AUGUR

Maximizing Analysis of Minimalized Datasets

Making the Most of a Data-Scarce Acquisition Environment

2024-05-14

Taylor Fountain, Obai Kamara

Presenter Bios

Taylor Fountain

- Cross-competency analyst at Augur Consulting
- DoD acquisition support through cost estimation and data analytics, with a focus on IT Business Systems
- Experienced w/ tool development, training at Augur Consulting
- Holds a B.S. in Mathematics from George Mason University

Obai Kamara

- Senior Technical Advisor at Augur Consulting
- Cost and performance management support to government acquisition programs
- Specializes in parametric estimation, machine learning techniques, and data visualization
- B.S. in Physics from Davidson College; M.S. in Business Analytics from The University of Virginia

Agenda

- Introduction
- Literature Review
- Simulation and Testing
- Leveraging SME Input
- Conclusion

Introduction

- The key to any defendable estimate is data informed analysis

- Most analytical techniques are dependent on large datasets
  - Large datasets required to reliably identify relationships within data
  - Not always possible for variety of reasons

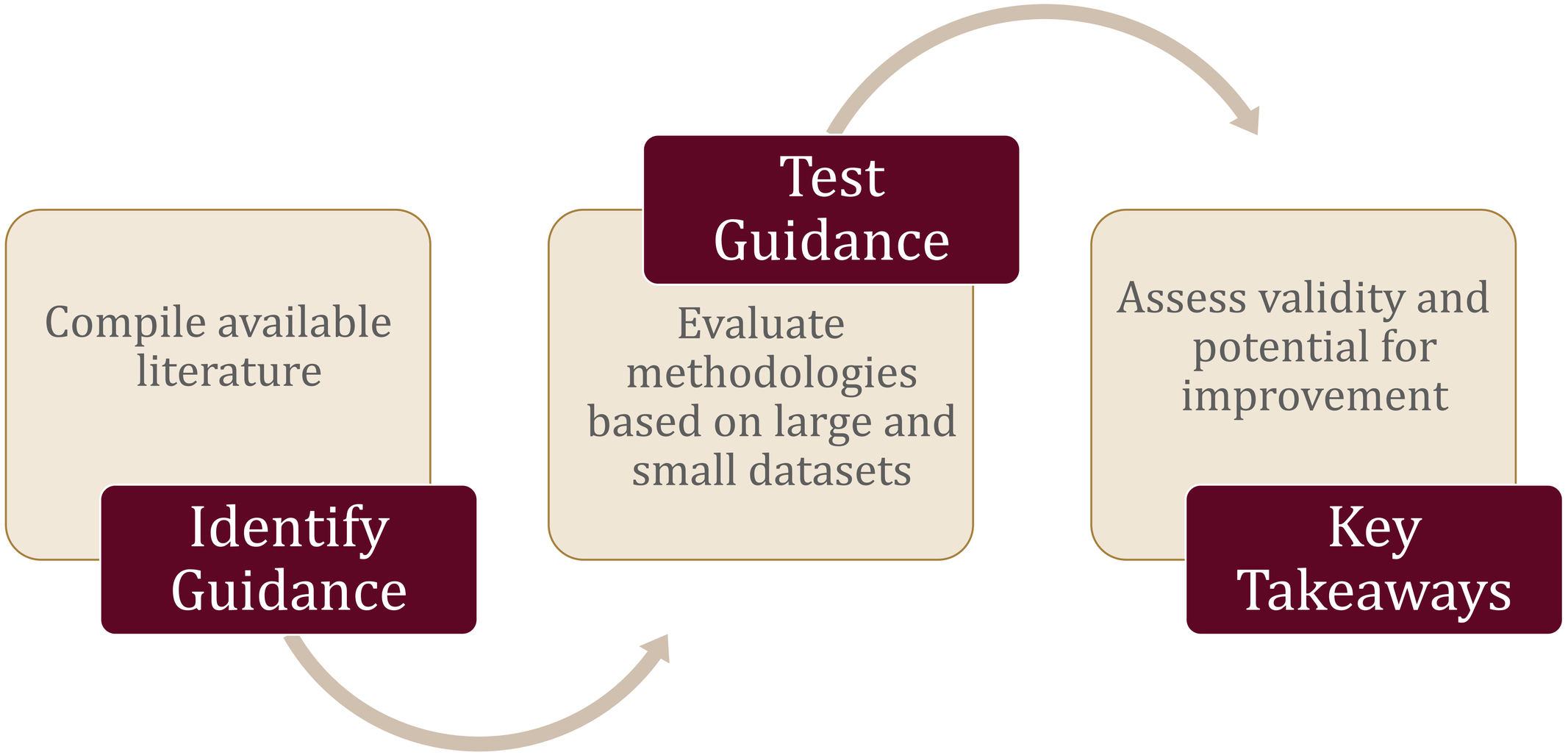
Vendor Proprietary

Unique projects

Data sharing concerns

- Goal: assess and overcome limitations of small datasets

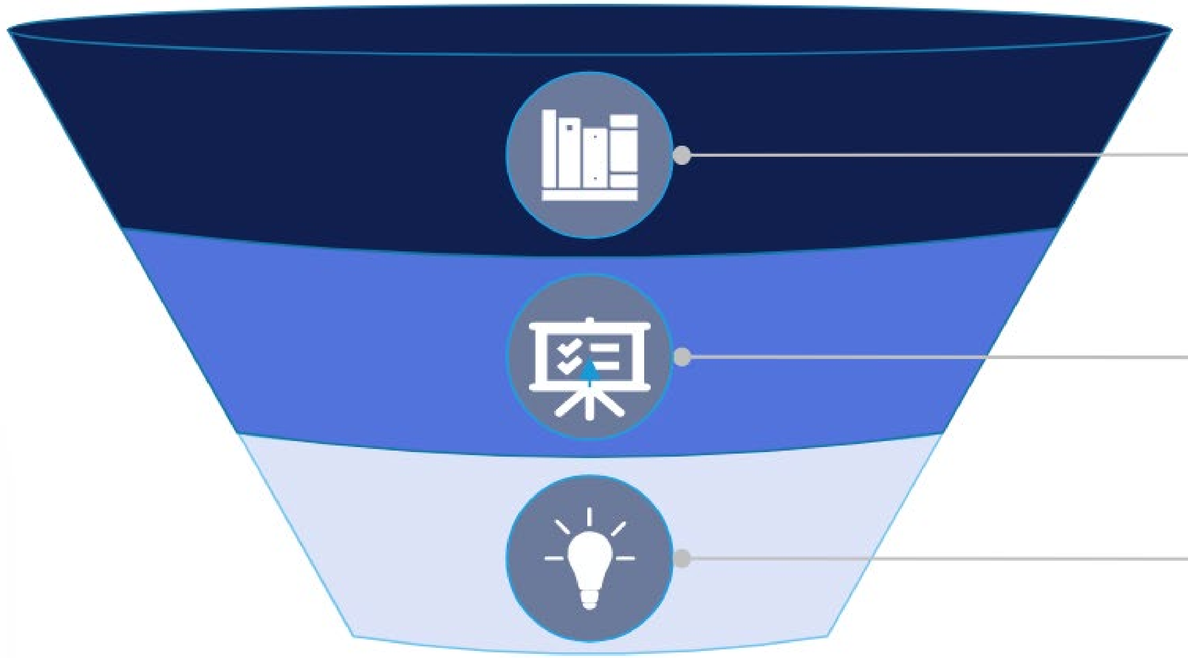
Approach


Identify Guidance
Compile available literature
Test Guidance
Evaluate methodologies based on large and small datasets
Key Takeaways
Assess validity and potential for improvement

Literature Review – Methodology

1. Reviewed industry guides and handbooks, including:
   - GAO Cost Estimating Guide, Joint Agency Cost Schedule and Risk Uncertainty Handbook, DOD CER Handbook, Navy Cost Estimating Guide
2. Searched ICEAA Archive for topics related to small data sets

Full list provided in backup



36 Total Presentations

Based on "limited" and "small" keyword search

13 Relevant Presentations

Papers or presentations focused on point estimate development or uncertainty implications

5 Provide Discrete Guidance

Research provided takeaways that could be applied to developing risk informed point estimates

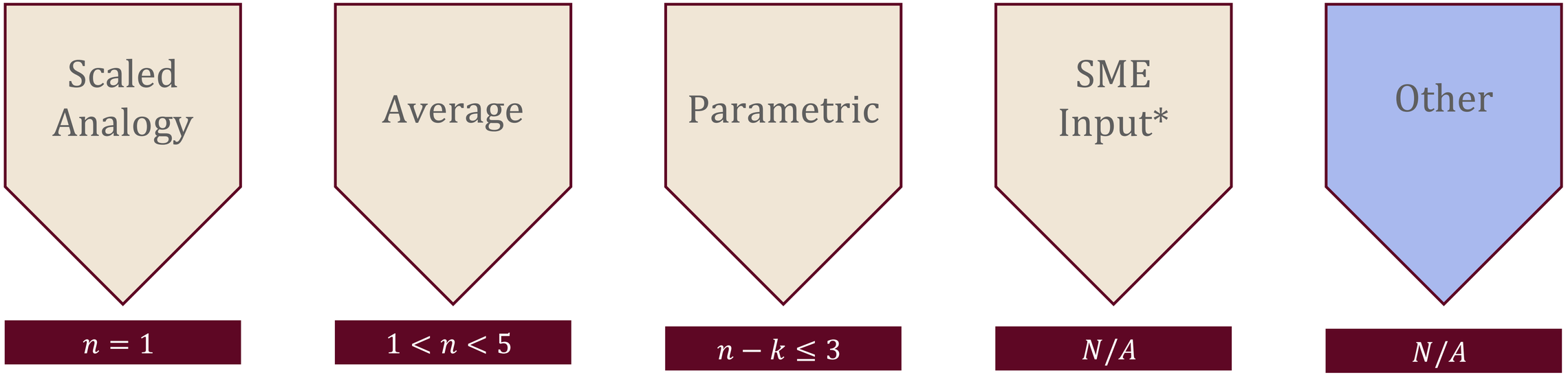
Literature Review – Results

Summary of Relevant ICEAA Presentations

Paper/Presentation	Author(s)	Year	Recommendation
Using Bayes' Theorem to Develop CERs – Extending the Gaussian Model	Christian Smart, David Jo	2023	Bayesian regression
The Unseen: Statistical Inference with Limited Data	Trevor VanAtta	2012	Infinity cropping - exclude wrong answer rather than searching for PE Gain scales - using scales to get better SME inputs
The Progression of Regressions	Jennifer Aguirre, Kyle Davis	2022	Maximize use of context Do not overfit Visualize CER Choose the best analogies
Fitting Absolute Distributions to Limited Data	Blake Boswell	2012	Use Decision on Belief (DoB) to determine optimal distribution based on 3 pt estimate
The Business Case for Bootstrapping: When You're Stuck with Incomplete Data, Here's How You Make it Work!	Brett Gelso, Glenn Grossman, Eric Druker	2010	Bootstrapping improves accuracy of CERs

Literature Review – Takeaways

- No clear definition of “small” within reviewed literature
  - Most papers did not define what was considered small (ranged from 5-30)
  - 30 observations generally considered minimum requirement for statistical significance
- Point Estimate Development

