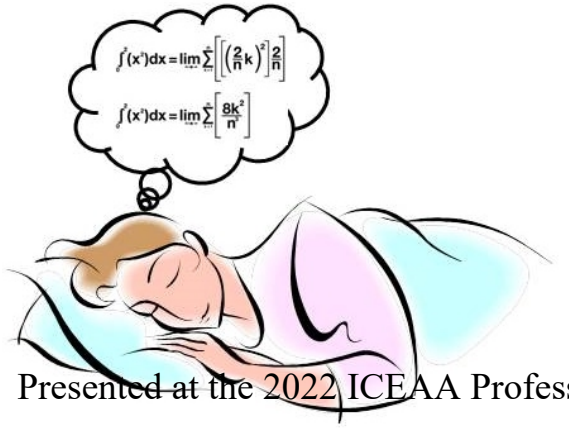
The Progression of Regressions

Jennifer Aguirre (AFLCMC/HBG, Quantech Services)
Kyle Davis (AFLCMC/HBG, Quantech Services)
Matt Hoffman (AFLCMC/HBGF)

The Progression of Regressions
Introduction

- Do you dream of having large, robust datasets that follow a standard regression form, but are woken up by the reality of dealing with limited data reporting and imperfect trends?
- Well fret no more! This brief will utilize a case study and its results to explore how various CER difficulties may be addressed including:
  - Working with a Limited Data Set
  - Determining Regression Form
  - Assessing Best Fit
  - Maximizing CER Applicability
  - Visualizing CER Results



The Progression of Regressions
Outline

- Case Study Background
- Data and Trend Analysis
  - Viewing Sets and Subsets of Data
- CER Creation
  - Selecting Forms
  - Exploring Functional Form
  - Considerations
- CER Statistical Analysis
  - Results
  - Result Segmentation
- Summary

The Progression of Regressions
Case Study Background

- Study finalized in **February 2021**
- Analyzes **16 command and control/IT software development projects** within various US Air Force directorates
  - Data includes Engineering Manufacturing Development (EMD) & Post Milestone-C Non-Recurring Engineering (NRE) efforts using a mix of agile and agile-like software development methodologies
  - Preponderance of data from projects executing within **FY16-FY20** timeframe
  - Collected and binned labor hours and labor cost into MIL-STD 881D WBS categories
  - Compared **PMP** (SW Development / Integration, HW NRE, Installation / System Integration, & On-Site SME Support) and **Acquisition Support** (Systems Engineering/Program Management, Data, and System Test & Evaluation) levels of effort
- Aims to provide a **consolidated database for metric selection** based on similar scope, development approach, and annual level of PMP effort as well as provide **several CERs for consideration**

- **The dataset is limited due to availability of program office data; our team is working to increase the number of data points included in the study as more program data becomes available**
- **The majority of data points come from programs that our team members have worked personally and therefore we have an intimate working knowledge of each data point**

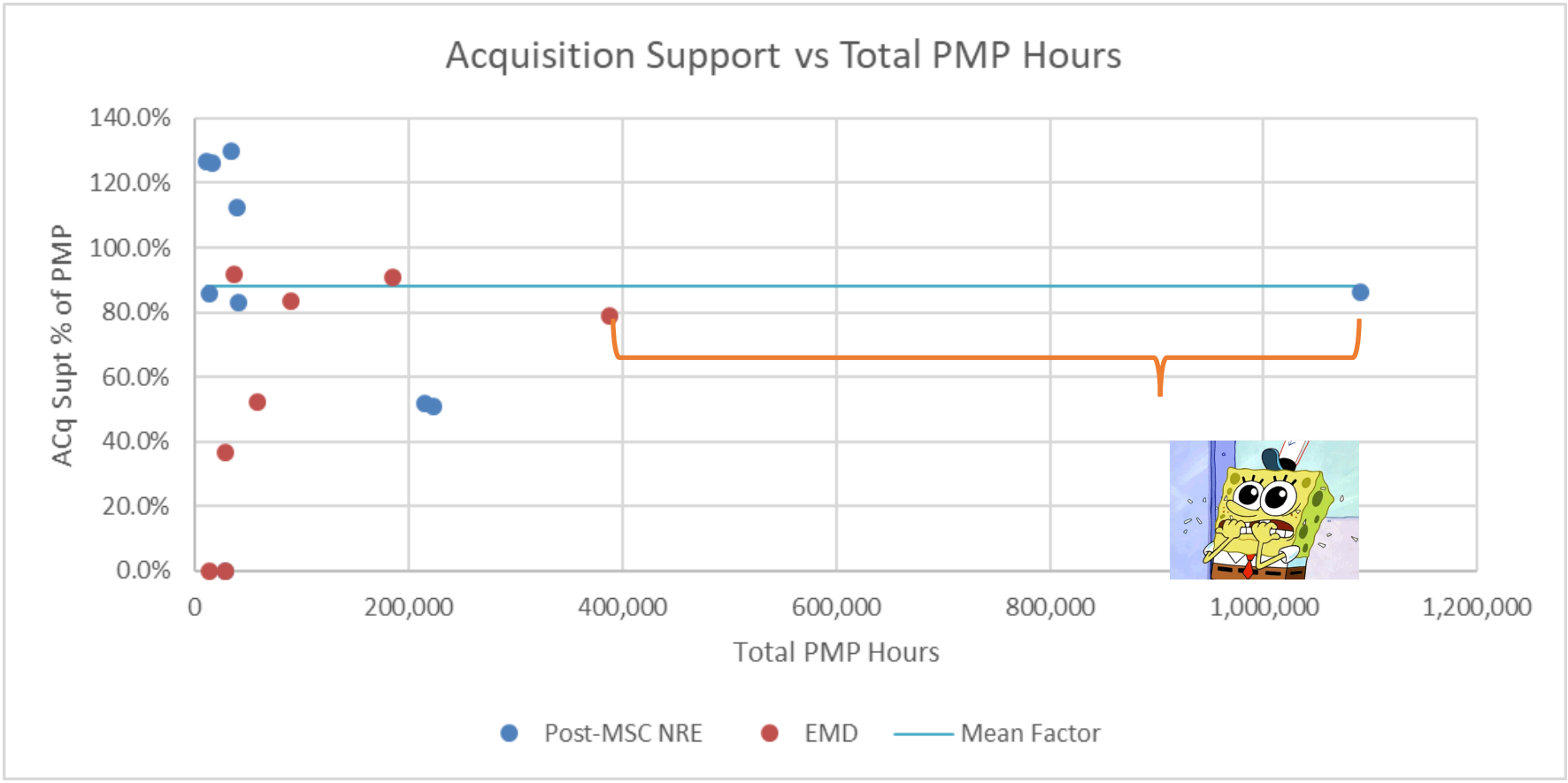
The Progression of Regressions
Case Study Background - Database

HB - IT/C2 Programs Acq Supt Study (EMD/NRE Programs) 2017 - Hours Based Analysis									
Project Name	Larger Program (Portfolio Level)	Phase	Total PMP Hours (SW Dev, HW NRE)	Total Acq Supt Hours (SE, PM, ST&E, Data)	Acq Supt % of PMP Hours	Total SE/PM + Data Hours	SE/PM + Data % of PMP Hours	Total ST&E Hours	ST&E % of PMP Hours
Project #1	Portfolio #1	Post-MSC NRE	222,720	113,280	51%				
Project #2	Portfolio #2	Post-MSC NRE	1,090,560	940,800	86%	762,240	70%	178,560	16%
Project #3	Portfolio #1	Post-MSC NRE	214,368	110,736	52%	94,512	44%	16,272	8%
Project #4	Portfolio #2	Post-MSC NRE	13,434	11,511	86%	10,183	76%	1,328	10%
Project #5	Portfolio #2	Post-MSC NRE	16,339	20,640	126%	18,895	116%	1,746	11%
Project #6	Portfolio #2	Post-MSC NRE	28,299	33,822	120%	22,846	81%	10,976	39%
Project #7	Portfolio #2	Post-MSC NRE	10,951	13,861	127%	11,377	104%	2,484	23%
Project #8	Portfolio #3	Post-MSC NRE	33,714	43,807	130%	28,774	85%	15,033	45%
Project #9	Portfolio #3	Post-MSC NRE	38,849	43,748	113%	27,306	70%	16,442	42%
Project #10	Portfolio #3	Post-MSC NRE	40,183	33,293	83%	19,732	49%	13,561	34%
Project #11	Portfolio #4	EMD	28,687	10,450	36%	10,450	36%	0	0%
Project #12	Portfolio #4	EMD	36,310	33,247	92%	27,458	76%	5,789	16%
Project #13	Portfolio #4	EMD	90,009	75,057	83%	37,022	41%	38,035	42%
Project #14	Portfolio #5	EMD	58,482	30,674	52%	26,320	45%	4,354	7%
Project #15	Portfolio #6	EMD	185,494	168,101	91%				
Project #16	Portfolio #6	EMD	387,946	306,807	79%				

- What set of data do you want to analyze? Should it be normalized?
- Analysis focuses on labor hours instead of dollars because it is skill mix agnostic
- Data considered at average annualized hours because programs do not have discrete endpoints

