating uncertainty. The regression equation estimates the mean of the y values, but usually an estimate for a new observation is desired.



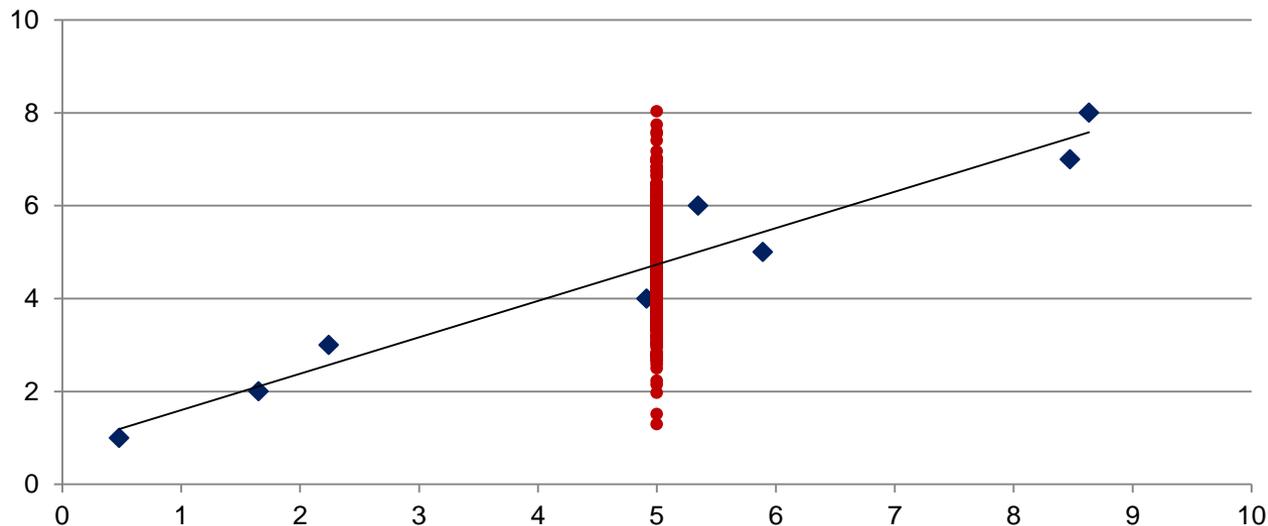
Output Comparisons: Model Uncertainty



Engineering
Cost
Office

This graph illustrates the results of including only model uncertainty in a simulation.

Modeling Uncertainty



It assumes that the input value for x is known absolutely at the time of estimating, which is not usually the case. It uses the regression equation to get one mean estimate and then uses prediction intervals to produce a range of y values for the sole x value.





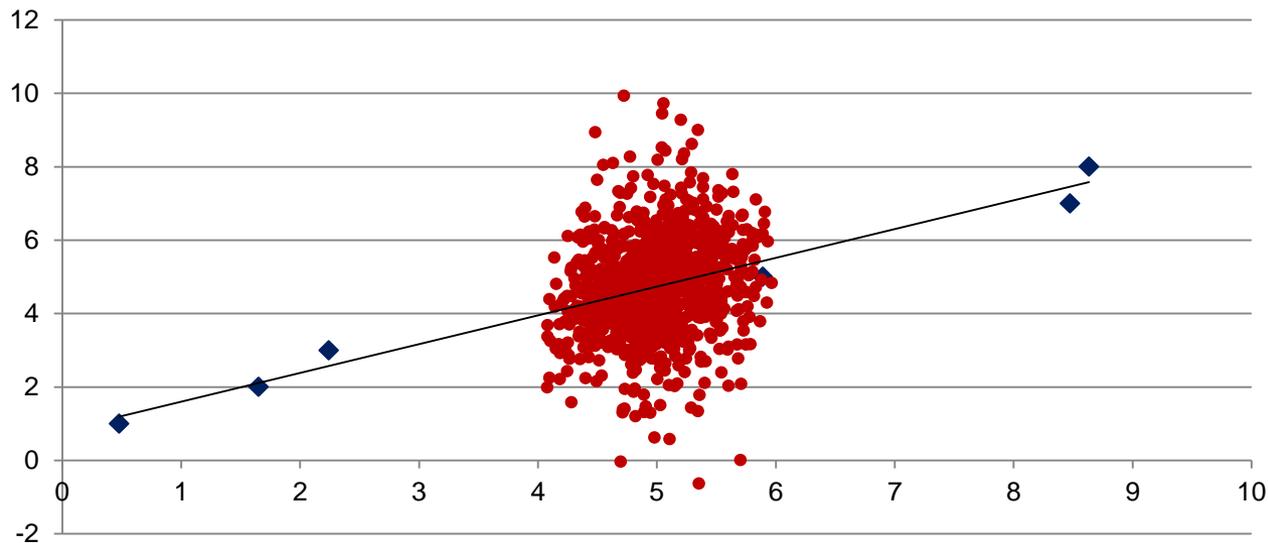
Output Comparisons: Input and Model Uncertainty



Engineering
Cost
Office

This graph illustrates the results of including both input and model uncertainty in a simulation.

Input and Model Uncertainty



This captures the widest and most accurate range of potential y values by including both types of uncertainty.