Volume 13 | Issue 2 | August 2026
ISSN: 2160-4746

Journal of Cost Analysis and Parametrics

Editor in Chief: David L. Peeler, Jr., CCEA®

ICEAA
International Cost Estimating & Analysis Association
www.iceaaonline.com

Journal of
Cost Analysis
and Parametrics

Volume 13 | Issue 2 | August 2026 ISSN: 2160-4746

Editor's Note

David L. Peeler, Jr., CCEA®

Thank you for opening the Journal of Cost Analysis and Parametrics (JCAP). We welcome and appreciate your interest and attention. Foremost, we thank all who have written, submitted, and revised pieces for publication; and we appreciate those who gave their time and mental energy to review the submissions. The scope of papers in this issue detail professional work and academic pursuits in software sustainment, historical learning curves, early-stage aircraft estimating, Bayesian statistical modeling, predictive EVM analysis, and professional workforce development. We hope you find something, if not everything, of interest.

We begin with an insightful look into the complex world of military software sustainment. In *An Analysis of Military Software Maintenance Data, Benchmarks and Trends*, James Reichwein, Lieutenant Timothy L. Nelson, and Robert Sherbs analyze over a thousand records from the Software Resources Data Report (SRDR) repository. Their work establishes crucial empirical benchmarks, revealing a median rate of 2.66 software changes per month at an effort of 147 hours per change, and highlights how workload varies significantly across different domains, languages, and contractors.

The second article explores the enduring nature of aircraft manufacturing. In *When History Rhymes: Cost Estimating Lessons from Seven Legacy Aircraft*, Brent M. Johnstone delves into the manufacturing histories of seven legacy aircraft, including the F-102, F-106, and P-3 Orion. He demonstrates that decades-old programs offer enduring lessons on how design changes, production gaps, rate fluctuations, and technology adoption create quantifiable impacts on the learning curve, providing a valuable framework for today's estimators.

Next, we move to the challenges of early-stage acquisition. First Lieutenant Austin J. Romanov, Dr. Robert D. Fass, Dr. Jonathan D. Ritschel, Major Michael J. Brown, and Shawn Valentine present *Parsimonious Cost Estimating Relationships for Early-Stage Aircraft Cost Prediction*. They develop a transparent and practical cost estimating relationship (CER) for use when detailed data is sparse. Focusing on modern aircraft, their final model demonstrates that two simple, observable variables – weight and stealth capability – can explain nearly 85% of the variation in recurring flyaway costs, offering a powerful tool for early affordability and trade-space analysis.

Article four is a deep dive into statistical methodology. In *Using Bayes' Theorem to Develop CERs – Extending the Gaussian Model*, Dr. Christian B. Smart provides a vital extension of his previous work on applying Bayesian methods to cost estimating, particularly in common low-data environments. The paper masterfully shows how to move beyond simplified assumptions by modeling unknown variance and incorporating the Student's t-distribution, resulting in more realistic and accurate predictive models for everything from satellite costs to program risk.

Anna B. Peters and Dr. Mark W. Hodgins provide a modern take on Earned Value Management in the fifth piece, *Advanced EVM Analysis & Management: Using Time Series Forecasts and Regression Models*. This article modernizes EVM by applying powerful time series forecasting techniques to newly digitized data from IPMDARs. The authors demonstrate how ARIMA and macroeconomic regression models can predict future cost and schedule performance, offering program analysts a critical forward-looking view to identify emergent risks and opportunities.

We close this edition with a look inward at the professional development of our workforce. *Perception vs Reality: Determinants of AFIT Graduate Education Pursuit in the Financial Management Officer Corps* by Capt Benjamin L. Ainsworth, Dr. Jonathan D. Ritschel, Dr. Robert D. Fass, and Major Michael J. Brown investigates the factors that shape an officer's decision to pursue a cost master's degree from the Air Force Institute of Technology (AFIT). Using survey data and structural equation modeling, the authors find that leadership support and perceptions of career progression – rather than structural barriers like service commitments – are the most significant drivers of an officer's intent to apply.

May you find something of interest in these pages and the overview a useful focus for your reading selections. Enjoy the pieces you choose to read. We trust you will find these topics helpful in your professional pursuits and satisfying to your personal interests.

An Analysis of Military Software Maintenance Data, Benchmarks and Trends

Mr. James Reichwein

Lieutenant Timothy L. Nelson

Mr. Robert Sherbs

Software sustainment constitutes a significant portion of lifecycle costs for modern Department of War (DoW) systems, yet it remains relatively unexplored in most cost estimating spaces. This study analyzes software maintenance activity using data from the Software Resources Data Report (SRDR) repository within the Cost Assessment Data Enterprise (CADE) to establish empirical benchmarks for sustainment workloads.

A dataset of 1,173 sustainment records from 2017–2025 was processed through a four-stage filtration methodology, producing a validated subset. Sustainment activity was normalized into monthly metrics to enable cross-program comparison and identify trends. The analysis focused primarily on Software Change Requests (SCRs) and associated labor hours, using robust descriptive statistics to account for skewed distributions and outliers.

Results indicate a median sustainment rate of approximately 2.66 SCRs per month and a median labor requirement of about 147 hours per SCR. Both metrics exhibit substantial variability, with mean values significantly influenced by infrequent, high-effort events and non-standard reporting methods. Additional findings show that sustainment activity varies across application domains, language, system size, and contractors, suggesting that no single model can adequately characterize all systems.

These results provide an initial, data-driven baseline for software sustainment and support the use of data-driven benchmarks in cost estimation. The study also highlights the need for enhanced SRDR standards to reduce variability and improve the reliability of future sustainment analyses.

INTRODUCTION

Modern DoW systems rely heavily on complex software capabilities to support operational effectiveness, decision-making, and system integration. Over the past several decades, software has evolved from a supporting component of weapon systems to a critical enabler of mission capability across domains including command and control, mission planning, intelligence processing, and embedded system control. As the size and complexity of software systems have increased, the effort required to sustain and maintain these systems throughout their operational lifecycle must be monitored and understood. Software sustainment activities – including implementing updates, correcting defects, integrating new capabilities, and ensuring interoperability with evolving hardware and operational environments – represent a significant portion of the total lifecycle cost of modern defense systems (Maurer et al., 2019). While extensive research and numerous parametric models exist to guide estimators in predicting software development schedules and costs, far fewer empirical models or open datasets have historically been available to analyze the post-production maintenance phase (Cannon, 2007).

Despite the importance of software sustainment, developing reliable cost estimates for sustainment activities remains a persistent challenge for analysts and program managers. Many existing cost estimation models focus primarily on the development phase of software systems, while sustainment activities are often more difficult to quantify due to variability in system architecture, operational demands, and reporting practices across programs. As a result, the defense cost analysis community has increasingly sought to report reliable empirical datasets that can provide insight into the real-world behavior of software sustainment workloads (Jones et al., 2015). Such data can support the development of improved Cost Estimating Relationships (CERs), benchmarking metrics, and analytical frameworks for evaluating sustainment resource requirements.

One of the primary sources of software development and sustainment data within the DoW is the SRDR system. SRDR submissions are required for many major acquisition programs and are intended to capture detailed information about software development effort, sustainment activities, system characteristics, and associated labor resources. These reports provide a structured mechanism for collecting software-related data across programs and contractors, creating a potentially valuable dataset for analysts seeking to understand patterns in software development and maintenance activities. Over time, SRDR submissions have accumulated into a substantial database covering a wide range of defense software systems across multiple application domains. However, an in-depth understanding of the software maintenance element of reporting is still less understood than its development counterpart.

The dataset analyzed in this study was derived from SRDR sustainment reports submitted by various defense programs and contractors. These records include information on SCRs, labor hours associated with implementing those changes, reporting timelines, programming languages, software size metrics, and system domain classifications. Collectively, these variables provide an opportunity to examine how sustainment workloads manifest across different types of software systems and operational environments. However, as with many large administrative datasets, the raw SRDR data requires careful processing and validation before it can be used for meaningful statistical analysis, such as the analysis DoW cost estimators would be interested in.

The purpose of this analysis is to examine sustainment activity patterns within the SRDR dataset and identify empirical metrics that can help characterize software maintenance workloads. Specifically, the study seeks to quantify how frequently SCRs occur, estimate the typical engineering effort required to implement those changes, and explore how sustainment activity varies across factors such as system domain, programming language, and software size. By translating raw reporting data into normalized and interpretable metrics, the analysis aims to provide insights that may inform future cost estimation and sustainment planning efforts.

To accomplish this objective, the analysis employs a structured methodology that includes data filtration, normalization of reporting durations, and descriptive statistical analysis. The filtration process removes records that are incomplete, duplicated, or logically inconsistent, producing a validated subset of the dataset suitable for analysis. Normalization techniques are then applied to account for differences in reporting periods across programs, enabling comparisons of sustainment activity on a consistent monthly basis. Finally, descriptive statistical methods are used to examine distributions, central tendencies, and variability across the key metrics of interest.

The results of this study provide an exploratory view into how software sustainment workloads appear within the SRDR dataset. While the findings are not intended to serve as definitive CERs, they offer preliminary benchmarks and observations that may be useful to analysts within the defense cost estimation and operations research communities. In particular, the analysis highlights patterns in software change activity, labor effort associated with sustainment tasks, and the influence of system characteristics on maintenance workloads.

By examining these patterns, the study contributes to a broader effort within the DoW to improve the empirical foundation for software cost analysis and lifecycle resource planning. As software continues to grow in both complexity and strategic importance, understanding the drivers of sustainment effort will remain an essential component of effective program management and long-term defense capability planning.

BACKGROUND

The DoW relies on the Cost and Software Data Reporting (CSDR) framework as a primary mechanism for collecting standardized cost, technical, and resource data across its acquisition programs. As seen in Figure 1, the CSDR is

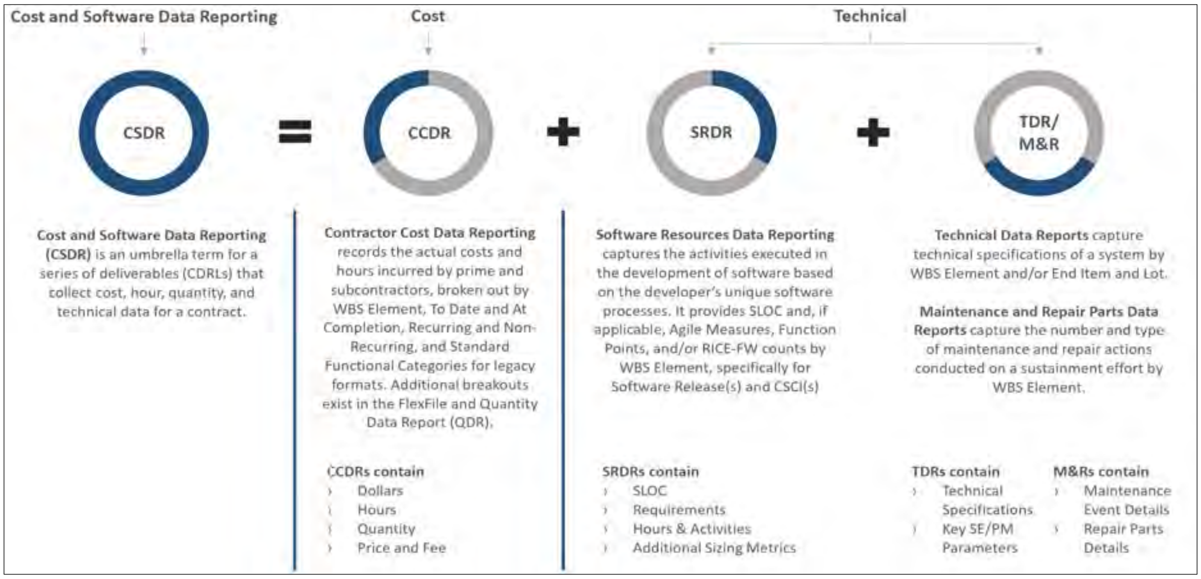


Figure 1: CADE's breakdown of CSDR Deliverables

composed of three principal elements: Contractor Cost Data Reporting (CCDR), which captures cost and production data for hardware and procurement activities; SRDR, which focuses on software-intensive development and sustainment efforts; and Technical Data Reports and Maintenance & Repair (TDR/M&R), which document system-level technical specifications and maintenance characteristics by Work Breakdown Structure (WBS) (Defense Acquisition University [DAU], 2023). The implementation thresholds and data governance policies for these deliverables are strictly regulated under formal Department cost collection mandates (U.S. Department of Defense, 2021).

Among these components, the SRDR serves as the central repository for software-related data, capturing key metrics such as Source Lines of Code (SLOC), software size, labor hours, and sustainment activities. This data is maintained within CADE, where it is accessible to government analysts for cost estimation and program evaluation purposes. The utilization of this historical archive is central to establishing standardized, defensible baselines across major defense programs (U.S. Department of Defense, 2020).

As modern defense systems have evolved, software has become the central enabler of mission capability, functioning as the central nervous system of weapon systems and operational platforms (Clark et al., 2017). Its role spans a wide range of domains, including command and control, intelligence processing, mission planning, and embedded system control. This growing reliance on software has shifted the cost structure of defense systems, with a significant portion of Total Ownership Cost (TOC) now occurring in the post-production phase. While sustainment costs vary by program, it is estimated that upwards of 80% of a system's software lifecycle costs are incurred during sustainment, where activities include defect correction, capability enhancement, system adaptation, and cybersecurity updates.

This shift has introduced a critical challenge for the DoW cost community. While established CERs exist for software development, comparable models for software sustainment remain underdeveloped. Sustainment activities are inherently different from development, as they are driven less by new code creation and more by the continuous implementation of SCRs, adaptation to evolving operational environments, and maintenance of legacy systems. Rather than relying strictly on static structural size metrics, contemporary defense frameworks emphasize monitoring the velocity of these software modifications and engineering interventions to accurately measure ongoing system support (Sams, 2011). These factors introduce significant variability and complexity, making sustainment costs more difficult to predict and manage.

The need for improved sustainment cost estimation is further underscored by broader cost and workload trends. Government Accountability Office (GAO) data indicates software sustainment costs nearly doubled from $429 million in 2012 to $828 million in 2018, with projected expenditures exceeding $15 billion across the Future Years Defense Program (FYDP) (Maurer et al., 2019). At the same time, the volume of sustainment activity has increased dramatically; for example, the Army experienced a 591% growth in software release requirements over a 14-year period (Lanham, 2013). These trends highlight the growing financial and operational significance of software sustainment and the risks associated with inadequate cost visibility.

Despite the availability of SRDR data, analysts face substantial challenges in leveraging this resource for sustainment cost estimation. Historically, SRDR sustainment reporting has been characterized by inconsistent data standards, incomplete labor reporting, and a lack of standardized definitions for maintenance activities. These hindrances have created a data fog that limits the ability of analysts to develop reliable benchmarks and defensible CERs. As a result, cost estimates for sustainment are often based on limited empirical evidence, increasing the risk of budget instability, cost growth, and reduced program readiness.

This study is motivated by the need to address these limitations by transforming raw SRDR sustainment data into a structured, smaller consistent subset. By filtering inconsistent records, normalizing reporting periods, and extracting key performance metrics, the research aims to establish an empirical foundation for understanding software sustainment workloads. Ultimately, this effort seeks to improve the transparency, consistency, and predictive capability of software sustainment cost estimation, enabling more effective resource planning and lifecycle management across the DoW.

PURPOSE

The primary purpose of this study is to establish a defensible, data-driven baseline for software sustainment costs by rehabilitating and analyzing the DoW's historical archive of SRDRs. While the SRDR repository contains over two decades of software development data, its sustainment-specific reporting has only matured over the last eight years. This relatively recent shift in reporting requirements has resulted in a dataset that is both nascent and inconsistent, particularly regarding the reporting of SCRs. This research seeks to bridge the gap between raw, unverified data and actionable thresholds for software sustainment.

Specifically, this study aims to achieve three core objectives:

1. Metric Isolation and Validation: To identify the specific variables, primarily Monthly Labor Hours and SCRs, that serve as the most reliable indicators of sustainment effort and to validate these metrics through a rigorous four-stage filtering process.
2. Benchmark Development: To derive medians and standard deviations for sustainment effort, providing analysts with initial thresholds to evaluate the reasonableness of contractor-submitted maintenance estimates.
3. Policy and System Reform: To propose concrete structural corrections to the current SRDR Data Item Description (DID) (DI-MGMT-82035A) and the CADE ingestion logic. These recommendations are designed to prevent future data degradation and ensure that the SRDR ecosystem evolves into an audit-ready asset.

Ultimately, this report serves as a foundational baseline for the DoW. By transforming a fragmented dataset into a curated portfolio of 524 records with consistently reported data, this study provides the quantitative evidence necessary to move the DoW away from subjective estimation and toward a standardized, repeatable, and defensible sustainment budgeting process.

SCOPE

The scope of this study is focused exclusively on post-production software sustainment records within the CADE/SRDR environment. The analysis initiated with a total population of 1,173 records spanning 2017–2025 and is current as of 23 September 2025.

The study's boundaries are defined by the following parameters:

- Data Source: Limited to contractor-submitted SRDRs via the CADE database.
- Metric Focus: Primary analysis is restricted to Monthly Labor Hours and SCR Generation Rates as the leading indicators of effort.
- Exclusions: This study excludes initial software development (build) data, hardware-centric maintenance, any records assumed to be active, and other records that failed the project's derived integrity gates.

ANALYTICAL APPROACH

DATA PROCESSING AND ANALYTICAL EXECUTION

The analysis conducted in this study relies on software sustainment data collected from CADE through the SRDR sustainment database. The SRDR database serves as the DoW's primary mechanism for collecting standardized software development and sustainment information across acquisition programs. Sustainment reporting focuses on ongoing software maintenance activities after initial system deployment, capturing metrics such as SCRs, labor hours, system size, programming languages, and release timelines.

The dataset used for this analysis represents a snapshot of the SRDR sustainment database current as of September 23, 2025. Because SRDR data is submitted by multiple programs and contractors, the raw dataset contains substantial variability in reporting quality, format consistency, and completeness. Consequently, the analytical process required a structured data preparation methodology designed to isolate a subset of records suitable for statistical analysis.

The objective of the methodological approach was to transform a heterogeneous administrative dataset into a normalized analytical dataset capable of supporting comparisons across programs with differing reporting structures and timeframes. This required several stages of filtration, standardization, and statistical processing before analysis could proceed.

To that end, the first phase of the methodology consisted of a structured data filtration pipeline designed to eliminate records that could introduce logical inconsistencies or statistical distortion. This

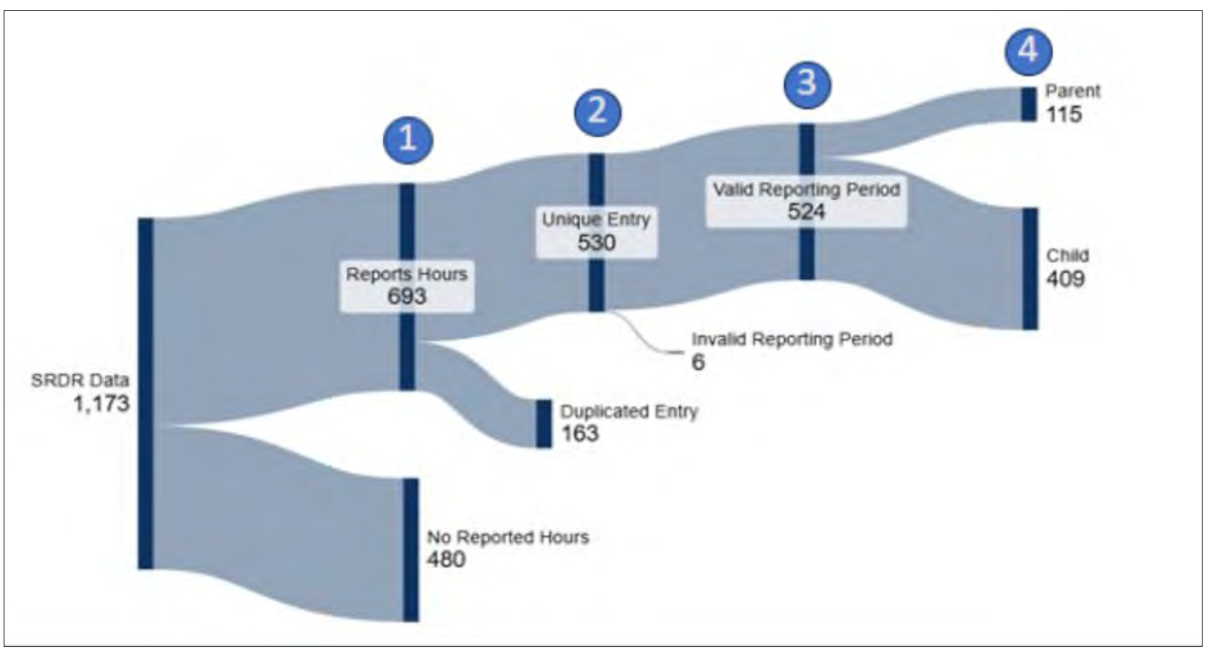


Figure 2: Sankey chart depicting filtration process from raw SRDR dataset to usable data points

pipeline was necessary due to the presence of incomplete, duplicated, or mathematically infeasible records within the raw SRDR dataset.

The filtration process seen in Figure 2 followed four primary validation steps:

1. Hours reported validation: Each record was evaluated to determine whether it contained reported sustainment labor hours. Due to proprietary restrictions and the absence of direct financial values within the archive, labor hours are utilized as the primary proxy of sustainment effort. Records missing reported labor hours were excluded from further analysis because they could not contribute to effort-based metrics.

2. Duplicate entry check: The dataset was evaluated for duplicate entries. Duplicate records can occur when identical software metrics are submitted multiple times across reporting periods or release submissions. When duplicates were detected, only the first instance of the record was retained to prevent inflation of statistical counts.

3. Valid reporting period: The dataset was screened for valid reporting timelines. To ensure the study is based on completed and verified effort, records lacking "Report As Of" or "Release End Date" markers were assumed to be active and were excluded to prevent the inclusion of incomplete data cycles. Additionally, records indicating zero-length or negative reporting durations were removed because these represent mathematical impossibilities within the context of release reporting. A valid release record must have a positive time interval between the start and end of the reporting period.

4. Parent/child WBS relation: Records were categorized based on WBS hierarchy, distinguishing between parent and child SCRs. Parent SCRs represent the highest-level change request within a WBS structure, while child SCRs represent subordinate changes associated with the parent request. Identifying this relationship helped prevent double counting of change activity during aggregation.

This filtration process produced a “final dataset,” a subset of records that passed all validation gates and could support defensible statistical analysis. This data subset is assumed to be statistically representative of broader governmental software sustainment trends. The parameters governing this clean-up process are derived from historical data collection boundaries and quality assessment frameworks established by OSD data governance offices (Lanham et al., 2018; Office of the Secretary of Defense, 2019). To further mitigate undesirable variation within this subset, extreme outliers representing non-standard program events such as catastrophic architectural failures were removed to ensure the stability of regression equations and to focus benchmarks on typical sustainment throughput.

Data Normalization

Following filtration, the next methodological step involved normalizing the dataset to allow meaningful comparisons across programs with differing reporting durations. SRDR submissions frequently cover reporting periods of varying lengths, ranging from several months to multiple years. Directly comparing cumulative values, such as total hours or total SCRs across these varying durations, would lead to misleading interpretations. To address this issue, the analysis developed a normalization framework that converts sustainment activity into standardized monthly metrics.

The normalization process begins by converting the total number of days within a reporting period into a normalized month to allow sustainment metrics to be compared regardless of reporting duration:

Normalized Month = Number of Days / 30.4375

Using this normalization approach, the analysis derived a monthly labor burn rate, defined as:

Burn Rate = Total Reported Hours / Normalized Months

This burn rate provides a standardized measure of sustainment effort intensity across programs and reporting periods.

Programming Language	Count
Ada	46
Ada 95	58
Assembly	4
C	9
C#	35
C, C++	53
C++	79
Fortran	2
Java	46
Jovial	12
OGPII Custom Scripting Language	6
VB	8
XML	11
Missing	22
N/A, None	131

Table 1: Standardized list of primary programming languages reported within SRDR data

Standardization of Programming Languages

Another methodological challenge involved the standardization of programming language reporting. The raw dataset contained more than 30 distinct language labels, many of which represented minor spelling variations or alternate naming conventions for the same programming language. Examples include variations such as C++, C Plus Plus, and other variations.

To ensure statistical consistency, these labels were consolidated into 15 standardized programming language categories as seen in Table 1. This consolidation prevented fragmentation of the dataset across redundant categories and allowed language-specific analysis to be conducted with meaningful sample sizes.

The analysis also examined whether software sustainment effort varied by primary programming language. A Kruskal-Wallis test comparing hours per SCR across programming language categories produced a chi-square statistic of 158.55 with a p-value less than 0.0001. These results provide statistical evidence that hours per SCR differ significantly among programming language groups, indicating that programming language is associated with variation in sustainment effort and is unlikely to be independent of the labor hours required per SCR. In cases where multiple programming languages were reported for a single record, the first listed language was designated as the primary language for analysis. This approach provided a consistent method for assigning each record to a single language category.

