Maximizing Analysis of Minimalized Datasets

Making the Most of a Data-Scarce Environment

Taylor Fountain, Obai Kamara

2024 ICEAA Professional Development & Training Workshop

February 20th, 2024

Abstract

Many techniques exist to determine parametric relationships within large datasets. While cost estimation relies heavily on identifying such relationships, a data-scarce environment, driven by factors such as vendor proprietary restrictions, security concerns, and the uncertainty of emergent technologies, is a common barrier in implementing these techniques. This topic will evaluate common methods to analyze minimalized datasets for developing defendable cost estimates and demonstrate the statistical impacts of their underlying assumptions.

Table of Contents

Table of Figures

Maximizing Analysis of Minimalized Datasets

1. Introduction

Leveraging historical data is crucial in developing impartial cost estimates. While many analytical methods exist to identify cost estimating relationships (CERs), most rely upon having a sufficiently large set of historical data. However, due to a combination of factors such as emergent technologies, vendor proprietary restrictions, and security concerns, gathering enough data points to execute more robust analytical methods is often not possible. Therefore, different techniques must be used to leverage every applicable point in a minimal dataset and develop a defendable cost position.

This paper will review existing guidance on selecting a point estimate in a data-scarce environment. The benefits and drawbacks of these different rules of thumb will be demonstrated through a series of Monte Carlo-based simulations.

2. Selecting a Point Estimate and Bounding Uncertainty

A data-scarce environment is not a new phenomenon in cost estimating, and rule-of-thumb guidance exists on selecting a data-informed point estimate from a limited set of historical data and applying risk and uncertainty to account for the unknown. However, these rules of thumb rarely delve into the quantitative implications of choosing one method over another. To develop an informed risk-adjusted cost position, understanding the impact and limitations of the choice of model is crucial.

A review of existing industry guidebooks, papers, and presentations yielded several results focusing on the development of a point estimate with a limited data set. The literature review was conducted using the search terms “small data” and “limited data” within the ICEAA archives resulting in 36 unique results. Of those 36 results 5 were found to provide specific recommendations for determining a point estimate and/or accounting for risk and uncertainty. In addition to research presented at ICEAA, a review of cost estimating handbooks was also completed.

The GAO Cost Estimating and Assessment Guide identifies three primary and three secondary methodologies for developing a point estimate. Of these, they recommend leveraging analogy and parametric-based methodologies when generating the point estimate with limited data, with subject matter expert input as a methodology of last resort. While various parametric techniques can be applied to small datasets (n<30), most of the reviewed documents recommended a focused attention to improving the quality of the data rather than provide recommendations for methods that can provide accurate results in a data scarce environment. A singular 2021 ICEAA paper “Assessing Regression Methods via Monte Carlo Simulations” was found to evaluate the impact of using various regression techniques on small vs large datasets and three additional papers/presentations highlighted specific techniques for developing the point estimate. Of note is the 2023 follow

on study to the 2014 ICEAA best paper award winning study recommending the use of Bayesian regression.

The most direct recommendations on how to estimate using small data were provided by the Joint Agency Cost Estimating Relationship (CER) Development Handbook. Specifically, the handbook recommended the following guidelines for choosing analytical methods:

$$n = 1 \text{ -> Scaled Analogy}$$

$$1 < n < 5 \text{ -> Scaled Analogy or Average}$$

$$n - k \geq 3 \text{ -> Parametric}$$

Where n = number of data points and k = number of independent variables

3. Simulating an Obscured Viewpoint - Methodology

When applying the above guidance to a minimal dataset, one consideration that may come up is how and how much the point estimate is affected by the historical data points the analyst has visibility into. However, the limited nature of a minimal data set makes observing outliers difficult if not impossible, leaving all analytical methods susceptible to a false confidence in the prior distribution. In the same vein, evaluating how sensitive the methods outlined above are to the choice of inputs would not be possible with a dataset where limited data exists. Therefore, a simulation-based approach was used to determine the sensitivity of the above methods to available inputs.

