INTRODUCTION

01 As the statistician George Box once said, “All models are wrong, but some are useful”

02 Development of cost models involves reliance on correct application of statistics

03 But - this is not a cut-and-dried mechanical process
   • Significant amount of judgment ("art") required
   • Many effective ways to develop a good cost model but no single best way

04 In this presentation we discuss the cost modeling process, and discuss a variety of considerations

05 To make the discussion concrete we discuss this topic in the context of a recent project, SEER for Space
Dave Featherstone, Professor of Biology/Neuroscience at the University of Illinois, argued that "art = science"

Both science and art are attempts to understand and describe the world around us.

Subjects and methods have different traditions and the intended audiences are different, but the motivations and goals are fundamentally the same.
ART vs. SCIENCE – FALSE DICHOTOMY (2)

Researchers from The UCLA Jonsson Comprehensive Cancer Center argue the same point:

Common misconception that art and science are vastly different - they never overlap

Creativity is as essential to the scientific process as it is to the artistic method

Artists and scientists share a curiosity for the unknown, an appreciation for the beauty of the worlds they explore and an interest in creating something new

Source: http://newsroom.ucla.edu/stories/unexpected-intersection-art-science-UCLA-Jonsson-Cancer-Center

This talk will discuss the art and science of cost modeling
The use of heuristics in cost modeling is an example of injecting artistic ideas into cost modeling.

**Some Examples:**

- Dollars per kg (or lb) to estimate the cost of different classes of space missions (Hamaker 2001)
- Assumed CER b values (or slopes) in y = a (Kg)^b CERs
- Nonrecurring to Recurring cost ratios, including the Rule of 166
- The use of beta distributions to time phase cost
- NASA Technology Readiness Levels
THE RULE OF 166

Heuristic that allocates total acquisition cost for bus and some payload costs into nonrecurring and recurring categories

The name comes from the total of 166 one obtains if one assumes the values in the top table

Example: Bus DDT&E = $100, Bus TFU = $22, Instrument DDT&E = $33 and Instrument TFU = $11
DATA COLLECTION

Data are the foundation of cost estimates

Researchers at Google found that “simple models and a lot of data trump more elaborate models based on less data” (Halevy, Norvig, and Pereira 2009)

Need sound, quantitative data
• Cost
• Technical
• Programmatic

Analysts spend the majority of their time developing techniques and honing tools, when the most important focus should be on the quality and quantity of data.
DATA COLLECTION (2)

• How much data? Lots! For a regression model statisticians recommend at least 50 data points with an additional 10 data points for each variable.

• Obtaining this much data can be a challenge even for satellites which have more data points than most defense systems.

• When you have small amounts of data, consider using Bayesian methods (Smart 2014).
DATA COLLECTION (3)

Relevancy of data - multiple dimensions, two are:

APPLICABILITY

- Data should be analogous to the system you want to estimate
- If non-analogous, do not use

CURRENCY

- Moore's law is the observation that the number of transistors in a dense integrated circuit doubles about every two years
- Need to use current data but also need to balance against availability – average launch date of NASA launch vehicles used in the Crew and Space Transportation CER is 1979
DATA COLLECTION (4)

What to do if you are missing data:

1. Leave out the data point
2. Not use that particular variable in the CER
3. Use imputation to make educated guesses to use more data points

Variety of methods for this, including:
- Average value for similar data points, where similarity could be determined by k-nearest neighbors
- Regression analysis to predict the missing value
For the SEER-Space cost model, we collected CADRe data for bus subsystems, instruments (including data processing units and OTAs), and integration.

1. We only collected data for NASA systems.

2. We only collected data for the last twenty years.

3. Assumed that data for earth-orbiting and planetary missions can be used together for CER development – can include a dummy variable (aka binary/indicator variable) to account for the difference in the model.
DATA WRANGLING

One of the most laborious parts of the cost modeling process is the wrangling of the raw data into a consistent work breakdown structure.

INCONSISTENT REPORTING

INCONSISTENT FORMATTING

DATA ARE MISSING (Error of Omission) OR WRONG (Error of Commission)

HOWEVER...
DATA WRANGLING (2)

Using CADRe data helps expedite this…

HOWEVER THERE ARE STILL A VARIETY OF FORMATS AND LEVEL OF DETAIL

SOME AMOUNT OF ALLOCATION AND RE-BUCKETING IS STILL REQUIRED
DATA NORMALIZATION

For some older missions adjusted to account for full-cost accounting, as applicable

Split all costs into nonrecurring and recurring costs

Normalized cost to a constant base year using latest NASA inflation guidance

For multiple units, normalized recurring costs to a theoretical first unit cost
MODEL FORMS AND VARIABLE SELECTIONS

It's common to use the power-law equation in cost estimating $Y = aX^b$, works for single-variable and multiple-variable equations.

This is based on the observation that as the physical size of a product increases, its cost (both development cost and unit production cost) also increases but at a slower rate than size (Hamaker 2008).