Total Source of Lines of Code

The SRDR DID, specifically the DI-MGMT-82035A document, outlines in section 3.3.2.6.3 the actual delivered size of the software maintained by the prime and subcontractor and defines the appropriate source code categories. Table 2 outlines the defined source code categories to be provided in the SRDR data (U.S. Department of Defense, 2017).

SOURCE CODE CATEGORY	DEFINITION
NEW CODE	New code shall be partitioned into human-generated which is source code that is created manually by a programmer and auto-generated code.
REUSED CODE WITH MODIFICATION	Modified code is defined as pre-developed code that can be incorporated in the software component with a significant amount of effort, but less effort than required for newly developed code.
REUSED CODE WITHOUT MODIFICATION	Code reused without modification is code that has no design or code modifications. However, there may be an amount of retest required.
CARRYOVER CODE	Code developed in previous releases that is carried forward into the current release and delivered (also called Internal Reuse).
AUTO-GENERATED CODE	Auto-generated code is source code created automatically by software tools such as AI assistants, compilers, or modeling tools rather than being written manually by a developer
GOVERNMENT FURNISHED CODE	Indicate the amount of Government Furnished Code delivered to the contractor. If the Government did not provide the contractor any code, enter 0.

Table 2: Types of SLOC and their definition as defined in DI-MGMT-82035A

This study uses the following formula to calculate total source lines of code (TSLOC):

TSLOC = SLOC New Human Gen Actual To Date (ATD) + SLOC New Machine Gen ATD + SLOC Reused w Mod ATD + SLOC Reused w/o Mod ATD + SLOC Carryover w Mod ATD + SLOC Carryover w/o Mod ATD + SLOC Gov Furnished w Mod ATD + SLOC Gov Furnished w/o Mod ATD

The TSLOC equation used in this analysis is valuable because it provides a comprehensive measure of total software size by incorporating not only newly developed code but also reused, modified, and carried-over components. This approach reflects the reality of modern software systems, where a significant portion of functionality comes from existing code rather than entirely new development. Since software size is considered a primary driver of cost and effort in most estimation models, capturing these distinctions improves the accuracy and comparability of results. While the concept of adjusting SLOC to account for reused and modified code is widely accepted across the software industry, the specific TSLOC formulation used in SRDR is not universal. Nevertheless, utilizing empirical sizing aggregates in this manner aligns with established parametric cost-estimation frameworks designed to isolate baseline software maintenance effort drivers (Clark & Madachy, 2015).

Statistical Methods

Given the presence of extreme variability within the dataset, the analysis relied heavily on robust descriptive statistics rather than simple averages. Specifically, the study emphasized metrics such as:

- Median values
- Interquartile Ranges (IQR)
- Distributional shape
- Standard deviation

Median values were used as the primary measure of central tendency because they are less sensitive to extreme outliers than arithmetic means. This approach helps provide a more representative measure of typical sustainment behavior within the dataset. Software maintenance metrics across highly heterogeneous, multi-contractor portfolios frequently demonstrate these non-linear characteristics and extreme scale variances, necessitating robust, distribution-free statistical treatments (Sharma & Chaudhary, 2023).

Descriptive statistical analysis was used to explore several key sustainment metrics, including:

- SCR generation rates
- Labor hours per SCR
- Sustainment effort across software domains
- Staffing levels relative to system size
- Sustainment effort by programming language
- Contractor reporting variation

These metrics collectively form the basis of the results and findings presented in the following sections.

ANALYSIS RESULTS

This study conducted an exploratory analysis of software sustainment activity using data derived from the CADE SRDR database. The research was motivated by the need to better understand the empirical characteristics of software sustainment workloads, particularly in the absence of widely accepted, data-driven CERs for post-deployment software maintenance.

The analytical approach focused on transforming a heterogeneous administrative dataset into a structured and interpretable set of sustainment metrics. The research emphasized transparency in data conditioning, careful normalization of reporting periods, and the use of robust statistical techniques to account for skewed distributions and outliers. Rather than attempting to construct a predictive cost model, the study aimed to characterize observable sustainment behavior and identify patterns that could inform future cost estimation practices.

The resulting analysis is best understood as a descriptive and exploratory examination of sustainment activity, designed to answer key questions, such as:

- How frequently do software systems undergo change during sustainment?

- How much labor effort is typically required per change?
- How do sustainment characteristics vary across system domains and sizes?
- What level of variability exists across programs and reporting entities?

These questions guided the structure of the analysis and informed the selection of metrics presented in the results.

Key Findings: Change Activity

The descriptive statistics as shown in Table 3 for completed SCRs per month reveal a distribution that is highly skewed, widely dispersed, and dominated by a small number of extreme values. One of the central findings of the analysis relates to the rate at which software systems undergo change during sustainment. After normalizing for reporting duration, the data indicates that software systems experience a median rate of approximately 2.66 SCRs per month. This highlights the continuous nature of software sustainment, where mature systems require regular updates due to defect correction, cybersecurity needs, evolving operational requirements, and system integration.

Completed SCRs per Month	
Mean	14.67
Standard Error	2.06
Median	2.66
Mode	115.90
Standard Deviation	39.65
Sample Variance	1572.36
Range	242.21
Minimum	0
Maximum	242.21
Sum	5429.51
Count	370

Table 3: Descriptive Statistics for Completed SCRs per Month

The mean value of 14.67 SCRs per month is significantly higher than the median, indicating a strong positive skew. This result further supported by the large standard deviation (39.65) and wide range (0 to 242.21), which demonstrate substantial variability across reporting periods. As seen in Figure 3, most observations cluster at relatively low levels of activity, while a small number of periods exhibit extremely high change volumes. Within defense software reporting frameworks, these periodic surges are statistically recognized as indicators of major block upgrades, consolidated capability releases, or the administrative resolution of accumulated maintenance backlogs (Lanham et

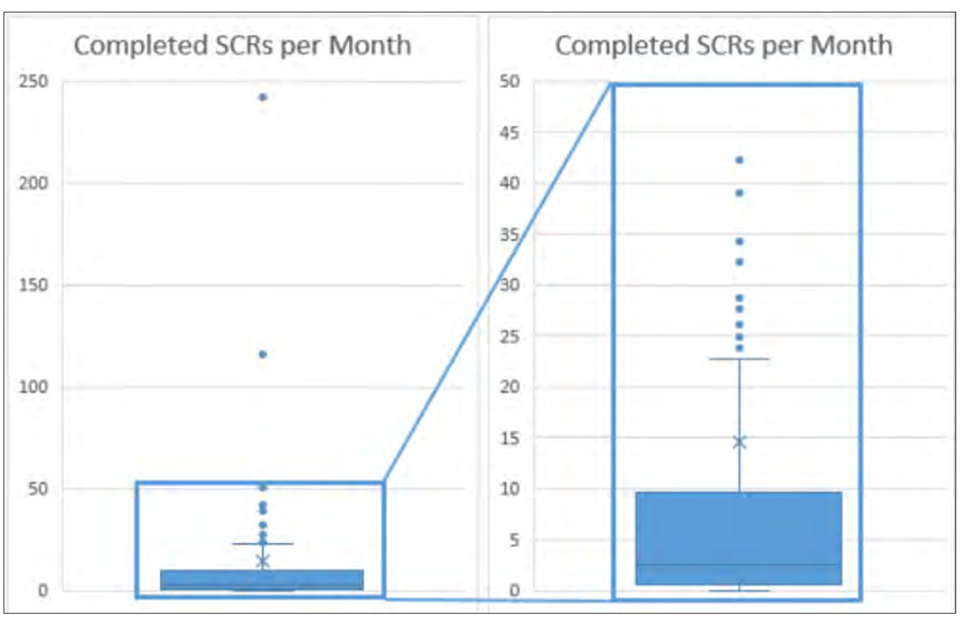


Figure 3: Box and whisker plot of Completed SCRs per Month depicting high skew due to outliers

Hours per SCR	
Mean	2305.46
Standard Error	603.23
Median	146.89
Mode	210
Standard Deviation	11634.75
Sample Variance	135367491.9
Kurtosis	107.35
Skewness	9.35
Range	161185.5
Minimum	0
Maximum	161185.5
Sum	857631.54
Count	372

Table 4: Descriptive Statistics for Hours per SCR

al., 2018). These spikes likely correspond to major software updates, system overhauls, or the resolution of accumulated backlogs.

This finding is particularly important for cost estimation because it shows sustainment activity cannot be accurately represented by average values alone. The episodic and highly variable nature of SCR activity means that reliance on mean-based metrics may overstate typical workload levels, at least based on findings within this dataset. Instead, cost estimators may wish to emphasize median values and consider the potential for infrequent but high-impact surges in activity when planning resources and budgets.

Key Findings: Labor Effort per Change

A second major finding concerns the labor effort required to implement individual software changes. Similar to Completed SCRs per Month, the descriptive statistics as shown in Table 4 for labor hours per SCR indicate a distribution that is extremely skewed, highly variable, and dominated by a small number of very large values.

The analysis identified a median value of approximately 147 labor hours per SCR, which represents the central tendency of effort across the dataset and provides a meaningful benchmark for sustainment planning. This outcome suggests even routine software changes require a substantial level of effort, reflecting the complexity of modern systems and the need for testing, integration, and validation.

The mean value of 2,305 hours per SCR is nearly 16 times greater than the median, indicating a strong positive skew driven by extreme outliers. Figure 4 graphically depicts this skewness and highlights the IQR, showing the middle 50% of the data lies between 11.4 hours (first quartile) and 594.6 hours (third quartile). This spread is further supported by the very large standard deviation (11,634.75), extreme range (0 to 161,185.5 hours), and high skewness (9.35), all of which confirm the presence of a long-tailed distribution with significant outliers. These extreme values likely represent major system modifications, large-scale integrations, or complex updates requiring extensive engineering effort.

This substantial variability highlights that not all SCRs are equivalent in scope or effort. Some changes require minimal work, while others demand hundreds or even thousands of labor hours depending on system complexity, change type, and program-specific processes. As a result, treating all SCRs as uniform units introduces uncertainty into the analysis. However, in the absence of more

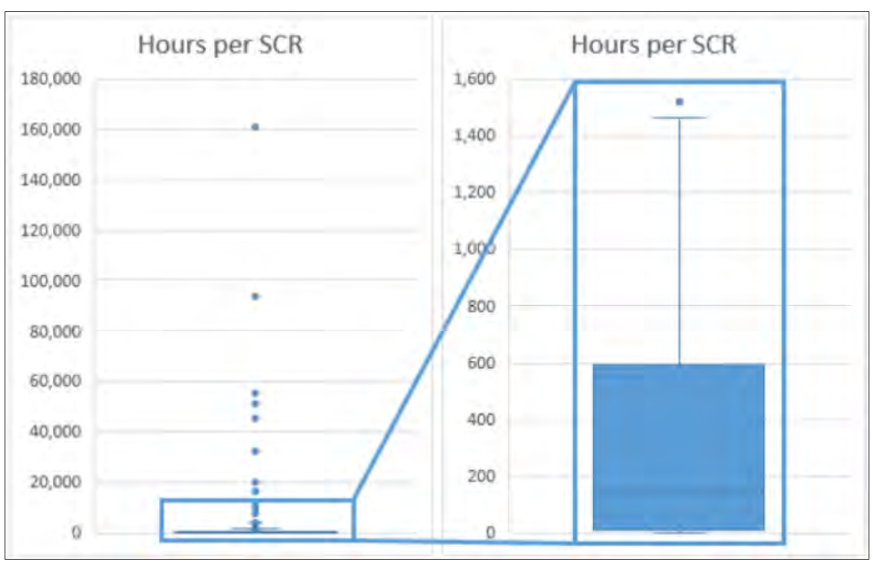


Figure 4: Box and whisker plot of Hours per SCRs depicting high skew due to outliers

detailed classification data, this simplifying assumption provides a practical basis for high-level analysis. Importantly, these findings reinforce that mean-based metrics alone can be misleading and that median values, along with an understanding of variability, are critical for accurately characterizing sustainment effort.

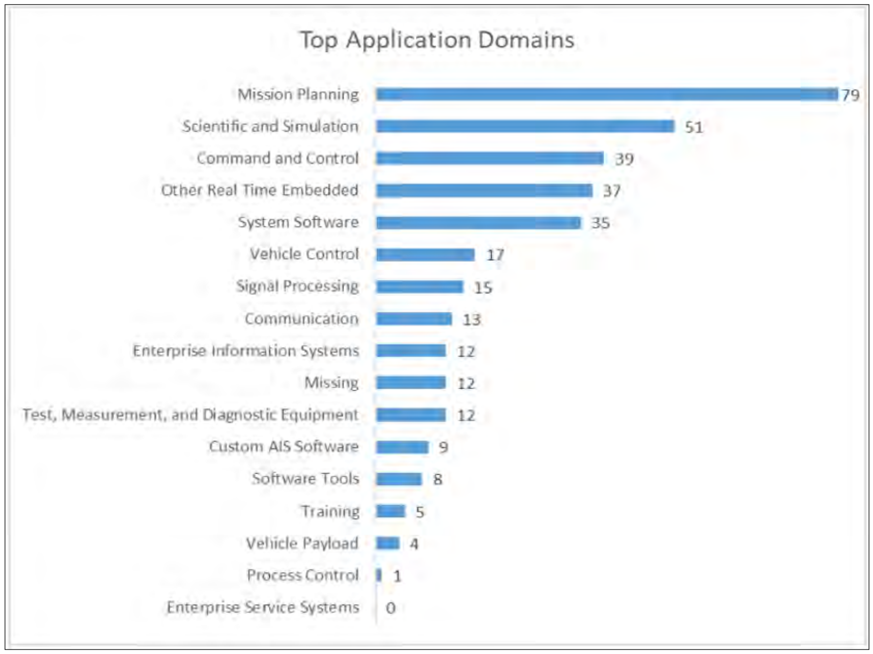


Figure 5: Application Domain count for analyzed subset

Domain-Level Observations

There are several necessary terms to understand for domain -level observations: A super domain is a broad top-level grouping used to organize related reporting metrics or data fields based on software type and purpose (See Appendix A). A domain is a smaller more specific grouping native to the SRDR reporting data. As broad groupings of related application domains with similar software characteristics and purposes, super domains enable analysts to compare sustainment activity across different types of systems such as real -time, engineering, or information systems without losing consistency or context. This organization is valuable when working with large, heterogeneous datasets found in SRDR. By grouping related domains together, super domains may allow analysts to identify patterns, trends, and differences in sustainment behavior at an aggregate level, potentially improving the reliability of statistical analysis and cost estimation.

The analysis also examined how sustainment activity varies across software super domains, including engineering systems, real-time systems, automated information systems, and support software. A Kruskal-Wallis test comparing SCR rates across super domains yielded a chi-square statistic (χ^2) of 14.84 with a p-value of 0.0051. This result provides evidence that SCR rates differ significantly among software super domains, indicating that sustainment activity is associated with system type and is unlikely to be independent of domain classification.

Similarly, a Kruskal-Wallis test examining the total number of implemented changes across software super domains produced a statistically significant result ($\chi^2 = 58.21$, $p < 0.0001$). This finding indicates that the volume of sustainment activity varies significantly across domain categories, providing additional evidence that software sustainment workloads are influenced by system characteristics and cannot be adequately represented by a single uniform model.

As shown in Figure 5, engineering and real-time systems dominate the dataset in both the number of records and the overall volume of sustainment activity. For example, within the real-time domain, Command and Control (39 records) and Other Real-Time Embedded systems (37 records) represent substantial portions of the data, along with additional contributions from Vehicle Control (17 records), Signal Processing (15 records), and Communication systems (13 records). Similarly, the engineering domain includes significant representation from Scientific and Simulation systems (51 records) and System Software (35 records), with additional contributions from Test, Measurement, and Diagnostic Equipment (12 records).

Automated information systems are also well represented, primarily driven by Mission Planning systems (79 records), which constitute the single largest application domain in the dataset, along with Enterprise Information Systems (12 records) and Custom AIS Software (9 records). In contrast, support software domains are less prevalent, with relatively small counts for Software Tools (8 records) and Training systems (5 records).

These distributions highlight meaningful variation in how sustainment activity is concentrated across domains. The prominence of real-time and engineering systems, many of which operate in mission-critical or performance-sensitive environments, suggests inherently higher sustainment demands. Conversely, the comparatively lower representation of support systems may indicate different maintenance profiles or lower overall sustainment intensity. Taken together, these domain-level differences reinforce that a single, uniform sustainment cost model may not be appropriate across all system types, and that domain-specific adjustments are likely necessary to improve the accuracy and defensibility of sustainment effort estimates.

Relationship Between Software Size and Staffing

Another important aspect of the analysis involves the relationship between software size and sustainment staffing levels. As illustrated in Figure 6, the data indicates a positive relationship between TSLOC and the number of personnel assigned to sustainment activities.

This relationship aligns with established intuition and prior research suggesting larger systems require more resources to maintain. However, the analysis also reveals significant variability around this relationship, indicating that software size alone does not fully determine sustainment effort.

The analysis also examined the relationship between software size and sustainment staffing levels. A linear regression analysis was performed using Total Staff (Full-Time Equivalents) as the predictor variable and TSLOC as the response variable. Results indicate a statistically significant positive relationship between staffing levels and software size ($F = 94.36$, $p < 0.0001$). The estimated regression coefficient for Total Staff was 24,249 source lines of code per additional full-time equivalent ($t = 9.71$, $p < 0.0001$), suggesting that larger software systems generally require larger sustainment organizations.

While the relationship is statistically significant, the lack-of-fit test was also significant ($F = 29.20$, $p < 0.0001$), indicating that a simple linear relationship may not fully capture the complexity of the association between software size and staffing. This finding suggests that additional factors, such as system architecture, mission requirements, software complexity, and maintenance practices, may also

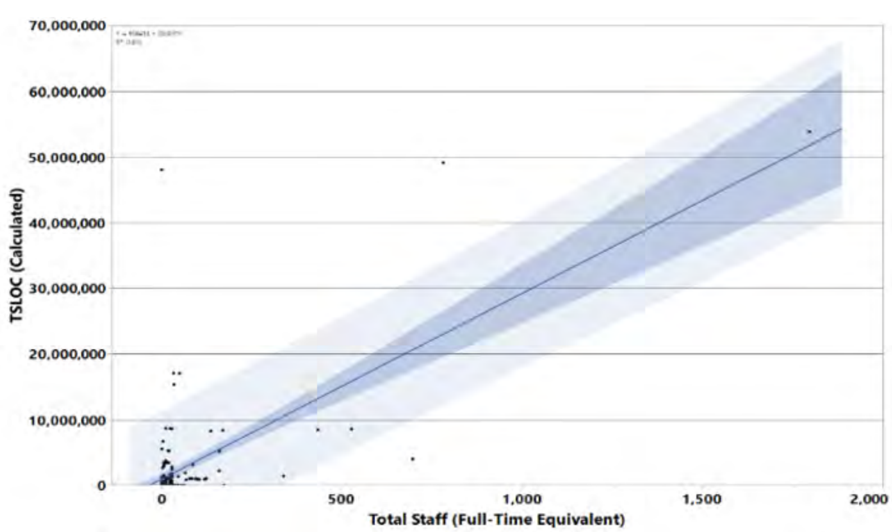


Figure 6: Relationship between Total Staff Full-Time Equivalent (FTE) and TSLOC

contribute to sustainment staffing requirements. Nevertheless, the results provide evidence that software size is an important factor associated with sustainment manpower levels and support the conclusion that sustainment effort varies according to system characteristics.

Contractor and Reporting Variability

The analysis also identified substantial variability across contractors and reporting entities. As illustrated in Figure 7, which presents a bar chart of changes implemented by reporting contractor alongside a box-and-whisker plot of total hours by reporting contractor, there are notable differences in both the volume of reported activity and the distribution of associated labor effort. Some contractors report significantly higher numbers of implemented changes, while others exhibit wide dispersion in total hours, indicating inconsistency in how sustainment effort is recorded and managed.

A Kruskal-Wallis test was conducted to evaluate whether sustainment activity differed among reporting contractors. For total labor hours, the test yielded a statistically significant result ($\chi^2 = 226.17$, $p < 0.0001$), indicating that reported sustainment effort varies significantly across contractors. Similarly, analysis of total implemented changes produced a statistically significant result ($\chi^2 = 184.72$, $p < 0.001$), demonstrating that the volume of reported sustainment activity also differs across reporting organizations. These findings reject the null hypothesis that contractors report comparable levels of sustainment effort and change activity, suggesting that contractor-specific factors influence reported sustainment workloads.

This variability highlights a broader challenge in the use of SRDR data for analytical purposes. Although the reporting framework provides a standardized structure, differences in reporting practices, interpretation of guidance, labor accounting methodologies, and implementation across programs may contribute to significant variation in reported results. Consequently, some of the observed differences may reflect not only true variations in sustainment workload but also inconsistencies in reporting behavior. These findings underscore the importance of continued efforts to improve data standardization, reporting guidance, and validation procedures within the SRDR framework in order to enhance the reliability and comparability of sustainment metrics.

Implications for Cost Estimation

The findings of this study have several important implications for software cost estimation within the DoW. First, the results establish empirical benchmarks for key sustainment metrics, including SCRs per month and labor effort per SCR, which

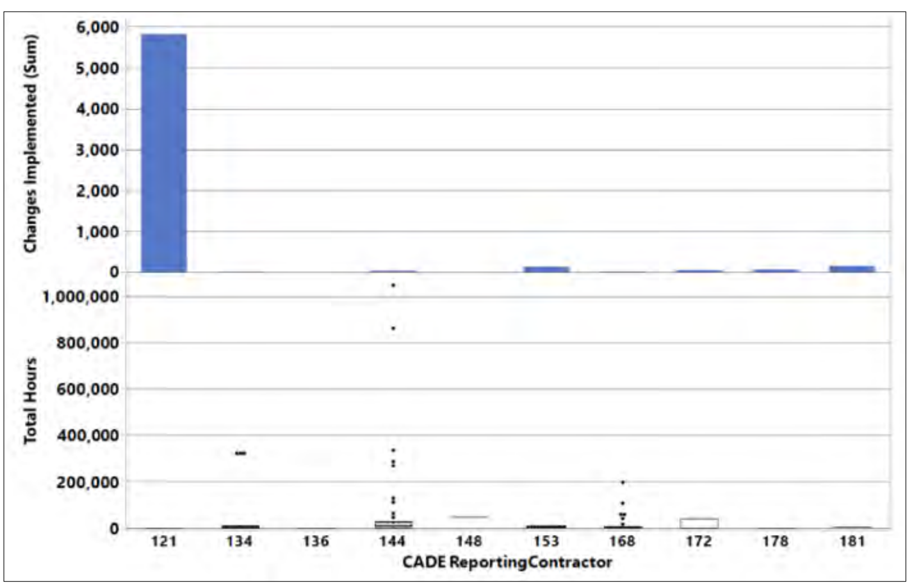


Figure 7: Changes Implemented and Total Hours by Reporting Contractor

can serve as valuable reference points for analysts developing and validating sustainment cost estimates.

Second, the analysis underscores the importance of explicitly accounting for variability and uncertainty in sustainment workloads. The wide distributions observed across the dataset indicate that reliance on single-point estimates may inadequately capture the full range of potential outcomes. Incorporating probabilistic methods or ranges may therefore improve the robustness and credibility of cost estimates.

Third, the results demonstrate that domain-specific and system-specific factors play a significant role in determining sustainment effort. Accordingly, cost estimation models should incorporate adjustments for relevant system characteristics, such as application domain, architecture, and operational context, rather than relying solely on generalized averages.

SENSITIVITY ANALYSIS

To evaluate the robustness of the study's findings, a sensitivity analysis was conducted across four primary vectors: data filtering constraints, distributional skew, uncertainty quantification, and reporting variability. This analysis ensures that the derived benchmarks reflect underlying sustainment behavior rather than artifacts of the analytical process.

Impact of Filtering and Duration Constraints

The study applied specific duration-based filters to exclude records with negative reporting periods or durations exceeding 2,500 days (indicative of entry errors or extreme aggregation). Sensitivity testing confirms that these exclusions had a negligible impact on central tendencies, including median SCR rates and burn rates. However, the filters significantly improved normalized metric integrity by preventing artificially low sustainment rates caused by undefined or inflated time horizons. This filtering confirms that the data salvage methodology enhances dataset reliability without introducing statistical bias.

Distributional Effects and Outlier Influence

The SRDR dataset exhibits a pronounced positive skew, where most records reflect low-intensity maintenance while a small subset captures large-scale modernization effort. A comparison of statistical approaches revealed that arithmetic means are highly sensitive to these high-change periods, resulting in inflated estimates.