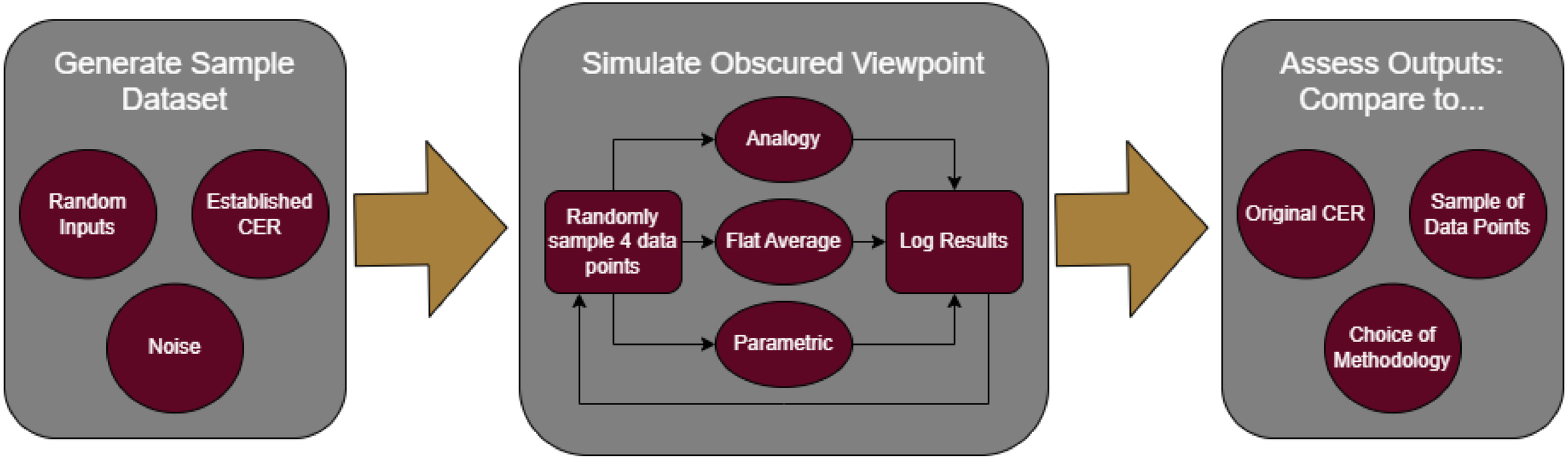


n = number of data points, k = number of independent variables

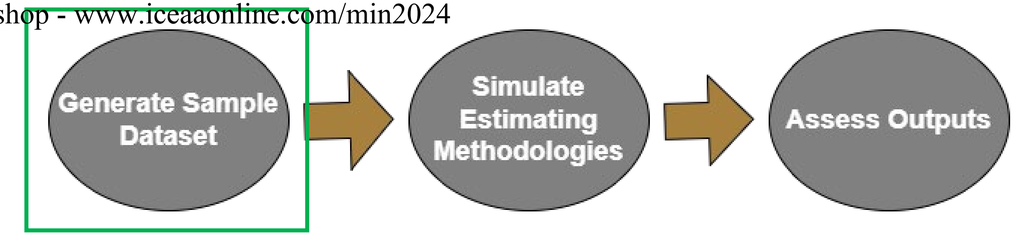
Simulating an Obscured Viewpoint

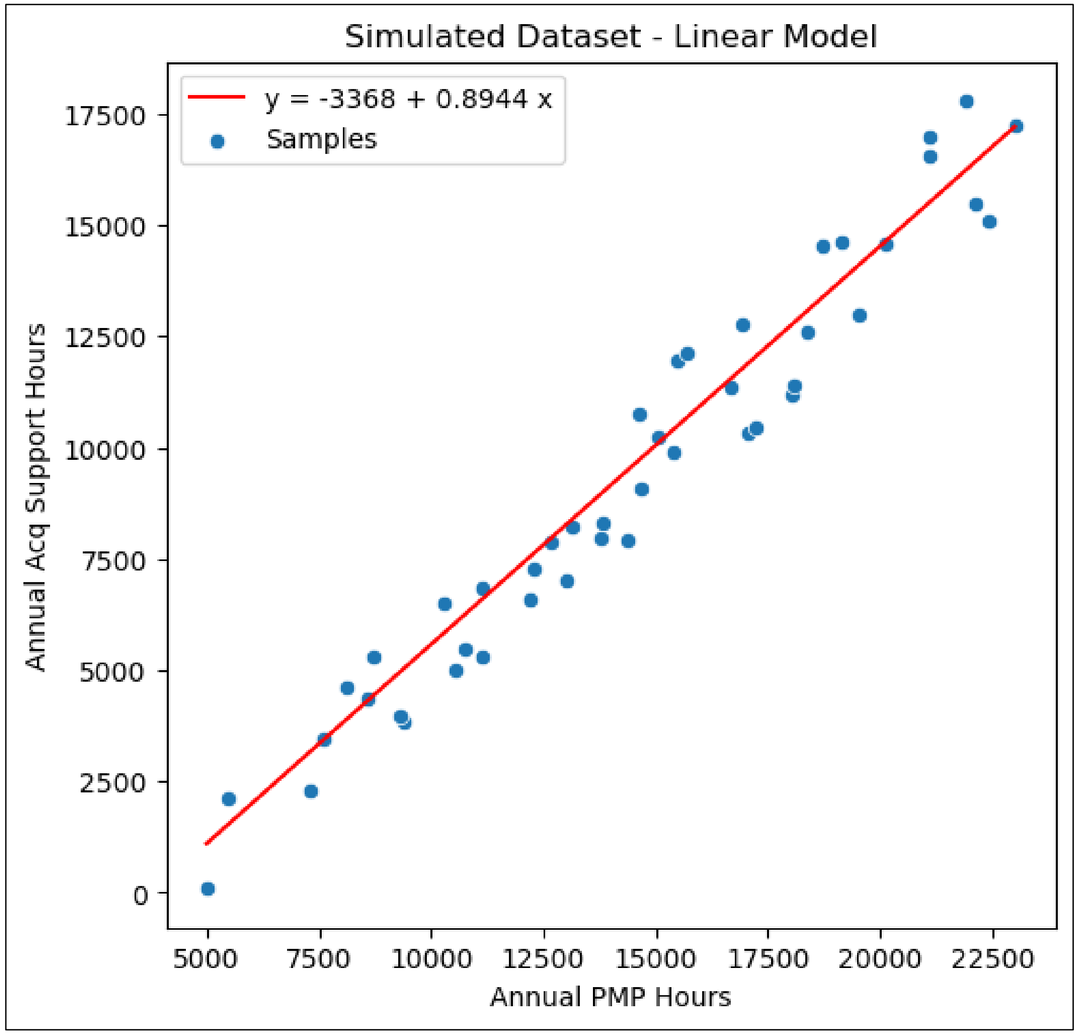
A Simulation-based Approach

- Analogous Programs > Data Points
- How do results vary depending on what data points are visible?



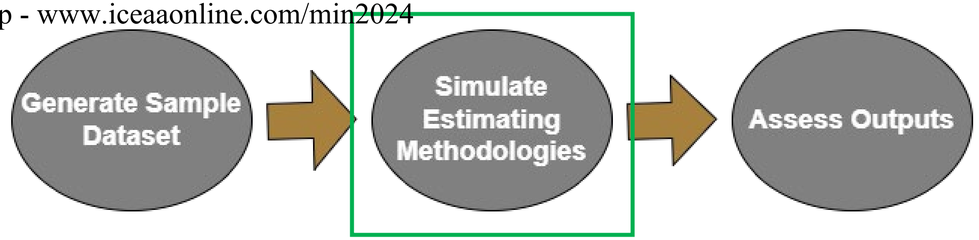
Simple Linear Model

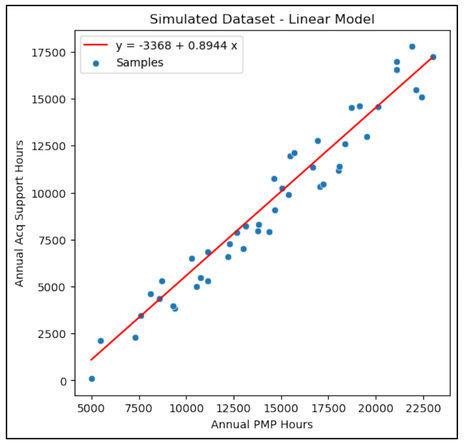




- Acquisition vs. PMP support hours for small USAF software development efforts
- Random sampling of inputs following normal distribution
- Generated expected value of CER
- Homoscedastic noise added to response to mimic natural variation

Analogy Methodology





Randomly select four data points

Random Data Points

- User Activity Monitoring (UAM) System 1 **8K PMP Hours**
- User Activity Monitoring (UAM) System 2 **1K PMP Hours**
- User Activity Monitoring (UAM) System 3 **19K PMP Hours**
- User Activity Monitoring (UAM) System 4 **5K PMP Hours**

"True" Project

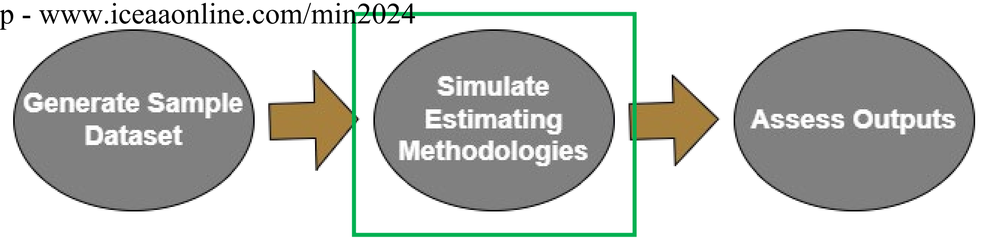
User Activity Monitoring (UAM) System **14K PMP Hours**

Represents one iteration of Monte Carlo simulation

Select data point closest to "True" value and repeat for 10,000 Iterations

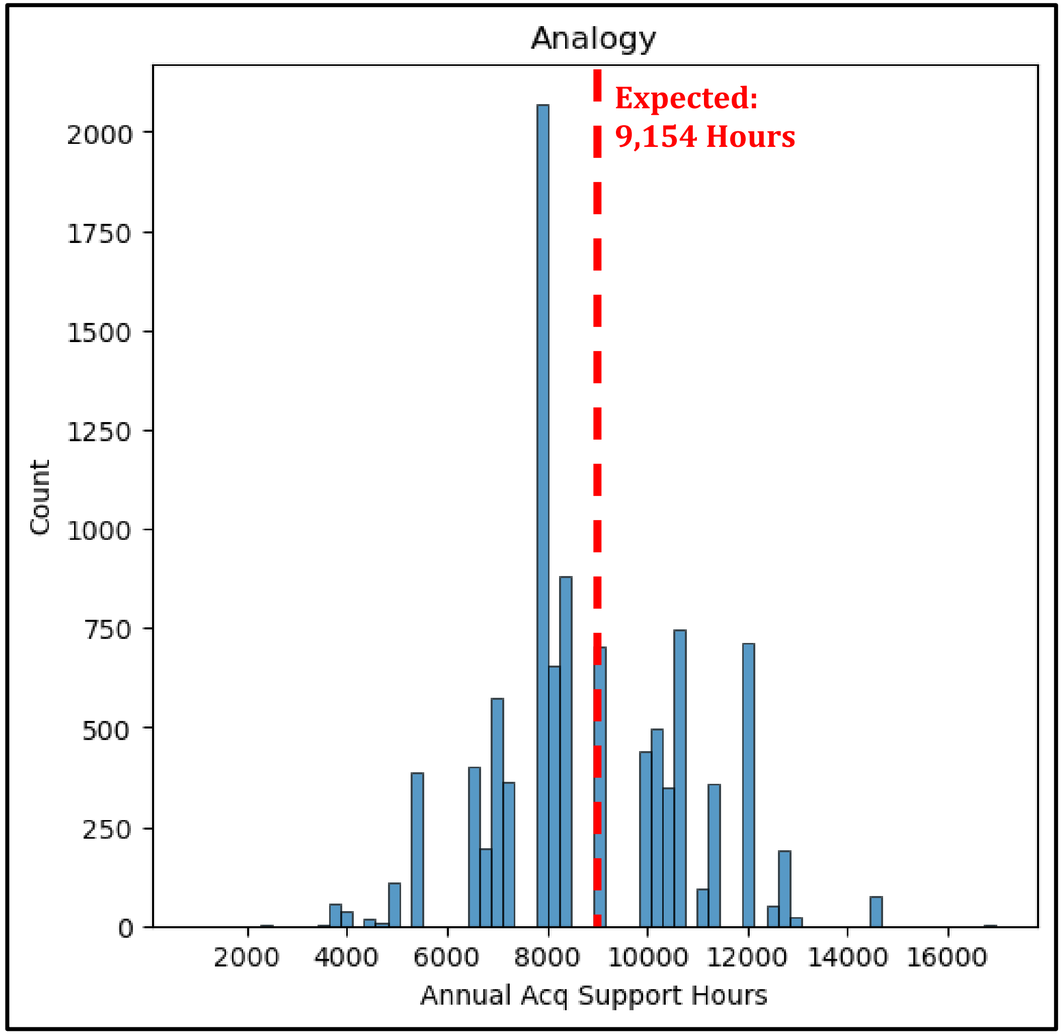
Generate probability distribution function & compare results against "true" value

Linear Model - Analogy

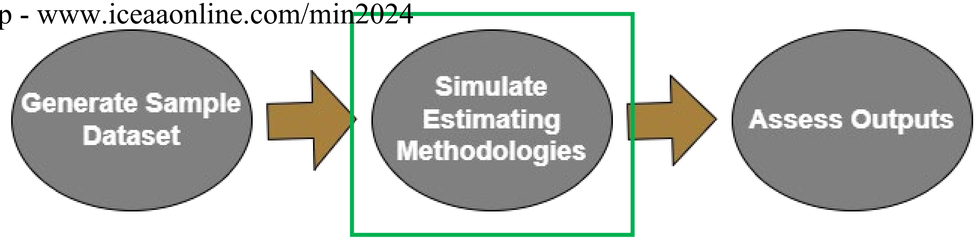


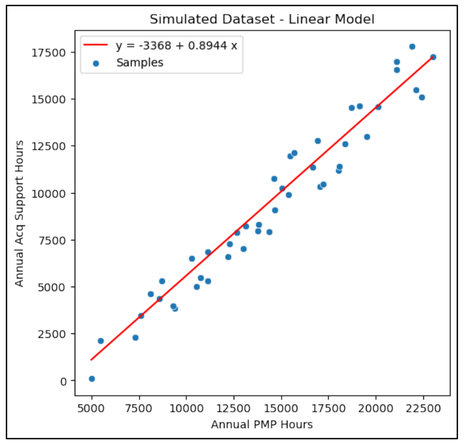
For observations (x,y) in sample, return y where x is closest to PMP hours of program being estimated

Inputs	
PMP Hours	14,000
Expected Acq. Support	9,154
Outputs	
Median Acq. Support	8,293
|Median - Expected|	861
Standard Deviation	1,974