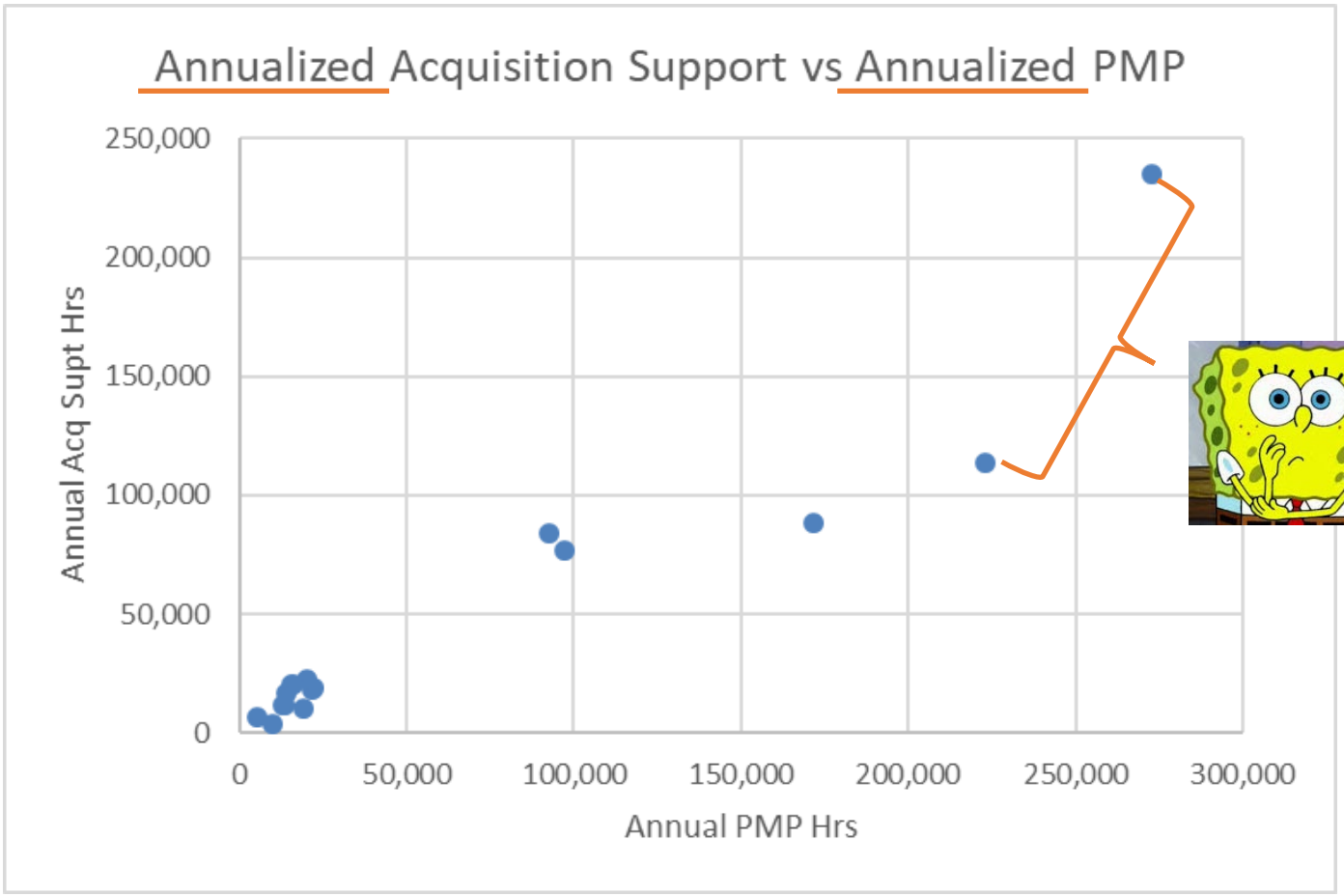
Data and Trend Analysis

The Progression of Regressions
Data and Trend Analysis



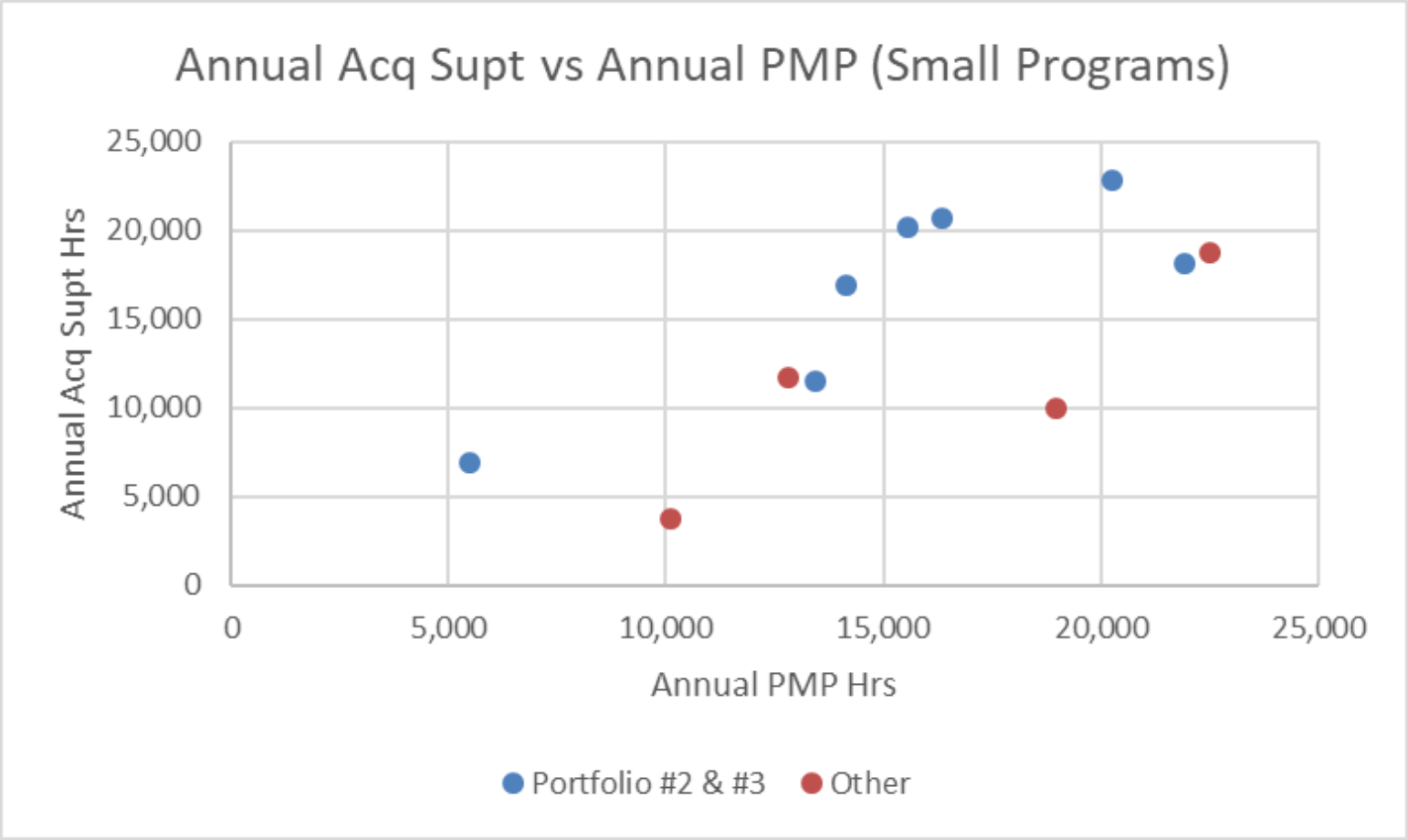
- It is important to view the data in a variety of different ways to see if any trends appear
- Acquisition Support is generally applied as a percent of PMP hours, however, here there is no distinguishable trend in Acquisition Supt percentage for programs of various size
- Project phase is separated as a possible cause for change in Acquisition Support; however, no distinguishable trend; must look into other ways to view the relationship as a clear connection cannot be found between Acquisition Support percent of PMP hours

The Progression of Regressions
Data and Trend Analysis



- **Another way we can look for trends is by comparing annual Acquisition Support hours against annual PMP hours**
- **Although total percentage can vary significantly, there does seem to be a trend among annual Acquisition Support and PMP hours**
- **Can we break this out to explain the variation even further by taking into account other factors?**

The Progression of Regressions
Data and Trend Analysis



- **Since we do not have many medium-to-large scale data points, we will see the most explanation in variation within the small scale data points**
- **Adding the portfolio as an explanatory variable does seem to have an effect on ratio of Acquisition Support hours to PMP hours – Portfolio 2 and 3 are within the same division**
- **Further exploration is needed for reasoning behind portfolio deltas**