Conversely, median-based metrics remained stable across varied data subsets, justifying the study's reliance on medians as the baseline for central tendency. To clarify the analytical framework: extreme outliers were intentionally retained in the primary descriptive dataset (including Table 4) to accurately preserve the full operational reality of the data. However, these extreme values were excluded only during specific model-fitting iterations to prevent them from disproportionately distorting regression trends. Consultation with a subject matter expert confirmed that retaining outlier data within the descriptive dataset provides critical context, and that these data points should remain a permanent component of our trend analysis.

To quantify predictive confidence, the analysis utilized 95% confidence intervals (CIs). While central relationships such as the correlation between system size and staffing are observable, the associated CIs are often wide. This variability indicates that while the benchmarks provide reliable portfolio-level trends, additional variables may be required for high-precision program-specific prediction. The width of these intervals serves as a transparent measure of the inherent uncertainty within current DoW software reporting.

Reporting Practices and Policy Interpretation

Sensitivity testing indicates that even under standardized DID guidance, interpretation of variance remains a significant source of uncertainty. Differences in how contractors define an SCR or aggregate labor hours primarily affect the spread (variance) of the data rather than the median values. This testing suggests the study's core benchmarks are resilient to individual reporting styles, though further standardization of SCR scope and labor categorization would diminish the observed noise in future datasets.

Overall Robustness

Across multiple stressors including adjustments to filtering gates and alternative statistical treatments, the core metrics remained stable. The stability of the median SCR rate and labor hours per SCR provides strong evidence that these findings are robust for use as baseline indicators. While application of these results must account for the inherent spread in the data, the sensitivity analysis confirms the study's conclusions are defensible and suitable for informing Department-level software sustainment estimates.

CONCLUSIONS AND DISCUSSION

The analysis into the DoW's SRDR environment sought to establish an empirical foundation for understanding post-production maintenance workloads. Using structured data filtering, normalization of reporting durations, and robust descriptive statistics, the analysis transformed a heterogeneous dataset into a defensible exploratory subset. Across the validated data, systems experienced a median sustainment rate of approximately 2.66 SCRs per month and a median of approximately 147 labor hours per SCR, indicating that operational software typically requires recurring and non-trivial engineering effort during sustainment. Both measures exhibited substantial dispersion and skew, reinforcing the importance of outlier resistant metrics when characterizing typical sustainment behavior.

The analysis further demonstrated that sustainment workloads vary meaningfully across system characteristics, including application domain, software size, programming language, and reporting contractor, confirming that sustainment effort cannot be adequately represented by a single uniform model. While the study highlighted limitations in the current SRDR reporting environment, such as inconsistencies in SCR definitions, labor accounting, and reporting timelines, sensitivity analysis showed the core findings remain stable across reasonable analytical assumptions. As a result, the identified median SCR rate and labor-hours-per-SCR measure provide defensible baseline benchmarks. Although not a definitive CER, these results support reasonableness checks, early sustainment planning, and future model development, while underscoring the need for improved SRDR reporting standards, refined DID guidance, and stronger validation practices to advance software sustainment cost analysis.

RECOMMENDATIONS

Recommendations for Revising the SRDR DID

A review of the SRDR DID (DI-MGMT-82035A) and DD Form 3026-2 indicates several revisions that would substantially improve the quality and analytical usefulness of software maintenance reporting. Although the DID states SRDR is intended to support credible cost, size, and schedule analysis, incomplete reporting and inconsistent definitions limit cross-program comparability. The DID should therefore require completion of key cost-driving variables such as labor hours, total effort, software size, release dates, and software change counts for all completed maintenance releases, with brief justification when a value is not applicable. In addition, maintenance activity definitions should be more tightly standardized through a common taxonomy (e.g., corrective, adaptive, perfective, preventive, cybersecurity response, testing and regression, and integration) and reinforced by explicit null-value logic requiring numeric entries, zeros, or N/A explanations for critical fields.

The DID should also strengthen attribution of labor to specific maintenance activities and enforce clearer record-maturity controls. Labor hours should be mapped not only to releases and WBS elements, but also to defined maintenance activity categories, enabling distinction between engineering change effort, testing, integration, cybersecurity response, and support activities. Completed releases should be explicitly separated from ongoing work through required start dates, end dates, and report-as-of dates. Finally, low-value descriptive fields that increase reporting burden without clear estimating benefit should be deemphasized in favor of variables with demonstrated predictive relevance. Collectively, these changes would move the Software Maintenance Report toward a more analytically disciplined and comparable sustainment dataset, better supporting empirical benchmarking, cost estimation, and lifecycle resource planning.

FURTHER ANALYSIS

While this study establishes an initial empirical baseline for software sustainment activity within the SRDR environment, it also identifies clear opportunities for extension. The primary contribution was the isolation of a defensible subset of records sufficient to derive preliminary benchmarks for SCR rates and labor effort. These benchmarks should be viewed as foundational, providing an empirical reference point rather than a final estimating construct.

Given improvements to reporting consistency particularly through refined DID guidance and stronger CADE validation, future research could expand beyond descriptive benchmarking to longitudinal analysis. Time-series data would enable examination of how sustainment workloads evolve within individual systems, across program phases, and among different classes of software-intensive systems.

Future work should also address variation in sustainment complexity. The strong skew observed in both SCR rates and labor hours suggests software changes differ substantially in scope and impact. Classifying SCRs by complexity, type, or mission relevance would improve the interpretability of labor-per-change metrics and strengthen subsequent predictive modeling efforts.

Collectively, these extensions would support progression from descriptive benchmarks toward a scalable, auditable, and predictive sustainment analytics framework. By establishing a credible baseline and identifying key data limitations, this study positions future efforts to expand the usable dataset, conduct longitudinal analysis, and ultimately develop robust software sustainment CERs to support improved lifecycle cost estimation.

References

Cannon, C. (2007, September 1). *Cost estimation of post production software support in ground combat systems.* https://fileserver-az.core.ac.uk/download/pdf/36697077.pdf

Clark, B., & Madachy, R. (2015). *Software cost estimation metrics manual for defense systems.* https://softwarecost.org/Software%20Cost%20Estimation%20Metrics%20Manual%20for%20Defense%20Systems.pdf

Clark, B., Miller, C., McCurley, J., Zubrow, D., Brown, R., & Zuccher, M. (2017, July 1). *Department of Defense software factbook.* https://www.sei.cmu.edu/documents/1273/2017_005_001_502661.pdf

Defense Acquisition University. (2023, January 29). *Cost and software data reporting (CSDR) for the acquisition community.* Retrieved from https://www.waru.edu/sites/default/files/202412/CSDR%20for%20the%20Acquisition%20Community_Live%20Training_DAU.pdf

Jones, C., et al. *(*2015, June 9). *How much does software maintenance cost?* https://www.iceaaonline.com/wp-content/uploads/2015/06/SW08-Presentation-Jones-Software-Maintenance.pdf

Lanham, N. (2013). *How I continued to stop worrying and love software resource data reports: ICEAA 2014 professional development & training workshop 1 software resource data report (SRDR) analysis.* https://iceaaonline.com/wp-content/uploads/2014/06/IT-5-Handout-How-I-Continued-to-Stop-Worrying-and-Love-Software-Resource-Data-Reports.pdf

Lanham, N., Russo, M., Strickland, D., Cipressi, R., Palmer, S., & Rudloff, C. (2018, February). *Department of Defense software resource data report verification and validation guide (Version 4.0).* https://cade.osd.mil/assets/docs/csdr/guidance/SRDR%20Verification%20and%20Validation%20(VV)%20Guide%20V4.0.pdf

Maurer, D., Newman, S., Czohara, L., Bagley, S., Boyles, S., Buquicchio, V., Lesser, A., Prybyla, J., Stavisky, A., & Weissman, C. (2019, February). *Weapon system sustainment: DoD needs to better capture and report software sustainment costs.* https://apps.dtic.mil/sti/html/trecms/AD1154364/index.html

Office of the Secretary of Defense, Cost Assessment and Program Evaluation. (2019, February 1). *The software resources data report implementation guidance (SRDR).* https://everyspec.com/DoD/DoD-PUBLICATIONS/SRDR_Implementation_Guide_01FEB2019_56198/

Sams, B. (2011, August 29). Deriving the cost of software maintenance for software intensive systems. https://apps.dtic.mil/sti/html/tr/ADA633891/index.html

Sharma, A., & Chaudhary, N. (2023). Prediction of software effort by using non-linear power regression for heterogeneous projects based on use case points and lines of code. https://doi.org/10.1016/j.procs.2023.01.138

U.S. Department of Defense, Cost Assessment and Program Evaluation. (2020). *Department of Defense cost estimating guide office of the secretary of defense cost assessment and program evaluation.* https://www.govinfo.gov/content/pkg/GOVPUB-D-PURL-gpo156710/pdf/GOVPUB-D-PURL-gpo156710.pdf

U.S. Department of Defense. (2021, May 7). *Cost and software data reporting* https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodm/500004p.pdf

U.S. Department of Defense. (2017). *Software resources data reporting: Development, maintenance and enterprise resource planning development reports, and data dictionary* (DI-MGMT-82035A).

U.S. Government Accountability Office. (2023). *Software acquisition: Additional actions needed to help DoD implement future modernization efforts.* https://www.gao.gov/products/gao-23-105611

U.S. Government Accountability Office. (2024, July 11). *IT systems annual assessment: DoD needs to strengthen software metrics and address continued cybersecurity and reporting gaps.* https://www.gao.gov/products/gao-24-106912

APPENDIX A – Super Domain Names and Definitions

The DoW SRDR Verification and Validation Guide define Super Domains as a group of application domains with similar software characteristics:

Super Domain	Description	Application Domain Mapping
Engineering	Engineering systems represent medium complexity software systems that process technical data or support engineering activities but are not strictly time-constrained like embedded real-time systems.	System Software
		Process Control
		Scientific and Simulation
		Test, Measurement, and Diagnostic Equipment
Real-Time	Real-time systems are the most complex category of DoD software because they must interact directly with hardware or physical systems while meeting strict timing constraints. These typically take the most time and effort due to very high reliability or hardware constraints.	Signal Processing
		Vehicle Control
		Vehicle Payload
		Other Real Time Embedded
		Command and Control
		Communication
		Microcode & Firmware
Automated Information Systems	Provides information processing services to humans or other applications. Allows authority to exercise control and access to typical business processes.	Mission Planning
		Enterprise Service Systems
		Custom AIS Software
		Enterprise Information Systems
Support	Support software assists with operator training, testing, or administrative tasks.	Training
		Software Tools

Table 5: Super Domain descriptions and Application Domain mapping

Mr. James Reichwein is a recent graduate of the United States Naval Postgraduate School (NPS), where he earned a Master of Science in Cost Estimating and Analysis in 2026 and a certificate in Cost Estimating and Analysis in 2022. He is a senior operations research analyst with DCS Corp currently working with the Air Force Research Laboratory (AFRL) specializing in modeling and simulation, data analysis, and parametric cost estimation on conceptual weapon systems.

Lieutenant Timothy L. Nelson, USN, is a recent graduate of the United States Naval Postgraduate School (NPS), where he earned a Certificate of Cost Estimating and Analysis and a Master of Science in Cost Estimating and Analysis. He is currently a Naval Aviator at Patrol Squadron Thirty (VP-30) of Jacksonville, FL.

Mr. Robert Sherbs earned a Masters of Science in Mathematics from University of Massachusetts in 2019 and a Certificate in Cost Estimation and Analysis from Naval Postgraduate School in 2025

When History Rhymes: Cost Estimating Lessons from Seven Legacy Aircraft

Brent M. Johnstone

Every historical aircraft program carries a narrative that can inform present-day cost estimating. By "reading" the manufacturing learning curves of seven legacy aircraft (F-102, F-106, B-58, F-111, L-188, P-3, and L-1011) we uncover recurring themes confirming the adage that history does not repeat itself, but it does rhyme. This study identifies quantitative and qualitative lessons from these legacy programs, revealing how design changes, delivery rate fluctuations, production gaps, end of program tail-up, and the knowledge transfer across multiple variants shaped cost curves. The analysis confirms that even decades-old programs can provide actionable insights for today's estimators, offering a framework to anticipate cost drivers, calibrate learning curve models, and improve forecasting accuracy for current and future aircraft estimates.

"It has been said that history repeats itself.
This is perhaps not quite correct;
it merely rhymes."
– Theodor Reik (Reik, 1965)

"History is more or less bunk."
– Henry Ford (Swigger, 2014)

Between these two extremes lies a paradox at the heart of cost estimating. Cost estimating relies on historical data to build relationships between requirements and costs. On the other hand, how relevant is the "old" history to efforts being estimated today? At the extreme lie two schools of thought. The first argues that the history of legacy programs can offer us a rich vein of analogous scenarios which are directly relevant to today's cost estimates. The second argues that today's manufacturing and technological environment is so different that any cost history more than a decade old is essentially useless.

This paper situates itself in the first camp. While today's manufacturing technologies, markets, and strategic environments differ substantially from fifty and sixty years ago, the reasons why costs change over time – learning benefits, redesign impacts, delivery rate fluctuations, production gaps, knowledge transfer both across and within programs – remain fundamentally the same. To Reik's point, history does not repeat itself, but it does rhyme. To this end, this paper presents the narrative history of seven historical aircraft – the Convair F-102 and F-106 interceptors, the Convair B-58 bomber, the Lockheed L-188 Electra / P-3 Orion family, the General Dynamics F-111 fighter/bomber, and the Lockheed L-1011 jetliner. In each case we "read" the manufacturing hours per unit history in relation to program history to get a better understanding of the mechanisms driving cost. This analysis demonstrates that even programs whose production ended decades ago can reveal recurring patterns useful to today's estimators.

Data Sources and Methodology

The programs chosen for this analysis had to meet several criteria. First, the historical aircraft cost data had to be available in sufficient detail to perform analysis. Second, the program cost data had to be sufficiently old so that there were no issues around release of proprietary data. The first suggested data from Lockheed Martin or heritage companies, while the second dictated programs which were long out of production, such as programs with first flights during the decades of the 1950s or 1960s (F-102, F-106, B-58, F-111), or lines of business in which Lockheed Martin no longer competes, such as the commercial aircraft business (L-188, L-1011).

Once chosen, the program cost data was reviewed as well as any contemporary documentation or analysis. In most cases, key program events could be identified and correlated with the observed cost impacts. In addition, aircraft histories and reference books were reviewed to see if these could provide additional insights into program cost drivers.

The seven aircraft histories follow. The cases are arranged chronologically by date of first flight.

Cases 1 & 2: The F-102 Delta Dagger / F-106 Delta Dart

F-102A Delta Dagger
Source: National Museum of the United States Air Force

F-102A Delta Dagger:

Manufacturer: Convair Division of General Dynamics (San Diego, Fort Worth)

Mission: Supersonic interceptor aircraft

Program Start: 1951

First Flight: 1953

Quantity Produced: 1,000

Year Production Ended: 1958

Program Background

The sound barrier was first broken by Chuck Yeager piloting the Bell X-1 in 1948. It was not long before the US Air Force envisioned a production supersonic aircraft capable of intercepting Soviet bombers targeting the United States. When the USAF released specifications for a supersonic all-weather, high-altitude interceptor, Convair proposed a delta wing design which eventually became the F-102A. In addition to supersonic speed, the F-102A would carry several new mission systems, prominently an electronic radar and fire control system. Air-to-air rockets and missiles would be carried in a large internal weapons bay, allowing an aerodynamically "clean" aircraft. (*Convair F-102A Delta Dagger*, undated; Mutza, 1999).

Despite their efforts, the F-102's development was difficult. "The Convair F-102 program experienced significant R&D delays, many of which stemmed from engine problems and the over ambitiousness of its planned automated avionics subsystems" (Lorell, 1998). Worse, early wind tunnel tests discovered that the aircraft would not be able to achieve supersonic speed. As Convair and the Air Force scrambled for a solution, independent government research indicated that a fuselage with a pinched waist (the so-called "coke bottle" or "Marilyn Monroe" configuration) would sufficiently reduce drag to enable supersonic flight, a principle later known as the "area rule" (Mutza, 1999).

To preserve program schedule, Convair and the USAF agreed to build ten aircraft in the subsonic configuration while redesign proceeded. The F-102 was planned with high concurrence between development and production. Production tools would be based on the initial design while several early units were built for flight test. Changes from flight test would be fed into the eventual production design, and USAF would commit to production before development was finished (AFLCMC, undated; Mutza, 1999). However, the fuselage redesign derailed the concurrency concept. Convair built 30,000 tools by October 1953. In the end, it scrapped 20,000 of them because of the redesign (Knaack, 1978b).

Waters ran smoother once the F-102 entered production. From a manufacturing build perspective, the F-102 was considered a "production man's dream." Convair's chief tool engineer, A.P. Higgins, envisioned the aircraft design as a series of easily handled subassemblies tooled to maximize manpower around the build article. F-102 production combined conventional build technology, heavy sheet metal presses, and half-shell construction methods new to the industry at the time (Mutza, 1999). Plans were also put into place for a two-seat trainer version designated the TF-102A. The primary difference was a redesigned forward fuselage to be built in Convair's Fort Worth facility and shipped to San Diego for final assembly.

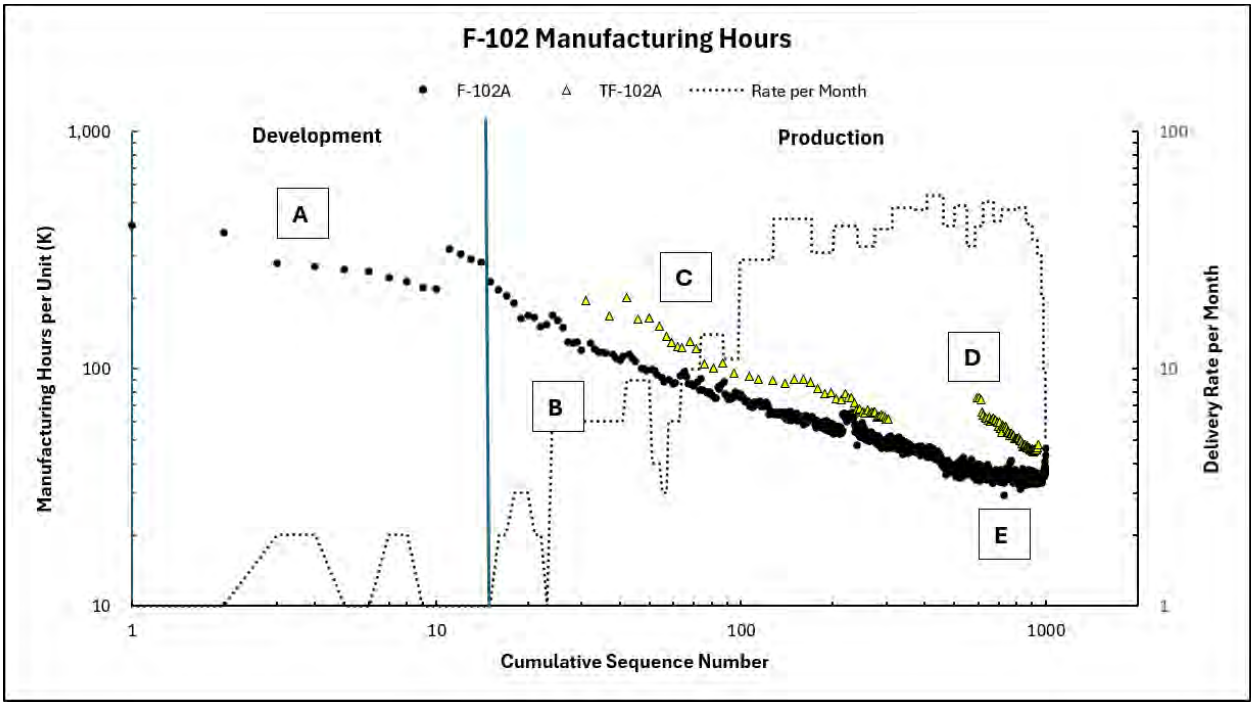


Figure 1. F-102A/TF-102A Manufacturing Hours per Unit

Learning Curve Analysis

From a learning curve perspective, the F-102A offers a little bit of everything: the classic S-shaped learning curve, redesign impacts, multiple variants, varying production rates, and a cost tail-up at the end of the program.

Figure 1 displays the F-102A and TF-102A manufacturing hours per unit (*F/TF-102 Estimating Handbook and Product Book*, undated). Five areas have been called out and are explained as follows:

A. **#1 through #14 (F-102A):** The initial ten F-102A aircraft came down an 83% slope. Unit 11 introduced the new "area rule" configuration, which introduced an 80% setback on the learning curve back to approximately unit 2.5. An internal F-102 Industrial Engineering (IE) analysis noted it "demonstrates a typical flat DT&E [Development Test and Evaluation] aircraft cost curve. The flat curve was caused by: * Slow production rate * Changing task and configuration * Tooling adjustments and alterations * Small lot sizes" (*F-16 A/B FSD – Production Manufacturing Estimating Rationale*, undated).

B. **#14 through #500 (F-102A):** The F-102A learning curve steepened to 75% as it transitioned into production. "The learning curve traditionally steepens sharply at the beginning of the production program," according to the IE analysis, "resulting from acceleration of the move rate, specialization of labor, improved tooling, stabilization of task, and mature planning, crew loads, etc." *(F-16A/B FSD – Production Manufacturing Estimating Rationale,* undated). Production rates accelerated rapidly from 1 to 2 aircraft per month in development to reach the planned maximum rate of 45 per month around unit 150. It's worth noting how incredibly quickly the F-102 moved through this phase – the 500th unit was delivered within 26 months of the first production delivery. By contrast, the F-35 – a modern, high-rate program -- delivered its 500th unit over 8 *years* after its first production delivery.

C. **#31 through #304 (TF-102A):** The TF-102A trainer first delivered at common sequence #31.

As expected, it demonstrated setback on the learning curve, but because of its high commonality with the F-102 (a unique forward fuselage but limited changes for the rest of the airframe) approximately a dozen units of learning were lost. Production continued until common unit 304, when it was put on hold to resolve a buffeting issue that surfaced during flight test due to the TF-102's peculiar canopy.

D. **#597 through #946 (TF-102A):** TF-102A two-seater production resumed after a six month gap. The buffeting problem was resolved by redesigning the canopy to improve airflow and increasing the size of the vertical tail (Mutza, 1999). Not surprisingly, the first TF-102A delivered after the production break shows a regression on the learning curve and higher hours.

E. **#501 through #1000 (F-102A):** The F-102A learning curve significantly flattened to 89%. IE analysis noted this flattening occurred concurrent with maximum delivery rates (the F-102A briefly reached 50+ deliveries per month) "when the task and crews stabilized, lot sizes became firm, and tooling augmentation was complete" *(F-16A/B FSD – Production Manufacturing Estimating Rationale*, undated). Also observable is a significant 10,000 hour per unit tail-up in the last units of F-102A build.

F-106 Delta Dart
Source: National Museum of the United States Air Force

F-106A/B Delta Dart:

Manufacturer: Convair Division of General Dynamics (San Diego, Fort Worth)

Mission: Supersonic interceptor aircraft

Program Start: 1955

First Flight: 1956

Quantity Produced: 340

Year Production Ended: 1961

Program Background

The F-106A was an outgrowth of the F-102A program. As early as 1951, it became apparent the F-102A could only partially meet the requirements of what the USAF considered its "Ultimate Interceptor." The resolution was to go forward with the F-102A but continue design work on an upgraded interceptor called the F-102B. Predictably the two designs diverged with extensive structural changes and a new powerplant. The new aircraft was redesignated as the F-106A (Mutza, 1999; *Convair F-106A Delta Dagger*, undated).

In December 1956, the same month the F-102A delivered its 201st aircraft, the first F-106A was delivered. Like its predecessor, the F-106A was a supersonic delta wing high altitude, all-weather interceptor but equipped with a more advanced fire control system (Knaack, 1978b). The F-106A was also two and a half feet longer and weighed 4,300 pounds more than its predecessor. Like the F-102A, it had an internal weapons bay to carry missiles and rockets. The F-106 also sported a two-seat trainer version, the F-106B, whose forward fuselage would be built in Fort Worth and shipped to San Diego for final assembly (*F/TF-102 Estimating Handbook and Product Book,* undated; *F-106A/B Estimating Handbook and Product Book,* undated).