Flat Average Methodology





Randomly select four data points

Random Data Points

- User Activity Monitoring (UAM) System 1 **8K Acq Hours**
- User Activity Monitoring (UAM) System 2 **1K Acq Hours**
- User Activity Monitoring (UAM) System 3 **19K Acq Hours**
- User Activity Monitoring (UAM) System 4 **5K Acq Hours**

Iteration Results

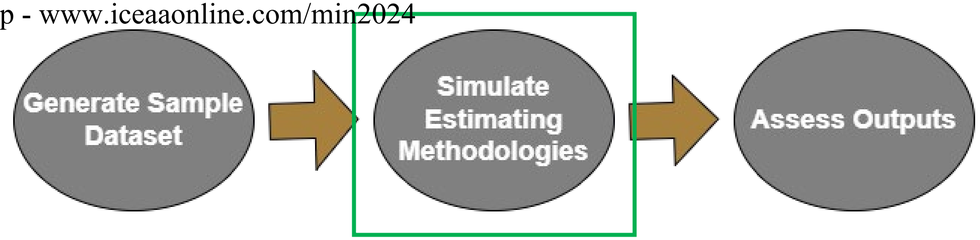
Average = 8.25K Acq Hours

Represents one iteration of Monte Carlo simulation

Average four data points and repeat for 10,000 Iterations

Generate probability distribution function & compare results against “true” value

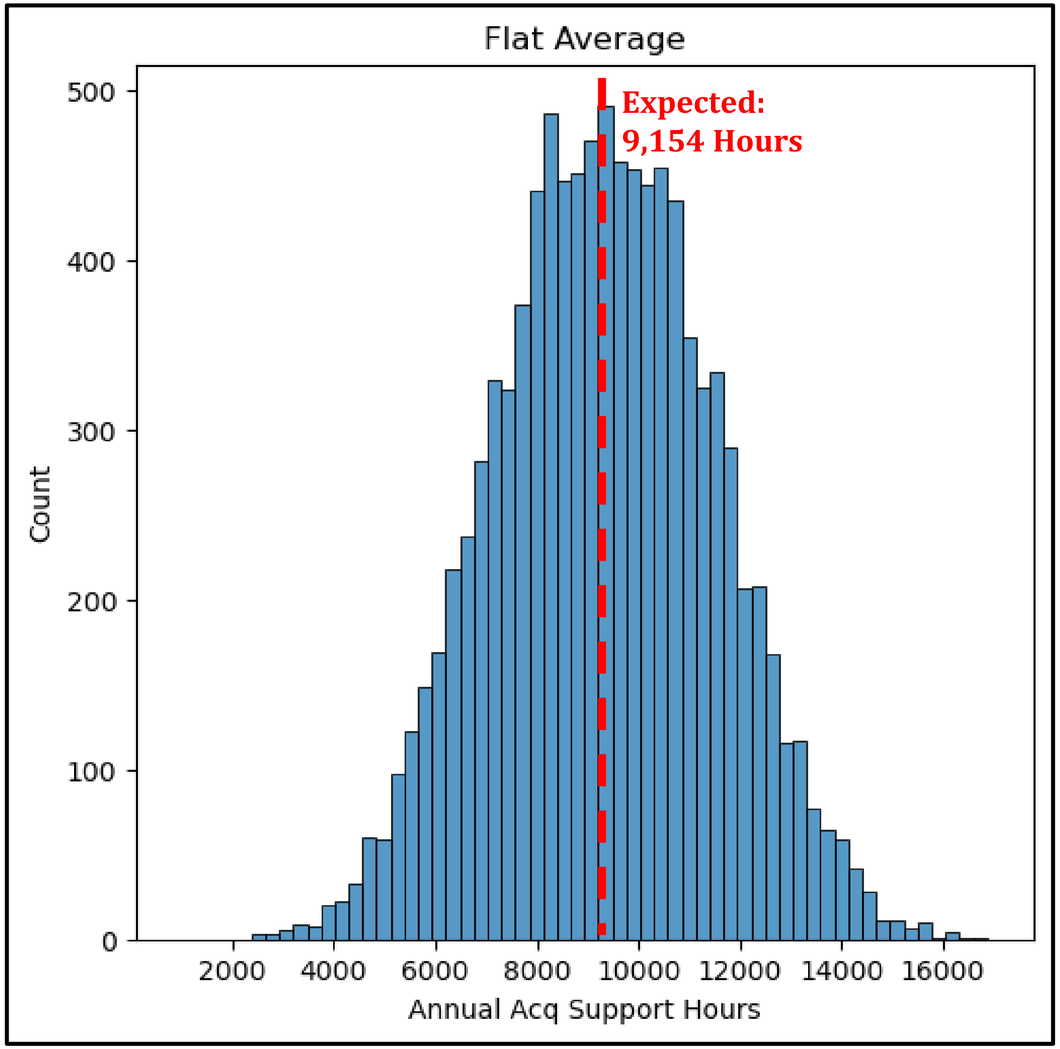
Linear Model – Flat Average



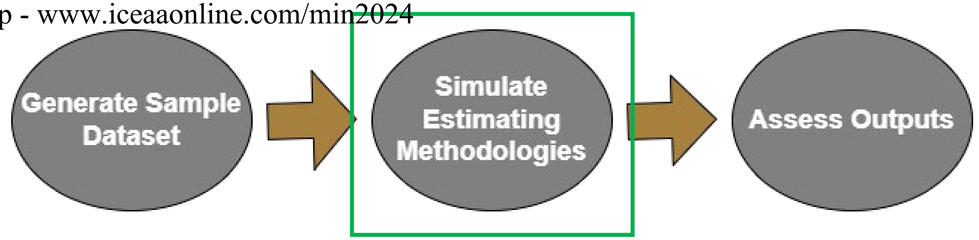
Average Acq. support for all observations in sample

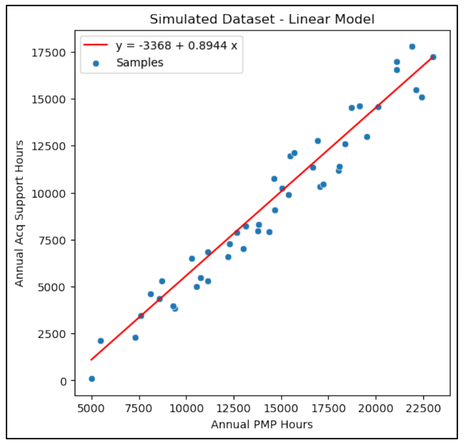
Assumes insufficient justification for weighting

Inputs	
PMP Hours	14,000
Expected Acq. Support	9,154
Outputs	
Median Acq. Support	9,309
|Median - Expected|	155
Standard Deviation	2,160



Parametric Methodology





Randomly select four data points

Random Data Points

- UAM System 1: 3K Acq Hours, 8K PMP Hours
- UAM System 2: 300 Acq Hours, 1K PMP Hours
- UAM System 3: 13K Acq Hours, 19K PMP Hours
- UAM System 4: 500 Acq Hours, 5K PMP Hours

Iteration Results

CER
Acq Support Hours = .75*PMP Hrs – 2040

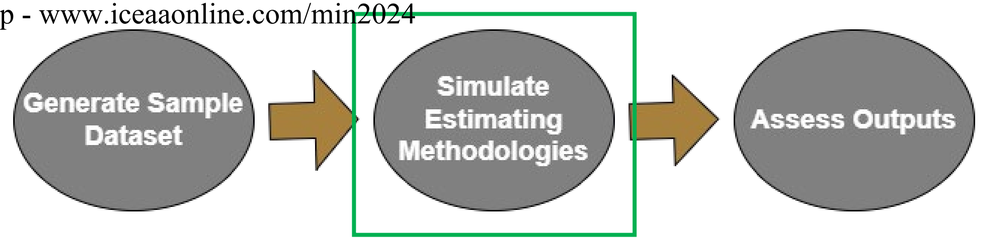
Output based on "True" project inputs
8.5K Acq Hours

Represents one iteration of Monte Carlo simulation

Predict acq. support hours based on CER generated from four randomly selected data points and repeat for 10,000 Iterations

Generate probability distribution function & compare results against "true" value

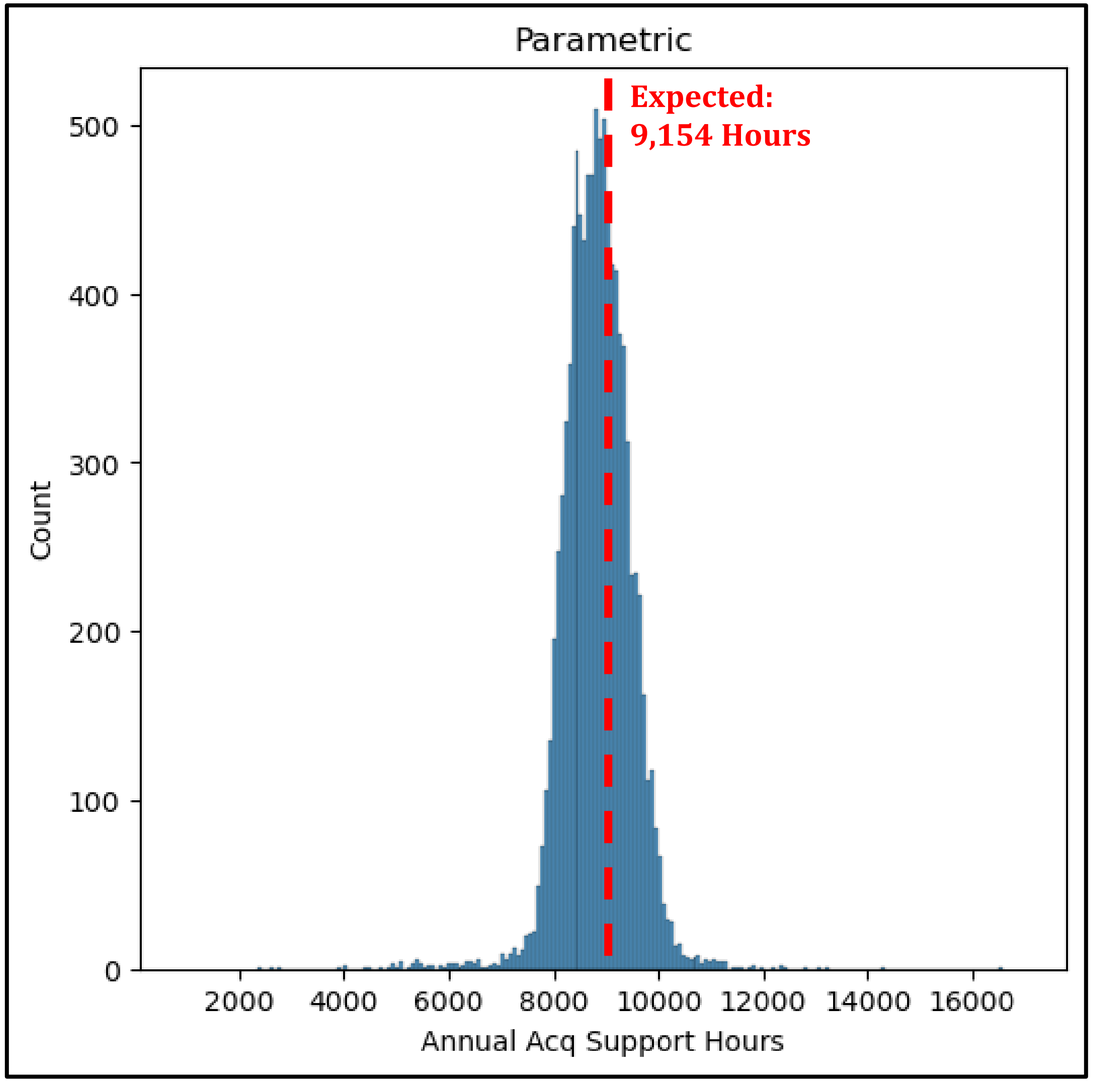
Linear Model – Parametric



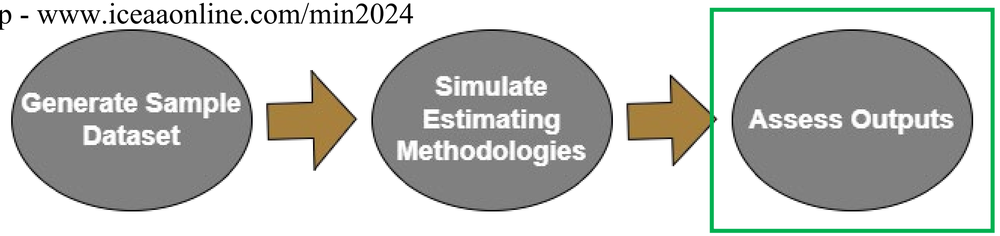
Results of OLS regression from all observations in sample

Assumes insufficient justification for WLS regression

Inputs	
PMP Hours	14,000
Expected Acq. Support	9,154
Outputs	
Median Acq. Support	8,831
|Median - Expected|	323
Standard Deviation	662