To evaluate the effects of a data-scarce environment on the results of a data-driven point estimate, the first step was to determine an expected result from a simulation. To do this, an existing simple linear CER and the applicable range of values for the explanatory variable was identified, such that,

$$\{f(x) = a * x + b + \varepsilon | x \in (l, h)\}$$

From there, we defined the input and expected response of our simulation as,

$$(x_{ref}, y_{ref}) = \left(\frac{l+h}{2}, f\left(\frac{l+h}{2}\right)_{|\varepsilon=0}\right)$$

The next step was to randomly generate a set of 45 tuples that follow the behavior of the existing CER. To do this, we used the open-source Python packages NumPy and SciPy. First a normally distributed sample of values $x \in (l, h)$ were generated. Next, the CER was applied to these values without noise. Lastly, noise was added to the response to simulate the natural variation of real-world data, consisting of random draws from the set:

$$N = \left\{ n = a * (.9 + x) \middle| a = \frac{f(l) + f(h)}{2}, x \in U(0, .2) \right\}$$

This results in a set of 45 data points that follow the trend of the reference CER, have an input and response that generally follow the normal distribution, and have responses with homoscedastic noise, denoted as

$$S_f = \{(x_i, y_i) | y_i = f(x_i)_{|\varepsilon=0} + n, n \in N\}$$

The next step was to simulate an analyst being able to see only four data points in S_f, and then applying each of the three methods outlined in the CER handbook for minimal datasets: an analogy, a flat average, and a full parametric regression. For each $(x, y)_f = s_f \subset S_f$, the point estimate for each methodology were defined as follows:

$$y_{Analogy} = y_{f[\text{argmin}\{|x_f - x_{ref}|\}]}$$

$$y_{Average} = mean\{y_f\}$$

$$y_{Parametric} = a_{s_f} * x_{ref} + b_{s_f}$$

Where a_{s_f} and b_{s_f} are the coefficient and intercept, respectively, of ordinary least squares regression for the set s_f.

This simulation is repeated for 10,000 s_f, and the results $(y_{Analogy}, y_{Average}, y_{Parametric})$ are logged for each iteration. For each method, the range of results are assessed both independently, relative to y_{ref}, and relative to each other to appraise their sensitivity to the explanatory variable.

Additionally, the tuples (a_{s_f}, b_{s_f}) are logged to assess the limitations of a full parametric regression across the entire range of applicable values for the explanatory variable.

While in a practical application, qualitative attributes of the program being estimated would be factored in to the choice of data points, this simulation operates under the assumption that all tuples of S_f represent programs sufficiently analogous to the one being estimated.

This methodology is outlined in Figure 1.

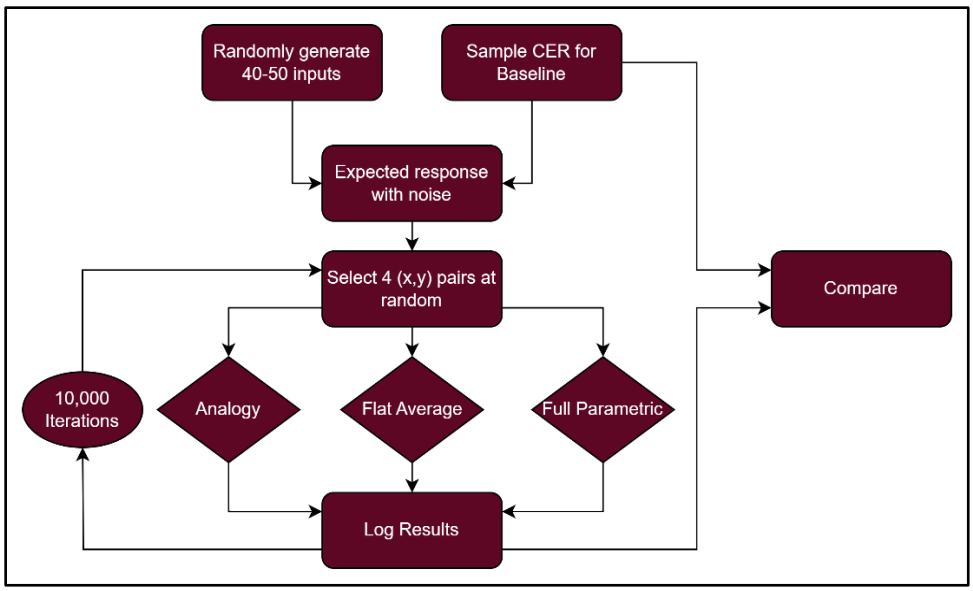

Figure 1: Diagram of Simulation

4. Results

4.1. Linear Model

In regression analysis, the choice of model is going to have a significant impact on the results. While there are many models to choose from, the first simulation was evaluated using a linear model. This was partly to develop a baseline for how the three methods compare independent of transformations applied to the data before comparison. Additionally, when working with minimal datasets, it is unlikely that inspecting the historical data points would provide sufficient insight into the shape of a CER.