Determined to identify problems before production began, the F-106A program had an unusually high number of test aircraft (31 A and 11 B models). Unfortunately, this precaution did not prevent the earliest F-106s from disappointing the USAF. The two prototype F-106A aircraft fell well short of Convair's estimates for acceleration and maximum speed. Convair proposed, and the USAF quickly

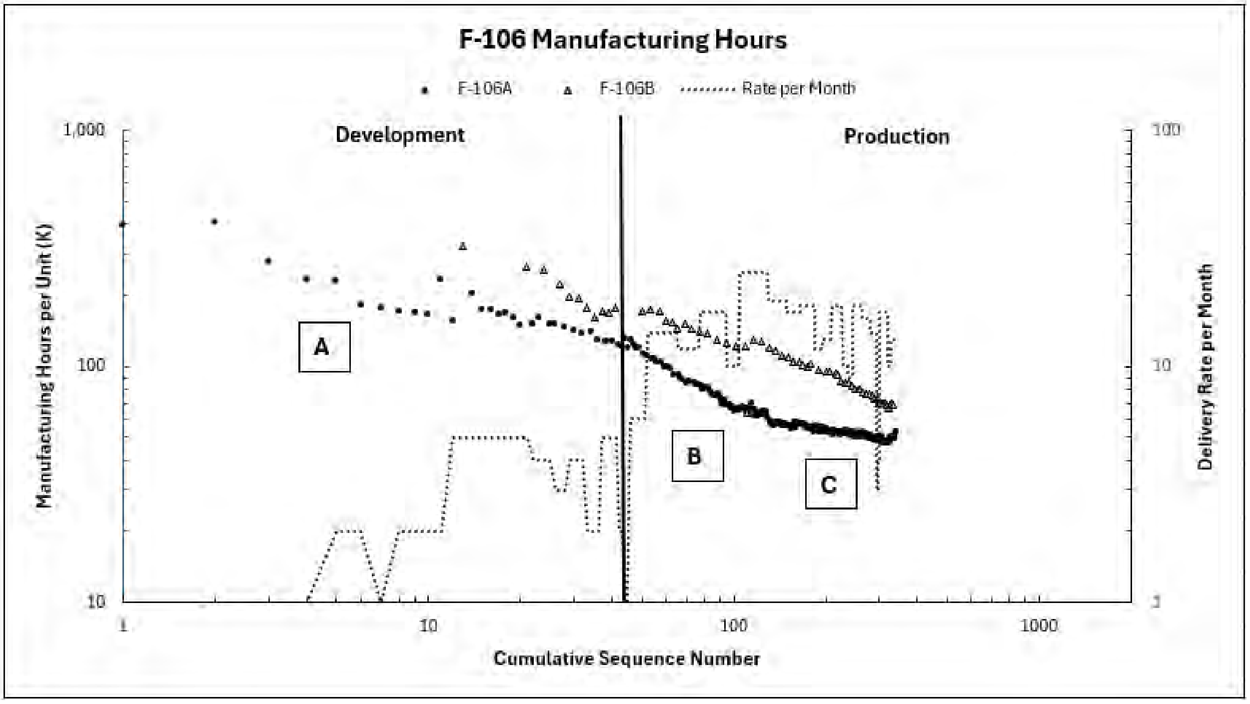


Figure 2. F-106A/B Manufacturing Hours per Unit

approved, changes to the F-106 inlet ducts to reduce aircraft drag. These changes were implemented at Unit #11. However, these engineering changes, combined with the development delays of the J-75 engine and the MA-1 fire control system, left the program in a fiscal crunch. Although the USAF was happy with the aircraft's eventual performance, it only had the funds to buy just over a third of the original planned 1,000 aircraft (Knaack, 1978b).

Learning Curve Analysis

Like F-102, the F-106 history can be split into several major areas. Figure 2 displays the F-106A/B manufacturing hours per unit (*F-106A/B Estimating Handbook and Product Book,* undated). Three areas have been called out and are explained as follows:

A. **#1 through #42:** As noted, performance issues drove a configuration change beginning at Unit #11 that forced a partial learning loss back to approximately unit #4. Development required an unusually high number of test aircraft, as noted. Overall development slope was 82%.

B. **#43 through #130**: After an upward initial cost "bump" at the start of production to incorporate engineering changes from the development program, the program experienced a steep 62% slope as the production rate exceeded 15 per month.

C. **#131 through #340**: As soon as September 1958, before the first production aircraft was delivered, it was apparent the F-106 fleet would be much smaller than originally planned. Cost improvement slowed to a 87% slope, although production rates remained around 15 per month. Interestingly, after adjustment for weight differences, the F-106A flattening occurred just as it approached a similar level of performance for the F-102A. [Reference Figure 3] The expected cost tail-up showed up as the last few F-106A aircraft were completed.

A question for derivative aircraft is often: How much learning gain can we expect to be transferred over to the new system? In the case of the F-106A, the configuration changes appear to be extensive enough that the F-106A began at virtually the same T-1 hours per pound as the F-102A after adjustment for weight differences, as shown in Figure 3.

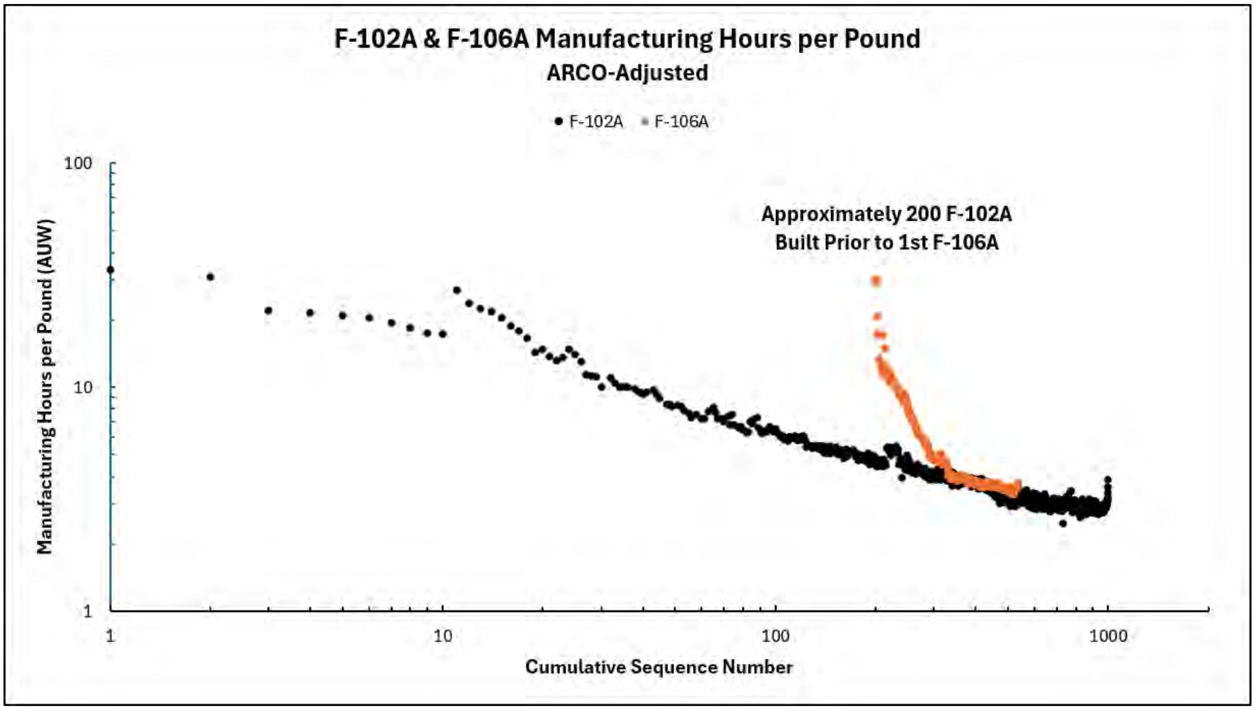


Figure 3. F-102A/F-106 Manufacturing Hours per Pound (Aircraft Unit Weight)

Lessons from the F-102/F-106:

- The F-102 and F-106 both display the classic S-curve learning curve shape across the product life cycle (relatively flat development, steep early production, flattening across later production).
- Design changes such as the redesign of the F-102A for the area rule or the TF-102A canopy for the buffeting issue drive setbacks on the learning curve. Major redesigns produce a quantifiable loss of learning before recovery to the underlying curve.
- Exclusive of redesign impacts, concurrency can impact program costs in other ways such as the widescale scrappage of tools on the F-102A. The replacement tools also require proofing, and the errors discovered corrected with negative impacts on manufacturing learning.
- Unique variants, such as two-seater configurations, show a significant degree of learning transfer from the baseline model, but at the same time show higher hours per unit / hours per pound across the length of the production program.
- A production gap in the build of a unique variant produces a loss of learning and a regression on the learning curve, as seen through the TF-102A production gap.
- The evidence on learning transfer between follow-on aircraft, e.g., learning transfer between the F-102A and F-106A, is mixed. The F-106 shows virtually no learning gain from the preceding F-102 build in terms of first unit hours. Other aircraft later in this paper will provide different results.
- Tail-up impacts are evident in the final lots of both programs, adding a roughly 10,000 hour per-unit penalty as the line wound down.

B-58 Hustler
Source: Wikipedia Commons

Case 3: The B-58 Hustler

Manufacturer: Convair Division of General Dynamics (Fort Worth)

Mission: High altitude / high speed nuclear bomber

Program Start: 1955

First Flight: 1956

Quantity Produced: 116

Year Production Ended: 1962

Program Background

The B-58 poses an interesting study in what happens when a new, cutting-edge manufacturing technology is introduced to the factory floor – with a significant number of problems to be resolved. But first some background explanation is required.

The B-58 Hustler's mission was to fly at high altitudes and supersonic speed into Soviet airspace and deliver a nuclear weapon over target. No USAF jet bomber before or since was as small or fast. And it presented a peculiar set of design challenges. For one thing, its maximum speed was Mach 2, more than twice the speed of the subsonic B-47 and B-52 jet fighters. And its subsystems were the most highly advanced of its time -- "one of the largest collections of vacuum tubes and mechanical analog machinery ever conceived and fabricated by man" (Hall, 1981).

Attaining such a high speed required a small aircraft, which created a challenge to carry enough fuel to meet the required range. And the advanced equipment was *heavy*. The Sperry AN/ASQ-42 bombing-navigation system alone required a 1,200 pound on-board analog computer. To carry this equipment, achieve and maintain Mach 2 speed, and meet its required range, the B-58 would need to dedicate 55% of its takeoff weight to fuel (Hall, 1981). That meant the airframe structure would need to be as light as possible.

Convair's engineers believed the solution lay in bonded honeycomb sandwich panels. These panels were a radical departure from the traditional monocoque riveted metal construction used since the 1930's. A half-inch thick fiberglass or aluminum honeycomb core was sandwiched between two aluminum metal skins each a millimeter thick, and adhesively bonded. The assembled panel was then cured in an oven. The resulting part was stiff and strong, retained its surface smoothness in all flight conditions -- and most important, was lightweight (Robinson, 1967; Hall, 1981; Miller, 1985). An illustration of a honeycomb sandwich structure and its layup are illustrated in Figures 4 and 5.

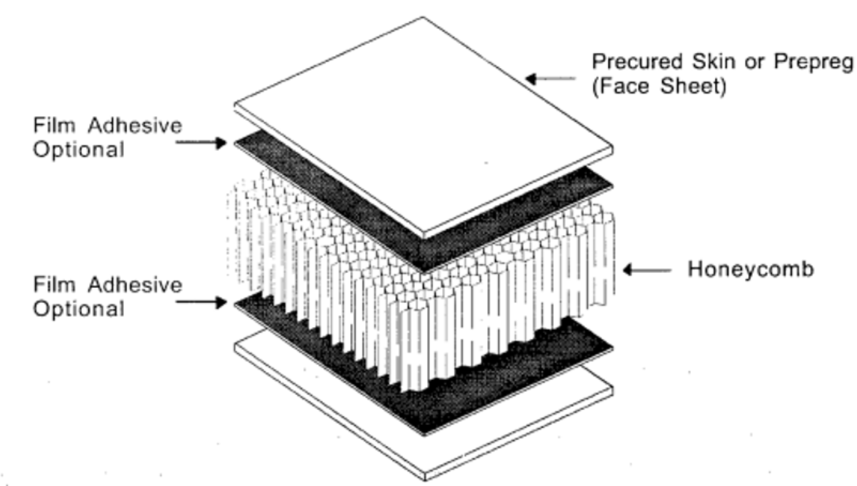


Figure 4. Honeycomb Sandwich Panel (Price, 1997)

But bonded honeycomb had rarely been used on aircraft programs. Convair's massive B-36 bomber used fiberglass honeycomb on some select fuel cell support panels (Joyce, 2003). But the B-58 would go far beyond that. 90% of the wing and 80% of the fuselage skins were constructed from honeycomb panels (Hall, 1981). "[T]he entire outer skin of the wing is attached by adhesive alone without the aid of mechanical fasteners," wrote two General Dynamics engineers. "The entire surface covering of the B-58 is composed of bonded structure, except for the elevons and the aft portion of the nacelles and pylons." Those areas were made of a brazed stainless steel sandwich intended to withstand the heat of the jet engines (Smith, et al., 1962). But brazing stainless steel aerospace-quality parts was also a relatively unfamiliar manufacturing technique in the 1950s.

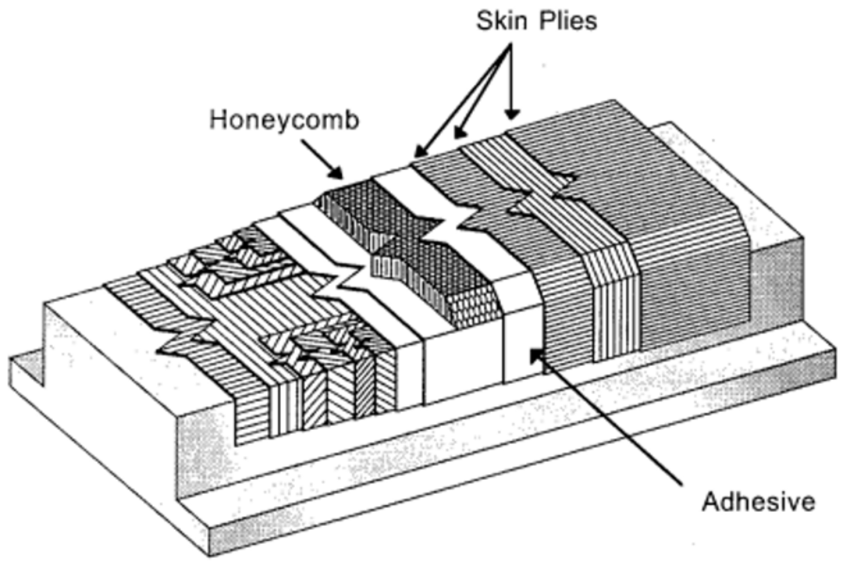


Figure 5. Typical Honeycomb Sandwich Lay-Up (Price, 1997)

Satisfied with its radical design, Convair pushed forward into the tooling and build phases. But from the beginning, these new processes created engineering and tooling problems. A 1963 Convair Tooling department report notes that extensive development was required for production of bond forms – needed to hold the sandwich construction in place during cure, since a panel once cured could not be bent or deformed. Development of the cure cycle itself was trial and error. There were difficulties in grinding and machining the honeycomb core. Fabrication of the bond form tooling took more time than originally estimated. Brazing also created extensive development issues (*Tooling Historical Facts and Figures*, 1963). On the factory floor, special facilities for bonding had to be built. "Absolute cleanliness is essential for solid bonding," noted a B-58 analyst, "and the department of the Convair plant where this was done was known as the 'hospital section'" (Robinson, 1967). This is even true for today's bonding processes: as a current day commentary notes: "The entire bonding process must be carefully controlled in order to ensure the reliability of the bonding assembly. With a bit of carelessness here, a moment of inattention there, even an inadvertent fingerprint, the entire end product of everyone's efforts may have to be discarded" (Price, et al., 1997).

Learning Curve Analysis

These tooling and engineering problems conspired to create an extraordinarily high first unit cost…nearly a million and a half touch labor hours for the first aircraft. The B-58 was 43 hours per pound at first unit, over twice that of other USAF bombers built during the 1950s (B-47, B-52, B-57, B-66) (Levenson, 1965).

As we shall see, the development phase of an aircraft typically shows relatively flat learning curve slopes. This is driven by several factors: a high number of engineering changes, incomplete tool families, high scrap and rework costs, low production rates, part setup costs allocated over small lot sizes, and unique configurations for development aircraft depending on their test mission (e.g., aerodynamics, flutter, mission systems, etc.). The B-58 provides a surprising exception to this rule. During the development phase, the B-58 exhibited a 73% slope.

Why such a steep slope? An IE post-mortem analysis attributed it to the "[h]ighly developmental nature of design pushing the state of the art and the

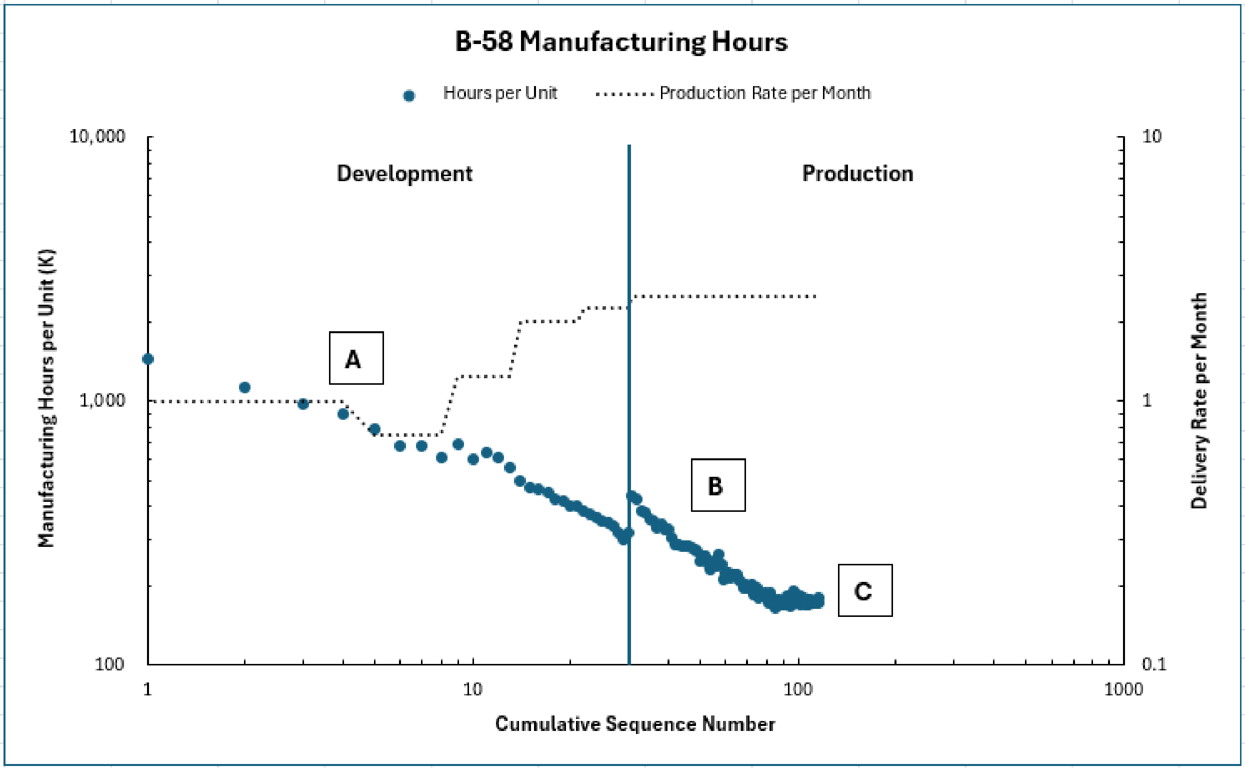


Figure 6. B-58 Manufacturing Hours per Unit

unit cost higher initially" (*F-111/B-58/B-36 Curve Slope Criteria,* undated).

The resulting high first unit hours meant the possibilities for subsequent cost reduction were greater as the new bonding and brazing processes became better understood and tooling, engineering, and shop floor methods modified. This phenomenon of high T-1 hours together with steep learning curves has been observed by other learning curve students. This relationship between high hours at T-1 and subsequent steep learning curve slopes was first documented by RAND in the 1950s (Asher, 1956) and confirmed by subsequent studies (Sawyer*,* et al., 2002).

Figure 6 displays the B-58 manufacturing hours per unit (*B-58 Manufacturing Manhours by Department, Airplane, Lot & Contract*, undated). Three areas have been called out and are explained as follows:

A. **#1 through #30:** As noted, the development units had a surprisingly steep slope due to the high first unit cost and the subsequent recovery as Convair worked through the bonding and brazing difficulties. In addition, Fort Worth's IE department introduced crew loading at ship 14, giving each worker specifically preplanned work arranged in the most efficient sequence (*Crewloading*, 1979).

B. **#31 through #76:** Significant reconfiguration impacts occurred at the beginning of production resulting in a 13 shipset loss of learning (43% setback) with relatively quick recovery. Producibility improvements resulted in a redesign of shop floor tooling and fixtures. Initial production slope including recovery was 65%. Maximum production rate of three per month was reached at unit 48.

C. **#77 through #116:** Improvement slowed significantly to a 90% slope as termination of the program became apparent. Small tail-up activity was observed in the last production lot.

The reconfiguration impacts seen in Area B were caused by discoveries during the flight test program.

The original B-58 test plan called for 13 test aircraft. In the end, 30 test aircraft were used. Several test aircraft crashed during flight test, triggering subsequent redesign of the airframe structure. The worst crash occurred in 1959 when a test aircraft disintegrated in mid-air during a high-speed flight. Subsequent investigation revealed that the vertical stabilizer was too flexible and unable to withstand loads in certain flight conditions (Miller, 2011).

Bigger troubles were brewing for the B-58. Once aircraft entered the USAF inventory, the Strategic Air Command (SAC) found the B-58 to be expensive to fly and prone to high accident rates. Because of its high cost per flight hour, the cost of maintaining and operating two B-58 wings equaled that of six wings of B-52s. Worse, advances in Soviet air defenses made it doubtful that a high-altitude aircraft like the B-58 could survive to perform its mission (Hall, 1981).

SAC began to lose interest. Planned for a maximum production rate of six aircraft per month, the B-58 only reached half that rate (Miller, 1985). As it became apparent that the program was to end prematurely, the manufacturing hours per unit began to rapidly level out [see area C in Figure 6] as company investment in affordability dried up and the production line began to be put away.

Lessons from the B-58:

- High first-unit hours create a steep learning curve. The B-58's unprecedented bonding and brazing processes created a very high first unit cost. This high T-1 in turn produced a large "headroom" for cost improvement, resulting in atypically steep learning curve slopes during the development and early production phases.
- Major design changes act as learning curve reset events. Flight test driven redesigns caused measurable learning losses at the beginning of the production phase before recovery to the underlying slope.
- The prospect of program termination flattens the curve. As the B-58 neared cancellation, investment in tooling and process improvements stopped, causing the learning curve to approach 90% even before the 100th unit.

L-188 Electra
Source: Wikimedia Commons

Cases 4 & 5: The L-188 Electra / P-3 Orion

L-188 Electra:

Manufacturer: Lockheed (Burbank)

Mission: Turboprop commercial jetliner

Program Start: 1955

First Flight: 1957

Quantity Produced: 170

Year Production Ended: 1962

Program Background

The L-188 Electra story is a relatively short and unhappy one. It does, however, prove that even a failed aircraft can have a second life.

By the late 1950's, Lockheed's most successful commercial airliner, the Constellation, had run its course. Unlike its competitors Boeing and Douglas Aircraft, Lockheed believed that the economics of short domestic routes justified a new turboprop aircraft. Lockheed jumped into the market with its L-188 aircraft.

At first the L-188 program performed quite well. Its development program moved swiftly, and first flight was 8 weeks ahead of schedule. Design, build, certification and first delivery took place in 26 months. The real problems developed, however, once the aircraft went into service. Less than two years after its first delivery, L-188s crashed in Texas and Indiana. Investigators determined the crashes were due to in-flight structural failure of the engine mount. Under certain circumstances, a condition called "whirl mode" created excessive stress leading to eventual wing failure. Lockheed redesigned the mounts and undertook an expensive retrofit program to modify existing aircraft (Boyne, 1998). But passengers and airlines had lost interest in the L-188. Orders dried up and by 1962 it was apparent that the commercial program was no longer viable.

P-3 Orion
Source: Lockheed Martin

P-3 Orion:

Manufacturer: Lockheed (Burbank, Palmdale, Marietta)

Mission: Antisubmarine (ASW) patrol aircraft

Program Start: 1957

First Flight: 1959

Quantity Produced: 628 (excluding 21 CP-140 derivative aircraft)

Year Production Ended: 1995

Even before the Texas and Indiana crashes, however, Lockheed had been searching for other market opportunities for the L-188. One came from the US Navy, searching for a replacement airframe for the Lockheed P-2V Neptune and Martin P5M flying boat. The Navy's proposal specifications dictated a tight development schedule and strict per unit cost requirements. Only an off-the-shelf aircraft adapted to the mission would meet those needs (Rhodes, 2012).