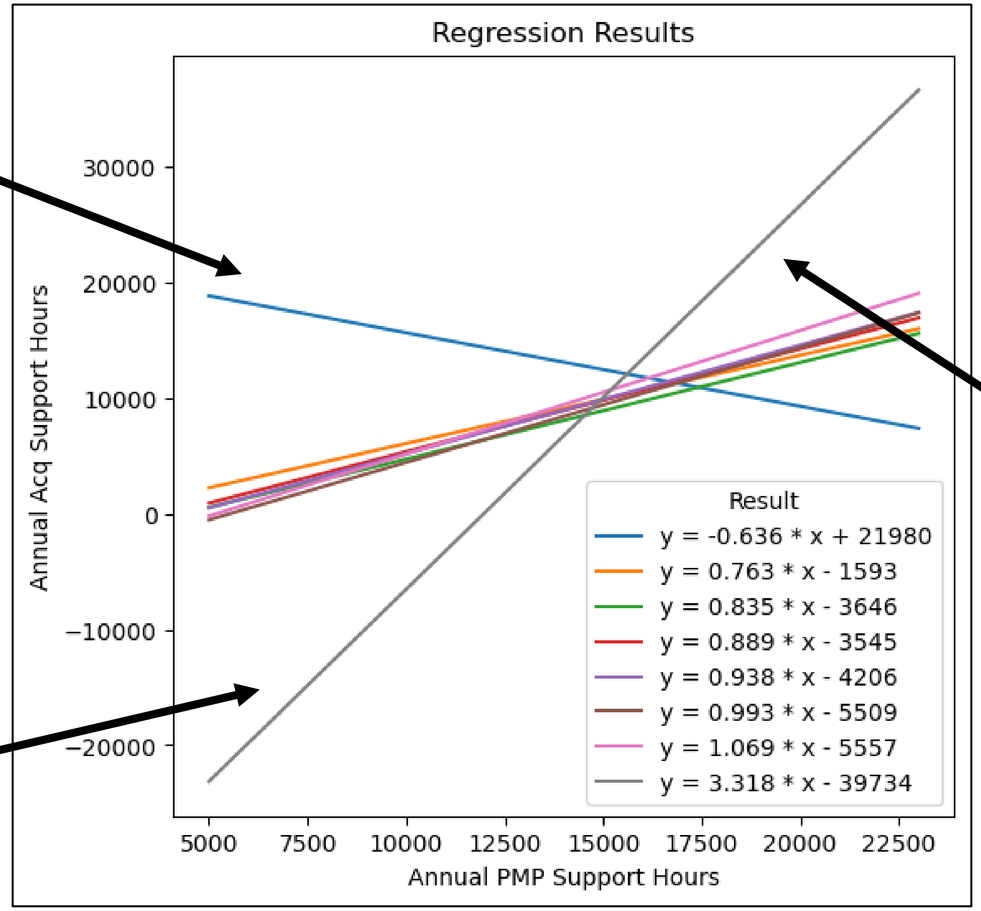
Linear Model - Results



Methodology	Observations	Benefits	Limitations
Analogy	Median << Expected Wide spread, discrete	Quality over quantity	Assumes 100% causal relationship between variable, response
Flat Average	Median ≈ Expected Wide spread, continuous	Assumes all observations are equally likely	
Parametric	Median ≈ Expected Low spread, continuous	Most quantitatively informed	Potential for unrealistic results

Leveraging Domain Knowledge

Less acquisition support for larger efforts?

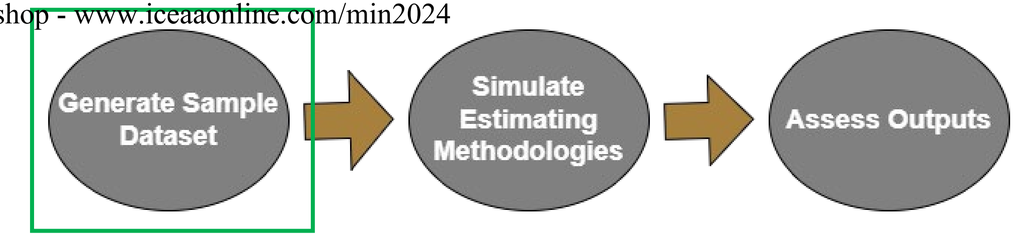


3 hours of Acquisition Support per hour of PMP?

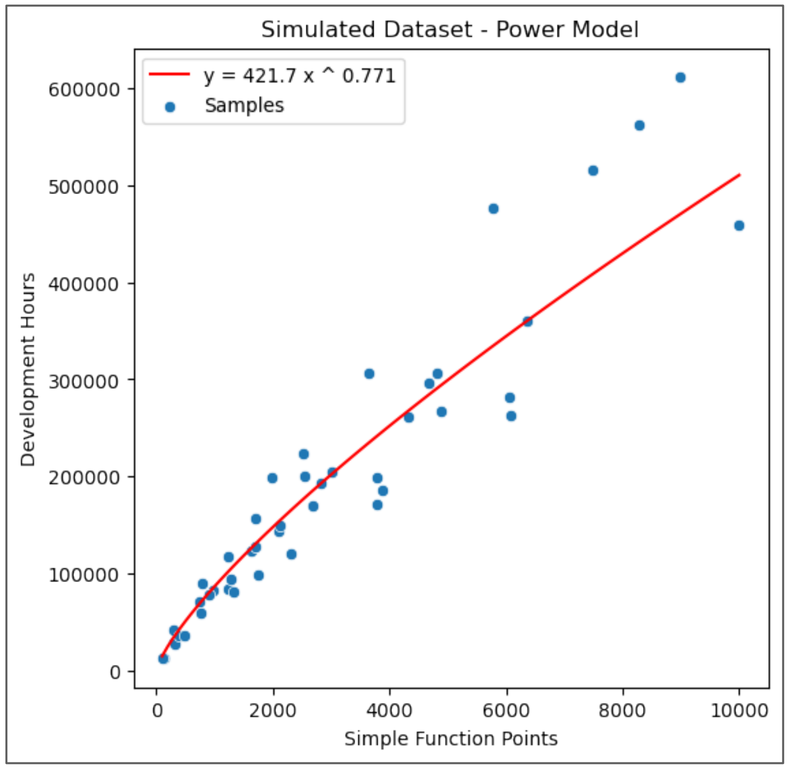
Negative values in the reference range?

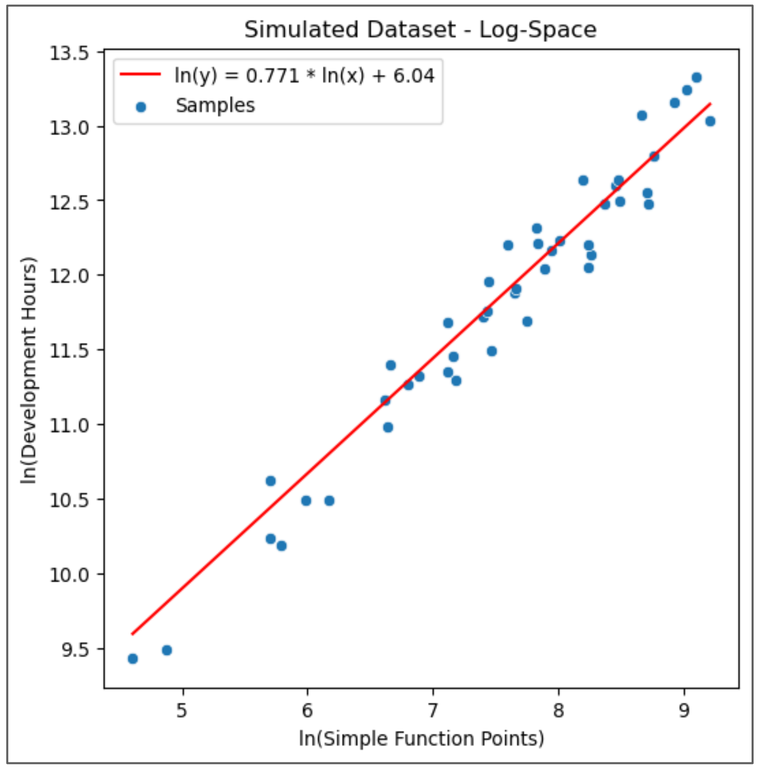
Data alone is not the answer!

More Complex Models

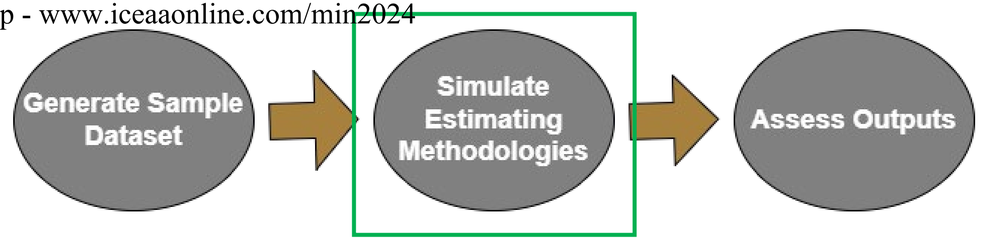


- Real world is more complex than simple linear relationships
- Development hours for full-scale software development efforts as a function of simple function points



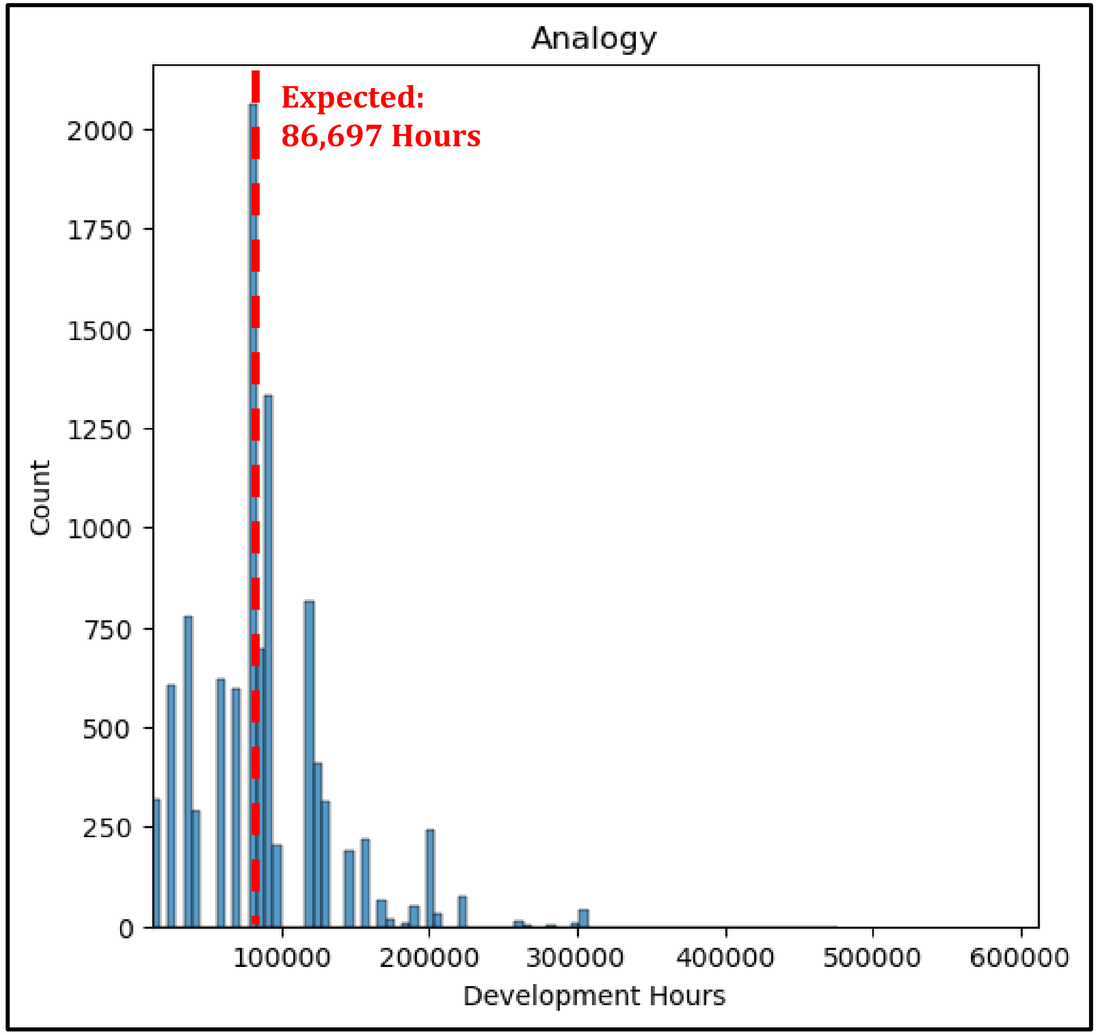


Power Model - Analogy

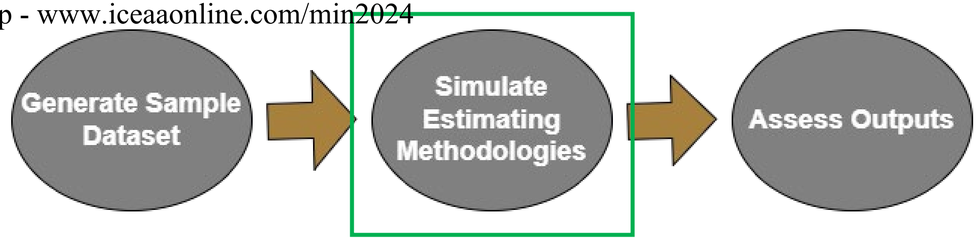


For observations (x,y) in sample, return y where x is closest to Development Hours of program being estimated