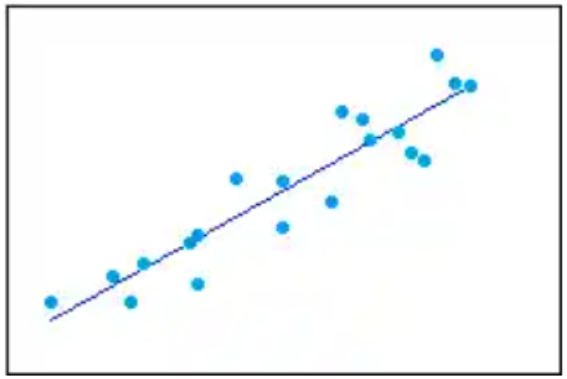
CER Creation

The Progression of Regressions
CER Creation – Selecting Forms

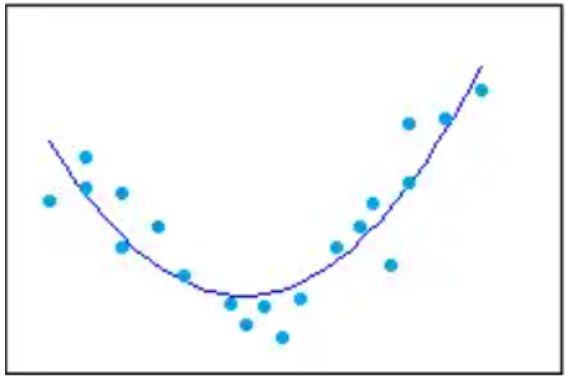
Most Common Regression Forms

Polynomial Regressions

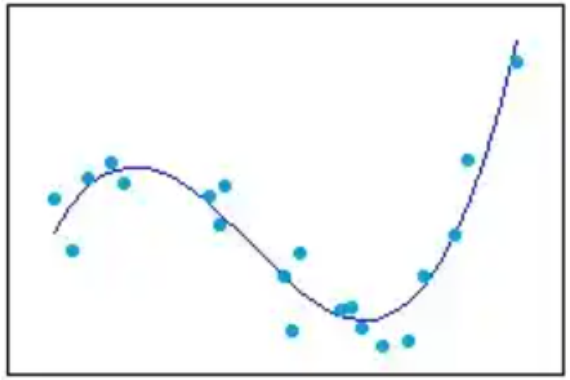
Linear



Quadratic

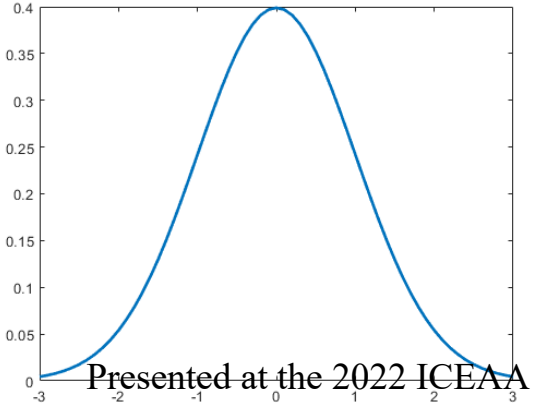


Cubic



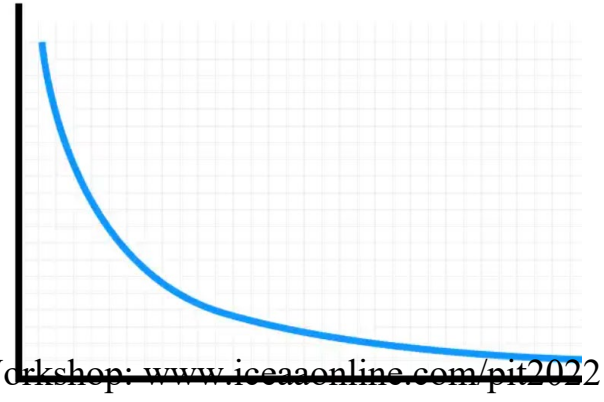
Non-Polynomial Regressions

Beta



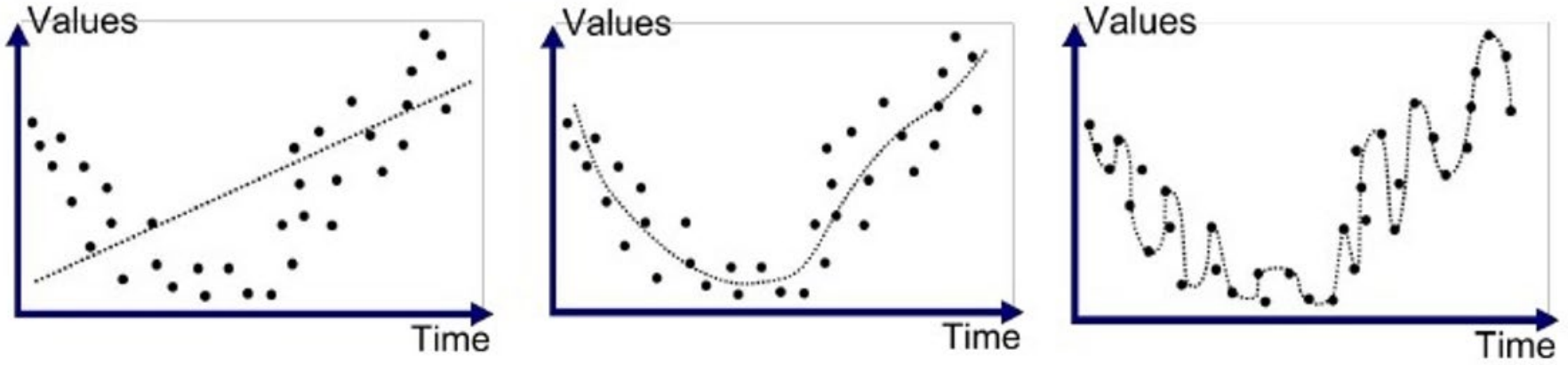
Log Linear

Learning

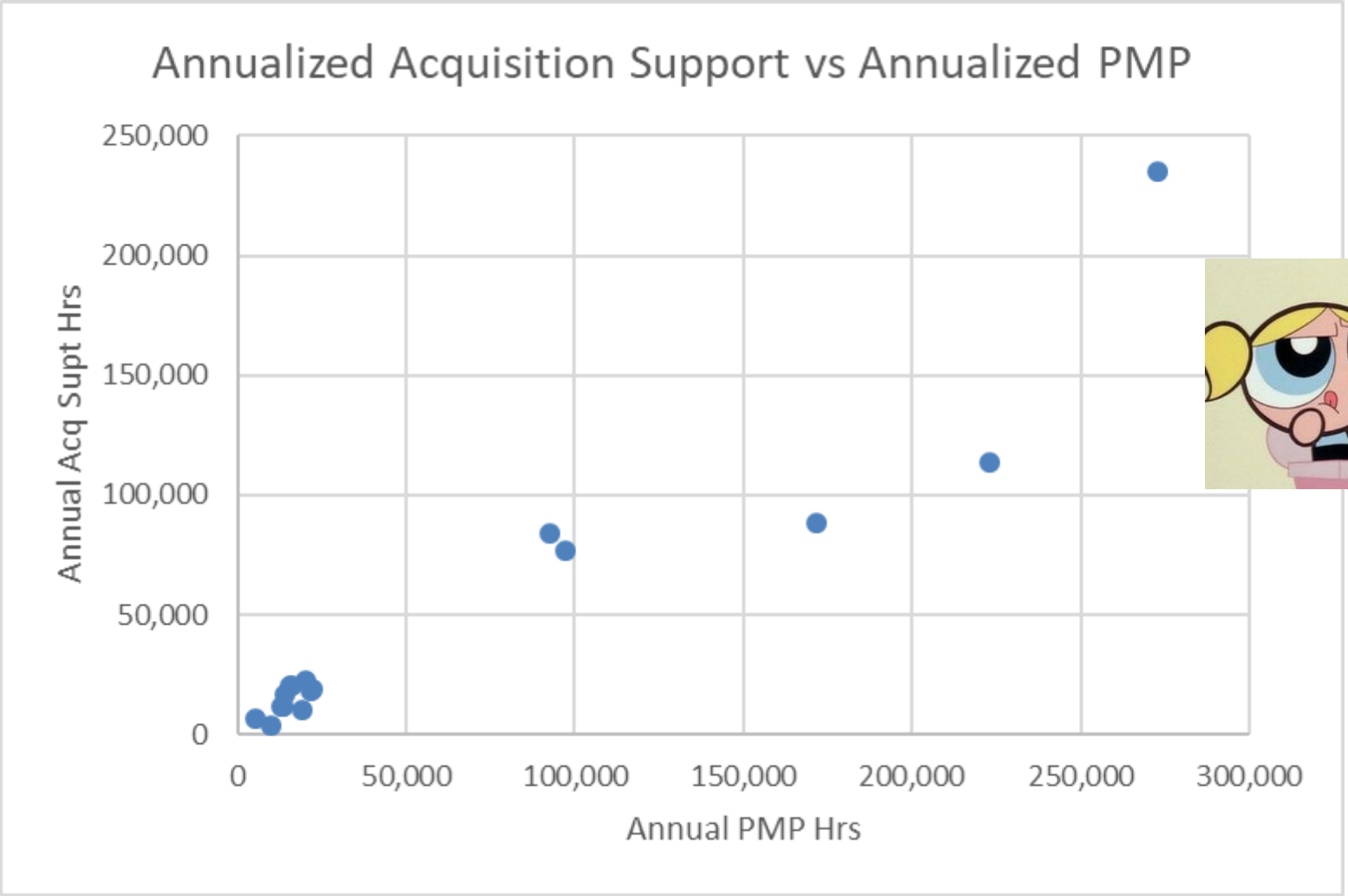


The Progression of Regressions
CER Creation – Selecting Forms

- When selecting the best fitting regression form, be careful not to over fit the model
- CER should follow the shape of the data, but should be careful not to move to fit every data point
- There is a delicate balance, especially when working with small datasets, as all data points drive the shape when data is limited
- As a result, it is important to make sure the narrative of the CER being fit makes sense