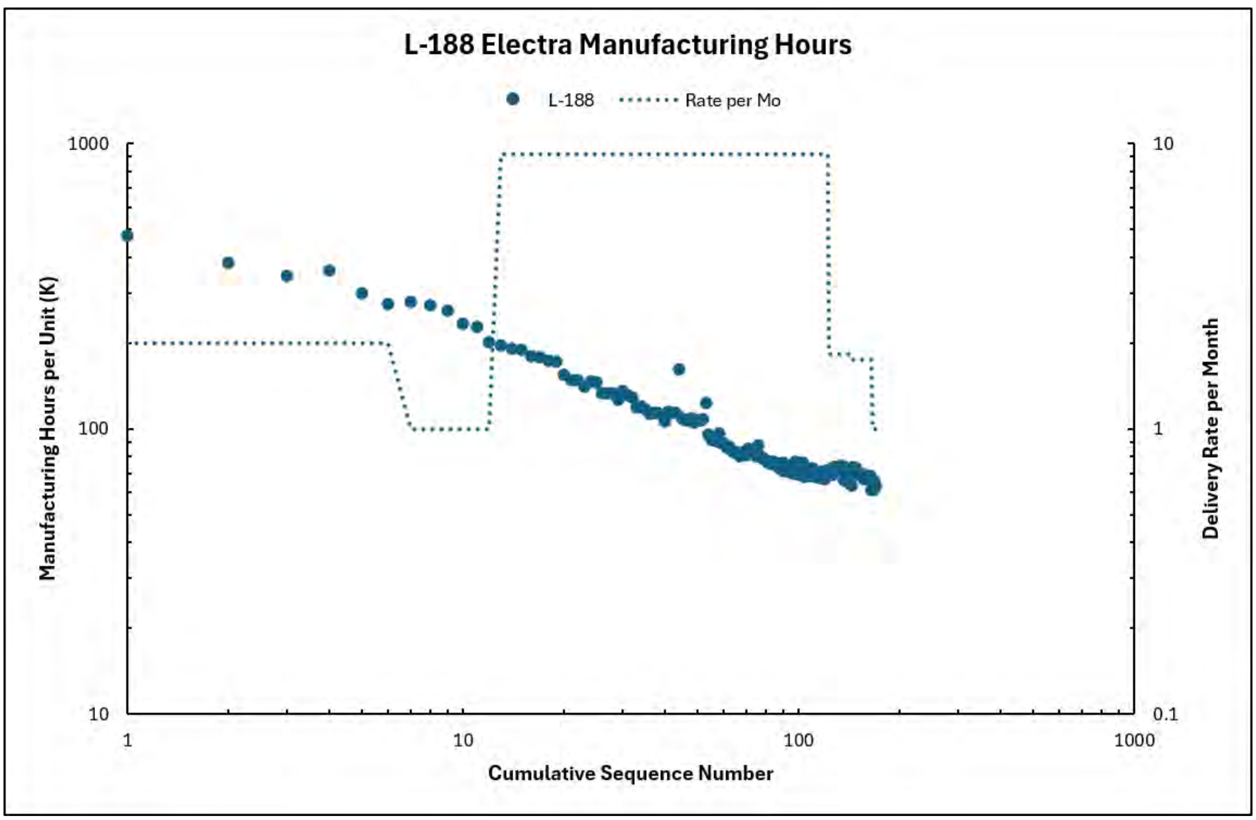


Figure 7. L-188 Manufacturing Hours per Unit

Lockheed decided a modified L-188 configuration would be a perfect match, retrofitting an Electra production aircraft with a mockup of a Magnetic Anomaly Detector (MAD) boom at the rear of the aircraft (Boyne, 1998). Lockheed incorporated the whirl mode design fix into the P-3 and convinced the Navy it would experience no further issues. The Navy was sufficiently impressed and Lockheed prevailed in the competition over Martin and Convair (Rhodes, 2012).

Learning Curve Analysis

Lockheed's performance on the L-188 development and production programs was relatively unremarkable, as shown in Figure 7.

Of more interest is the transition from the L-188 to the P-3. From a learning curve perspective, the L-188 / P-3 history offers several interesting scenarios. First, we can ask: How much learning was transferred between the L-188 and the P-3? Figure 8 shows that P-3 deliveries began just as the last L-188s were completed. It also shows that P-3 experienced learning gain such that its first unit cost was equal to the L-188 at approximately unit #5. If this gain seems small, recall that the P-3 had major structural modifications with an entirely different avionics system with corresponding impacts to electrical and plumbing subsystems. At the same time, it is this apparent learning transfer from the L-188 that gives a unique flavor to the early P-3A build experience.

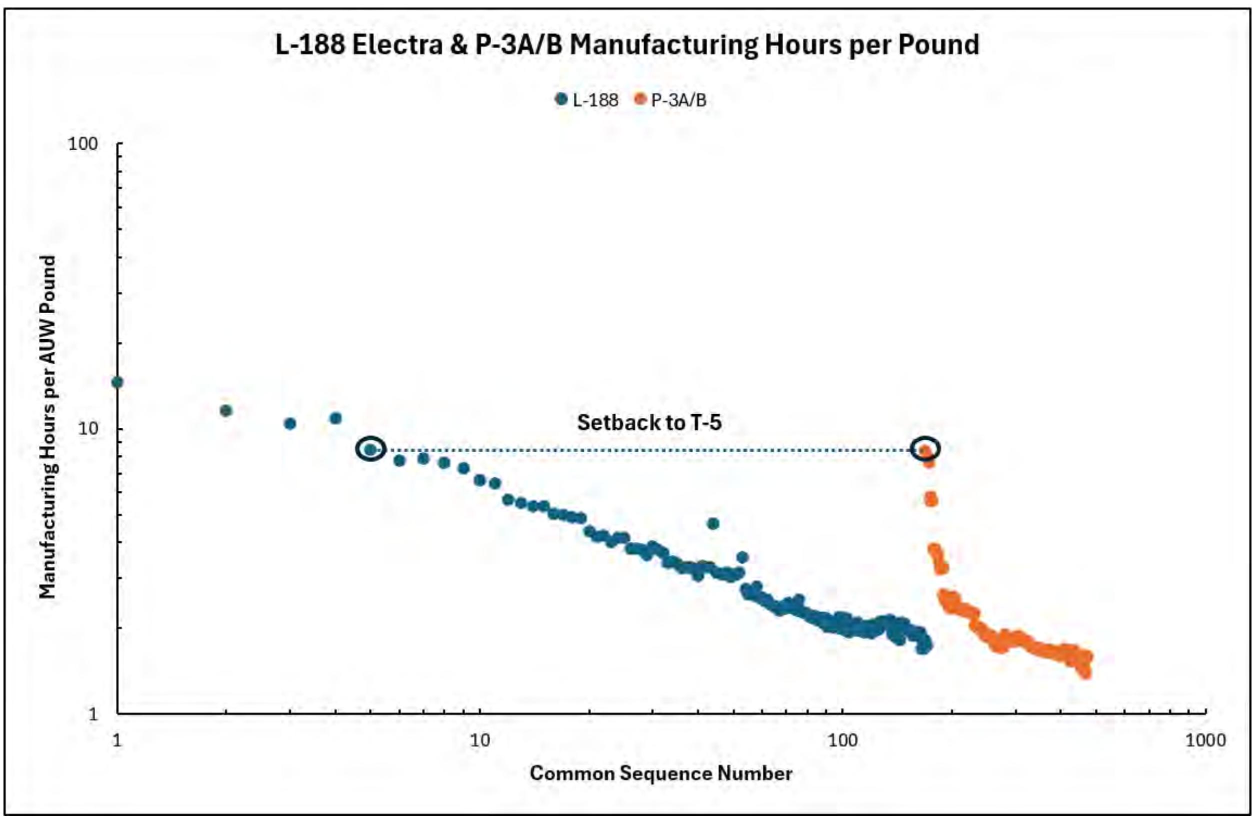


Figure 8. L-188 & P-3A/B Manufacturing Hours per Pound

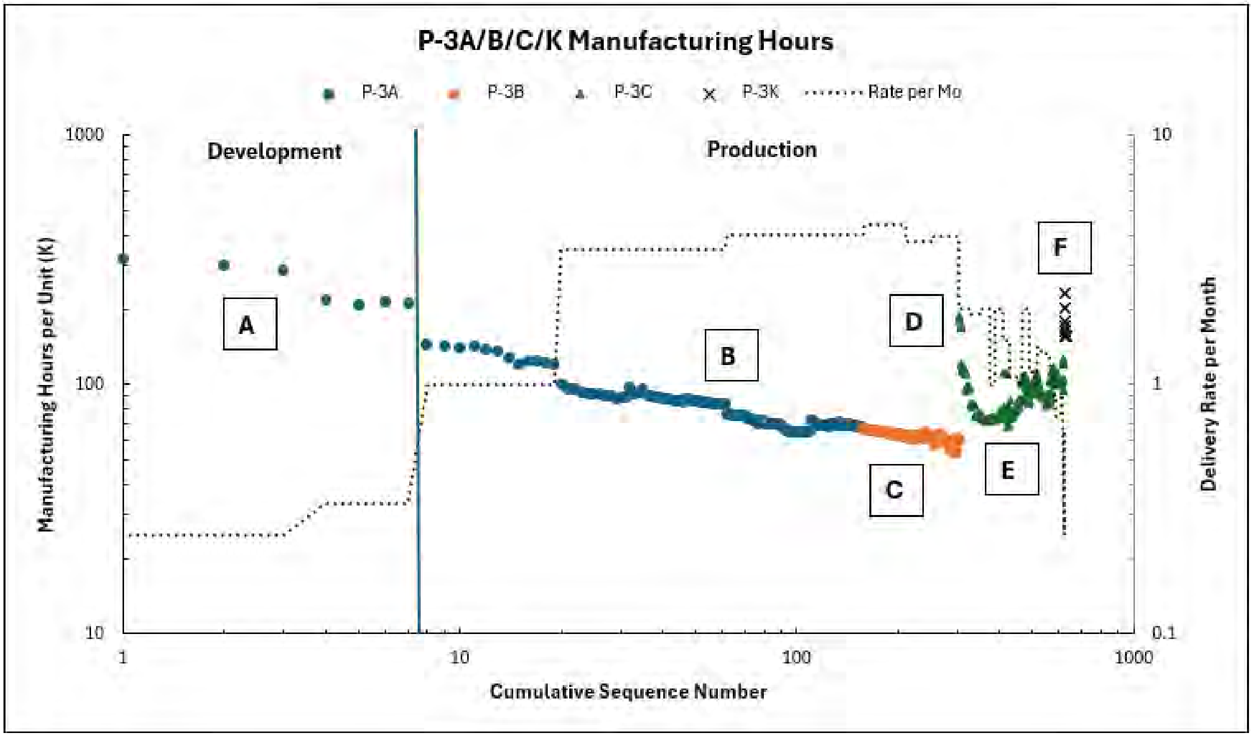


Figure 9. P-3 Manufacturing Hours per Unit

Figure 9 displays the P-3 manufacturing hours per unit (*Model P-3A/B Fact Sheet*, undated; *Model P-3C Fact Sheet*, undated; *Model P-3K Fact Sheet*, undated). Six areas have been called out and are explained as follows:

A. **Units 1-7:** Early P-3A development units showed an 84% slope. This phase is best understood as a mix of new development as well as retention of prior learning on the L-188 program. Reference Figure 8.

B. **Units 8-164:** P-3A production continued the 84% slope. As Figure 8 makes clear, this can be understood as the P-3 returning to the prior underlying curve established by its L-188 predecessor. Production rates reached the planned peak of approximately 4 per month.

C. **Units 165-302**: P-3B production began. Changes from the A to B models were relatively minor with no noticeable upward shift in the hours per ship at the point of introduction. However, the slope began to flatten to an 86% slope.

It is after the 300th unit of the P-3 that things get interesting. Figure 10 shows a larger view of these phases.

D. **Units 303-573:** Introduction of the P-3C. Unlike the B model, the C represented a major jump in aircraft capability. Walter Boyne writes the P-3C was the "first digital, computer-based, software-controlled avionics system developed by Lockheed." This involved both structural and systems changes (Boyne, 1998). The C model showed a substantial setback on the learning curve, regressing almost 300 units back to unit #7. While making a strong recovery back to the prior performance, the P-3C improvement stalled out as a series of production rate fluctuations between one and two aircraft per month took hold.

E. **Units 573-620**: By the early 1980s, Lockheed began to consider shuttering the Burbank facility where P-3 had been built due to the obsolescence of its physical plant and equipment. In 1983 the decision was made to transfer the line to Lockheed's other California facility in Palmdale. This line transfer created

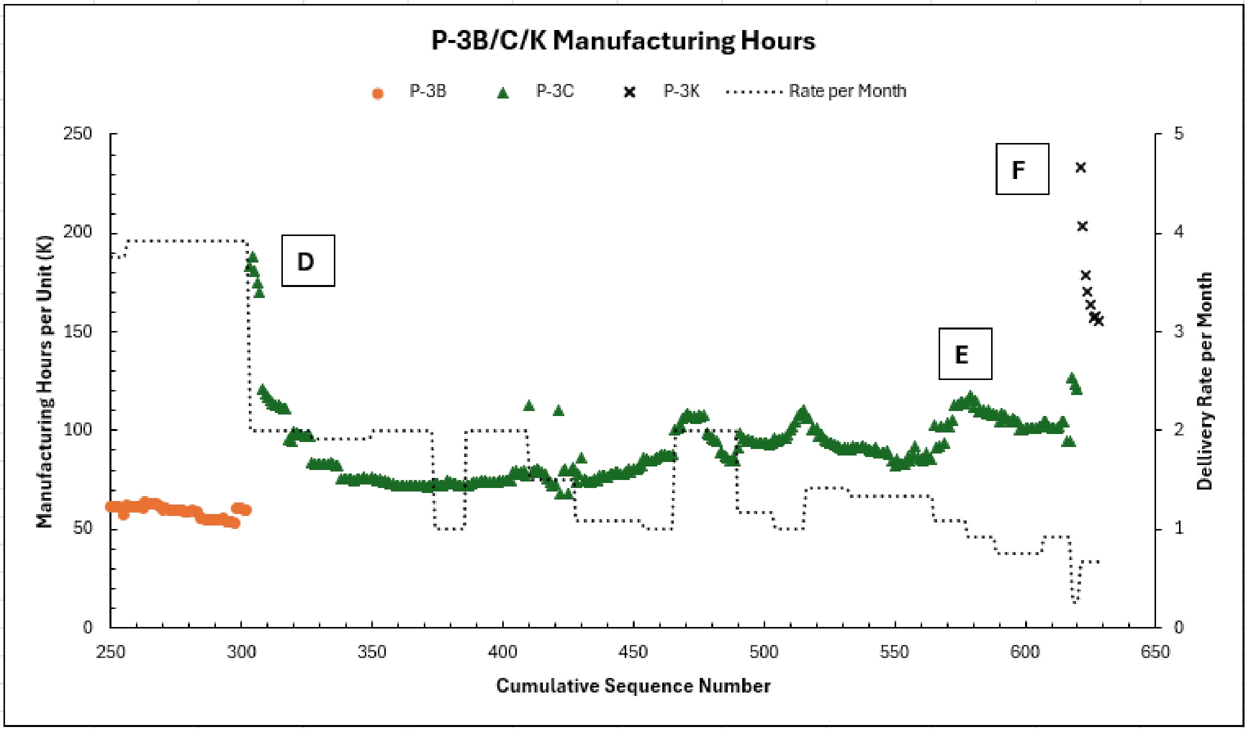


Figure 10. P-3 Manufacturing Hours per Unit (Unit 250 & on)

loss of learning due to the exit of personnel with P-3 experience. Production rates continued to drop below one aircraft per month with a relatively slow rate of improvement before spiking up for what was presumed to be the end of the program.

F. **Units 621-628:** Lockheed planned to complete the P-3 production run in Palmdale. However, after Palmdale production had shut down, an order was received for eight additional P-3 aircraft from South Korea. Lockheed decided to build these additional aircraft not in California but in its Marietta, Georgia facility where the C-130 was (and still is) built. The Korean P-3s represented not only a second production line transfer but the impacts of a two-year production gap. Predictably, the results were disastrous with the build hours regressing all the way back up the learning curve to virtually the beginning of the program. While Lockheed was able to move back down the learning curve relatively quickly, the last Korean aircraft was still 50,000 hours per unit higher than the typical Palmdale aircraft. Lockheed attempted to find other P-3 buyers, but when none materialized it shuttered the Marietta line, closing the production program for good.

As it transpired, the Electra design proved to have a much longer life as a military aircraft than a commercial effort. Lockheed stopped the L-188 after 170 aircraft were built. The P-3 on the other hand sold over 600 aircraft over 30 years of production. In all, 14 countries operated the P-3 (Boyne, 1998).

Lessons from the L-188/P-3:

1. Some learning transfer can occur in the adaptation of a commercial product to a military purpose. The P-3 picked up approximately five shipsets of learning from the L-188 despite the P-3's structural modifications and new avionics suite.

2. New variant introductions can act as "reset" events for learning. The introduction of the P-3C with its major upgrades wiped out earlier learning gains and forced a prolonged recovery period. Even a relatively minor change like the P-3B created a modest flattening in the learning curve.

3. Production rate volatility negatively influences learning slopes. Steady production at maximum

rate allows learning to continue, while frequent rate oscillations or a production rate dropping below one per month flattens learning curve slopes.

4. Production line moves have major cost impacts. The P-3 is unusual in that it experienced not one, but two such moves in its lifetime. The move to Palmdale introduced a loss of learning with limited recovery, while the second move to Marietta – when combined with a two-year production gap – caused a near-total loss of accumulated experience, pushing unit hours back to the P-3's initial levels.

F-111 Aardvark
Source: Lockheed Martin

Case 6: The F-111 Aardvark

Manufacturer: General Dynamics (Fort Worth, San Diego)

Mission: Supersonic tactical fighter / bomber

Program Start: 1962

First Flight: 1964

Quantity Produced: 562 (excluding 2 partially complete F-111K)

Year Production Ended: 1976

Program Background

Perhaps the most controversial procurement in the history of the US Defense Department, the F-111 began as a joint program between the US Air Force and US Navy. The choice of General Dynamics (GD) as winner of the Tactical Fighter Experimental (TFX) competition immediately prompted a nine-month congressional investigation (*TFX Contract Investigation*, 1964).

The F-111 was the first true multi-role military aircraft. Its moveable wings could be swept backward during flight for maximum speed. Extended forward, the wings allowed short takeoff and landings. It could also fly at supersonic speeds at both high and low altitudes, and its terrain-following radar allowed it to avoid detection while hugging close to the ground. Its long range meant it could fly from the United States to Europe without refueling (Lockheed Martin, 2020). At its 1964 rollout, Defense Secretary Robert McNamara stated, "For the first time in aviation history, we have an airplane with the range of a transport, the carrying capacity and endurance of a bomber, and the agility of a fighter pursuit plane" (Knaack, 1978b).

One technical challenge that could not be overcome was building an aircraft suitable for both the Air Force and the Navy. The services fought bitterly over the requirements, and neither was happy with the results. Eventually, seven F-111Bs were built and tested but the Navy terminated its participation in the program in 1968 (Miller, 1982). Not coincidentally, there was not another joint Air Force/Navy program until the F-35 Lightning II three decades later (Reece, 2011).

While the first F-111A test aircraft rolled out two weeks ahead of schedule, this success belied the difficulties faced by the development program. The immediate problems were engine malfunctions and aircraft weight. The TF30 engine compressor stalled at high speeds and angles of attack. Engineering traced the problem to an incompatibility between the engine and aircraft inlet (Miller, 1982). Multiple F-111 inlet configurations were developed but it was not satisfactorily solved until the final F-111 configurations (Knaack, 1978b). In addition, the aircraft was significantly overweight. USAF requirements grew the F-111A gross weight to 92,000 pounds – 30,000 pounds over plan. GD introduced a Super Weight Improvement Program (SWIP) but was only partially successful in reversing weight gains (Miller, 1982).

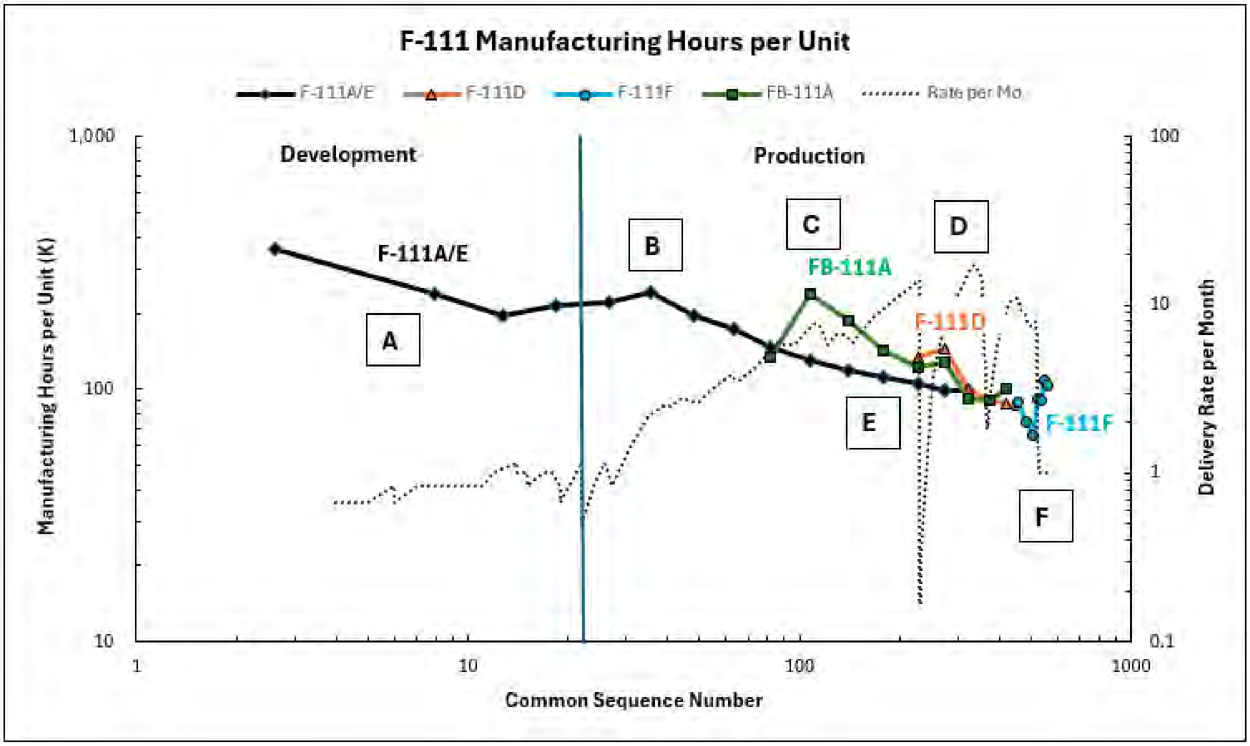


Figure 11. F-111 Manufacturing Hours per Unit by Lot (Fort Worth Only)

Learning Curve Analysis

Figure 11 displays the F-111 manufacturing hours per unit (*F-111A Manufacturing Program Overview;* Johnstone, 2007). The F-111 history is by production lot, rather than unit, so the plotted values in Figure 11 represent the lot midpoints. Six areas have been called out and are explained as follows:

A. **Lots 1-4 (Common #1 through #23):** Early F-111A development Lots 1-3 alone show a 77% slope but Lot 4 represented the introduction of the SWIP weight reduction configuration along with other design changes. These changes resulted in an approximately ten-unit learning loss. The SWIP configuration also forced scrap and replacement of tooling. The overall development slope including SWIP impacts was 81%.

   An internal F-111 Industrial Engineering analysis commented: “Considerable learning gain from the B-58 program…Some developmental tasks relating to high strength steel welding” (*F-111/B-58/B-36 Curve Slope Criteria*, undated). This refers to development of the troubled D6AC steel wing pivot and carry through structure.

B. **Lots 5-11 (Common #24 through #156):** F-111A production begins. In Lot 6 the first reconfiguration of the inlet (designated Triple Plow I) was introduced with an observable increase in hours. Other changes were introduced such as the addition of a weapons bay gun and reconfiguration of the weapons bay doors. At Unit #98, the program reached eight aircraft per month (*F-111/B-58/B-36 Curve Slope Criteria*, undated). However, in Lot 10, the UK F-111K program was terminated in mid-build, forcing a major version rescheduling.

C. **Lots 9-17 (Common #73-439, FB-111 #1-76):** The FB-111, a nuclear strategic bomber version, is introduced in Lot 9 with significant structural changes including a stretched fuselage, extended wing spans, stronger undercarriage and landing gear, and larger fuel tanks. New systems were introduced as well, including a terrain-following radar, digital computers and other upgrades. It also introduced an initial version of the redesigned Triple Plow II inlets (Knaack, 1978b). The introduction of the FB-111 resulted in an observable learning loss and its build hours were significantly higher than the F-111A over most of its build.

D. **Lots 13-18 (Common #204-#467, F-111D #1 -96):** The F-111D was introduced in Lot 13, featuring the new and revolutionary Mark II avionics system. Unfortunately, the avionics suppliers struggled with cost overruns and schedule delays. Integration of the individual components was difficult; one supplier complained that the original specifications were simply unachievable. Delays in receiving the Mark II avionics meant only one F-111D was delivered in June 1970 and no more for another year. When deliveries finally resumed, the F-111D program was years behind schedule (Knaack, 1978b).