This is reflected in the $b$ value (or “slope” parameter of the CER in log-space).

Other model forms are possible, but care should be taken to avoid overfitting (given enough model forms, one will fit the data you have extremely well just by chance, but this will not generalize well).
MODEL FORM EXAMPLE

For the data in the graph with nine data points:

- Linear fit has $R^2 = 41%$
- 8th degree polynomial is a perfect fit
- One is a perfect (over) fit but that does not mean you should use it for estimating
For SEER for Space, we used the power equation for most subsystems and instruments.

In our experience power forms have worked well for other models, and they fitted the data well for this project, so we used them extensively.

In some cases when the cost of the system was low, we used linear equations (e.g., data processing units):

- When the cost is low, it has a limited range as the lower bound is equal to zero.
- Log transformations compress the spread of the data, so low cost items are typically not nonlinear.
The primary CER variable that we use is weight (aka mass).

This is not a causative driver of cost, merely a scaling parameter.

Thus our models are not based on cost drivers – weight “no more ‘drives’ cost than your dog can drive your car” (Prince 2016).

However these parameters explain historical variation well and we typically have estimates of weight early in a program’s lifecycle so we continue to use these parameters.
VARIABLE SELECTION (2)

Another controversial topic is the use of subjective parameters such as heritage (or its reciprocal, percent new design).

The amount of new design is strongly correlated with the nonrecurring cost of a program. Heritage is typically overrated early in a program’s lifecycle, can lead to underestimation.

The subjective nature can lead to misestimation when compared to the historical cost (“dark side of the force” – Prince 2016).

However, there is a need by customers to discern the impact of new design on cost, so we have made the decision to include new design as a parameter in SEER for Space.
SEER for Space includes weight and new design as parameters.

We followed a disciplined process of developing new design values to avoid trying to explain cost variation with this parameter.

We also included other programmatic parameters, including AO/Directed; Earth-orbiting/Planetary; more than one sponsor; international involvement; extensive testing; number of instruments; number of active instruments; and design life.

To avoid issues due to multicollinearity used a collector variable in the regression for all programmatic parameters.
The number of instruments and number of active instruments is only applied to the spacecraft bus CERs.

All instrument CERs except for the telescope CER include power as a driver.

In addition to the common factors, there are several CERs that include unique technical independent variables:

- Telescope – Mirror diameter is the primary driver, weight is not an independent variable.
- Bus subsystems – varies by subsystem.
MODERN REGRESSION METHODS

• The oldest method still in common use for developing power-form CERs is log-transformed ordinary least squares linear regression (LOLS)

• However this method has been criticized over the years for multiple reasons (Book 2012):
  ▪ Transformation causes equation to be optimal for “log-dollars”
  ▪ Result is biased (estimating the median vice the mean)

• Other methods such as Minimum Unbiased Percentage Error (MUPE) and Zero-bias Minimum Percent Error (ZMPE) have been recommended as alternatives to log-transformed linear regression

• However, statistical analysis of residuals for spacecraft CERs has shown that the best for the residuals is the lognormal distribution
MODERN REGRESSION METHODS (2)

- Developed a new CER method (Smart 2017) that has the advantages of log-transformed linear regression without the transformation issues
- Used solver to minimize a maximum likelihood equation that is optimal for lognormal residuals
  - Source: C. Smart, “Cutting the Gordian Knot: Maximum Likelihood Regression of Log Normal Error,” presented 2017 ICEAA conference; this method is also included in the U.S. DoD CER Handbook
  - Also known as the MRLN method (“Merlin”)
- Similar to LOLS but unbiased
- No transformation required
- Also set logical constraints on the values for the coefficients
Nate Silver in The Signal and the Noise, calls overfitting the “the most important scientific problem you’ve never heard of.”

Overfitting is confusing noise with signal.

If the fit to the historical data is too loose, it is underfit; if too tight, it is overfit.

Overfitting is appealing because the fit statistics look great – high $R^2$s, low standard errors, etc.

Overfitting much more common in practice than underfitting.

John von Neumann – “with four parameters I can fit an elephant, and with five I can make him wiggle his trunk.”
CROSS-VALIDATION

The practical impact of overfitting is that the model will not generalize when used outside of the sample you have used to develop your model.

To avoid this we have used cross-validation for all CERs developed for SEER-Space.

Split the data set into multiple partitions, do the testing over multiple small partitions, and average the results.
Among many, many other things, Leibniz conceptualized the expression “the best of all possible worlds”.

The expression explains the concept of free will (as opposed to a clockwork world).
Adapting this optimism to cost estimating, one can argue that underestimating projects in order to “sell them” is the best of all possible worlds since to estimate projects accurately, precludes their existence (i.e., they would not “sell” if their true cost was known) (Hamaker 2011)

Counter argument is that good cost models also offer tools to guide projects toward an affordable solution
REFERENCES

REFERENCES (2)