Inputs	
Simple Function Points	1,000
Expected Dev. Hours	86,697
Outputs	
Median Dev. Hours	82,938
|Median - Expected|	3,759
Standard Deviation	45,084



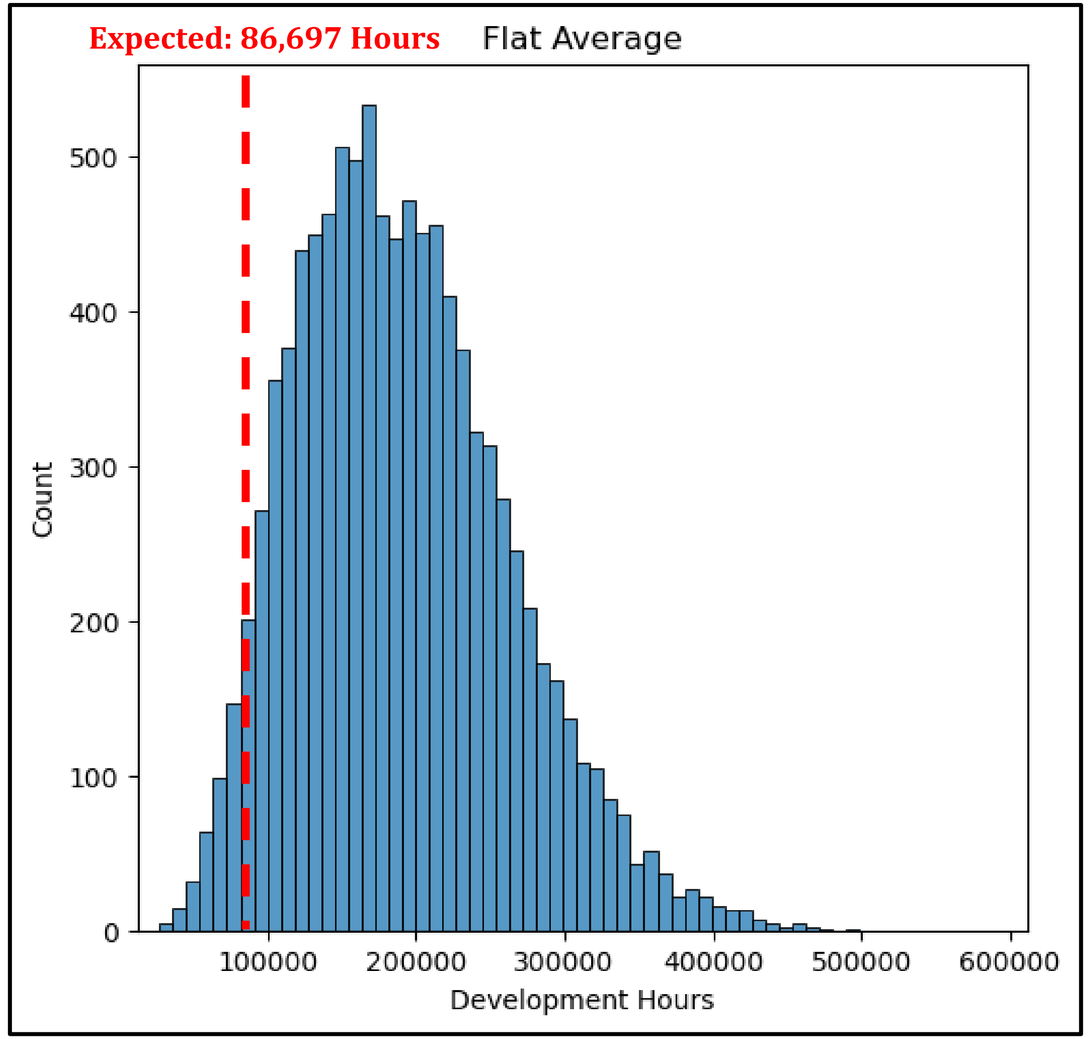
Power Model – Flat Average



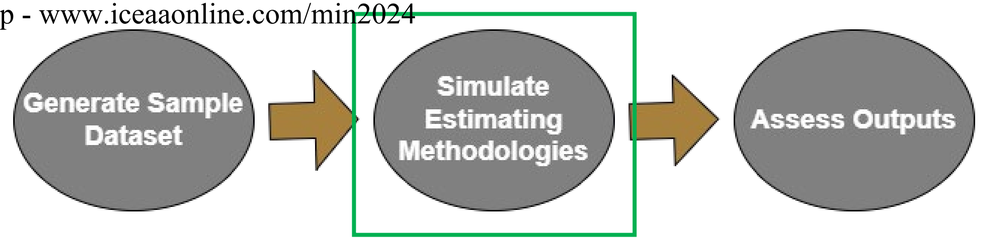
Average Dev. hours for all observations in sample

Assumes insufficient justification for weighting

Inputs	
Simple Function Points	1,000
Expected Dev. Hours	86,697
Outputs	
Median Dev. Hours	182,968
|Median - Expected|	96,271
Standard Deviation	71,187



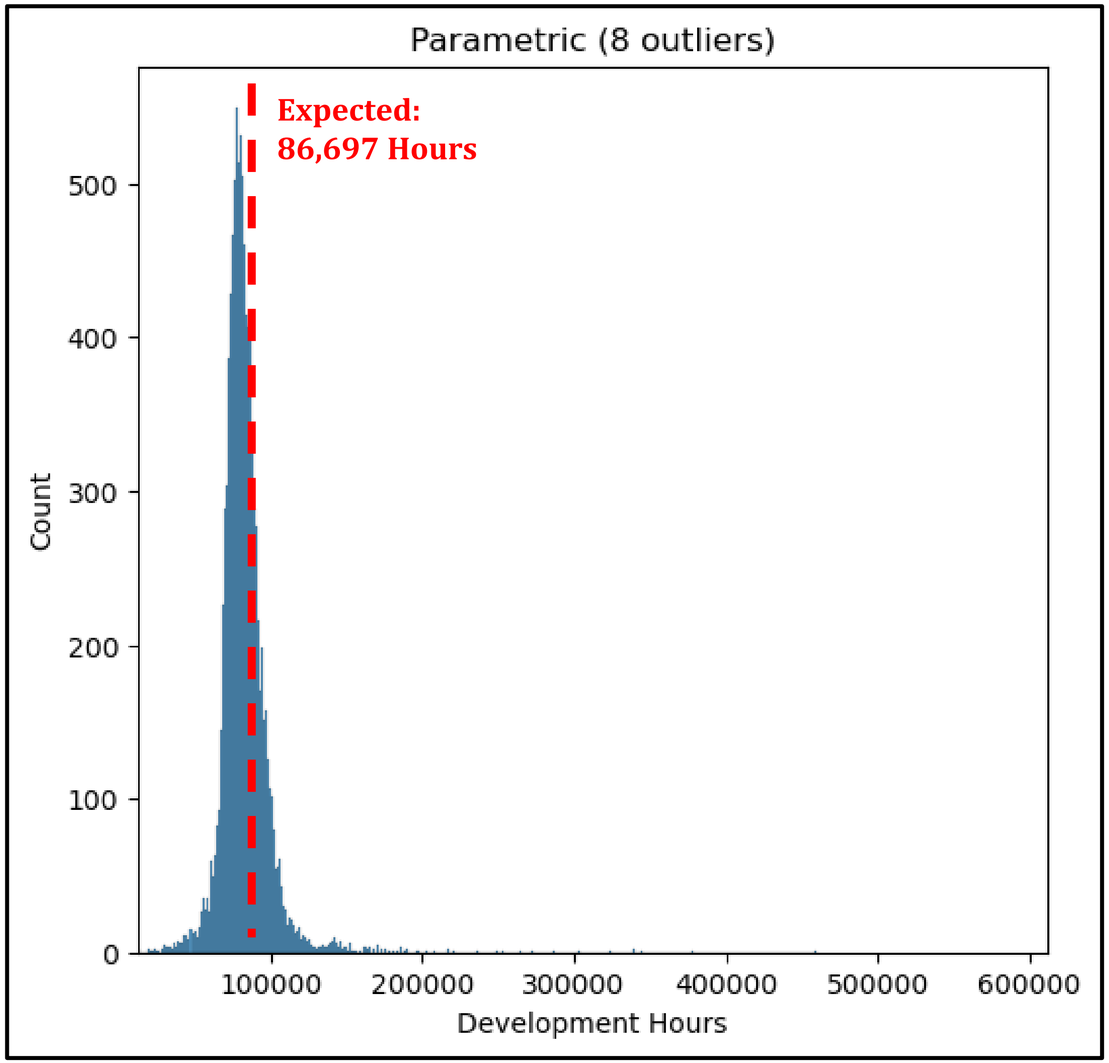
Power Model – Parametric



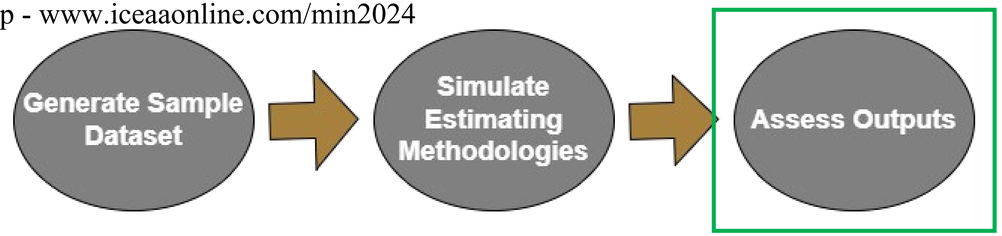
Results of OLS regression from all observations in sample

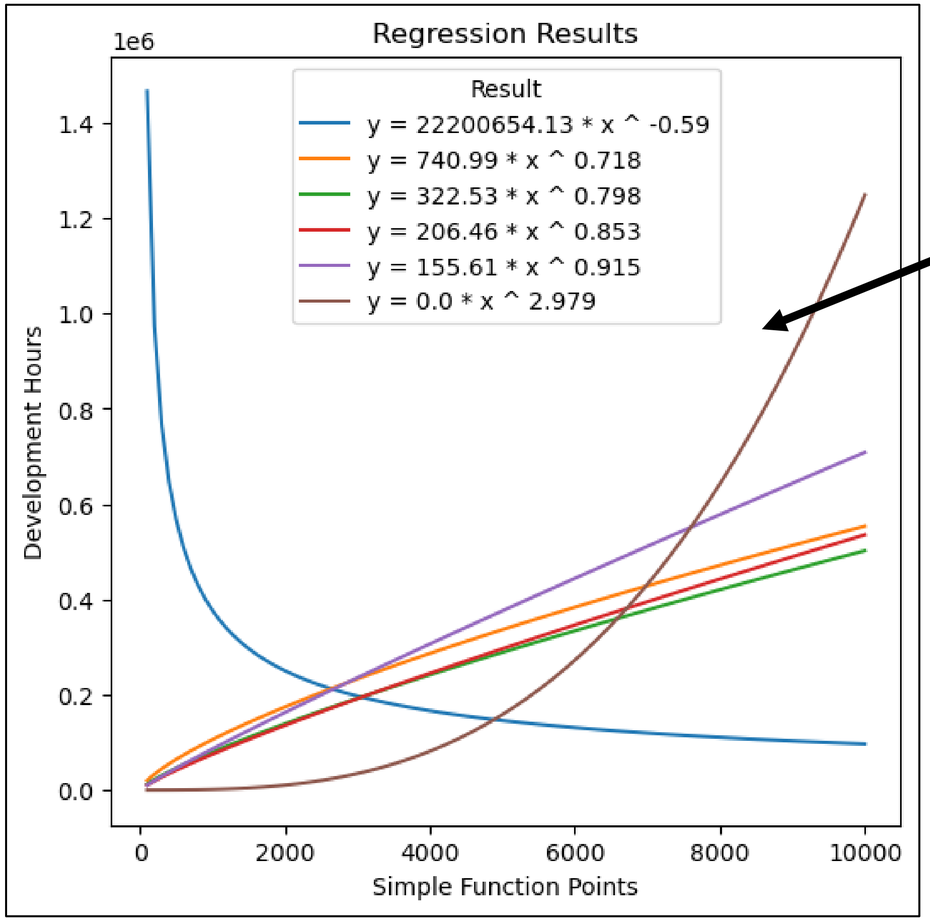
Assumes insufficient justification for WLS regression

Inputs	
Simple Function Points	1,000
Expected Dev. Hours	86,697
Outputs	
Median Dev. Hours	80,373
|Median - Expected|	6,324
Standard Deviation	17,084