E. **Lots 12-20 (Common #157-516):** Design problems plagued the F-111. The wing carry-through and pivot were built from D6AC alloy steel, whose properties were not fully understood. The wing carry-through experienced a fatigue test failure at Lot 12 at less than 50 percent of design life. This forced line inspections of the wing fitting and accompanying rework. In Lot 14, a delivered aircraft crashed due to failure of the D6AC steel forged wing pivot fitting. Subsequent investigation showed a one-inch surface crack had developed during fabrication and was missed repeatedly during nondestructive inspection. The F-111 fleet was grounded pending USAF investigation and deliveries ground to a halt with only two deliveries from January to July 1970 (Miller, 1982, Flores-Lamb, 2019). When deliveries resumed, more stringent acceptance standards were installed.

These issues worked to slow down the rate of cost improvement, particularly factoring in the

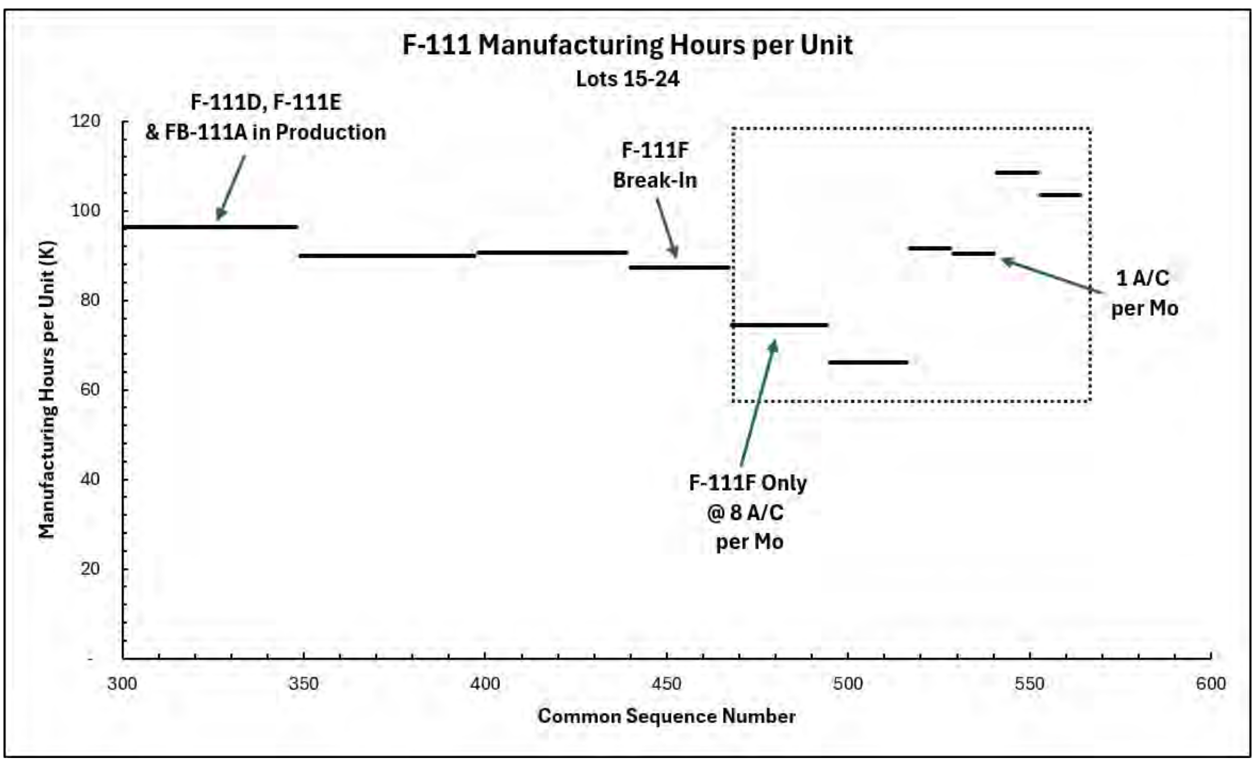


Figure 12. F-111 Manufacturing Hours per Unit, Lots 15-24 (Fort Worth Only)

impact of the FB-111A and F-111D variants. During this same period, the F-111E model was introduced in Lot 13 with the production variant of the Triple Plow II inlet and other minor upgrades.

F. **Lots 18-24 (Common #440 through #564).** Lot 18 introduced the F-111F with a stripped-down Mark II avionics systems and a variety of engineering improvements. By the subsequent lot, production on all other variants ceased. Interestingly, Lots 19 and 20 represent the all-time best F-111 cost performance with a substantial lot-over-lot drop in hours, as seen in Figure 12. This was attributed to good performance, the ability to focus on a single variant, and the use of leftover D model assets. Lot 21, on the other hand, showed a substantial slowdown in F-111 production rates, dropping to 1 per month. Contemporary documents show the F-111 had reached a "warm line" staffing level such no further personnel reductions can be made, or the program would be unable to retain all the technical skills needed for aircraft production (*F-111 Aircraft, Total Factory/Field Operation Contract 8260-13403-1130A-1130B Lot Cost*, undated).

Lessons from the F-111:

- Learning curve slopes are not static; they fluctuate with design-change intensity. Each major redesign (SWIP, Triple Plow inlet, FB-111A introduction) produced measurable learning loss.
- Variant proliferation is a major cost driver. Each new variant (FB-111A, F-111D, F-111F) reset or slowed the learning curve because of changes in tooling, planning, and worker knowledge.
- Weight reduction programs such as SWIP introduced substantial rework. SWIP created a ten-unit learning loss in early development as well as tooling replacement.
- Exotic materials can introduce substantial rework. The F-111's selection of D6AC steel forced eventual redesign of the wing carry-through, created delivery delays, and forced stricter inspections.
- Production rate and staffing levels influence cost performance. The increased production rate to eight aircraft per month in the early production lots helped maintain a strong learning curve despite design changes. On the other hand, the rate slowdown beginning in Lot 21 to one aircraft per month hit "warm line" minimum staffing levels resulting in substantially increased hours per unit.

L-1011 Tri-Star
Source: Wikimedia Commons

Case 7: The L-1011 TriStar

Manufacturer: Lockheed (Palmdale, Burbank)

Mission: Turbojet commercial airliner

Program Start: 1968

First Flight: 1970

Quantity Produced: 250

Year Production Ended: 1984

Program Background

After the market failure of the L-188 Electra, Lockheed began looking for another opportunity to reenter the commercial airline market. Unfortunately, Boeing and Douglas Aircraft had established a firm duopoly on the US turbojet market. Lockheed believed, however, there was an unserved niche in the market for a passenger widebody medium-range aircraft. True, Douglas was working on its new DC-10 aimed at the same market. But Lockheed had confidence that an offering with advanced technology (a new and more efficient engine, fly-by-wire technology, automated landing capability, a below-deck galley) would carry the day (Boyne, 1998).

In the end, this proved to be the first miscalculation of many. Lockheed chose Rolls Royce as its engine manufacturer, then watched as Rolls declared bankruptcy in early 1971 (*Business: Lockheed's Rough Ride with Rolls-Royce*, 1971). L-1011 build stopped, and thousands of employees were furloughed in Burbank and Palmdale. Lockheed was forced to ask for a $250 million loan guarantee from the US government, which it received in late 1971. This delayed delivery of the first aircraft to Eastern Airlines by six months and gave the DC-10 a competitive lead in a niche market that was too small for both companies to prosper.

The L-1011 pushed forward, weathering a downturn in orders as airlines struggled to deal with higher energy prices in the 1970's. Even in good times Lockheed could not make money; when orders surged in 1978 and 1979, the sudden and unanticipated demand meant Lockheed had to pay higher-than-expected wages and materials prices. At last, Lockheed threw in the towel and exited the commercial aviation business, with its last L-1011 delivered in 1984. In all, Lockheed lost $2.5 billion (then-year) on the L-1011 program with an average loss of $10 million per aircraft. This was despite an aircraft which the airlines found reliable and efficient (Greenwald, 1981; Boyne, 1998).

Learning Curve Analysis

During its production run, the L-1011 produced four different variants:

- Type 1 – first production model
- Type 100 – common fuselage with Type 1, new center fuel tank, extended range
- Type 200 – common fuselage with Type 1, identical to type 100 except new upgraded engines
- Type 500 – an extended range aircraft, Type 500 had a shortened fuselage (14 feet), increased wingspan, and improved systems (Lockheed L1011 TriStar, 2025).

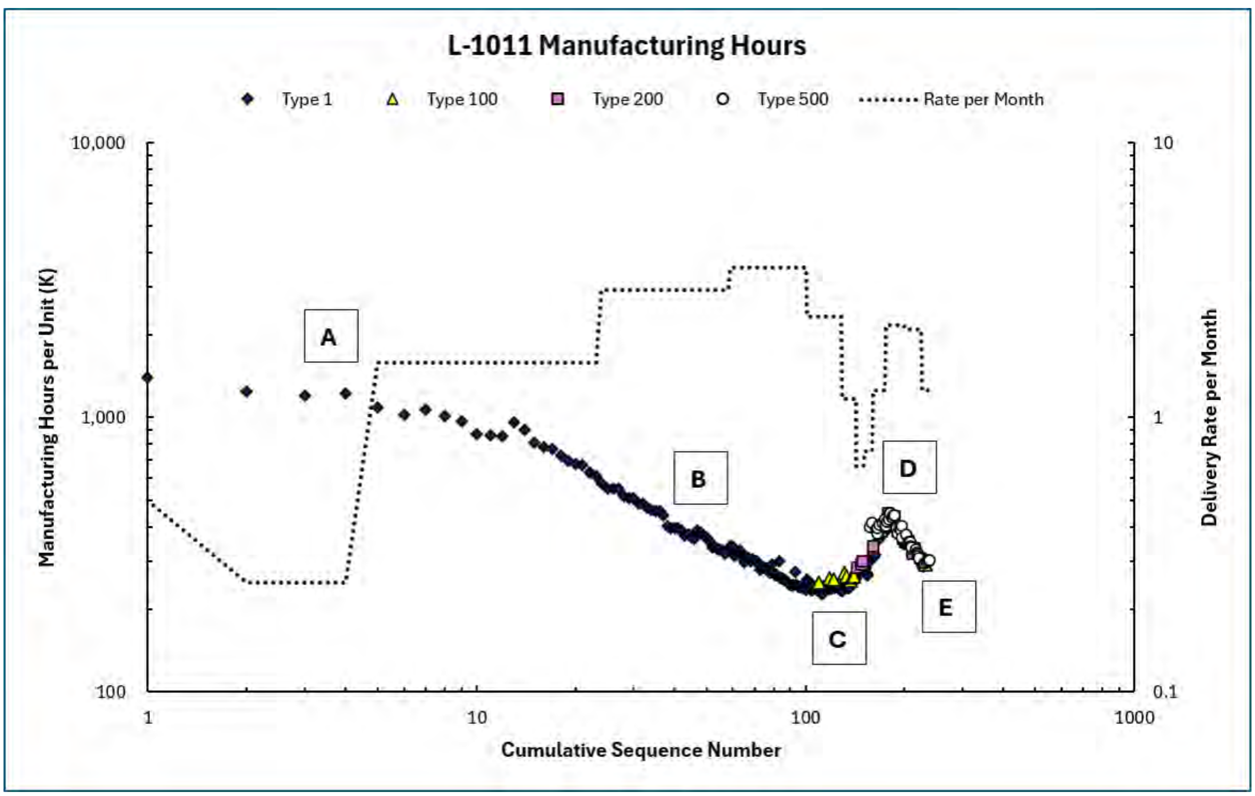


Figure 13. L-1011 Manufacturing Hours per Unit (Palmdale/Burbank)

Stanford economist Lanier Benkard used the L-1011 as his primary data source on his paper on organizational forgetting, and many of his conclusions will be repeated here. Regarding the cost impacts of the different L-1011 variants, Benkard concluded the hours per unit impacts between the Type 1, 100, and 200 were not statistically significant – not surprisingly since all three versions share the same fuselage. However, he did find a cost impact associated with the Type 500 with its unique fuselage. Based on his analysis, Benkard established that the work between the L-1011-1/100/200 and the L-1011-500 was 73% common, the rest being unique to the Type 500 (Benkard, 2001).

Figure 13 displays the L-1011 manufacturing hours per unit. We will use the data set used in Benkard's published analysis of the L-1011 which covers the first 238 aircraft by individual aircraft. This leaves out the last 12 aircraft – data for these aircraft is incomplete and therefore omitted (Benkard, 2001). Five areas have been called out and are explained as follows:

A. **Units 1-12**: Five aircraft were used for the flight development program before being eventually refit and delivered to airlines (Birtles, 1998). The flat slope continued beyond these five aircraft, persisting through unit 12. The reasons for this 87% slope are not identified but they are likely related to the slow recovery of engine deliveries of Rolls Royce and the rehire of personnel after the work stoppage forced by the Rolls bankruptcy.

B. **Units 13-100:** The L-1011 learning curve sharpened steeply at this point to a 64% slope as the production rate increased to over three per month.

C. **Units 101-130:** After the 100th unit, L-1011 hours per unit flattened out as production rates decreased to two per month. Although the Type 100 variant comes online, this appears to have little impact.

It is after the 100th unit of the L-1011 when the unit cost first flattens and then rises sharply. Figure 14 shows a larger view of these phases.

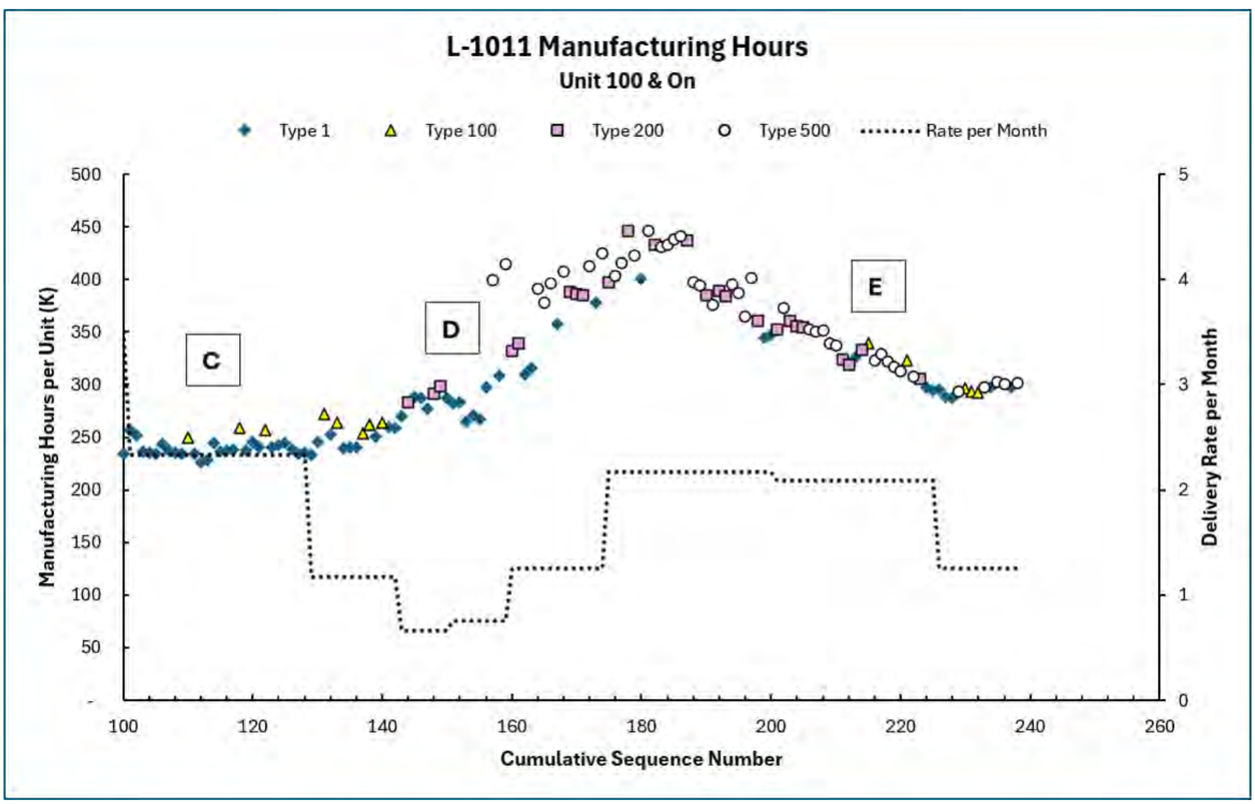


Figure 14. L-1011 Manufacturing Hours, Unit 100 & On

D. **Units 131-180:** This most remarkable feature of this phase is the continued reduction in production rates. This drop can be traced back to the 1973 Arab oil embargo and the continued fuel crisis of the mid-1970s as the Organization of Petroleum Exporting Countries (OPEC) exploited its cartel power to raise oil prices. Airlines around the world, faced with rising fuel prices and declining air traffic, throttled back on new orders. The L-1011 declined from 41 deliveries in 1973 to as few as eight aircraft five years later (Birtles, 1998; Boyne, 1998).

It is this precipitous decline in production rates that attracted Benkard to study the L-1011 experience as the primary source on his paper on organizational learning. Benkard wrote:

> Since an aircraft firm's experience is embodied in its workers, it seems likely that turnover and layoffs may lead to losses of experience….Experience may depreciate in times of falling production rates, since typically such times are accompanied by layoffs. During subsequent increases in production, the firm is often unable to acquire the same workers that it formerly released and must completely retrain new ones. Normal rates of employee turnover may also lead to experience depreciation during periods of constant production, particularly if turnover rates are high (Benkard, 2001).

This phenomenon is compounded by the union contracts frequently in use in the aircraft industry:

> As hourly employees attain seniority, they have the option of requesting an upgrade into a higher-level job classification if one should become available. At the same time, the company is bound to fill positions from within the existing ranks whenever possible, even if that means retraining. Of course, if an opening is filled from within, then the position vacated by the worker upgraded is subject to the same rules. Similarly, when there are layoffs, the higher-seniority workers can bump down lower-seniority workers, and so forth down the ranks. Whenever employment changes, there is a domino effect throughout the company that has been known to affect as many as ten positions for just one job opening. This "bumping" has the greatest impact on increases in employment, since it could mean that the company must retrain existing workers in their new positions as well as the usual training of new workers. Under "bumping," turnover rates are high, which leads to greater organizational forgetting (Benkard, 2001).

Lastly, it should be noted that this period also sees the introduction of the Type 500 variant which can be expected to increase labor hours due to reconfiguration and accompanying loss of learning. Even after controlling for the impact of the Type 500 introduction, however, Benkard still concluded that the impact of employee turnover was the most significant factor in the increase of hours per unit during this time.

E. **Units 181-238**: The L-1011 deliveries bottomed out in 1977 and 1978, then increased as airline orders picked up and Lockheed returned to two aircraft per month. Hours per unit began to decline as new workers hired to address this surge in orders became more proficient as well as the production line's adjustment to the peculiarities of the Type 500 version. However, production deliveries after 1981 slumped again, falling to one per month and the observed hours per unit flatten out.

The original plan was for the L-1011 to be delivered at a maximum rate of 10 per month. (Birtles, 1998). In fact, it never came even close, peaking at less than four per month and frequently operating much less than that, as we have seen.

Lessons from the L-1011:

1. Increases or decreases in production rates drive employee turnover and reassignments, which in turn have cost implications due to lost organizational learning.
2. Major engineering reconfigurations, such as the introduction of the Type 500, increase unit cost and drive loss of learning.
3. External supply shocks (Rolls Royce bankruptcy, financial distress, 1970s fuel crisis) forced Lockheed to cut output, lay off workers, and later rehire, each step resetting the learning curve.

Conclusions

We began this paper by asking: Can cost history from older, long-shuttered programs teach today's cost estimators anything useful? Our examination of these seven programs has revealed many situations – configuration changes, production gaps, swings in delivery rates, new technologies, to name a few – which seem familiar. If in fact history does not repeat, but instead rhymes, what overarching lessons can we draw from our seven case studies to assist today's estimators?

1. Use a multiple leg learning curve framework to separate the product lifecycle into distinct phases (development, early production, later production) and assign a unique slope to each phase. In all cases except the L-188, a single slope learning curve across all program phases would have been a poor fit to the data.

2. Treat major design or variant changes as a learning reset event – quantify the proportion of work that is new and set it back on the learning curve to capture the associated loss of learning. We saw numerous examples in our case studies such as the F-102 "area rule" reconfiguration, the F-111A SWIP and Triple Plow inlet reconfigurations, and the P-3C redesign.

3. Production rate effects should be explicitly modeled. As the L-1011, F-111 and P-3 delivery rate swings indicate, production rate increases or decreases can create turnover, new job assignments and force retraining. This cost impacts can be particularly acute in a "warm line" environment when critical skills personnel must be retained even at rock-bottom delivery rates. Labor skill and seniority impacts should be considered especially during ramp-up or ramp-down phases.

4. New technologies that push the state of the art can create high first unit costs combined with atypically steep learning curve slopes as seen in the case in the B-58. This can be an exception to the typical rule that development slopes are relatively flat, and slopes do not steepen until the program enters early production.

5. Production line transfers, as seen in the P-3 case study, can erase a substantial fraction of accumulated experience. When combined with a years-long production gap, virtually all the learning experienced to date can be lost. Line transfers should be modeled with a steep learning loss penalty with a subsequent reset period.

6. Learning transfer between an older program and its derivative or follow-on is not guaranteed. The F-106 showed little to no learning from its predecessor, the F-102; and learning gain from the L-188 to the P-3 was limited to a relatively small number of ships. Do not assume full carry -over.

7. Capture tail-up effects in the final lots – as programs wind down (F-102, F-106, F-111, B-58, P-3), hours per unit typically rise sharply. Include a "tail-up" multiplier for the last lots of production.

By embedding these seven principles into cost estimating models, analysts can more accurately forecast labor hours, unit costs, and schedule risk for both current production and new aircraft development projects.

References:

Air Force Life Cycle Management Center (AFLCMC). (2024, September.) *This Week in AFLCMC History – September 23-30, 2024.* Retrieved from https://www.aflcmc.af.mil/NEWS/Article-Display/Article/3915525/this-week-in-aflcmc-history-september-23-30-2024/#:~:text=At%20its%20roots%2C%20Cook%20and,to%20the%204th%20Generation%20fighters.

Asher, H. (1956). Cost-Quantity Relationships in the Airframe Industry. Santa Monica, California: RAND Corporation.

B-58 Manufacturing Manhours by Department, Airplane, Lot & Contract. (Undated). Fort Worth, TX: General Dynamics Fort Worth Division.

Benkard, C. L. (2001). Learning and Forgetting: The Dynamics of Aircraft Production. *American Economic Review* 90(4): 1024-1053.

Birtles, P. (1998). *Lockheed L1011 Tri-Star.* Osceola, WI: MBI Publishing Company.

Boyne, W. J. (1998) *Beyond the Horizons: The Lockheed Story*. New York, NY: St. Martin's Press.

"Business: Lockheed's Rough Ride with Rolls-Royce." (1971, February 15). *Time.*

"Convair F-102A Delta Dagger." National Museum of the United States Air Force. Retrieved from: https://www.nationalmuseum.af.mil/Visit/Museum-Exhibits/Fact-Sheets/Display/Article/198057/convair-f-102a-delta-dagger/.

"Convair F-106A Delta Dart." National Museum of the United States Air Force. Retrieved from: https://www.nationalmuseum.af.mil/Visit/Museum-Exhibits/Fact-Sheets/Display/Article/196408/convair-f-106a-delta-dart/.

Crewloading: Manufacturing Work Control Technique. (1979, October.) Fort Worth, TX: General Dynamics Fort Worth Division.

F-16A/B FSD-Production Manufacturing Estimating Rationale. Fort Worth, TX: General Dynamics Fort Worth Division.

F/TF-102 Estimating Handbook and Product Book. (Undated.) San Diego, CA: General Dynamics Convair Division.

F-106A/B Estimating Handbook and Product Book. (Undated.) San Diego, CA: General Dynamics Convair Division.

F-111 Aircraft, Total Factory/Field Operation Contract 8260-13403-1130A-1130B Lot Cost. (Undated). Fort Worth, TX: General Dynamics.

F-111A Manufacturing Program Overview. (Undated) Fort Worth, TX: General Dynamics.

F-111/B-58/B-36 Curve Slope Criteria. (Undated). Fort Worth, TX: General Dynamics Fort Worth Division.

Flores-Lamb, J. and Motes, D. T. III. (2019, June) "Historical Rationale for Design Restrictions on D6AC Steel in US Aerospace Structures." Defense Systems Information Analysis Center. Retrieved from: https://dsiac.dtic.mil/technical-inquiries/notable/historical-rationale-for-design-restrictions-on-d6ac-steel-in-u-s-aerospace-structures/#:~:text=D6AC%20is%20considered%20a%20medium,is%20listed%20in%20Table%201.

Greenwald, J. (1981, December 21). "Catch a Falling TriStar." *Time*.

Hall, R. C. (1981, November-December.) "The B-58 Bomber: Requiem for a Welterweight." *Air University Review* 33(1): 44-56. Retrieved from: https://babel.hathitrust.org/cgi/pt?id=mdp.39015011368860&seq=3

Johnstone, B. M. (2007, August). *F-111 Production – Total Hours by Lot.* Unpublished Study.