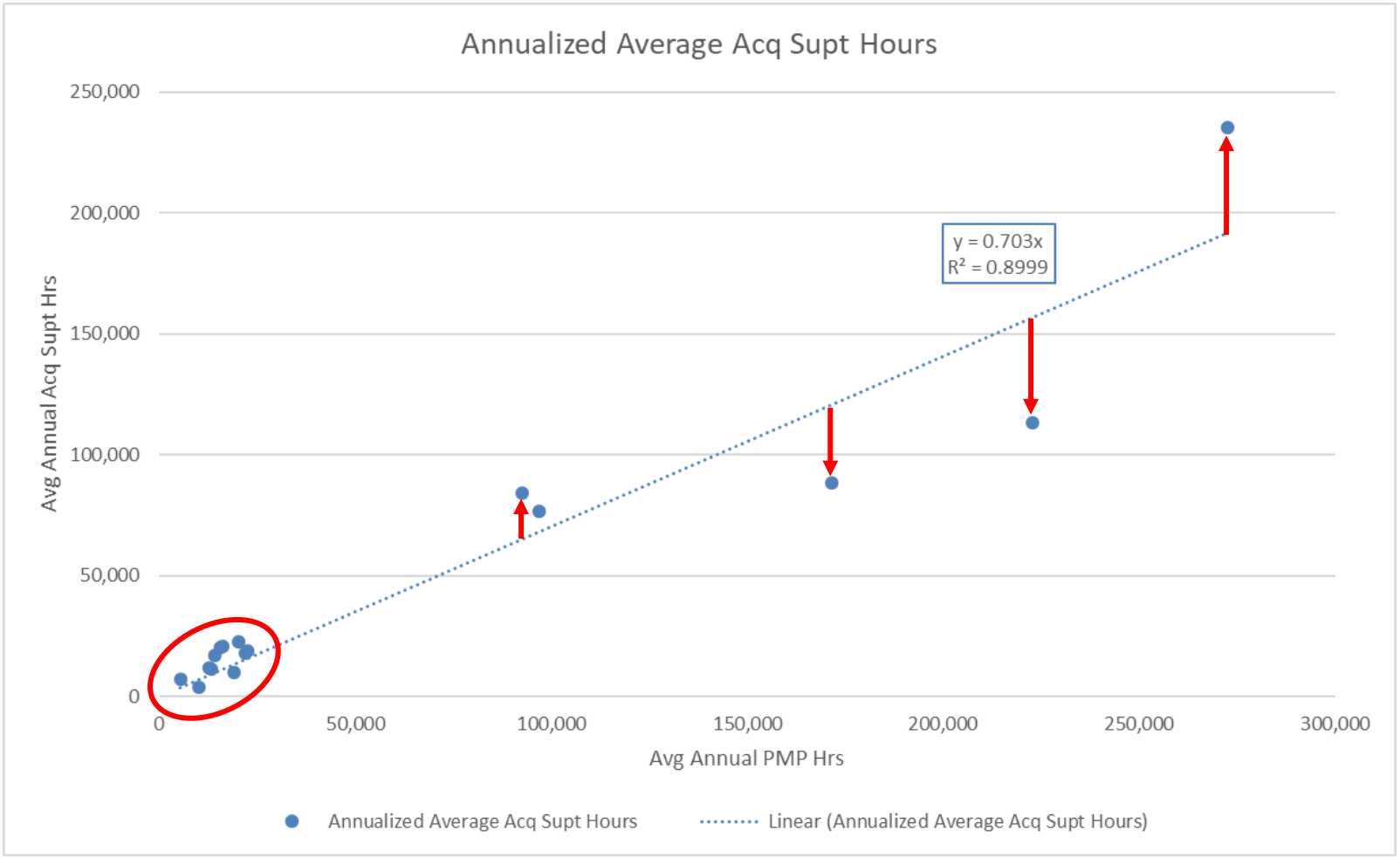


The Progression of Regressions
CER Creation – Exploring Functional Form



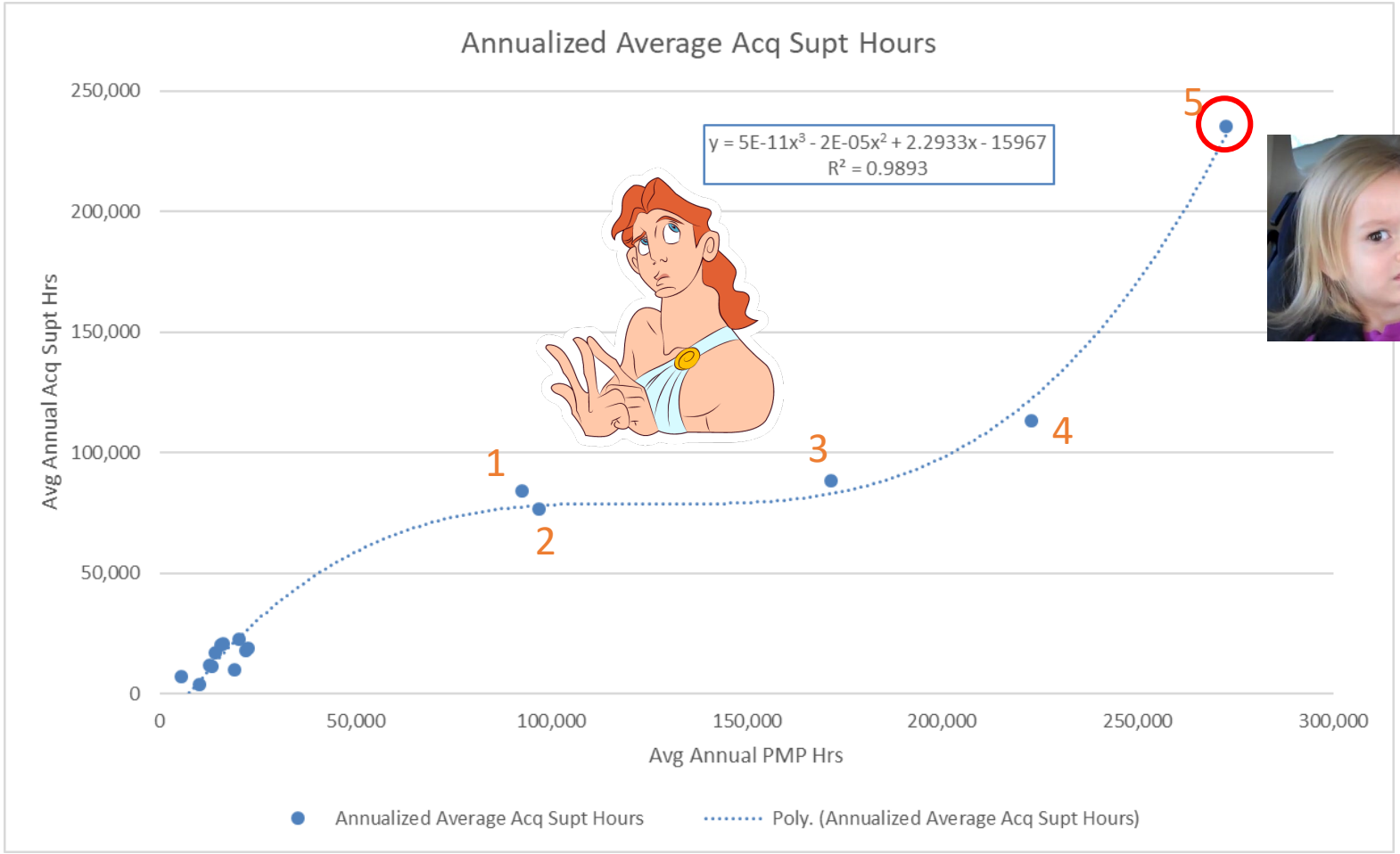
- **The plotting above shows the average annualized Acquisition Support hours against average annualized PMP hours**
- **When looking at this chart, what form of regression would you try to fit first?**

The Progression of Regressions
CER Creation – Exploring Functional Form



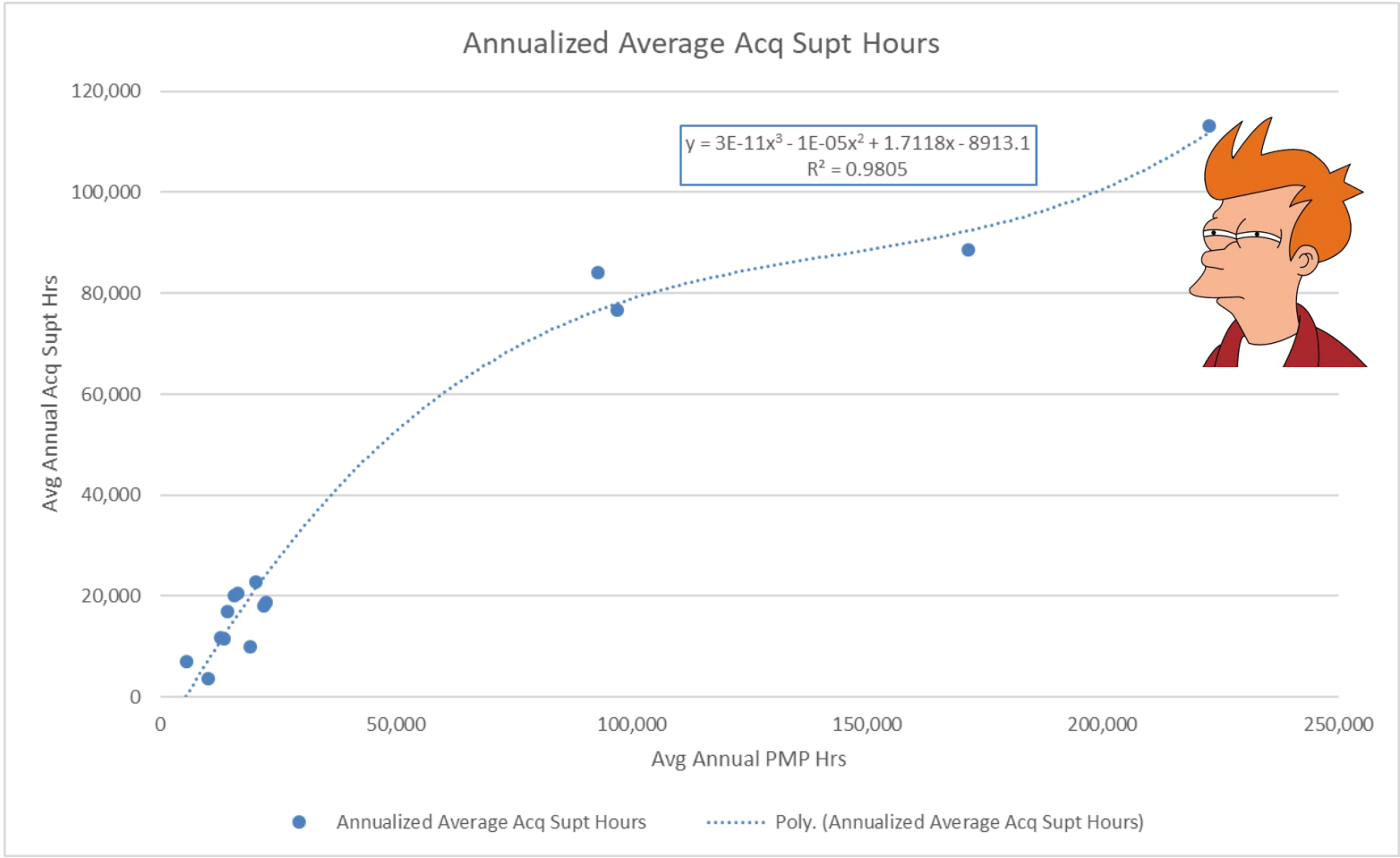
- **If we fit a linear regression, it will be too heavily influenced by the medium-to-large scale data points and will skew the results**
- **The regression has a high r-squared, but it under predicts the small scale programs and has a high level of error for the medium-to-large scale programs**