False Prior Distributions

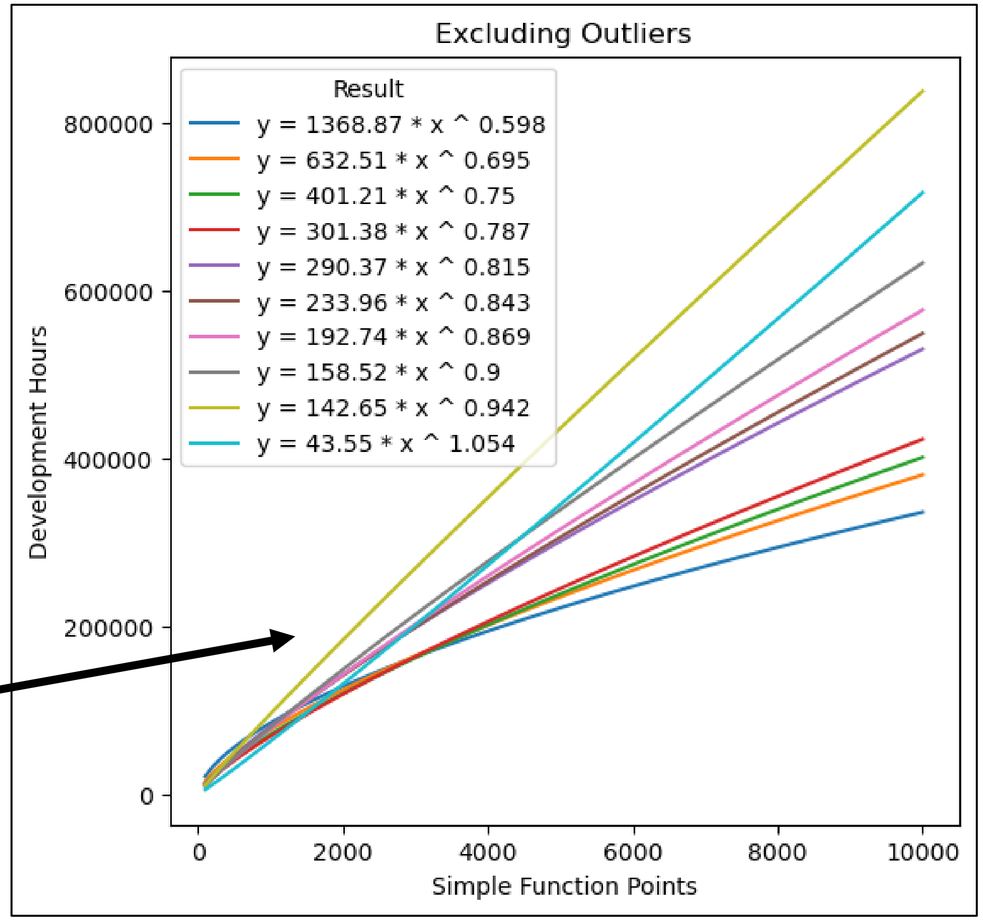




Contextually unrealistic

More complex model = more sensitive to inputs

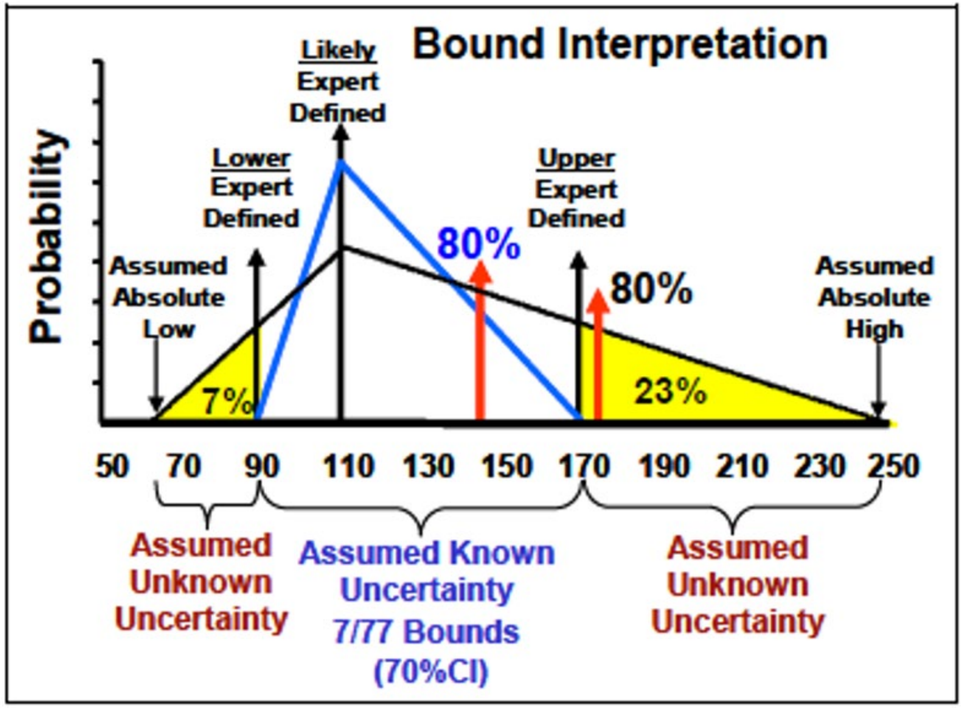
Potentially unreliable



Leveraging SME Input: Defensibility Without Data

Leveraging SME Inputs

- Generally viewed as a last resort estimating methodology
  - Informed by experience rather than data
- SME inputs typically applied using three-point estimate
- JACSRUH recommends adjusting uncertainty for skew



What is the effect of applying these rules of thumb?

Testing Rules of Thumb

1. Simulate different assumptions for triangular distributions in Python
2. Compare effects on median, coefficient of variation, shape of distribution
3. Normalize inputs: low of 0, high of 1, most likely < .5
   - Generalizable via linear transformations

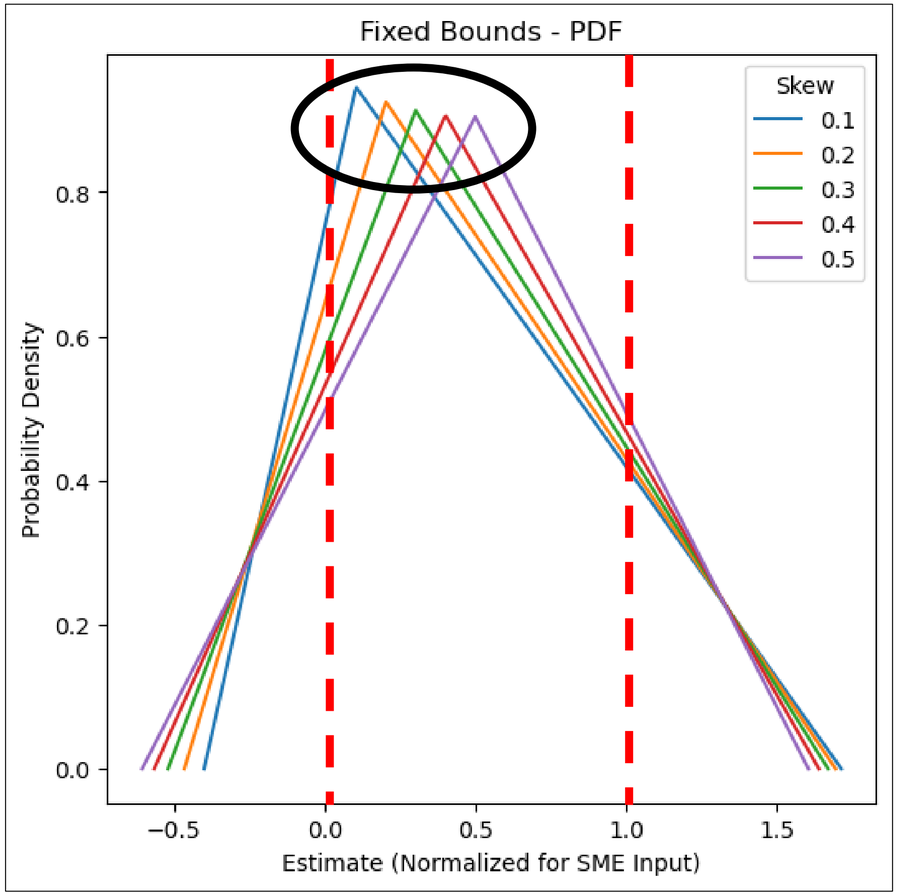
Attribute	Fixed Bounds	Weighted Bounds
Skew = $\frac{Most\ Likely - Low}{High - Low}$	Independent	Independent
CL of Low Estimate	15%	Skew * 30%
CL of High Estimate	85%	70% + Skew * 30%

Fixed Bounds

Overestimates skew for more extreme samples

Quantity of known + unknown uncertainty depends on skew

Sample Skew	Low %CL	High %CL	Median	CV	Skew of Results	Absolute Range
0.010000	0.15	0.85	0.401972	0.946169	0.233011	2.040047
0.014949	0.15	0.85	0.402133	0.947256	0.237305	2.045437
0.019899	0.15	0.85	0.402274	0.948227	0.241503	2.050298
0.024848	0.15	0.85	0.402506	0.949088	0.245587	2.055231
0.029798	0.15	0.85	0.402682	0.949852	0.249602	2.059516
...	...	...	...	...	...	...
0.480202	0.15	0.85	0.493878	0.905586	0.491094	2.210579
0.485152	0.15	0.85	0.495458	0.904863	0.493239	2.211139
0.490101	0.15	0.85	0.496974	0.904126	0.495477	2.211209
0.495051	0.15	0.85	0.498428	0.903375	0.497808	2.210791
0.500000	0.15	0.85	0.499987	0.902647	0.500017	2.210947



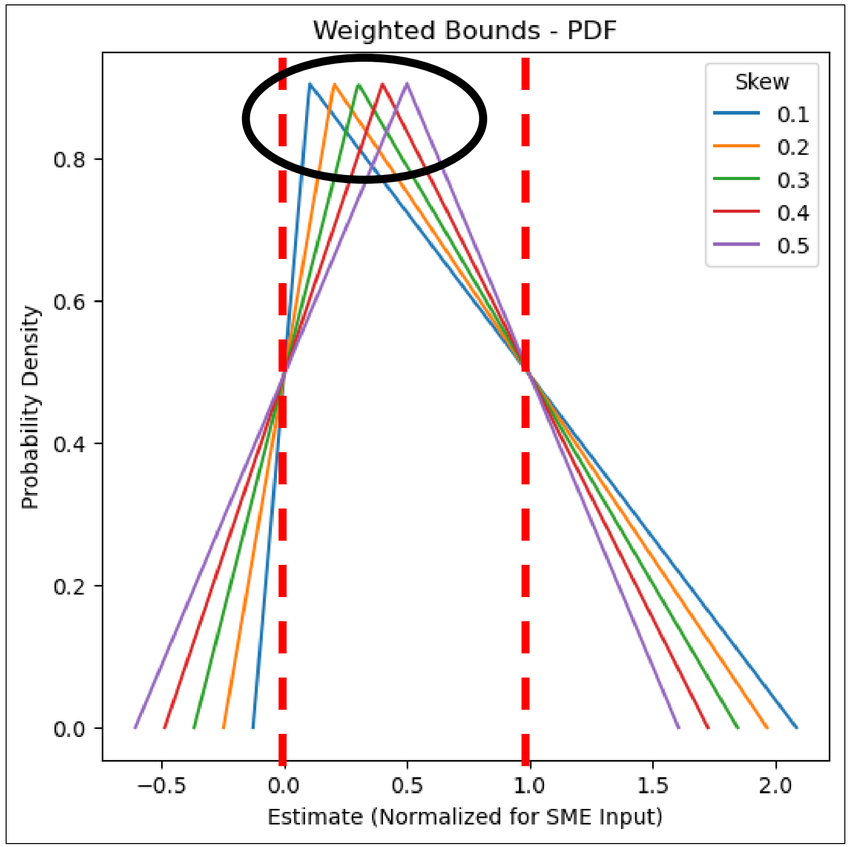
probability densities of the SME inputs vary with skew

Weighted Bounds

Skew of sample passes on to skew of results*

Quantity of known + unknown uncertainty is independent of skew*

Sample Skew	Low %CL	High %CL	Median	CV	Skew of Results	Absolute Range
0.010	0.003000	0.703000	0.642528	0.708149	0.010008	2.208296
0.015	0.004485	0.704485	0.641001	0.708694	0.014953	2.210266
0.020	0.005970	0.705970	0.639196	0.709255	0.019898	2.211256
0.025	0.007455	0.707455	0.637286	0.709834	0.024843	2.211852
0.030	0.008939	0.708939	0.635328	0.710431	0.029788	2.212250
...	...	...	...	...	...	...
0.480	0.144061	0.844061	0.502265	0.886244	0.480232	2.210811
0.485	0.145545	0.845545	0.501657	0.890244	0.485177	2.210846
0.490	0.147030	0.847030	0.501074	0.894311	0.490122	2.210880
0.495	0.148515	0.848515	0.500517	0.898445	0.495067	2.210914
0.500	0.150000	0.850000	0.499987	0.902647	0.500012	2.210947