The first simulation was conducted on a CER relating annual PMP support hours to the level of acquisition support required for small USAF software development programs (Aguirre et al, *The Progression of Regressions*). The CER is defined as follows:

$$\{f(x) = .8944 * x - 3368\ hrs + \ \varepsilon | x \in (5{,}000\ hrs, 23{,}000\ hrs)\}$$

Based on the methodology defined in section 3, this CER yields a reference point of

$$(x_{ref}, y_{ref}) = (14{,}000\ PMP\ hrs, 9{,}153.6\ Acq\ Hours)$$

45 normally distributed values between 5,000 and 23,000 PMP hours were generated and provided as inputs to the CER. Random values of noise between $\pm 1{,}609$ hours were added to the response to generate the set S_f, which is illustrated in Figure 2 below.

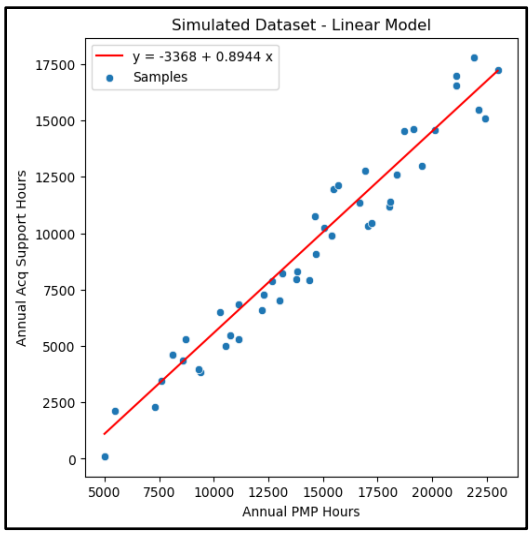

Figure 2: Randomly Generated Dataset for Linear Regression

10,000 simulations of an obscured viewpoint were then applied to this set, which each iteration resulting in a point estimate for the analogy, flat average, and full parametric estimating methods. By plotting the histogram and cumulative distribution function (CDF), some initial observations can be made.

First looking at the results for the analogy method, shown in Figure 3, we can see that the distribution of results is discrete and not continuous; this makes sense as the choice of point estimate is limited by the priors. This is not necessarily a problem; if sufficient risk and uncertainty are applied, the estimate can still adequately capture a developing program's potential to deviate from the analogy.

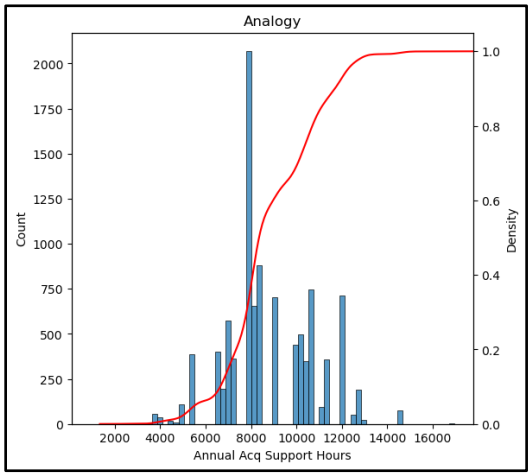

Figure 3: Simulation of an Analogy-Based Point Estimate for a Linear Model

The second observation is that most of the results are clustered at ~8,000 acquisition support hours, which deviates a notable amount from y_{ref}. While this is an incidental finding due to the noise in S_f decreasing the responses where the PMP hours are close to

x_{ref}, it demonstrates an over-confidence in the relationship between the input and the response for the chosen analogy.