Joyce, P. "Sandwich Structures." (2003) Retrieved from: https://www.usna.edu/Users/mecheng/pjoyce/composites/Short_Course_2003/13_PAX_Short_Course_Sandwich-Constructions.pdf

Knaack, M.S. (1978a). *Post-World War II Bombers: 1945-1973, Washington, D.C.: Office of Air Force History.* Retrieved from https://apps.dtic.mil/sti/tr/pdf/ADA209273.pdf.

Knaack, M.S. (1978b). *Post-World War II Fighters: 1945-1973, Washington, D.C.: Office of Air Force History.* Retrieved from https://books.google.com/books?id=xDOU991tA0sC&pg=PA158&source=gbs_toc_r&cad=2#v=onepage&q&f=false

Levenson, G. S. (1965) *Manufacturing Labor Hours for Aircraft Airframes*. Santa Monica, CA: RAND.

"Lockheed L1011 TriStar." (2025) Retrieved from: https://en.wikipedia.org/wiki/Lockheed_L1011_TriStar.

Lockheed Martin. (2020, October). *F-111*. Retrieved from https://www.lockheedmartin.com/en-us/news/features/history/f-111.html#:~:text=The%20F%2D111%20is%20the,as%20little%20as%202%2C000%20feet

Lorell, M., Leavaux, H. P. (1998). *The Cutting Edge: A Half Century of US Fighter Aircraft R&D.* MR-9390-AF. Santa Monica, CA: RAND. Retrieved from https://apps.dtic.mil/sti/tr/pdf/ADA354082.pdf.

Miller, J. (1982). *Aero Series Vol. 29 General Dynamics F-111 "Aardvark."* Fallbrook, CA: Aero Publishers, Inc.

Miller, J. (1985). *Aerograph 4 Convair B-58*. Arlington, TX: Aerofax, Inc.

Miller, J. (2011, June). *B-58A 55-664 Accident Report: The Last Formal Act*. Retrieved from: http://sftentx.org/North_Texas_Chapter_SFTE/Event_Photos/Pages/Jay_Miller_Paper.html

Model P-3A/B Fact Sheet. (Undated). Palmdale, CA: Lockheed Martin.

Model P-3C Fact Sheet. (Undated). Palmdale, CA: Lockheed Martin.

Model P-3K Fact Sheet. (Undated). Palmdale, CA: Lockheed Martin.

Mutza, W. (1999). *Convair F-102 Delta Dagger*. Atglen, PA : Schiffer Military History.

P-3 Orion Research Group – The Netherlands. (2013). Retrieved from: https://www.p3orion.nl/history.html

Price, T.L., Dalley, G., McCullough, P.C., Choquette, L. (1997, April). *DOT/FAA/AR-96/75 Handbook: Manufacturing Advanced Composite Components for Airframes.* Federal Aviation Administration: Washington, DC. Retrieved from https://apps.dtic.mil/sti/tr/pdf/ADA325978.pdf.

Reik, T. (1965). *Curiosities of the Self: Illusions We Have about Ourselves.* New York, NY: Farrar, Straus & Giroux.

Rhodes, J. (2012, August 22). *Code One.* "P-3 Orion Anniversary." Retrieved from: https://www.codeonemagazine.com/p3_article.html?item_id=107

Robinson, D. H. (1967.) *The B-58 Hustler.* New York, NY: Arco Publishing Company

Reese, B. L. (2011) *Development of the TFX F-111 in the Department of Defense's Search for Multi-Mission, Joint-Service Aerial Platforms*. Master's Thesis. U. S. Army Command and General Staff College. Retrieved from: https://apps.dtic.mil/sti/tr/pdf/ADA547500.pdf.

Sawyer, J; Smouse, R.; Tetrault, P. (2002, January). *Airframe and Avionics Cost Improvement Relationships*. Arlington, VA: Joint Strike Fighter Program Office.

Smith, L.M. and Rogers, C. W. (1962.) "Bonded Bomber - B-58." *SAE Transactions* (70): 477-486.

Swigger, J. (2014). *"History Is Bunk": Assembling the Past at Henry Ford's Greenfield Village.* Amherst, MA: University of Massachusetts Program.

TF-102A Nose Program History. (1963, October.) Fort Worth, TX: General Dynamics.

TFX Contract Investigation. (1964) Hearings before the Permanent Subcommittee on Investigations of the Committee on Government Operations, United States Senate, Eighty-Eighth Congress. Washington, DC: Government Printing Office 10 parts.

Tooling Historical Facts and Figures. (1963.) Fort Worth, TX: General Dynamics.

Younossi, O., Kennedy, M., Graser, J. (2001.) *Military Airframe Costs: The Effects of Advanced Materials and Manufacturing Processes.* MR-1370. Santa Monica, CA: RAND.

Photo Credits:

All photographs are in the public domain and free of copyright restrictions.

B-58: https://commons.wikimedia.org/wiki/File:B-58_Hustler.jpg

F-102: https://www.nationalmuseum.af.mil/Visit/Museum-Exhibits/Fact-Sheets/Display/Article/198057/convair-f-102a-delta-dagger/

F-106: https://www.torch.aetc.af.mil/About/Fact-Sheets/Display/Article/364668/dissecting-the-delta-dart/

F-111: https://www.lockheedmartin.com/en-us/news/features/history/f-111.html#:~:text=The%20F%2D111%20is%20the,as%20little%20as%202%2C000%20feet

Henry Ford: https://commons.wikimedia.org/w/index.php?search=henry+ford&title=Special%3AMediaSearch&type=image

L-1011: https://commons.wikimedia.org/wiki/File:Lockheed_L-1011-1_Tristar,_Lockheed_JP5893645.jpg

L-188: https://commons.wikimedia.org/wiki/File:Varig_Lockheed_L-188A_Electra_Groves-1.jpg

P-3: https://www.lockheedmartin.com/en-us/news/features/history/p-3-orion.html

Theodor Reik: https://commons.wikimedia.org/wiki/File:Theodor-Reik.jpg

Brent Johnstone is a LM Fellow at Lockheed Martin Aeronautics Company. He has 38 years' experience in the military aircraft industry, including 35 years as a cost estimator. He has worked on the F-16 program and Advanced Development Programs and has been for 29 years the lead Production Operations cost estimator for the F-35 program. He has a Master of Science from Texas A&M University and a Bachelor of Arts from the University of Texas at Austin.

Parsimonious Cost Estimating Relationships for Early-Stage Aircraft Cost Prediction

First Lieutenant Austin J. Romanov

Dr. Robert D. Fass

Dr. Jonathan D. Ritschel

Major Michael J. Brown

Shawn Valentine

Accurate early-stage cost estimates are important as the Department of War (DoW) and U.S. Air Force (USAF) pursue advanced aircraft under strict affordability pressures (Office of the Secretary of War, 2025). We develop practical and parsimonious cost estimating relationships (CERs) for recurring T100 flyaway cost using variables that are observable during the earliest phases of concept development. Using airframe specifications from Air Force Life Cycle Management Center (AFLCMC) and weight statements from Cost Assessment Data Enterprise (CADE), we evaluate physical dimension and capability-based variables to determine their relative influence on cost. Focusing on aircraft with a first flight in or after 1970, the final model relies on only two early-stage predictors – weight and stealth capability – to explain 84.6% of the variation in recurring T100 flyaway cost. Rather than expanding the model with additional predictors that are typically unavailable or unreliable during concept development, the final specification emphasizes transparency and early observability. The results demonstrate that simple, decision-relevant inputs can support credible early-stage cost estimation. These inputs can inform affordability-driven trade studies, such as those between stealth and non-stealth capabilities, as well as requirements setting and acquisition decisions as early as Milestone A.

Introduction

Accurately estimating aircraft cost in the earliest stages of acquisition remains a persistent challenge, largely because traditional CERs depend upon detailed subsystem definitions or performance specifications that are not yet well-defined during concept development. Early design decisions must therefore be made in data environments that are too sparse to support many established parametric approaches. This exploratory study investigates whether physical dimensions and observable capability attributes – variables available well before detailed engineering data emerge – can provide a credible basis for early-stage cost prediction. By developing and comparing several parsimonious model forms, the study offers flexible tools aligned with the data constraints of early acquisition contexts.

The U.S. Air Force (USAF) and Department of War (DoW) face increasing pressure to deliver advanced capabilities both rapidly and affordably in preparation for potential near-peer conflicts (Office of the Secretary of Defense, 2022; Congressional Research Service, 2024). In this environment, accurate cost data and estimation plays a critical role in acquisition decision-making (Reese et al., 2021), ensuring budget realism, and enabling effective program and portfolio-level planning (U.S. Government Accountability Office, 2020). Developing CERs – statistical models that relate system cost to physical and performance-based characteristics – is a cornerstone of rigorous cost analysis and forecasting.

Recent acquisition reforms have institutionalized affordability and cost realism as central considerations in defense procurement. The Weapon Systems Acquisition Reform Act

(WSARA) (2009) and subsequent Better Buying Power (BBP) initiatives strengthened expectations for credible cost estimates and improved acquisition discipline, even though concerns about affordability and cost growth predated these reforms. More recent DoW acquisition reform guidance places renewed emphasis on improving cost estimation practices as a means of achieving better acquisition outcomes, including more informed early trade-offs, reduced cost growth, and greater program stability (Office of the Secretary of War, 2025). Collectively, these reforms underscore the growing importance of early, accurate life-cycle cost estimation.

Although all CERs rely on historical data, this dependence presents challenges when applied to emerging technologies or platforms with novel performance characteristics. Rojo et al. (2023) incorporate a binary indicator equal to one for aircraft with first flights between 1944 and 1968 to account for generational differences in cost data. While this approach captures broad shifts across historical eras, such binary indicators may not fully reflect the underlying design and capability attributes that drive cost in next-generation systems. Cisneros et al. (2024) demonstrate that an expanded range of predictor variables including design attributes, performance, and categorical variables can improve model relevance for modern platforms with advanced capabilities. Together, these studies highlight a central tension in CER development: historical data remain indispensable as the empirical basis for estimation, yet sustaining relevance for modern systems requires careful selection of predictors that reflect evolving design and capabilities.

Across several decades of cost-analysis research, studies have employed CERs using T100 recurring flyaway cost as the dependent variable (Rojo, et al., 2023; Hildebrandt & Sze, 1986). T100 reflects the projected cost of producing the 100th aircraft, a point at which early learning-curve inefficiencies have largely stabilized. As a result, the T100 metric provides a consistent basis for comparing recurring airframe production costs across platforms and is widely regarded as a reliable, decision-relevant measure for estimating production cost behavior. Rojo et al. (2023) further demonstrate the value of this approach by developing Lot Mid-Point (LMP) and Average Unit Cost (AUC) models to estimate T100 flyaway cost for early-stage acquisition analysis. Building on this foundation, our analysis develops new CERs using T100 flyaway cost data and detailed airframe specifications obtained from the Air Force Life Cycle Management Center (AFLCMC).

Background

Department of Defense Instruction (DoDI) 5000.73 and statutory milestone certifications mandate the use of credible, decision-relevant cost analyses to support acquisition planning, PPBE execution, and long-term program stability (Department of Defense, 2024; U.S. Government Accountability Office, 2020). Because early baselines exert lasting influence over downstream affordability (Jones et al., 2014; Office of the Secretary of Defense, 2025), the Office of the Secretary of War (OSW) requires early estimates to be realistic, defensible, and comprehensive. Achieving this standard requires blending quantitative rigor with professional judgment, as these analyses are conducted under conditions of incomplete design definition (Office of The Secretary of Defense, 2022). Consequently, early-stage credibility rests not only on analytical rigor, but on utilizing cost relationships that are observable, interpretable, and appropriate to the program's current phase (Department of Defense, 2024).

The Analysis of Alternatives (AoA) represents the most consequential early decision point in the acquisition lifecycle, as cost estimates developed during this period frequently serve as the primary inputs informing affordability assessments, initial program budgets, and contracting strategies (Office of the Secretary of Defense, 2022). At this stage, analysts are typically constrained to a narrow set of observable design attributes, relying on historical analogies and high-level system characteristics rather than detailed subsystem specifications. The AoA Cost Estimating Handbook cautions that reliance on poorly grounded data or overly complex modeling assumptions during this phase can establish a "faulty foundation," committing the Department to acquisition paths based on inaccurate financial expectations (Office of the Secretary of Defense, 2022). In response, the Director of Cost Assessment and Program Evaluation (DCAPE) emphasizes sufficiency reviews that prioritize transparent model structure, consistent work breakdowns, and defensible cost drivers that can be traced to available technical information (Office of the Secretary of Defense, 2022).

Cost estimation supports the defense acquisition process and extends to broader strategic planning activities such as wargaming, capability development, and budget prioritization. Cost estimating relationships play a central role in this effort by providing analysts with a repeatable, data-driven framework for generating estimates early in the acquisition lifecycle, before detailed cost outcomes can be observed. Accurate and timely cost estimates are essential to informed decision-making at all levels of Air Force leadership (Reese et al., 2021).

Prior CER Studies

Early efforts to develop aircraft CERs established an empirical foundation by demonstrating that fundamental system specifications – primarily weight and speed – could forecast production costs with considerable accuracy (Levenson, Boren, Tihansky, & Timson, 1972; Large, Campbell, & Cates, 1976). However, as aircraft designs advanced, the factors driving cost necessarily evolved. The introduction of advanced composites in the 1970s marked a profound shift in aircraft production technology, introducing significant new complexities in design and manufacturing that altered traditional cost relationships (Rogers, 1992).

To account for these technological shifts, subsequent models expanded beyond basic physical dimensions to incorporate variables capturing evolving manufacturing practices, advanced materials, and structural breaks in production (Hess & Romanoff, 1987; Younossi, Kennedy, & Graser, 2001). While this literature underscores the importance of integrating detailed technical attributes as programs mature, it also highlights a critical gap: highly complex, subsystem-dependent models are often incompatible with the severe informational constraints of early concept development."

Building on these approaches, Rojo et al. (2023) apply log-linear models to derive CERs that incorporate drivers such as *empty weight*, *speed*, and *mission type*. Their models achieve strong predictive performance relative to untransformed linear specifications but also include additional variables such as *legacy status*, *stealth*, and a *cohort* binary variable representing unique statistical groupings. Although these factors enhance explanatory power, the emphasis on model fit comes at the expense of parsimony and broad applicability.

Beyond traditional continuous cost drivers such as *empty weight*, *wingspan*, and *thrust*, stealth technology imposes substantial additional costs. Because stealth entails specialized provisions, advanced coatings, and unique structural materials, it significantly increases manufacturing complexity and recurring production expenses. The necessity of accounting for these impacts is well-documented: Younossi et al. (2001) report that stealth requirements drive major increases in engineering and tooling, while more recent studies by Light et al. (2024) and Rojo et al. (2023) confirm its statistically significant impact on flyaway cost models. Consistent with this prior research, we incorporate *stealth capability* as a binary variable. This approach effectively captures categorical design features that cannot be represented on a continuous scale while preserving overall model simplicity (Rojo et al., 2023; Hu & Smith, 2021).

Robust CERs should meet three criteria. First, they should consist of relevant and practical cost drivers grounded in engineering judgment and supported by subject matter experts (U.S. Government Accountability Office, 2020). Second, when estimated using regression-based methods, the statistical properties of the estimation approach must be evaluated and documented to ensure valid inference (Wooldridge, 2018). Third, CERs should employ variable representations that reflect the underlying nature of design decisions, including the use of categorical (binary) indicators for discrete capability choices and continuous variables for scalable performance attributes, when appropriate (Cisneros, et al., 2024; Rojo, et al., 2023). These criteria provide the foundation for empirical analysis in this study. Building on these principles, the present study tests models that integrate physical dimension, performance-based, and categorical cost drivers for aircraft T100 flyaway cost estimation while emphasizing a parsimonious, transparent, and broadly applicable approach.

Possible Explanatory Variables			
Variable Type	*Variable Name*	*Number of Observations*	*Description*
Binary	Fighter	54	Fighter Type Airframe
Binary	Bomber	11	Bomber Type Airframe
PRESS R^2	Transport	18	Transport Type Airframe
Binary	UAV	4	UAV Type Airframe
Binary	Trainer	5	Trainer Type Airframe
Binary	Recon	2	Recon Type Airframe
Binary	Electronic	3	Electronic Type Airframe
Binary	Patrol	2	Patrol Type Airframe
Binary	Rotary Wing	2	Rotary Wing Type Airframe
Binary	Unassigned	1	Unassigned Type Airframe
Binary	Stealth	4	Airframe with Stealth Capability
Binary	Cohort	5	Cohort Used in Rojo (2023)
Performance	Max Takeoff Weight	65	Max Takeoff Weight of Airframe
Performance	Max Speed	67	Max Speed of Airframe
Performance	Ceiling	45	Maximum Flying Ceiling of Airframe
Performance	Combat Radius	44	Combat Radius of Airframe
Performance	Weight Capability	65	Max Takeoff Weight of Airframe - Empty Weight of Airframe
Performance	Range	49	Range of Airframe
Physical	Height	80	Height of Airframe
Physical	Length	85	Length of Airframe
Physical	Wing Span	85	Wing Span of Airframe
Physical	Wing Area	76	Wing Area of Airframe
Physical	Empty Weight	85	Empty Weight of Airframe

Table 1: Possible Explanatory Variables

Methodology

A distinctive trait of this research is its emphasis on parsimony. While many existing CERs rely on detailed subsystem specifications, propulsion parameters, or manufacturing attributes defined late in a program's lifecycle, this study focuses on cost drivers available during early concept formulation. Taking an exploratory approach, we initially scan a broad set of easily observable physical dimensions (e.g., height, length, wingspan) and performance measures (e.g., weight, speed, thrust) to map out their relative explanatory power. Our analysis also tests binary metrics, such as stealth, to capture the cost disparities associated with categorical design features. Based on this open exploration, our final model specification demonstrates that a limited subset of these early-observable predictors can capture meaningful variation in recurring T100 flyaway costs. By grounding our specifications in these parsimonious variables, this approach provides cost estimators with a transparent, robust framework that balances analytical rigor with practical applicability across diverse aircraft classes.

Data Collection

We analyze data provided by AFLCMC that include recurring T100 flyaway costs along with selected specifications, performance measures, and physical characteristics for multiple mission-design-series airframes. The initial dataset contained a wide range of potential explanatory variables. To identify those most likely to influence recurring T100 flyaway cost, we first organized candidate variables according to their functional role in representing design and capability information. Specifically, variables were grouped as discrete capability indicators (e.g., mission type or categorical capabilities), physical dimensions (including empty weight), and continuous performance measures. This organization facilitates a systematic assessment of how different forms of design information contribute to variation in recurring flyaway costs for the 100th production unit.

The number of observations reported in Table 1 differs from the original AFLCMC dataset, since additional observations were incorporated using weight statements from the Air Force's Cost Assessment Data Enterprise (CADE). Multiple weight statements exist for each airframe and mission-design series within CADE; when available, the statement closest to the 100th production unit is selected. To supplement missing data, we also used official Air Force and Navy fact sheets. These sources provided additional information on length, height, wingspan, wing area, empty weight, and maximum takeoff weight.

One constructed measure we include in the analysis is *weight capability*, defined as the portion of an airframe's Maximum Takeoff Weight (MTOW) that remains after accounting for its Empty Weight (EW). Operationally, this measure reflects the maximum capacity that can be allocated to fuel, munitions, personnel, or other payloads. While we considered other individual measures like speed, range, or specific payload limits, these metrics vary heavily depending on the specific mission profile. Weight capability was ultimately chosen because it acts as a steady, all-in-one indicator of an aircraft's size and capability. By calculating weight capability as the difference between MTOW and EW, our analysis captures a performance-relevant characteristic that integrates both design and mission considerations. Compared to weight alone, weight capability provides a more mission-integrated indicator of an aircraft's usable capacity and serves as a capability-rooted driver of recurring flyaway cost.

Data Normalization

All cost data are normalized to Constant-Price (CP) 2024 dollars using an aircraft-specific Producer Price Index (PPI). This normalization removes escalation effects, including both general inflation and real price change, allowing historical program costs to be expressed on a common base-year basis. By removing time-dependent price effects, this approach ensures that observed cost differences and estimated relationships reflect variation in underlying program costs rather than escalation-driven effects across aircraft developed in different decades.

Data Inclusion/Exclusion

Our study develops and compares several model specifications to examine the relationship between aircraft dimension or capability metrics and recurring T100 flyaway cost. Our analysis proceeds in stages, beginning with bivariate log-log models that isolate the individual effects of key physical dimension variables. These are followed by multivariate specifications that integrate physical dimensions, performance metrics, and categorical capability indicators to assess joint explanatory power. Tables 2 and 3 summarize the key inclusion and exclusion criteria used to construct the analytical datasets used in our model construction.

In addition to these conceptual and data-availability criteria, preliminary statistical screening was conducted to inform variable reduction during model development. Stepwise

Data Inclusion/Exclusion: Airframe Scaling Model		
Inclusion/Exclusion Criterion	*Number of Observations*	*Notes*
Airframes with Cost Data Provided	100	Airframes which had provided cost data from AFLCMC
Airframes with Height Available from Weight Statement(s)	80	Airframes for which either height was provided by AFLCMC or height was located in weight statement(s) closest to the 100th production unit
PRESS R^2	85	Airframes for which either length was provided by AFLCMC or length was located in weight statement(s) closest to the 100th production unit
Airframes with Wingspan Available from Weight Statement(s)	85	Airframes for which wing span was provided by AFLCMC or wing span was located in weight statement(s) closest to the 100th production unit
Airframes with Wing Area Available from Weight Statement(s)	76	Airframes for which wing area was provided by AFLCMC or wing area was located in weight statement(s) closest to the 100th production unit
Airframes with Empty Weight Available from Weight Statement(s)	85	Airframes for which empty weight was provided by AFLCMC or empty weight was located in weight statement(s) closest to the 100th

Table 2: Data Inclusion/Exclusion: Airframe Scaling Model

Data Inclusion/Exclusion Strategic Capability Model		
Inclusion/Exclusion Criterion	*Airframes Excluded*	*Airframes Remaining*
Airframes with Cost Data Provided		100
Airframes with Stealth Data Provided	14	86
PRESS R^2	2	84
Airframes with Max Takeoff Weight Provided/Available from Official Fact Sheets	20	64
Exclusion of Training Airframes	5	59
Exclusion of Airframes Produced before 1970	32	27
Total Airframe Used In CER		27

Table 3: Data Inclusion/Exclusion Strategic Capability Model

regression procedures were used solely as an exploratory screening tool to assess the relative explanatory contribution of candidate variables and to identify combinations warranting further consideration. This procedure was not used to define final model specifications. Instead, final variable selection was guided by engineering logic, early-stage observability, coefficient stability, and interpretability within an early acquisition decision context. This approach ensures that statistical screening supports, rather than replaces, analytical judgment.

Our analysis develops two complementary model specifications to assess early-stage aircraft cost behavior: an *Airframe Scaling Model* that emphasizes geometric scale and a *Strategic Capability Model* that focuses on a limited set of early-stage capability drivers. We estimate two models because early-stage cost analysis can begin under different information conditions. The physical dimension bivariate models and *Airframe Scaling Model* applies when only physical dimension information from early drawings or layouts is available and approximates cost using observable measures of scale such as wingspan and wing area. The *Strategic Capability Model* applies to early trade studies in which a small number of mission-relevant requirements can be specified and explains recurring cost variation using a minimal set of early available capability drivers. Together, the two models provide complementary approaches for transparent and practical cost estimation prior to Milestone A.

For inclusion in our *Airframe Scaling Model*, each airframe must have the following data available: (i) recurring T100 flyaway cost, (ii) wingspan, and (iii) wing area. Airframes missing one or more of these required inputs are excluded.

Aircraft are included in the *Strategic Capability Model* only if they met four criteria: (i) recurring T100 flyaway cost was available, (ii) empty weight was documented in CADE contractor weight statements, (iii) MTOW was documented in CADE or service-level fact sheets, and (iv) stealth capability was recorded as a valid binary entry. We constructed the weight capability variable only when both weight measures are verified to ensure consistency and comparability across platforms. “Empty weight” refers to the value listed on the original contractor weight statement. When the contractor weight statement for the 100th production unit is unavailable, we source the value from the contractor weight statement closest in calendar date. If only annual data are available, we use the nearest available year. When MTOW is missing from the AFLCMC dataset, we obtain the value from official Department of the Air Force or Department of the Navy fact sheets. The AFLCMC dataset provides stealth capability as a binary indicator (true/false), and we introduce no additional observations. We exclude any airframe missing one or more required fields – *recurring T100 flyaway cost*, *empty weight*, *MTOW*, or *stealth capability* – from the analysis.

For the *Strategic Capability Model*, in addition to data availability, a primary inclusion criterion was established to ensure the dataset represents a technologically homogenous set of modern airframes. Based on the findings of Rogers

(1992), this analysis excludes all aircraft with a first flight date prior to 1970. This chronological boundary corresponds to the beginning of the widespread application of advanced composite materials in primary aircraft structures. By focusing the analysis on aircraft developed within this modern materials era, the resulting CER avoids distortions from older manufacturing paradigms and is more relevant for estimating the costs of contemporary and future systems.