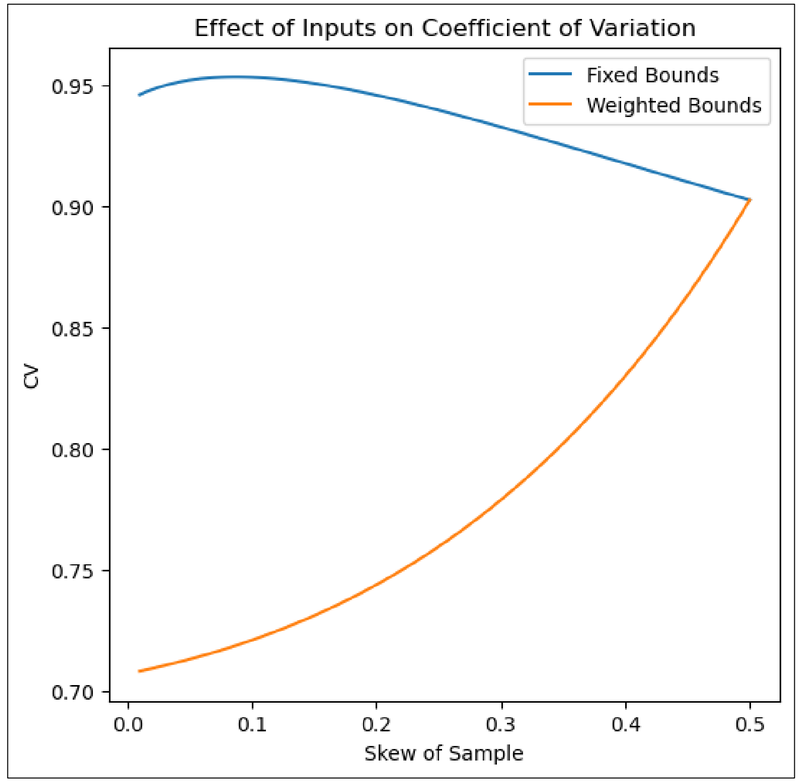


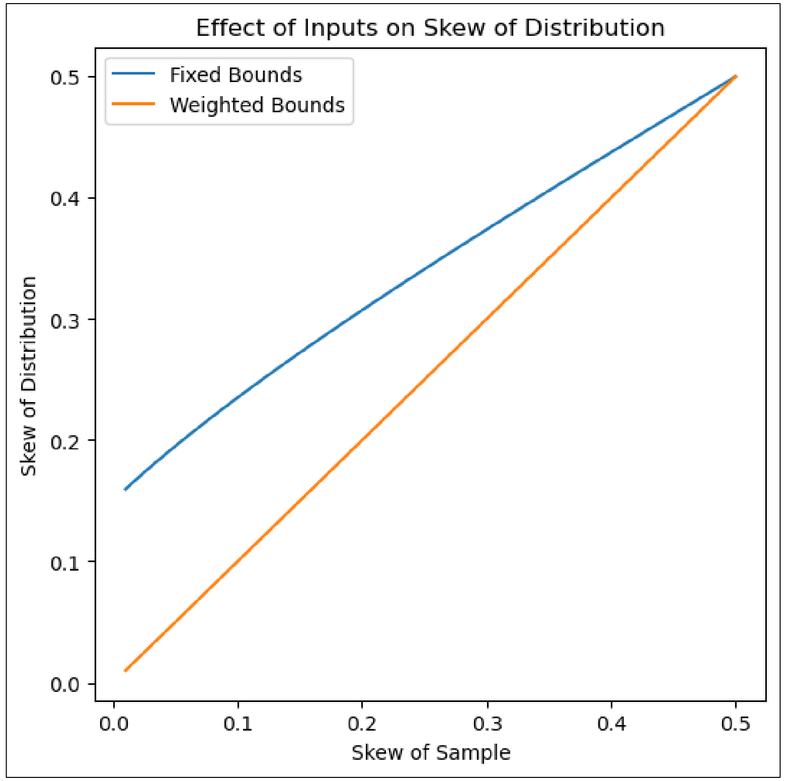
Fixes the probability densities of the SME inputs

* Variation due to floating point error, stochastic methods

Comparing the Results

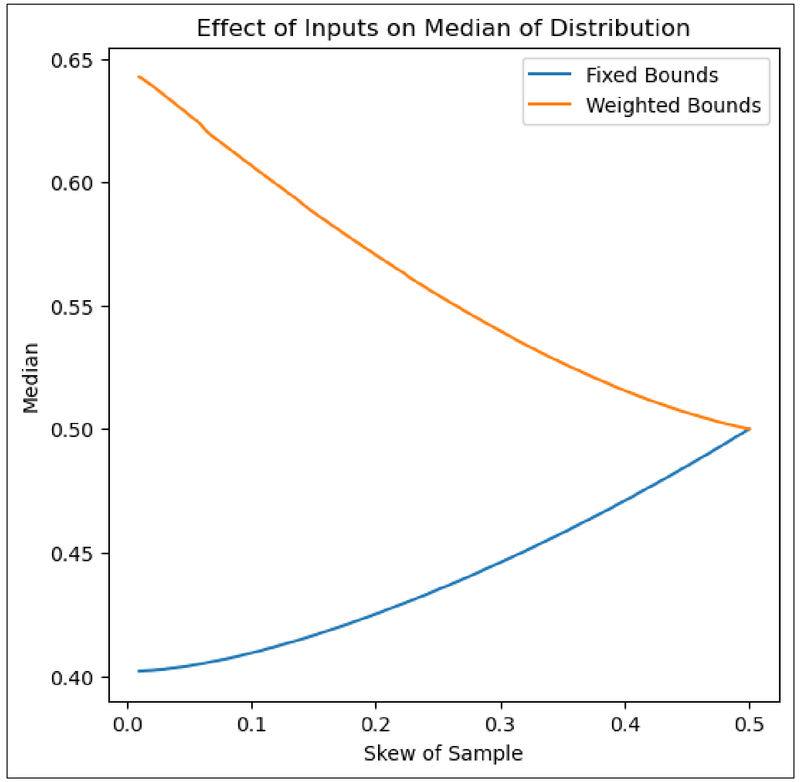
Weighting bounds creates a spread more informed by the inputs





Weighted bounds pass skew of sample on to skew of distribution

Fixed bounds underestimate risk of high estimate on the 50%CL



Conclusions

- Data scarcity is a constant limitation in government cost estimating
- Simulation based tests reinforce industry guidance for limited data

Number of Data Points	Estimating Method	Takeaways
$n = 1$	Analogy	High uncertainty, prioritize data quality
$1 < n < 5$	Average	All data points equally likely, useful early in acquisition life cycle
$n - k \leq 3$	Parametric	Most quantitatively informed, can suffer from false priors and spurious effects
No Data	SME Input	Not directly backed by data, must adjust for skew

- Qty of available data points dictates potential estimating methods
  - Must understanding the limitations of each to select the appropriate choice
- Assess results both practically and mathematically to ensure effective use of limited data

Backup

Bibliography

1. Aguirre, J. et al. (n.d.). The Progression of Regressions. In www.iceaaonline.com. 2022 International Cost Estimating and Analysis Association Professional Development and Training Workshop. https://www.iceaaonline.com/wp-content/uploads/2022/06/AM07-Aguirre-The-Progression-of-Regressions.pdf
2. Naval Center for Cost Analysis, (2018). Joint Agency Cost Estimating Relationship (CER) Development Handbook. https://www.asafm.army.mil/Portals/72/Documents/Offices/CE/CER%20Develop ement%20Handbook.pdf
3. Rosa, W. et al. (n.d.). Let's Go Agile! Data-Driven Agile Software Cost and Schedule Models. In www.iceaaonline.com. 2022 International Cost Estimating and Analysis Association Professional Development and Training Workshop. https://www.iceaaonline.com/wp-content/uploads/2022/06/SA09-Rosa-Lets-Go-Agile.pdf
4. Schiavoni, M., & Bearce, R. (n.d.). Assessing regression methods via Monte Carlo Simulations. In www.iceaaonline.com. 2021 International Cost Estimating and Analysis Association Online Workshop. https://www.iceaaonline.com/wp-content/uploads/2021/06/ANA02-ppt-Schiavoni-Assessing-Regression-Methods.pdf
5. Smart, C., & Jo, D. (n.d.). Using Bayes' Theorem to Develop CERs – Extending the Gaussian Model. In http://www.iceaaonline.com/. 2023 International Cost Estimating and Analysis Association Professional Development and Training Workshop. https://www.iceaaonline.com/wp-content/uploads/2023/06/AM06-Smart-Using-Bayes-Theorem-to-Develop-CERs-paper.pdf
6. U.S. Air Force, "Cost Risk and Uncertainty Analysis Handbook," U.S. Air Force, Hanscom Air Force Base, MA, 2007.
7. U.S. Government Accountability Office, GAO-20-195G Cost Estimating and Assessment Guide, Washington, DC: U.S. Government Accountability Office, 2020.
8. U.S. Government (2015). Joint Agency Cost Schedule Risk and Uncertainty Handbook (JA CSRUH). https://cade.osd.mil/Files/CADE/JA%20CSRUH%20Final%2012Mar2014%20With%20Signatures%2011May2015.pdf

Literature Review - References

Title	Author(s)	Summary	Year	Track	Search
Mining for Cost Estimating Relations from Limited Complex Data	Mark Jacobs, Shawn Hayes	NASA's robotic Earth and space science missions cover a diverse group of projects addressing complex science objectives that include challenging implementation approaches. Progress applying Principle Component Analysis techniques covering project management, systems engineering, mission assurance, integration & test, and spacecraft subsystems is described. Supporting data analysis efforts include a large detailed set of technical and programmatic input candidates that are analyzed to identify the primary spacecraft cost drivers.	2016	Government Processes	"Limited data"
Early Stage Cost Estimating for Radars and Sensors	Jeremy Goucher	The majority of costs for programs are locked in even before a program enters production, which makes accurate early stage cost estimates vital for effective resource management and program success. This study proposes a method for analyzing combat system cost prior to a complete requirements description. The data set includes old, new, big, small, ground, sea, air, domestic, and foreign systems. The result is a model that requires limited data and is widely applicable.	2016	Acquisition	"Limited data"
Developing a Dynamic Expense-Volume-Profit Model to Determine Break-Even point	William Kent	Analysts must be able to determine when a program will be profitable. This insight is used to drive contract negotiation, pricing, and strategy. Many factors need to be evaluated: expenses, revenue, and volume of sales. These factors must be determined based on limited historical data and extrapolated over a time horizon. This paper introduces a framework for analyzing these factors using learning curves, regression, and financial statement analysis to determine break-even.	2016	Business Case Analysis	"Limited data"
Right Sizing Earned Value Management for Your Project	Gordon Kranz	Earned Value Management (EVM) is a program management tool that provides data indicators that can be used on all programs to enable proactive decision making throughout the program lifecycle and facilitate communication across the program team. Each program has unique attributes that should be considered when determining program management and reporting requirements, including, but not limited to, contract size and type, scope of work, complexity, risk, technology maturity, and resource requirements. ...	2014	Cost Management	"Limited data"
Estimating Cloud Computing Costs: Practical Questions for Programs	Kathryn Connor	Cloud computing has garnered the attention of the Department of Defense (DoD) as data and computer processing needs grow and budgets shrink. In the meantime, reliable literature on the costs of cloud computing in the government is still limited, but programs are interested in any solution that has potential to control growing data management costs. ...	2014	Information Technology	"Limited data"
Bayesian Parametrics: Developing a CER with Limited Data and Even Without Data	Christian Smart	This paper discusses Bayes' Theorem, and applies it to linear and nonlinear CERs, including ordinary least squares and log-transformed ordinary least squares.	2014	Parametrics	"Limited data"
NASA's Phasing Estimating Relationships	Chad Krause, Erik Burgess, Darren Elliott	Cost and schedule estimating in support of budget formulation is limited when cost phasing is not considered. As a result, NASA'a Office of Evaluation (OE) Cost Analysis Division (CAD) initiated a review of historic mission funding profiles for the purpose of corroborating current phasing profiles and optimizing future budgeting performance. Actual expenditures by year, technical parameters, and programmatic information were compiled and normalized from NASA's extensive library of CADRe (Cost Analysis Data Requirment) documents for programs since 1990. ...	2014	Space	"Limited data"
The Unseen: Statistical Inference with Limited Data	Trevor VanAtta	Objective measurements of probability are often unavailable, and most significant choices under risk require an intuitive evaluation of probability.' -Nobel Laureates Daniel Kahneman and Amos Tversky. What are the odds of rolling a sum total of seven when tossing two dice? What is the probability of red turning up after a spin of a European roulette wheel? Most analysts, given a little time and a calculator, could answer these two questions with exact precision. For both of these questions, there is only one true correct answer. Such is the nature of probability analysis for questions that are decompositional (all possible outcomes can be determined), frequentistic (the experiment can be repeated an infinite number of times), and algorithmic (the results can be measured with numbers)...	2012	Risk	"Limited data"
Fitting Absolute Distributions to Limited Data	Blake Boswell	The choice of probability distributions is a critical component for cost risk and uncertainty modeling. When data is available, distribution fitting techniques, such as Goodness of Fit (GoF) tests and Information Criteria (IC), can be applied to determine distributions that accurately describe potential cost realizations; however, with limited data GoF tests and IC based methods provide little or no insight into the best distribution choice. Therefore, when data is limited it is standard practice in cost risk and uncertainty modeling to solicit expert opinion in the construction of triangular distributions with vertices representing the best case, typical, and worst case scenarios...	2012	Risk	"Limited data"