The Progression of Regressions
CER Creation – Exploring Functional Form



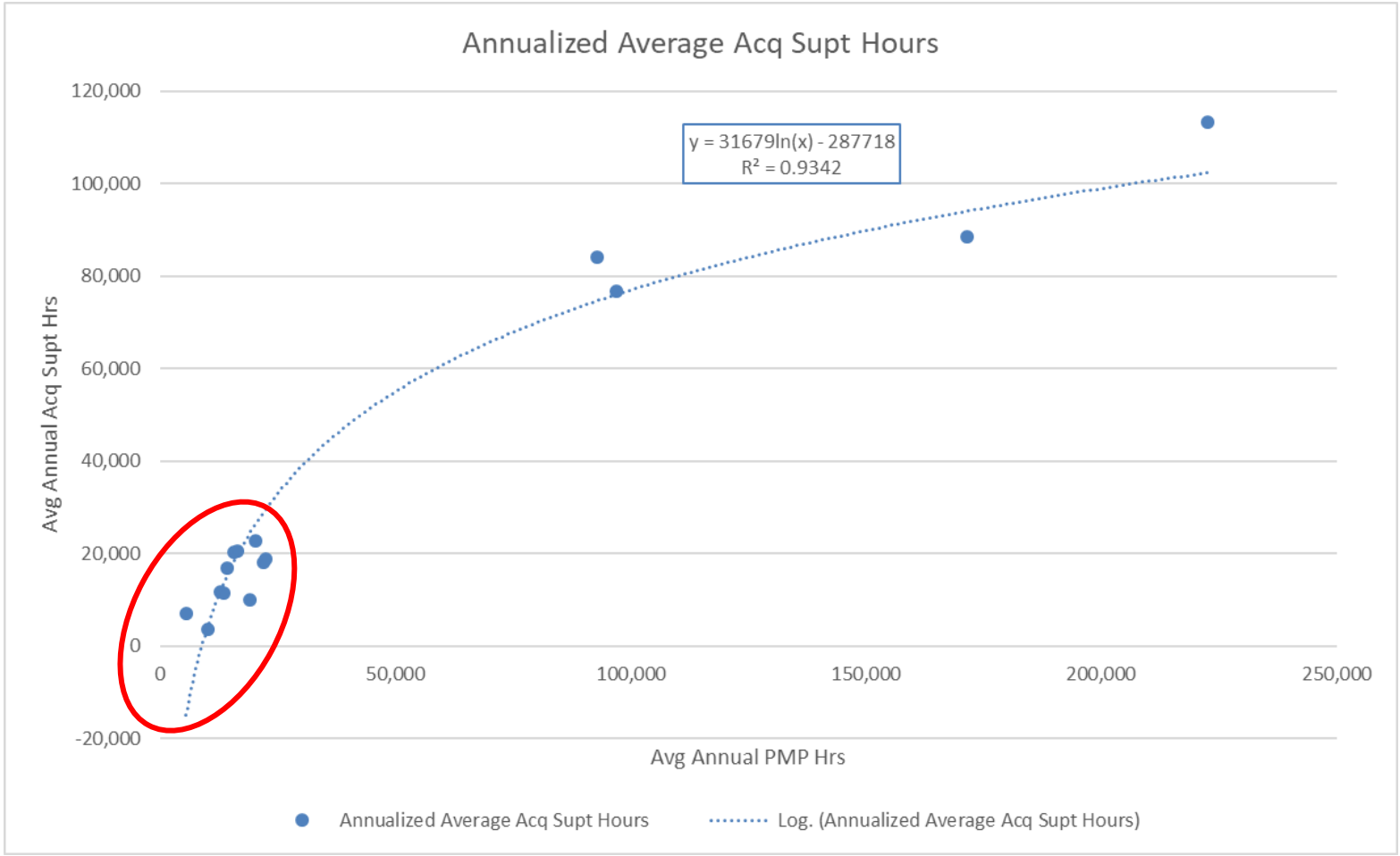
- **However, if we fit a polynomial regression, in particular a cubic regression, we allow for the curve to flatten out for the mid-range data points and increase again for the large scale program**
- **There are only a few medium-to-large scale data points driving this trend so we need to:**
  - **Identify any potential outliers to ensure shape is still cubic with their removal**
  - **Ensure the interpretation of the cubic form is rational given the context of the data set**

The Progression of Regressions
CER Creation – Exploring Functional Form



- **Largest data point is a gov't led integration effort leveraging numerous contracts and enterprise level acquisition support on top of ART level acquisition support activities – may be an outlier**
- **Even with the removal of the largest data point, the cubic trend still seems to be a strong fit**
- **The cubic trend upward is more subtle when we exclude the potential outlier and looks fairly similar to a log linear regression. That should be tested as well to ensure best fit is found.**

The Progression of Regressions
CER Creation – Exploring Functional Form

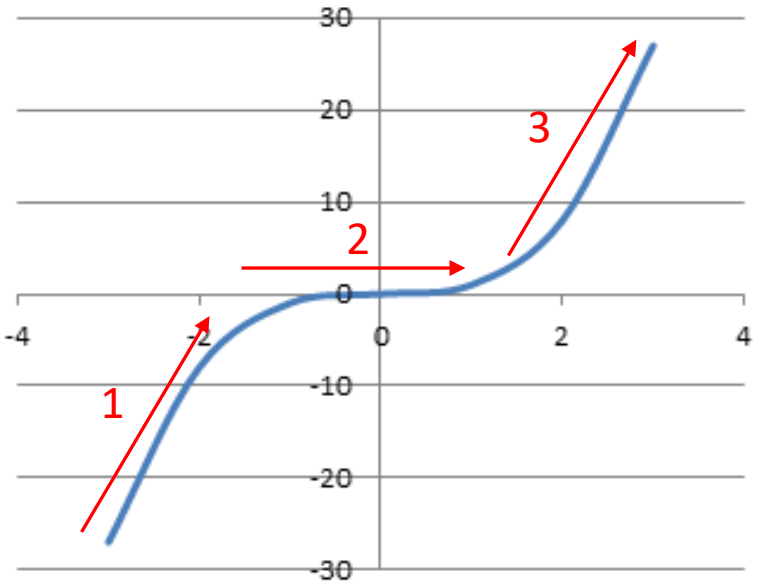


- **Although the regression looks somewhat log linear when we remove the potential outlier, the trend line does not fit the dataset as well as the cubic form**
- **The regression goes negative within predictive range and has a higher error for small programs**
- **Since the data appears to follow a cubic trend, we need to ensure the rationale for a cubic CER is sound**

The Progression of Regressions
CER Creation – Exploring Functional Form

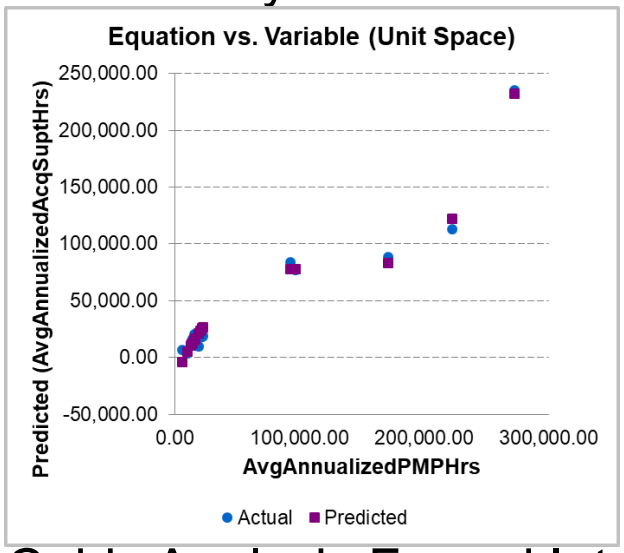
Cubic Model Interpretation:

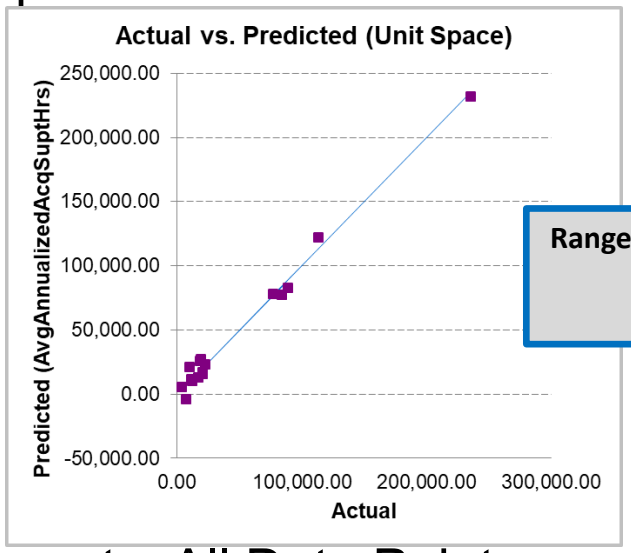
1. Suggests Acquisition Support increases at one rate for smaller programs
2. After a certain program size, teams hit a level of efficiency and steady out – more PMP can be added without need for additional Acquisition Support
3. Once the program passes a certain size threshold, the Acquisition Support will once again need to increase, potentially at a different rate from that of small programs