Moving on to the results of the flat average methodology, seen in Figure 4, the distribution of possible results resembles a continuous normal distribution with a wide spread. The peak of the curve appears to fall ~9,000 hours, which is in line with y_{ref}. However, the spread of data indicates it would be likely to over- or under-estimate the level of acquisition support required.

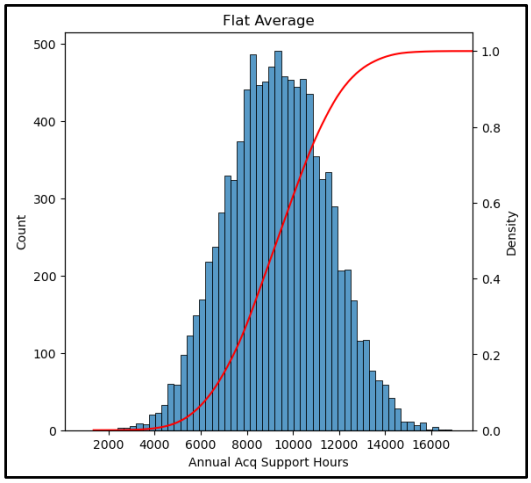

Figure 4: Simulation of an Average-Based Point Estimate for a Linear Model

Observe that the flat average considering every point of the visible subset to be equally likely and is independent of the PMP hours of the program being estimated. Practically, this can be beneficial – if the program is early enough in the acquisition life cycle that the PMP hours cannot be reliably estimated, a point estimate for the acquisition support hours can still be developed.

For this same reason, if response has an unusual residual from the trend of the greater dataset, the effect of that noise will be less pronounced than in an analogy, as it will be offset by other points in the sample.

However, a flat average is sensitive to outliers in inputs. If the visible subset contains any values of the explanatory variable that are abnormally small or large due to factors that cannot be observed through market research, a flat average will consider them more applicable to other programs than they are. In this case, the point estimate could reflect the program being just as likely to require 14,000 PMP hours as 23,000 (and, by extension, 9,000 acquisition support hours as 17,000), which may not be realistic.

Lastly, the results for a full parametric regression are shown in Figure 5. The distribution of results resembles a continuous normal distribution with a narrow spread. Like the flat average, the results peak close to y_{ref}; however, the results rarely fall out of the 7,000-to-11,000-hour range, which differs significantly from the flat average methodology. This

indicates a high likelihood of the predicted value aligning with the trend of the larger dataset.

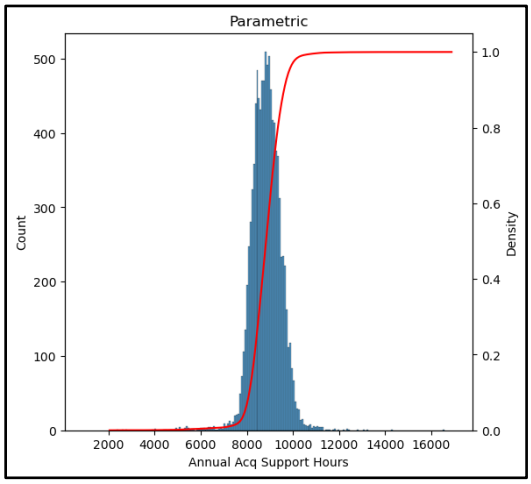

Figure 5: Simulation of a Regression-Based Point Estimate for a Linear Model

However, there are still some results that fall at extreme ends of the spectrum. While this can be accounted for in the flat average by an independence from the PMP hours, the parametric accounts for x_{ref} yet still exhibits these extremes. Since the explanatory variable is being controlled for and neither the explanatory variable nor response were transformed for the regressions, it can be concluded that the variation in the results is fully attributable to the noise in the response.

While these extreme outliers are not always readily identifiable with a minimal dataset, there are some instances where domain knowledge can be used to evaluate the coefficient and intercept of the regression and rule out inaccuracies. Figure 6 plots results of the regression analysis performed in the simulation across the entire range of valid PMP hours.