Literature Review - References

Title	Author(s)	Summary	Year	Track	Search
Use of Weibull Failure Rates	Virginia Stouffer	Often one of the largest costs of an operational phase, estimating reparable costs remains more an art than science. A number of different probability distributions are commonly assumed to underlie failure rates for predicting future reparable requirements. In the case where a limited amount of data is available on demonstrated failure rates, the Weibull can be a versatile and useful estimating tool. This paper demonstrates how observed failure rate data from a time-limited test can be adapted to a Weibull distribution for predicting future failure rates. ...	2009	Risk	"Limited data"
Using Bottoms-Up Cost Estimating Relationships in a Parametric Cost Estimation System	Dale Masel, Robert Judd	The parametric approach to manufacturing cost estimation is preferable because of the limited amount of part data required to generate an estimate. However, the limited data means that the estimate cannot achieve the accuracy of a bottoms-up estimating system. A hybrid cost-estimating system has been developed and implemented to estimate the cost of jet engine components. The system utilizes a bottoms-up estimating approach for the actual cost estimates, but appears as a parametric system to the user. The inputs needed for bottoms-up cost- estimating relationships (CERs) are calculated from user-specified parameters based on relationships derived between part parameters and feature attributes.	2007	Models	"Limited data"
Developing Risk Adjusted Results with Limited Data	Gabriela Rohlk	This paper expands on work by Dr. Stephen Book, MCR by addressing additional distributions in performing quick-risk analyses. We provide an overview of the purpose and methodology and then develop the equations for implementation. Given a simple point estimate, this methodology, when combined with some assumptions, allows the decision maker to estimate the high confidence cost/benefits. The methodology is developed for three different assumed statistical distributions (Triangular, Normal & Log-Normal). Additionally a high-level discussion of cost/benefits risk analysis will be covered.	2007	Risk	"Limited data"
Beyond Regression: Applying Machine Learning to Parametrics	Kimberly Roye, Christian Smart	Cost estimating has relied primarily upon regression analysis for parametric estimating. However, regression analysis is only one of many tools in data science and machine learning, and is a small subset of supervised machine learning methods. In this paper, we look at a variety of methods for predictive analysis for cost estimating, including other supervised methods such as neural networks, deep learning, and regression trees, as well as unsupervised methods and reinforcement learning.	2019	Machine Learning	"Small data"
Agile Software Development Cost Factors: A Case Study	Blaze Smallwood	The lack of data on government agile software development programs has made estimating costs for new agile development programs challenging. This paper seeks to address this challenge through a case study of several completed DoD agile projects with cost, schedule, and performance data. Several relevant metrics will be examined, including cost per story point, cost per requirement, scope growth rates, impacts of team size changes on velocity and productivity, and various others.	2018	Agile	"Small data"
An Empirical and Visual Tale of a Cross-Country Bicycle Adventure	Rick Collins, Maggie Dozier, Orly S. Olbum, Paul Lanier Hardin III	The 'road' from Anacortes, Washington to Bar Harbor, Maine is paved with amazing landscapes, small towns, interesting people and cold beer. This 74 day cross-country cycling trip was an ideal opportunity to collect data that might explain daily riding speed. This presentation (and companion paper) paper describes the journey and post-ride analysis of the data using influence diagrams and constrained optimization (via Excel Solver) and will hopefully inspire others to get on a bike and experience the 'power' of cycling.	2018	Comprehensive Perspectives	"Small data"
Learning Rate Sensitivity Model	Timothy P. Anderson, Nichols F. Brown	In space cost estimates, learning curves are used to estimate cost of small quantity acquisitions. Recently, spacecraft providers have started proposing unprecedentedly large constellations. The authors have developed a methodology to test assumptions about learning rates versus proposed cost estimates, providing a data-driven assessment of whether a proposed learning rate/cost combination is feasible or even likely, and further describes the learning rate would be necessary in order to meet a proposed cost estimate.	2018	Methods & Models	"Small data"
Expanding the Range of Your Data: A Small Ships Case Study	Kathleen Hudgins, Robert Nehring, Elizabeth Koza, Anna Irvine	While the Navy has comprehensive Operating and Support (O&S) data for current Navy ships, there are a limited number of smaller boats available for inclusion. With increasing technological advances and a renewed effort to reduce personnel, smaller boats are receiving increased interest. This paper explores using Coast Guard O&S data to supplement Navy data for use in estimating O&S costs for smaller boats. Topics include data sources, normalization, and comparisons of data between the services.	2017	Operating & Support	"Small data"
Seven Degrees of Separation: The Importance of High-Quality Contractor Data in Cost Estimating	Crickett Petty	The popular notion that any two people are linked to one another on average via a chain with "six degrees of separation" is based on a relatively small sample. Using massive data sets, researchers have since discovered that the average degrees of separation is closer to seven. This highlights the need for complete and thorough data in any analysis, and cost estimating in particular. Despite its importance the collection of high-quality contractor cost data is challenging. Processes for meeting these challenges are presented.	2016	Methods & Models	"Small data"

Literature Review - References

Title	Author(s)	Summary	Year	Track	Search
Early Stage Cost Estimating for Radars and Sensors	Jeremy Goucher	The majority of costs for programs are locked in even before a program enters production, which makes accurate early stage cost estimates vital for effective resource management and program success. This study proposes a method for analyzing combat system cost prior to a complete requirements description. The data set includes old, new, big, small, ground, sea, air, domestic, and foreign systems. The result is a model that requires limited data and is widely applicable.	2016	Acquisition	"Small data"
Data Collection for Agile Projects	Blaze Smallwood	Agile software development projects produce a unique set of metrics, such as points and velocity, that can provide interesting insight into project progress. However, collecting this data requires specialized mechanisms, since no established standards exist, like SRDRs. This paper will discuss the types of data that are useful to collect for agile software development projects and mechanisms that have been used to collect them for several government projects.	2016	Software & IT	"Small data"
Should You Care About How Good An Estimation Process?	Alain Abran	This presentation is based on the author's 2015 book on Software Project Estimation (Wiley & IEEE Press) and illustrates how organizations (including in the car industry) having collected their own data using international standards have built their own estimation models and developed a key competitive advantage through improved software estimation capabilities. Examples will include data from a large European car manufacturer and a small software organization using Agile.	2016 Int'l	Risk Analysis	"Small data"
Estimating Software Development Costs for Agile Projects	Blaze Smallwood, Omar Mahmoud	Addressing the need to more rapidly develop and field capabilities for the warfighter, more and more software-centric DoD programs are transitioning towards an industry trend called "Agile" software development. While "Agile" software development is geared towards producing usable software products more rapidly than traditional waterfall or incremental methods, it also requires more flexibility with managing requirements. The main challenge this has created for program managers is figuring out how to effectively manage scope, cost, schedule, and performance in this flexible, fast-paced development environment in which requirements are more fluid. In turn, cost estimators have been challenged to develop new data collection approaches and estimating methodologies to more effectively estimate software costs for these "Agile" programs...	2013	Information Technology	"Small data"
Learning Curve Analysis of Small Data Sets - Spacecraft Bus Cost Improvement Analysis	Brian Welsh, James York	Recent initiatives to implement cost savings across DoD have produced a number of potential opportunities. Within the Space Portfolio one such opportunity is to acquire satellites more efficiently through block buys and limiting production breaks. To quantify these opportunities, we can analyze historical cost experience. This paper describes three approaches used to pool cost improvement data from multiple programs, allowing for regression of all data simultaneously. The approaches are (1) scaling based on weight, (2) normalizing to a common theoretical first unit, and (3) the use of program binary ("dummy") variables...	2013	Methods and Models I	"Small data"
Improving Cost Estimates in a Medical Acquisition Environment	Mark Russo, Kimberly Wallace	Estimating costs for drug development has proven difficult over the years. Difficulties stem from the overlapping of the Department of Defense (DoD) acquisition process and the Food and Drug Administration's (FDA's) review process, small data pool, industry close hold of cost data, the inherent difficulties in the drug development process, and the lack of using best practices in cost estimating. Over the past couple of years the medical cost community has been working towards improving their cost estimates. These efforts included the establishment of a standard drug development work breakdown structure to ensure that data could be collected and compared in a standard way. A medical cost model was then created in order to use the available data to develop program estimates...	2013	Methods and Models II	"Small data"
"The Answer is 5": Observations on Cost & Schedule in Small Defense Programs	Brian Fersch, Wesley Tate, Colleen Leonard	Cost & schedule overruns have been an enduring issue in government acquisitions. Our largest programs are typically dissected and their flaws illuminated with 20/20 hindsight. DoDs smaller programs, though comprising the majority of the budget and quantity of programs, do not get this same attention. Over the course of the past year this study team has collected historical & current cost and schedule data on numerous programs that has been collected at similar points in each programs acquisition lifecycle. The end result is a database that allowed the team to examine how cost estimates and schedule expectations changed from year to year for the same events...	2012	Earned Value Management Track	"Small data"

Literature Review - References

Title	Author(s)	Summary	Year	Track	Search
Enhancing Excel-Based Cost Models with PivotTable Reporting	Blaze Smallwood, Omar Mahmoud	A major challenge facing DoD cost estimators is developing Program Life Cycle Cost Estimates (PLCCEs) that serve as more than just an acquisition reporting "check-in-the-box" document. The ideal is to simultaneously create a dynamic, flexible tool that the program office can use to support their day-to-day costing needs, from answering a wide-range of cost-related data calls to integrating PLCCE outputs into their Program, Planning, Budgeting, and Execution (PPBE) processes. For DoD cost estimators that utilize Microsoft (MS) Excel to create their PLCCE models, a valuable tool that can be leveraged to achieve these goals is Excel's PivotTable reporting capability...	2012	Methods and Models I	"Small data"
Review of Three Small-Satellite Cost Models	Melvin Broder, Eric Mahr, Daniel Barkmeyer, Erik Burgess, Wilmer Alvarado, Samuel Toas, Gregory Hogan	The problem of estimating costs for small satellites is more vexing that it would seem. Small satellites, as described here, generally weigh less than 1,000 pounds, and are sometimes much smaller and significantly different from much of what exists now. Costing these kinds of satellites is not simply a matter of scaling down from larger systems. It requires incorporation of new data sets and consideration of new modeling methods. The challenge of addressing small?satellite development is important, considering the emerging priority for developing much smaller, modular or interactive space systems...	2010	Estimating	"Small data"
The Business Case for Bootstrapping: When You're Stuck with Incomplete Data, Here's How You Make it Work!	Brett Gelso, Glenn Grossman, Eric Druker	The purpose of inferential statistics is to reliably extend sample characteristics to a population. Mild assumptions such as the Central Limit Theorem and Identically and Independently Drawn (IID) estimators ensure that, based on a given level of precision, samples will be representative of the population from which it was drawn. However, when samples are small, most standard test statistics will be insignificant, because sample characteristics may fail to adequately capture variation in the population. Parametric Bootstrapping is an alternate approach which resamples the underlying distribution of the sample in order to estimate population characteristics. The objective of this presentation is to illustrate how to use Parametric Bootstrapping for Cost Estimating with small samples to enhance defensibility of cost estimates...	2010	Methods	"Small data"
Detecting Anomalous Cost Data in an Integrated Data Warehouse	Lawrence Brown	Data quality is always a key issue when using historical data to predict future costs. Often a small data error in the historical data can lead to much more significant errors in predicted costs. The problem becomes more acute if data is integrated from several sources, the sources systems have evolved, the source systems were not designed to support analysis, or extensive business rules are required to allocate or assign attributes to the data. This paper discusses systematic techniques, used by the Air Force Total Ownership Cost (AFTOC) program, to find anomalous data...	2010	Methods	"Small data"
Cost-Risk Analysis of Satellite Bandwidth Services	Sam Bresnahan	The purpose of this paper is to demonstrate application of basic risk- analysis techniques to a real-world cost estimating problem. Each year the U.S. Marine Corps must budget millions of dollars for satellite bandwidth services. Communication via satellite is frequently required during operations in theater, disaster relief, and any situation in which a secure communication infrastructure is not already present. The USMC is developing an Expeditionary Command and Control Suite (ECCS) that allows small teams of soldiers to establish secure satellite voice and data links with headquarters...	2007	Risk	"Small data"

Sampling Methodologies

Consider the generate dataset for CER $f(\mathrm{x})$, $S_f = \{(x_i, y_i) | y_i = f(x_i) + \varepsilon\}$

For each $(x, y)_f = s_f \subset S_f$, the point estimate for each methodology evaluated at x_{ref} is defined as:

Methodology	Description
Analogy	$y_{Analogy} = y_f[\mathrm{argmin}\{\lvert x_f - x_{ref}\rvert\}]$
Flat Average	$y_{Average} = mean\{y_f\}$
Parametric	$y_{Parametric} = a_{s_f} * x_{ref} + b_{s_f}$ Where a_{s_f}, b_{s_f} are the coefficients of OLS regression on s_f