The Progression of Regressions
CER Creation – Considerations

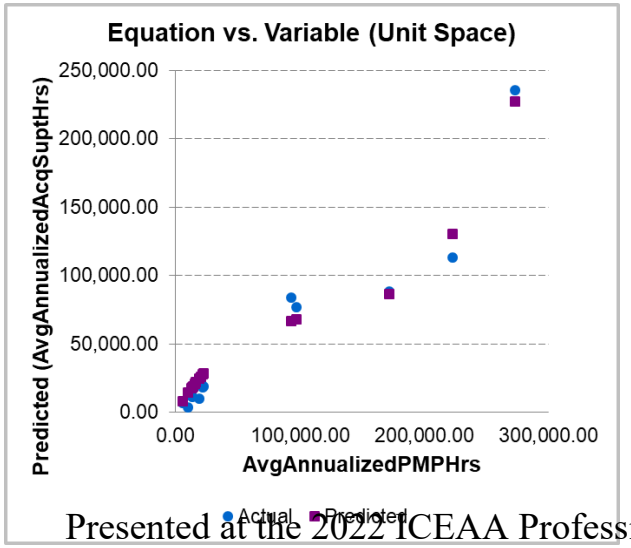
- Cubic Analysis Free Intercept – All Data Points

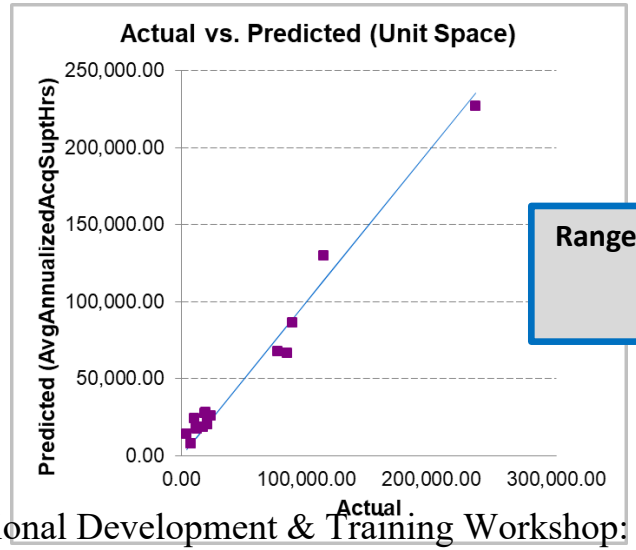




Range = ~5,000 to ~273,000 PMP Hours
Mean = 64,259 PMP Hours
R-Squared = 98.93%

- Cubic Analysis Forced Intercept – All Data Points

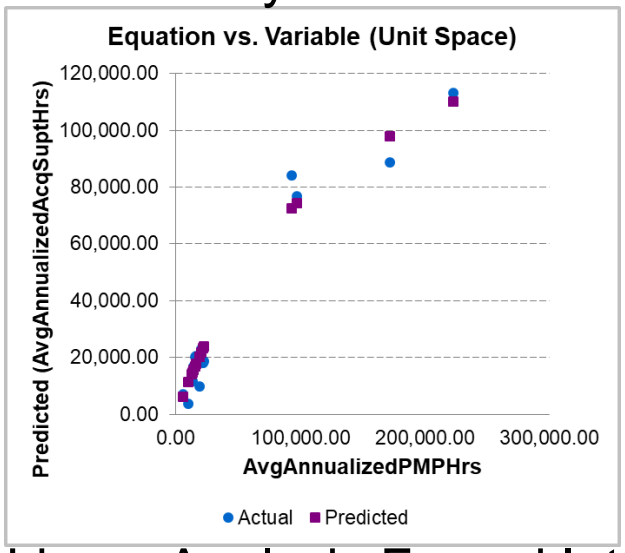


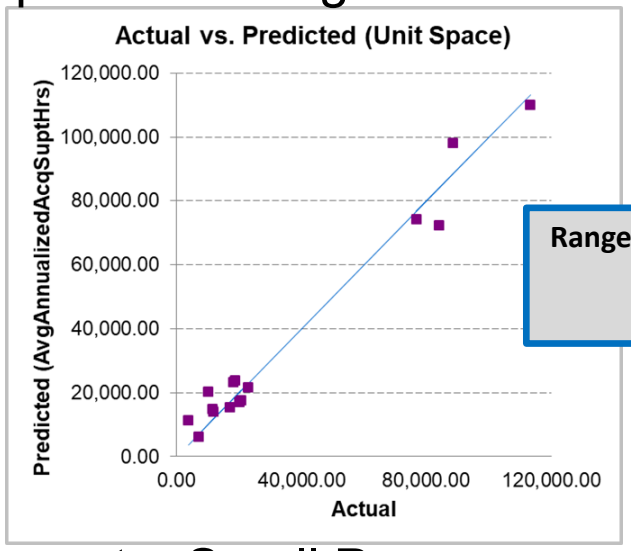


Range = ~5,000 to ~273,000 PMP Hours
Mean = 64,259 PMP Hours
R-Squared = 97.58%

The Progression of Regressions
CER Creation – Considerations

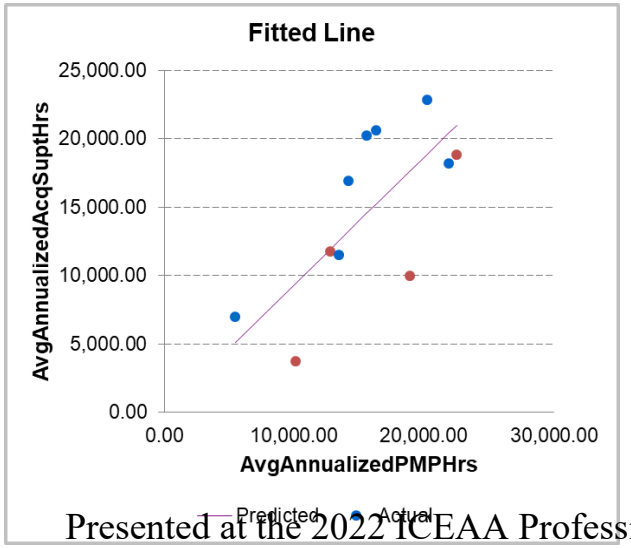
- Cubic Analysis Free Intercept – Excluding Potential Outlier

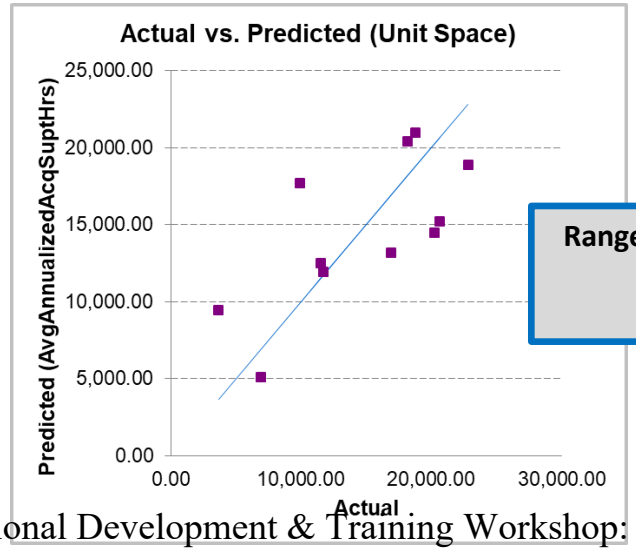




Range = ~5,000 to ~223,000 PMP Hours
Mean = 50,367 PMP Hours
R-Squared = 98.05%

- Linear Analysis Forced Intercept – Small Programs





Range = ~5,000 to ~23,000 PMP Hours
Mean = 15,596 PMP Hours
R-Squared = 97.24%

CER Statistical Analysis

The Progression of Regressions
CER Statistical Analysis – Results

Comparison of CERs	Rule of Thumb	All Data - Cubic	All Data - Cubic (Forced int)	Excl Outlier - Cubic	Excl Outlier - Cubic (Forced Int)	Small w Portfolio - Linear	Small - Linear (Forced Int)
Low Range Value	N/A	5,476	5,476	5,476	5,476	5,476	5,476
Mean Range Value	N/A	64,259	64,259	50,367	50,367	15,596	15,596
High Range Value	N/A	272,640	272,640	222,720	222,720	22,502	22,502
Average Actual (Avg Act)	N/A	47,447	47,447	34,930	34,930	14,666	14,666
Standard Error (SE)	N/A	7,057	10,171	5,664	6,447	3,392	4,455
Root Mean Square (RMS) of % Errors	< 30%	53.2%	84.9%	46.7%	61.7%	26.5%	55.3%
Mean Absolute Deviation (MAD) of % Errors	< 25%	31.8%	46.8%	30.4%	33.2%	22.5%	35.3%
Coef of Variation based on Std Error (SE/Avg Act)	< 25%	14.9%	21.4%	16.2%	18.5%	23.1%	30.4%
Coef of Variation based on MAD res (MAD Res/Avg Act)	< 25%	10.5%	15.6%	11.6%	13.4%	18.2%	24.7%
Pearson's Correlation Coefficient between Act & Pred	> 70%	99.5%	99.0%	99.0%	98.7%	87.3%	70.3%
Adjusted R-Squared in Unit Space	> 60%	98.7%	97.2%	97.5%	96.8%	70.4%	48.9%
Adjusted R-Squared	> 60%	98.7%	97.2%	97.5%	96.8%	70.4%	92.1%

- **Forced intercept is used to counteract unrealistically low values at low end predictions (<10K annual PMP Hrs)**
- **Forcing the intercept does cause our overall fit statistics to suffer slightly**
- **Although the regression utilizing all of the data has the highest adjusted r-squared, it also has significantly higher error terms**
- **Small program CER has the best overall error statistics, but falls in a very low range of data**
- **For most programs, the cubic regression excluding the outlier will be the best fit, but special considerations should be taken into account for each program**