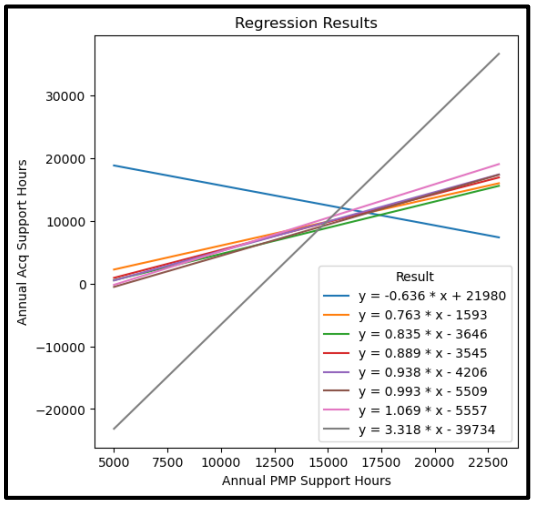

Figure 6: Sample Regression results, demonstrating outliers with unjustifiable coefficients

Consider the result of the regression yielding the equation $y = -0.636 * x + 21980$. While a result 13,076 acquisition support hours sounds high for a program with 14,000 PMP hours, the negative coefficient indicates the point estimate is not only high but unfounded, as the acquisition support hours would decrease should the program become larger in scope. Similarly, the result $y = 3.318 * x - 39734$ would indicate that every hour of PMP support requires over 3 hours of acquisition support, which is not defendable. Similarly extreme results would indicate that an analogy or flat average would be better to use.

Cross-checking with domain knowledge may not always be possible when the explanatory variable and response have a less intuitive relationship. However, if this is the case, it would not be advisable to construct a parametric relationship between the two variables based on a minimal dataset in the first place.

4.2. Power Model

While a linear regression model can provide insight into simple relationships between two variables, they are rarely reflective of the complexities of real-world relationships. While this can usually be accounted for by adding more explanatory variables to a model or increasing the degrees of freedom with polynomial regression, this may not be possible when data is not available. One common model used in regression analysis that requires neither more explanatory variables nor more degrees of freedom but does allow for modeling more complex relationships is a power model.

To adapt our experiment to account for a power model, we let:

$$\{\ln(f(\mathrm{x})) = a * \ln(x) + b + \varepsilon | x \in (l, h)\},$$

$$(x_{ref}, y_{ref}) = \left(e^{\frac{\ln(l)+\ln(h)}{2}}, f\left(e^{\frac{\ln(l)+\ln(h)}{2}}\right)_{|\varepsilon=0}\right)$$

And,

$$y_{Parametric} = b_{s_f} * {x_{ref}}^{a_{s_f}}$$

Where $\ln(a_{s_f})$ and b_{s_f} are the coefficient and intercept, respectively, of ordinary least squares regression for the set $\ln(s_f)$.

The next simulation was conducted on a CER relating simple function points of an Agile software development effort to the total development hours required. The CER is defined as follows:

$$\{\ln(f(x)) = .7708 * \ln(x) - \ln(421.68\ hrs) + \varepsilon | x \in (100\ SiFP, 10{,}000\ SiFP)\}$$

Based on the methodology defined in section 3, this CER yields a reference point of

$$(x_{ref}, y_{ref}) = (1{,}000\ SiPF, 86{,}573\ Dev.Hours)$$

45 log-normally distributed values between 100 and 10,000 were generated and provided as inputs to the CER. Random values of noise between $\pm.355$ ln (Dev. Hours) were added to the log of the response, then transformed to generate the set S_f.