The Progression of Regressions
CER Statistical Analysis – Result Segmentation

PMP Hours - Actual	Acq Supt - Actual	All Data vs Actual	All Data (FI) vs Actual	Excl Outlier vs Actual	Excl Outlier (FI) vs Actual	Small (Port) vs Actual	Small (FI) vs Actual
5,476	6,930	-157%	15%	-98%	-10%	15%	-26%
10,125	3,688	47%	284%	99%	206%	54%	156%
12,815	11,734	-11%	49%	-4%	20%	-31%	2%
13,434	11,511	1%	59%	6%	28%	31%	9%
14,150	16,911	-24%	13%	-22%	-8%	-7%	-22%
15,560	20,219	-24%	3%	-25%	-16%	-16%	-28%
16,339	20,640	-19%	5%	-21%	-14%	-14%	-26%
18,967	9,948	114%	149%	99%	105%	37%	78%
20,269	22,825	2%	15%	-6%	-5%	-7%	-17%
21,918	18,160	43%	54%	30%	28%	25%	12%
22,502	18,764	43%	52%	30%	27%	-11%	12%
92,747	84,050	-8%	-20%	-9%	-14%	-	-
96,986	76,702	2%	-12%	2%	-3%	-	-
171,494	88,589	-6%	-2%	4%	11%	-	-
222,720	113,280	8%	15%	-1%	-3%	-	-
272,640	235,200	-1%	-3%	-	-	-	-

For small data points, the intercept must be forced to avoid negative results

For large data points, forcing the intercept is not a concern

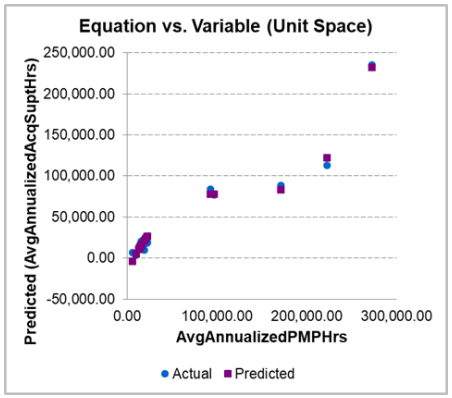
Legend
Actual Hours
Predicted v Actual ≥30%

- **Cubic regressions without forced intercept are unrealistically low even within the acceptable range of data**
- **Excluding Outlier Cubic Regression has the best overall fit although slightly high on low end and slightly low on high end; this is a great overall fit if not much is known about program**
- **If the project is known to be significantly small or significantly large, should consider using the Small Data CER forced int or All Data CER free int respectively**

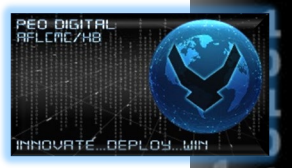
Summary

The Progression of Regressions
Summary

- Working with programmatic data is rarely a fairy tale
- You're often left working with small amounts of data or an imperfect spread of data points
- This case study is an ongoing analysis that we hope to refine with additional data points going forward which will be used to re-analyze the applicability of the proposed CERs
  - Emphasis is being placed on identifying additional medium-to-large scale data points for incorporation
- Analogies are great, but they are not the only option!
- We can still perform CER analysis that will provide us with better estimates than best guess analogies



The Progression of Regressions
Summary



- Although difficulties may arise when creating a CER, there are some strategies we can implement for combating some of the common issues that may arise:
  - **Limited Data Set**
    - Gather as much information on each data point as possible
    - View data in a variety of ways to see if a trend can be found
    - Adding in other explanatory factors may provide additional insight into data trends
    - Segmenting data may be necessary when determining cause of variation (small, medium, & large programs can have different explanations for their variation)
  - **Determining Regression Form**
    - Keep in mind most common regression shapes (Linear, Log-Linear, Learning, Beta, Quadratic, and Cubic)
    - Be careful not to over fit (ensure narrative of CER is rationalized)
    - Ensure potential outliers are analyzed and are not driving regression form
    - Analyze with and without forced intercept as well as analyze subsets of the data
  - **Assessing Best Fit**
    - Keep in mind that high r-squared does not always yield the best fit
    - Utilize rules of thumb for statistic results

The Progression of Regressions
Summary (Cont'd)

- Strategies Cont'd:
  - **Visualize CER Results**
    - Equation vs Variable in Unit Space
    - Plot of Actual vs Predicted in Unit Space
    - Compare predicted vs actual results and percent error across CERs to determine best fit at each subset of data (small, medium, large, and overall program sizes)
  - **Maximize CER Applicability**
    - Consider using different CERs for different scenarios
- Remember direct analogies may still be used over CERs if analogies can be made
  - Best practice is to include as much information and data points in the dataset as possible to allow for better analogy selection

Backup

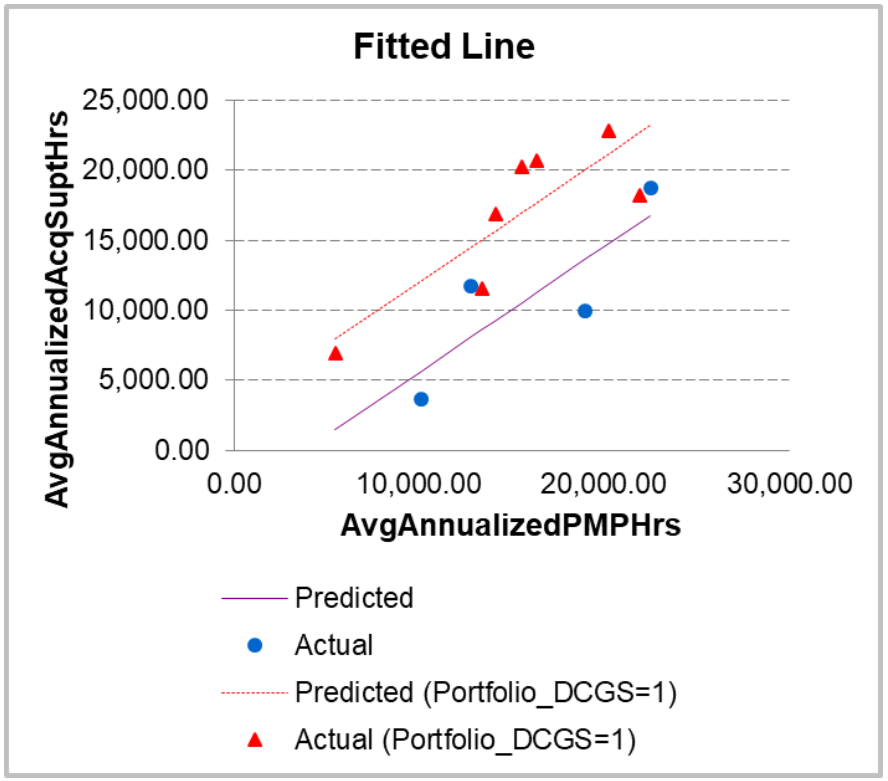
Dev/NRE Acquisition Support Study 2021

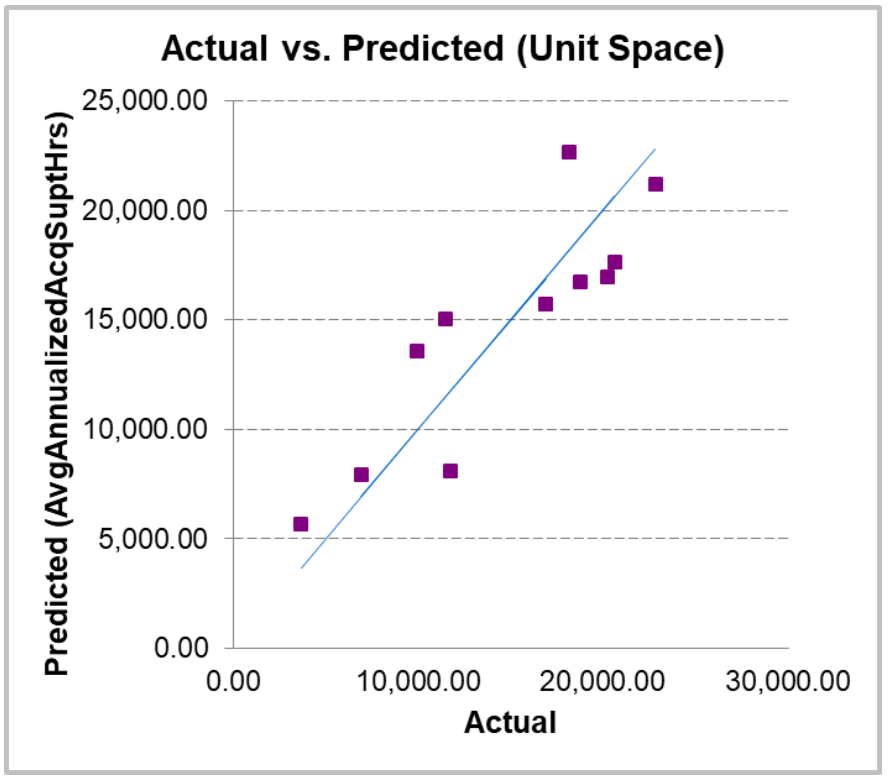
Annualized Average PMP Hours – CER

- **Linear Analysis – Small Programs, Portfolio Variable**

Range = ~5,000 to ~23,000 PMP Hours

Model Form:	Unweighted Linear model
Number of Observations Used:	11
Equation in Unit Space:	AvgAnnualizedAcqSuptHrs = (-3368) + 0.8944 * AvgAnnualizedPMPHrs + 6420 * Portfolio_DCGS