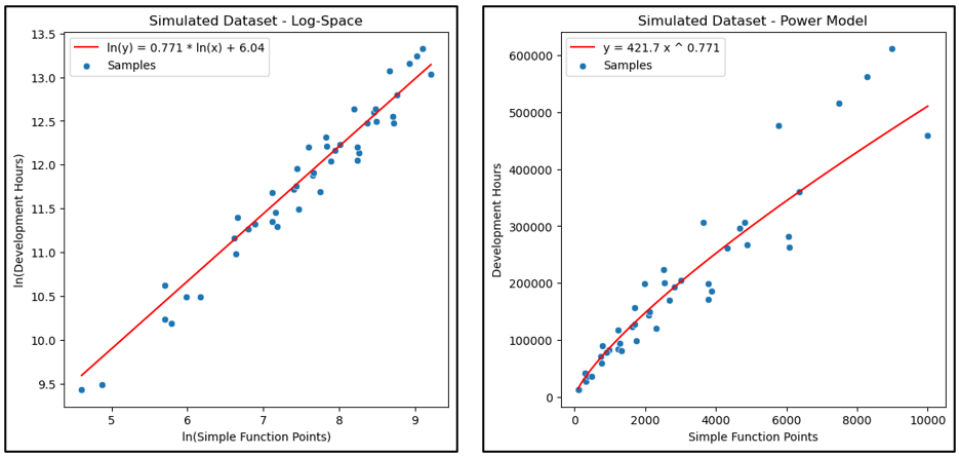

Figure 7: Randomly Generated Dataset for Power Regression

10,000 simulations of an obscured viewpoint were then applied to this set, which each iteration resulting in a point estimate for the analogy, flat average, and full parametric estimating methods. By plotting the histogram and CDF, some observations can be made.

Looking at Figure 8, we can see that, once again, that most of the results are clustered near y_{ref}. However, in the cases where only high-complexity outliers can be seen, this method massively overestimates the results, even more than the linear model. However, since this analogy is based on the simple function points of an Agile software development effort, it would be simple to disregard these results in practice based on the extreme difference in the explanatory variable.

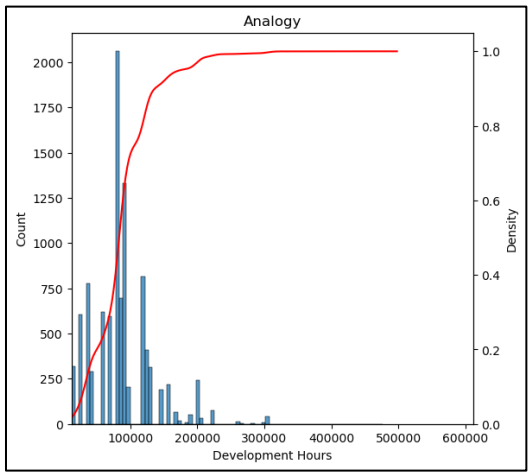

Figure 8: Simulation of an Analogy-Based Point Estimate for a Power Model

Looking at the result of the flat average method in Figure 9, we can once again see that the results of the simulation are continuous with a wide spread. However, unlike the results of

the linear model, the peak of the curve is noticeably greater than y_{ref}. With a mean and median of ~190,000 hours, most of the results would not yield a reasonable result.

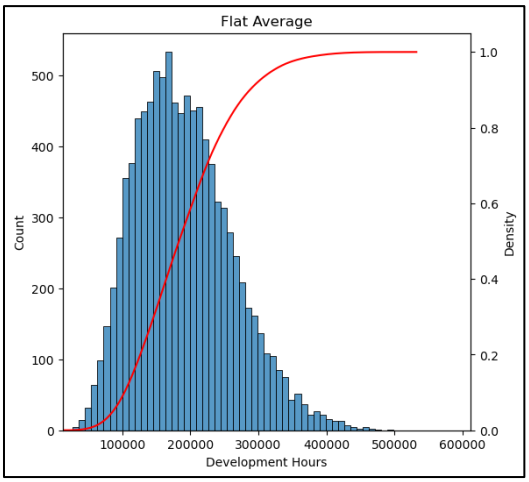

Figure 9: Simulation of a Flat Average-Based Point Estimate for a Power Model

This is because, as mentioned above, the flat average method is sensitive to outliers in the response. As a power model where the exponent is not equal to 1 or 0 is inherently skewed, the effect of these outliers is going to be more pronounced. While weighting techniques can be used to lessen these effects, developing an informed weighting method with a minimal dataset is rarely possible. A solution is to narrow down the dataset even further to a single data point and use that as an analogy and utilize the other data points to inform a risk distribution.

The result of the full parametric simulation is shown in Figure 10. Like the results with the linear model, the results have peak close to y_{ref} and have a low spread. However, eight of the results, not plotted below, were greater than 500,000 hours. Though in practice these points would be easy to discard, this highlights the increased sensitivity to noise compared to other methods, which is exaggerated with a more complex model.