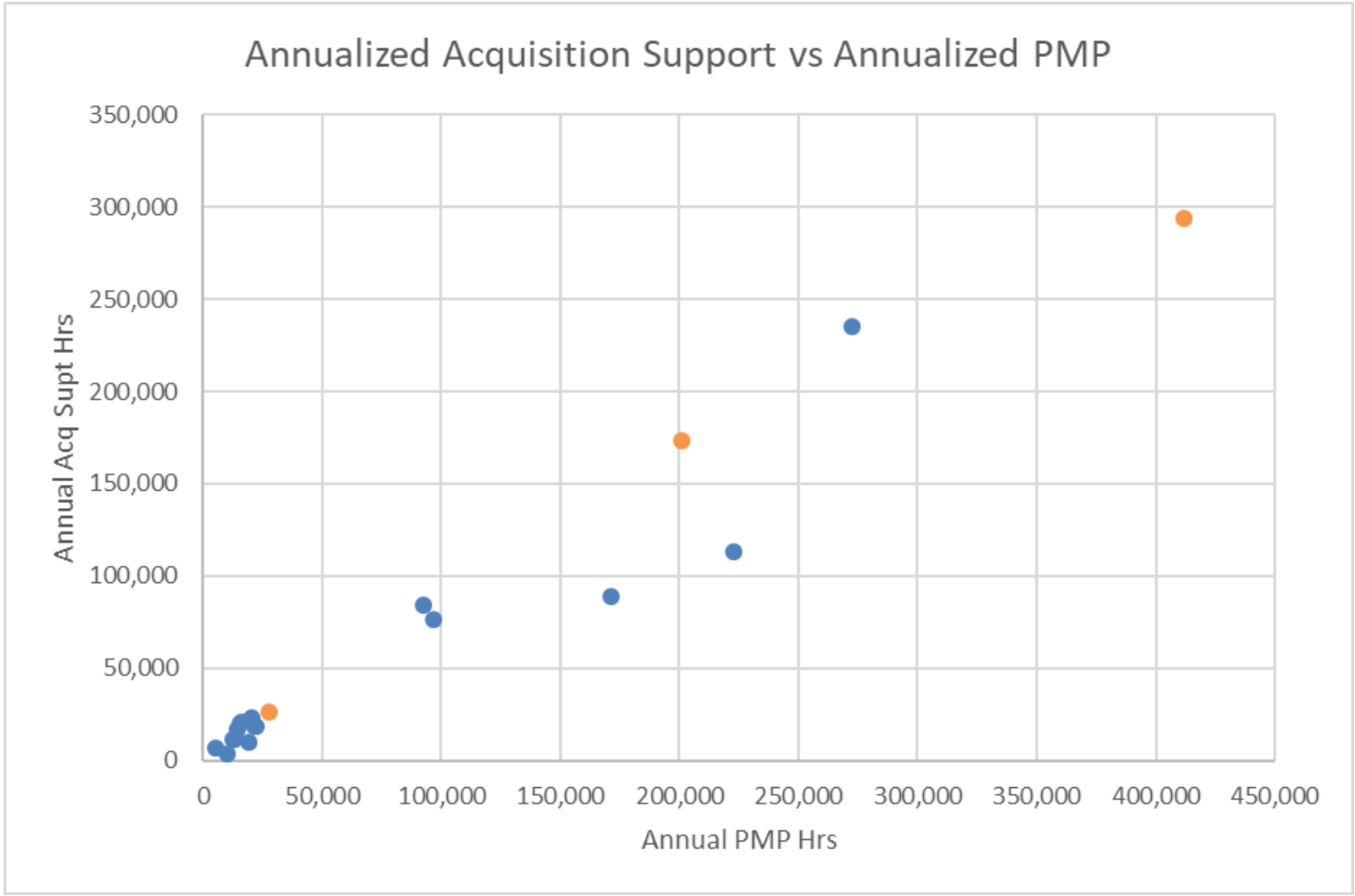
The Progression of Regressions
CER Analysis – Result Visualization / Segmentation

Legend
Actual Hours
Predicted v Actual ≥30%

PMP Hours - Actual	Acq Supt - Actual	Acq Supt - All Data	Acq Supt - All Data (Forced Int)	Acq Supt - Excl Outlier	Acq Supt - Excl Outlier (Forced Int)	Acq Supt - Small Data (Portfolio)	Acq Supt - Small Data (Forced Int)
5,476	6,930	-3,957	7,936	140	6,227	7,949	5,102
10,125	3,688	5,414	14,152	7,334	11,288	5,688	9,434
12,815	11,734	10,501	17,538	11,298	14,123	8,094	11,941
13,434	11,511	11,637	18,295	12,189	14,766	15,067	12,518
14,150	16,911	12,935	19,161	13,210	15,505	15,707	13,185
15,560	20,219	15,445	20,838	15,193	16,948	16,969	14,499
16,339	20,640	16,804	21,747	16,272	17,737	17,666	15,225
18,967	9,948	21,245	24,724	19,826	20,358	13,596	17,673
20,269	22,825	23,366	26,150	21,539	21,634	21,181	18,887
21,918	18,160	25,976	27,908	23,663	23,229	22,655	20,423
22,502	18,764	26,882	28,519	24,403	23,788	16,758	20,968
92,747	84,050	77,580	66,977	76,597	72,363	-	-
96,986	76,702	78,107	67,874	78,006	74,315	-	-
171,494	88,589	82,925	86,491	92,418	98,038	-	-
222,720	113,280	122,180	130,489	112,015	109,988	-	-
272,640	235,200	231,856	227,561	-	-	-	-

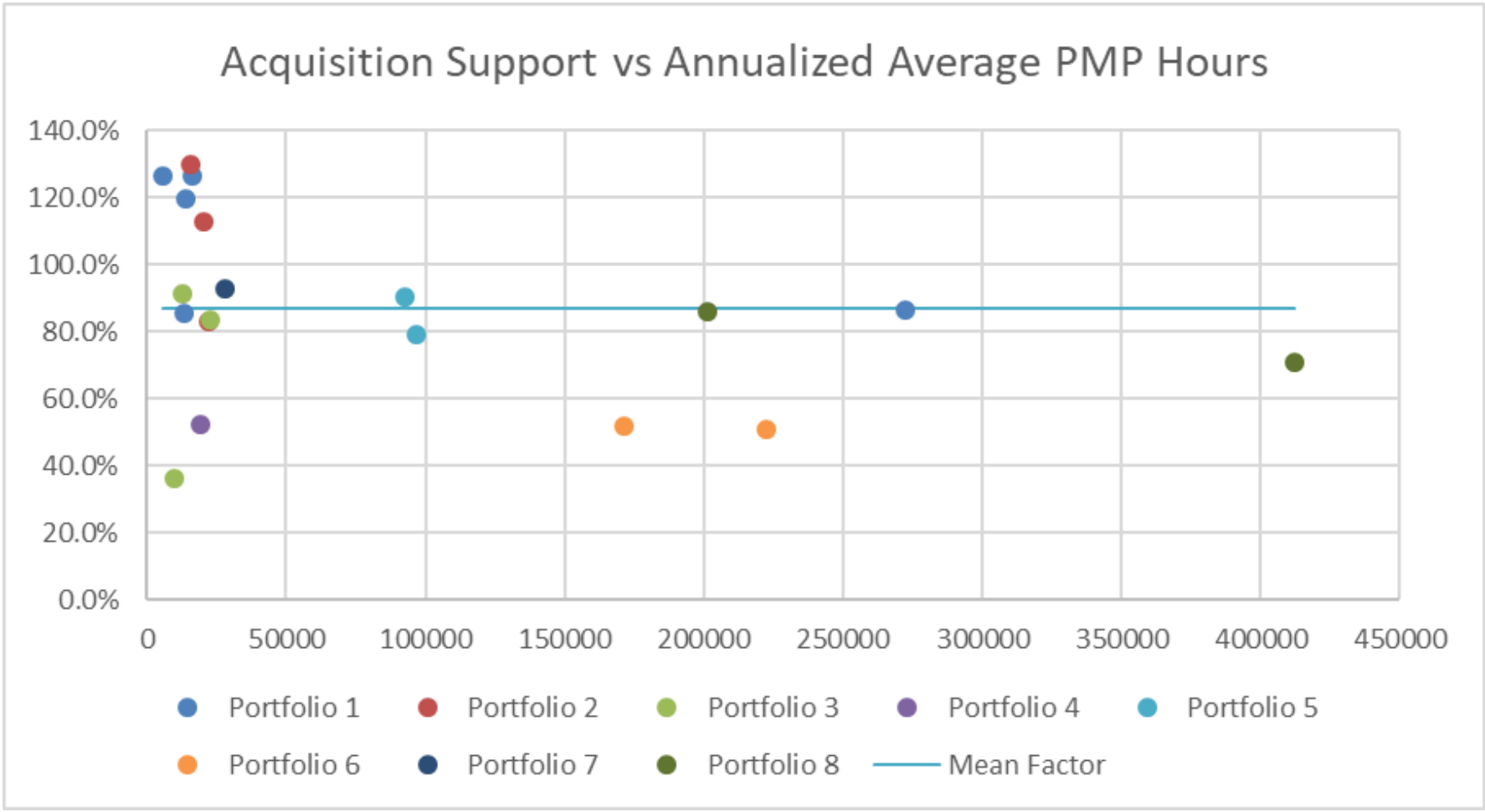
- **Cubic regressions without forced intercept are unrealistically low even within the acceptable range of data**
- **Excluding Outlier Cubic Regression with a forced intercept has the best overall fit although slightly high on low end and slightly low on high end**
- **Great overall fit if not much is known about program, but if the project is known to be significantly small or significantly large, should consider using the Small Data CER forced int or All Data CER free int respectively**

The Progression of Regressions
CER Form – Additional Data



- **Plotted view with 3 additional data points**
- **Middle and largest additions are projected PMP and Acq Supt as opposed to actual values**
- **Middle and largest additions are from another portfolio not previously included in study**

The Progression of Regressions
Data and Trend Analysis – Additional Data



- **Acq Supt % view with 3 additional data points**
- **Middle and largest additions are projected PMP and Acq Supt as opposed to actual values**
- **Middle and largest additions are from another portfolio not previously included in study (#8)**