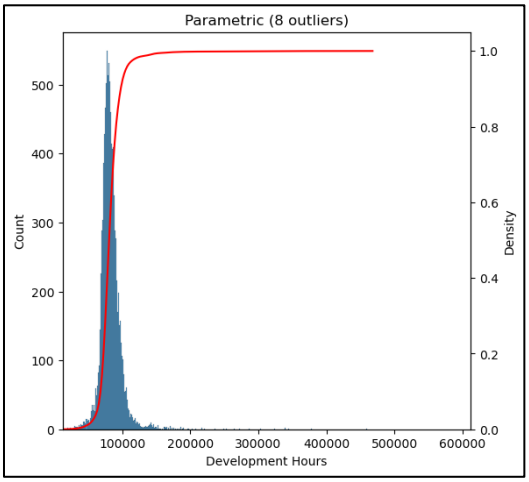

Figure 10: Results of a Full Parametric-Based Point Estimate for a Power Model

Despite the outliers, it appears these results are accurate. While this may be true for a low simple function point count, Figure 11 shows that this does not hold for larger values. Even when contextual and statistical outliers are removed, the range of possible results varies dramatically when the simple function point count is greater than 2,500.

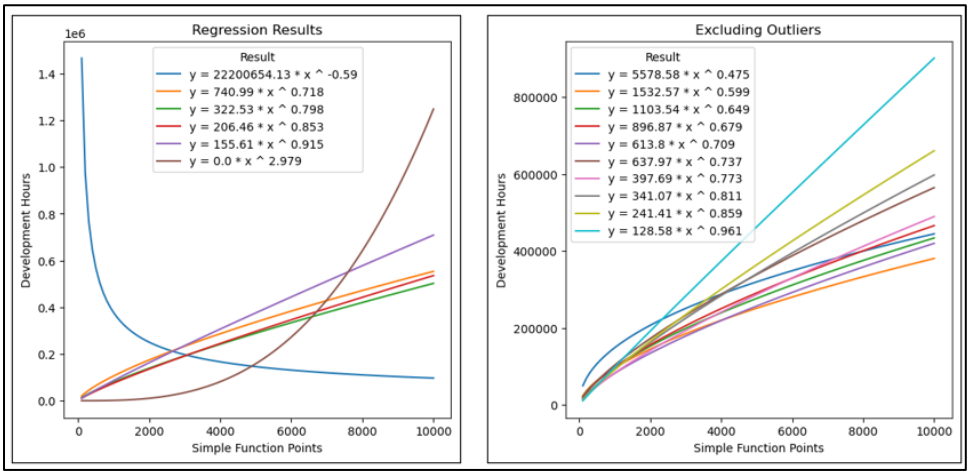

Figure 11: Sample results of Regression Analysis for a Power Model

Note that the CER handbook recommends using weighted least squares over ordinary least squares for a power model. However, it is assumed for the purposes of this experiment that there is not sufficient knowledge of the prior distribution to inform a weighted least squares approach.

5. Conclusions

Many techniques exist to determine parametric relationships between variables in large datasets. However, when developing cost estimates, data scarcity can impede the effectiveness of these techniques. A review of existing guidance highlighted the limited guidance on developing data-informed point estimates from a minimal dataset. Further, the guidance that does exist rarely demonstrated the benefits and drawbacks of using one method over another. However, three main techniques were identified in the literature – analogies, flat averages, and parametric methods.

By leveraging Python and existing CERs, we were able to simulate how each of these methods fare when only a subset of the prior distribution is observed. Through this testing, we concluded that each method had its benefits and limitations when applied to a minimal dataset. Estimating by analogy makes for a defendable estimate when little data is available but implies confidence in the relationship between the explanatory variable and response, which may be false. Estimating with a flat average can be done without quantitative inputs of the effort being estimated but is sensitive to outliers in the response of the sample. Finally, estimating with parametric methods has the potential to return accurate results, but is sensitive to noise in the observed sample and can create false confidence. The

drawbacks of all these methods are only emphasized as relationships between variables become more complex.

Overall, there is no way to advise which of these methods will work best for assessing a minimal dataset 100% of the time. However, being informed of the advantages and limitations of these methods allows cost analysts to not only develop defendable estimates, but also speak to the limitations created by data scarcity. By being able to navigate these known unknowns, analysts can make the most of ever observation in a data-scarce environment and maximize the analysis of minimal datasets.

Appendix A: References

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%20Development%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. 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