Five trainer aircraft are excluded from the analysis as they were identified as statistical outliers and are not representative of the combat and combat-support airframes central to this study. The excluded airframes are the T-45TS Goshawk, T-6A Texan II, T-38A Talon, T-3A Firefly, and T-39A Sabreliner. These aircraft, whether propeller-driven (T-6A, T-3A), early-generation supersonic (T-38A), or specialized for non-combat roles (T-45TS, T-39A), reflect design priorities, production scales, and cost structures that diverge sharply from the rest of the dataset. Including these aircraft would distort coefficient estimates and reduce the model's validity for its intended purpose. Their exclusion ensures that the regression is not unduly influenced by atypical cases and that the results remain interpretable for the core set of operational airframes.

Model Specification

To make the estimated relationships explicit and reproducible, we summarize the regression specifications used in the analysis within both Table 4 and the corresponding regression equations. We estimate three related sets of models: (i) bivariate physical-dimension models that examine the individual relationship between recurring T100 flyaway cost and key scalar measures, (ii) a multivariate physical-dimensions model that jointly captures geometric scale effects, and (iii) a final parsimonious capability-based model that relies on a limited set of early-available predictors. Equations (1) - (3) define the functional forms used in each case and provide the basis for the regression results reported in the following sections. All models are estimated using ordinary least squares with heteroskedasticity-robust standard errors to provide inference that is robust to potential non-constant error variance given the wide variation in aircraft size and cost represented in the sample.

$$\ln(T100_i) = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (1)$$

Where $T100_i$ denotes recurring T100 flyaway cost for airframe i, X_i denotes a physical dimension evaluated in separate bivariate models (height, length, wingspan, wing area, or empty weight), and ε_i is an error term.

$$\ln(T100_i) = \beta_0 + \beta_1 \ln(Wingspan_i)\, \beta_2 \ln(WingArea_i) + \varepsilon_i \quad (2)$$

$$\ln(T100_i) = \beta_0 + \beta_1 \ln(WeightCapability_i) + \beta_2 Stealth_i + \varepsilon_i \quad (3)$$

Model Specifications		
Model	*Dependent Variable*	*Independent Variable(s)*
Physical Dimension (univariate)	*Ln(Recurring T100 Flyaway Cost)*	*Height, Length, Wingspan, Wing Area, Empty Weight, Ln (Height), Ln (Length), Ln (Wingspan), Ln(WingArea), Ln(EmptyWeight) (evaluated individually)*
Physical Dimension (multivariate)	*Ln(Recurring T100 Flyaway Cost)*	*Ln(Wingspan), Ln(WingArea)*
PRESS R^2	*Ln(Recurring T100 Flyaway Cost)*	*Ln(WeightCapability), Stealth*

Table 4: Model Specifications

Where

$$WeightCapability_i = MTOW_i - EW_i$$

and

$$Stealth_i = 1\ (stealth), 0\ (non - stealth).$$

Because aircraft cost data are non-negative and highly right-skewed, direct estimation in untransformed specification can violate key assumptions of ordinary least squares regression. We proceeded through model development iteratively, beginning with fully linear specifications used to assess scale effects and guide subsequent functional-form selection. Visual diagnostics indicated improved variance behavior and interpretability under logarithmic transformation. To address these issues, we employ logarithmic transformations that reduce skewness, stabilize variance, and improve model fit (International Society of Parametric Analysts, 2008). Log-log specifications are applied when both cost and predictors exhibit multiplicative relationships, producing elasticity-based interpretations that describe the percentage change in cost associated with a one percent change in an explanatory variable (Wooldridge, 2018). These functional forms are widely used in cost analysis and support valid statistical inference across heterogeneous aircraft classes (Rojo, et al., 2023; International Society of Parametric Analysts, 2008).

Statistical Analysis

We first investigate the explanatory power of physical dimension variables for the airframes in the dataset. Because the physical scale of an airframe directly determines its material and labor requirements, these predictors represent logical cost drivers. Before settling on a functional form, we evaluate height, length, wingspan, wing area, and empty weight by estimating bivariate regressions against recurring T100 flyaway cost to identify which physical characteristics most strongly correlated with recurring production cost and how that relationship behaves.

The results show these variables do not exhibit a simple linear relationship with cost. Consistent with prior research, modeling airframe cost in natural logarithms provides a better fit due to the non-negative, right-skewed distribution of cost data and the multiplicative nature of cost growth (U.S. Government Accountability Office, 2020; Wooldridge, 2018). Additionally, performance-related characteristics such as weight often demonstrate diminishing marginal returns: additional investment in these attributes improves capability only up to a point, beyond which the incremental gain per dollar declines because of inherent design and operational constraints. Logging the predictors as well addresses this non-linearity directly, yielding a log-log specification in which coefficients are interpreted as elasticities. We therefore estimate all simple regressions in log-log form.

Regression Diagnostics

To evaluate the statistical assumption underlying OLS regression, we conducted a comprehensive set of diagnostic tests on model residuals. Normality of residuals is assessed using the Shapiro–Wilk test, supported by histograms and Q -Q plots to examine distributional shape and tail behavior. While strict normality is not required for valid inference when using heteroskedasticity-robust standard errors, these diagnostics provide insight into residual behavior and potential model misspecification.

Homoskedasticity is examined using the Breusch-Pagan test and residual-versus-fitted plots to identify non-linearity or unequal variance. Inference throughout the analysis relies on heteroskedasticity-robust standard errors, mitigating sensitivity to violations of constant-variance assumptions. Potential outliers and influential observations were detected through Cook's distance, with values exceeding 0.5 flagged for further review (Weisberg, 2013).

Multicollinearity among predictors is assessed using Variance Inflation Factors (VIFs), which measure the degree to which each independent variable is linearly explained by the others.

Model robustness is additionally evaluated using the Prediction Sum of Squares (PRESS) R^2 statistic, a cross-validation measure that assesses

Descriptive Statistics for Bivariate Physical Dimensions						
Variable	*N*	*Mean*	*Median*	*Standard Deviation*	*CV*	*IQR*
T100 CP$K	85	$52,147.21	$26,562.12	$65,734.49	1.26	$45,906.41
Length (ft)	85	75.75	60.2	43.5	0.60	25.02
PRESS R^2	80	$52,208.24	$24,107.05	$67,626.91	1.3	$43,283.27
Height (ft)	80	21.75	16.23	12.38	0.57	6.76
T100 CP$K	85	$52,147.21	$26,562.12	$65,734.49	1.26	$45,906.41
Wingspan (ft)	85	74.56	53	52.41	0.70	62.8
T100 CP$K	76	$53,275.45	$24,107.05	$68,905.58	1.29	$41,753.68
WingArea (sq ft)	76	1107.06	547.4	1389.77	1.26	664.28
T100 CP$K	85	$53,819.70	$28,319.27	$66,153.30	1.23	$47,488.13
Empty Weight (lb)	85	52700.58	28473	70475.74	1.34	28883.55

Table 5: Descriptive Statistics for Bivariate Physical Dimensions

predictive accuracy and guards against overfitting. This diagnostic provides a cross-validation-based assessment of predictive capability.

Results

The empirical results identify which specific design and capability attributes drive variation in recurring T100 flyaway costs. We evaluate these relationships through a structured sequence of analyses: descriptive statistics for both the dependent and independent variables; a series of bivariate regressions estimated in log-log form; and two multivariate model specifications motivated by capability-based and physical-dimension frameworks.

Descriptive Statistics

A preliminary review of the descriptive statistics in Table 5 indicates substantial dispersion in *recurring T100 flyaway costs* across model specifications. Depending on the matched sample used for each bivariate relationship (N ranging from 76 to 85), the Coefficient of Variation (CV) for T100 cost ranges from 1.23 to 1.30, indicating the standard deviation exceeds the mean in every specification and reflects considerable relative variability in flyaway costs. The distribution is consistently right-skewed, as evidenced by the mean exceeding the median across all samples. These patterns are expected given each matched dataset spans aircraft with widely varying physical dimensions and weight characteristics, resulting in substantial heterogeneity in production costs.

Table 6 reports descriptive statistics for the variables included in the *Airframe Scaling Model* (N = 76). *Recurring T100 flyaway cost* exhibits substantial dispersion, with a coefficient of variation of 1.29, indicating the standard deviation exceeds the mean and reflects considerable relative variability in production costs across the sample. The distribution is right-skewed, as evidenced by the mean exceeding the median. Among the physical predictors, *wingspan* displays moderate relative variability (CV = 0.74), while *wing area* exhibits pronounced dispersion (CV = 1.26), with variability exceeding the mean. The substantial spread in wing area reflects the

Descriptive Statistics for Independent Variables (Airframe Scaling Model)						
Variable	*N*	*Mean*	*Median*	*Standard Deviation*	*CV*	*IQR*
T100 CP$K	76	$53,275.45	$24,107.05	$68,905.58	1.29	$41,753.68
Wingspan (ft)	76	71.17	44.75	52.49	0.74	40.53
PRESS R^2	76	1107.06	547.40	1389.77	1.26	664.28

Table 6: Descriptive Statistics for Independent Variables (Airframe Scaling Model)

Descriptive Statistics for Independent Variables (Strategic Capability Model)						
Variable	*N*	*Mean*	*Median*	*Standard Deviation*	*CV*	*IQR*
T100 CP$K	*27*	*$87,762.37*	*$53,690.14*	*$88,647.54*	*1.01*	*$77,132.73*
Weight Capability	*27*	*89512.76*	*39,527*	*111676.39*	*1.25*	*55,857.15*
PRESS R^2						

Table 7: Descriptive Statistics for Independent Variables (Strategic Capability Model)

presence of both compact tactical aircraft and large transport or bomber platforms, contributing to heterogeneity in the sample.

Table 7 presents descriptive statistics for the parsimonious specification (N = 27). In this reduced sample, flyaway cost continues to demonstrate high dispersion (CV = 1.01) and right -skewness. *Weight capability* exhibits even greater relative variability (CV = 1.25), indicating substantial heterogeneity in payload-related capacity across aircraft types. The magnitude of dispersion in both cost and weight capability reinforces the presence of scale effects within the dataset and further motivates the use of logarithmic transformation in subsequent regression analysis.

In addition to these distributional characteristics, preliminary visual diagnostics – including residual and scale-based plots from the initial bivariate regression fits – suggested increasing dispersion at higher levels of aircraft size and cost, consistent with scale-driven effects commonly observed in aircraft cost data. While formal testing does not detect heteroskedasticity, these diagnostics, combined with the substantial right-skewness and relative dispersion observed in the predictors, motivate the use of log-log transformations in subsequent regression analysis. Logarithmic specifications improve interpretability, stabilize scale effects, and better align the models with cost -estimating theory.

The Physical Dimension and Airframe Scaling Models

Our analysis first explores the possibility of predictive power bivariately for each of the physical dimensions available in our data set. Each relationship is estimated using the log-log equations (4) - (8). While all the physical dimension variables are statistically significant, we prioritize those with the greatest correlation for further consideration in multivariate CER development, informed by both bivariate performance and subsequent variable-screening procedures. In Table 8, we summarize the statistical performance of each dimension variable as it relates to the natural logarithm of recurring T100 flyaway costs.

$$\ln(T100) = 2.7367 + 1.7845 \ln(Length) \tag{4}$$

$$\ln(T100) = 5.3782 + 1.6322 \ln(Height) \tag{5}$$

$$\ln(T100) = 5.9820 + 1.0397 \ln(Wingspan) \tag{6}$$

$$\ln(T100) = 4.3826 + 0.9014 \ln(WingArea) \tag{7}$$

$$\ln(T100) = 1.3696 + 0.8665 \ln(EmptyWeight) \tag{8}$$

By exponentiating the estimated log-log bivariate specifications seen in Table 9, we express the results as closed-form CERs that relate recurring T100 flyaway cost directly to physical dimensions of the airframe in their original units of measure. In this form, the estimated coefficients are readily interpretable in terms of proportional or semi-elastic cost responses, enhancing their applicability during early stages of aircraft design. Most physical dimensions exhibit greater-than-proportional cost effects, indicating nonlinear scaling of cost with aircraft size; wing area is a notable exception, displaying a comparatively smaller elasticity with respect to recurring flyaway cost.

To assess the explanatory power of individual physical airframe characteristics, each bivariate regression is estimated in log-log form, yielding simple and easy-to-use CERs suitable for early

Bivariate Regression Results for Physical Dimensions - Natural Logarithm					
Variable	*Intercept*	*Coefficient*	*P-Value*	R^2	*Interpretation*
Ln Length (ft)	2.7367	1.7845	<0.001	0.508	For every 1% increase in aircraft length, recurring T100 flyaway cost is associated with an approximately 1.78% increase.
Ln Height (ft)	5.3782	1.6322	<0.001	0.383	For every 1% increase in aircraft height, recurring T100 flyaway cost is associated with an approximately 1.63% increase.
PRESS R^2	5.9820	1.0397	<0.001	0.324	For every 1% increase in wingspan, recurring T100 flyaway cost is associated with an approximately 1.04% increase.
Ln Wing Area (sq ft)	4.3826	0.9014	<0.0001	0.526	For every 1% increase in wing area, recurring T100 flyaway cost is associated with an approximately 0.90% increase.
Ln Empty Weight (lb)	1.3696	0.8665	<0.0001	0.631	For every 1% increase in empty weight, recurring T100 flyaway cost is associated with an approximately 0.87% increase.

Table 8: Bivariate Regression Results for Physical Dimensions - Natural Logarithm

lifecycle cost estimation. Among these models, empty weight, and wing area explain the greatest proportion of variation in the dependent variable (highest R^2) depending on the specification. Early in an airframe's development, when only preliminary engineering drawings or conceptual schematics may be available, these parsimonious CERs enable analysts to generate cost estimates using a single observable physical characteristic. This capability supports earlier and more frequent cost-informed decision-making as an airframe concept evolves from an initial design idea into a mature configuration.

Our next effort produced the multivariate CER result shown in Table 10, based on *wingspan* and *wing area*, using the specification shown in Equation (9) and the exponentiated equation for use in unit-space applications (10). Potential explanatory physical-dimension variables were initially evaluated using a stepwise regression procedure to identify a parsimonious subset of predictors that explain meaningful variation in the data while preserving interpretability. Stepwise regression was employed strictly as a screening tool rather than a model-defining mechanism. Final model specification was informed by engineering logic, coefficient stability, and statistical significance.

Exponentiated Bivariate Cost Estimating Relationships	
Predictor	*Log–Log CER*
Length (ft)	$T100 = 15.44 \cdot Length^{1.7845}$
Height (ft)	$T100 = 216.63 \cdot Height^{1.6322}$
PRESS R^2	$T100 = 396.23 \cdot Wingspan^{1.0397}$
Wing Area (sq ft)	$T100 = 80.05 \cdot WingArea^{0.9104}$
Empty Weight (lb)	$T100 = 3.94 \cdot EmptyWeight^{0.8665}$

Table 9: Exponentiated Bivariate Cost Estimating Relationships

Ultimately, *wingspan* and *wing area* emerged as the superior predictors because they act as holistic proxies for the aircraft's overall physical scale. Because aircraft dimensions generally scale in tandem to maintain aerodynamic stability and weight distribution, these two metrics may inherently capture the compounding cost impacts of corresponding increases in length and height, rendering those additional variables redundant.

Including both variables helps to prevent any omitted variable bias in our estimated coefficients. Though these two variables are highly correlated (0.92), their VIFs (6.85) are well below the conventional threshold of 10; this indicates moderate multicollinearity that does not materially affect coefficient stability or reliability (Wooldridge, 2018). In terms of cost scaling behavior, the model shows wing area functions as a highly stable, proportional driver within the CER. Conversely, changes to linear dimensions like wingspan introduce a more pronounced, volatile cost response, likely reflecting the compounding structural and design modifications required to support a wider span.

$$\ln(T100) = 4.4410 - 0.8052\ \ln(Wingspan) + 1.3945\ \ln(WingArea) \quad (9)$$

$$T100 = 84.86 \cdot Wingspan^{-0.8052} \cdot WingArea^{1.3945} \quad (10)$$

Regression Results for Airframe Scaling Model					
Variable	*Coefficient*	*Std. Error*	*z-Statistic*	*P-Value*	*95% CI*
Intercept	*4.441*	*0.598*	*7.421*	*<0.0001*	*[3.291,5.591]*
LnWingspan	*-0.8052*	*0.41*	*-2.002*	*0.045*	*[-1.594,-0.017]*
PRESS R^2	*1.3945*	*0.274*	*5.19*	*<0.0001*	*[0.868,1.921]*

Table 10: Regression Results for Airframe Scaling Model

As displayed in Table 11, this *Airframe Scaling Model* explains roughly 55% of variation in flyaway costs. We must also note the closeness of our adjusted R^2 and PRESS R^2 to our original value, indicating that our model does not suffer from an overfitting or artificial inflation to our coefficient of determination.

Goodness of Fit Metrics for Airframe Scaling Model	
Metric	*Value*
R^2	0.552
Adjusted R^2	0.540
PRESS R^2	0.515

Table 11: Goodness of Fit Metrics for Airframe Scaling Model

Airframe Scaling Model Regression Diagnostics

A suite of regression diagnostics was applied to verify that the multivariate *Airframe Scaling Model* satisfies the core assumptions of ordinary least squares and that the estimated relationships are not artifacts of model misspecification. Residual analysis indicated no significant departure from normality, and formal testing using the Shapiro-Wilk statistic ($p = 0.26$) supported this conclusion. Homoskedasticity was evaluated using the Breusch-Pagan test, which failed to reject the null hypothesis of constant error variance ($p = 0.18$), indicating no evidence of heteroskedasticity in the residuals.

Residual-fitted and scatter plot diagnostics were examined to assess functional form and scale behavior. These visual diagnostics revealed more stable variance patterns and clearer monotonic relationships under log-log specifications relative to untransformed alternatives, reinforcing the appropriateness of logarithmic functional forms for modeling scale-driven aircraft cost relationships.

Multicollinearity was assessed using VIFs. Although wingspan and wing area are highly correlated, their VIF values remained below conventional thresholds, indicating that multicollinearity does not materially inflate the variance of the estimated coefficients or undermine the precision of statistical inference. Examination of residual-versus-fitted plots confirmed that the log-log specification adequately captures the underlying functional form without systematic patterns indicative of misspecification. Finally, Cook's distance values indicated that no individual observations exert undue influence on the regression results.

Together, these diagnostics provide confidence that the *Airframe Scaling Model* is empirically well-behaved and suitable for predictive aircraft cost estimation. The results should be interpreted as descriptive associations rather than structural *ceteris paribus* effects, as physical dimensions likely proxy for broader design, capability, and complexity characteristics not explicitly modeled.

The Strategic Capability Model

Our final modeling effort produced the parsimonious CER results shown in Table 12, based on *weight capability* and *stealth capability*. The corresponding model specification is presented in Equation (11) with its exponentiated version provided in equation (12). These variables were selected for their early observability and

Multivariate Regrssion Results for Strategic Capability Model					
Variable	*Coefficient*	*Std. Error*	*z-Statistic*	*P-Value*	*95% CI*
Intercept	*2.8751*	*0.624*	*4.606*	*<0.0001*	*[1.652,4.099]*
LnWeightCapability	*0.7439*	*0.058*	*12.732*	*<0.0001*	*[0.629,0.858]*
PRESS R^2	*0.8567*	*0.116*	*7.359*	*<0.0001*	*[0.629,1.085]*

Table 12: Multivariate Regrssion Results for Strategic Capability Model

relevance to aircraft design decisions, providing a concise framework for evaluating recurring T100 flyaway costs.

$$\ln T100) = 2.8751 + 0.7439 \ln(\text{WeightCapability}_i) + 0.8567 \text{ Stealth}_i + \varepsilon_i \quad (11)$$

$$T100 = 17.73 \cdot WeightCapability^{0.7113} \cdot 2.36^{Stealth} \quad (12)$$

where

$$Stealth = 1\ (stealth), 0\ (non - stealth).$$

The *Strategic Capability Model* explains approximately 85% of the variation in recurring T100 flyaway costs across the aircraft in our dataset ($R^2 = .846$). The specification relies on only two predictors, *stealth capability* and *weight capability*, and was deliberately structured to emphasize simplicity, transparency, and applicability across program phases. We examined a broader set of candidate capability measures during model development, with stepwise regression used to inform variable reduction and assess relative explanatory contribution. Rather than dictating final model structure, the procedure served to narrow the field of plausible predictors. Final variable selection reflected analytical judgment, consistency with aircraft design principles, and the stability and interpretability of estimated effects in an early-stage cost estimation context. Both predictors are statistically significant ($p < .0001$), underscoring their robustness as cost drivers.

Additionally, because this modeling effort specifically excludes trainer-type aircraft (such as the T-XX series) from the dataset, these results should not be applied to T-type aircraft. Instead, the estimated relationships and predictive boundaries of the Strategic Capability Model are explicitly tailored to, and valid for, frontline military aircraft (such as fighters, bombers, etc.).

The results indicate stealth-capable aircraft are associated with substantially higher recurring T100 flyaway costs. In the exponentiated specification, the estimated stealth coefficient implies a multiplicative cost factor of 2.36, indicating that stealth aircraft are associated with nearly two and a half times higher recurring costs than comparable non-stealth platforms. Expressed in percentage terms, this corresponds to an increase of approximately 136 percent. This large premium likely reflects the continuous manufacturing demands of stealth designs, which require more expensive specialized materials and highly precise assembly processes throughout production. Weight capability is also positively related to cost, consistent with expectations regarding aircraft scale and payload capacity. Together, these findings suggest that recurring costs are strongly influenced by a combination of capability-driven design features.

As shown in Table 13, the model's PRESS R^2 and adjusted R^2 are closely aligned with the traditional R^2, reinforcing that the model is not overfit and

Goodness of Fit Measures for Strategic Capability Model	
Metric	*Value*
R^2	0.846
Adjusted R^2	0.833
PRESS R^2	0.792

Table 13: Goodness of Fit Measures for Strategic Capability Model

that its explanatory power remains stable under cross-validation. This consistency strengthens confidence in the model's reliability for cost estimation across diverse program phases.

In Table 14, we report the out-of-sample predictive performance of the *Strategic Capability Model*, which was assessed using Leave-One-Out Cross-Validation (LOOCV). This technique involves iteratively training the model on all but one aircraft and then using it to predict the cost of the single excluded aircraft, a process repeated for every airframe in the dataset. This robust validation method yields a Mean Absolute Percent Error (MAPE) of 37.83% and a Median Absolute Percent Error (MDAPE) of 32.51%.

The divergence between the mean and median errors is itself an important finding. It indicates that the higher MAPE is being influenced by a small number of aircraft for which the model produces larger-than-average prediction errors. The MDAPE of 32.51% suggests for a typical aircraft, the model's prediction error is closer to this lower value. While a mean error of nearly 38% may appear substantial, it is critical to evaluate this performance within the intended context of early-stage, pre-Milestone A cost analysis. At this preliminary phase, where design details are fluid and only high-level capability drivers are known, such models are not intended for budgetary precision but for establishing a reasonable and defensible data-driven baseline. Therefore, the model provides significant value by offering a transparent cost estimate in a low-information environment.

LOOCV Predictive Accuracy of the Strategic Capability Model	
Metric	*Value*
Mean Absolute Percent Error	37.83%
Median Absolute Percent Error	32.51%
PRESS R^2	

Table 14: LOOCV Predictive Accuracy of the Strategic Capability Model

Importantly, the model demonstrates that a high proportion of cost variation can be explained using variables which are both conceptually intuitive and available early in the acquisition process. This early availability enhances the model's usefulness as a practical decision-support tool for cost estimators, program managers, and analysts, by enabling credible estimates to be generated with minimal technical inputs and without reliance on proprietary contractor data. The balance between explanatory power and simplicity positions this model as both an academic contribution to the cost-estimating literature and a practical resource for informing affordability assessments, requirements setting, and comparative analysis of alternative system designs.

The Shapley R^2 decomposition aids interpretation of the model's explanatory structure by offering clearer insight into the relative contribution of each predictor beyond what standardized coefficients alone provide. This technique provides a principled method for partitioning the model's overall explanatory power among the predictors and estimating the proportion of variance attributable to each factor (Lipovetsky & Conklin, 2001). As seen in the results listed in Table 15, weight capability explains the majority of the model's variation, indicating that scale- and capacity-related characteristics account for most of the explanatory power in the parsimonious CER, while stealth contributes a smaller but meaningful share.

Multivariate Regression Shapley R^2 Decomposition		
Variable	*Shapley Decomposition*	*Variation Explained*
Stealth	*0.0837*	*9.89%*
LnWeightCapability	*0.7622*	*90.11%*
PRESS R^2		

Table 15: Multivariate Regression Shapley R^2 Decomposition

Strategic Capability Model Regression Diagnostics

We applied a suite of regression diagnostics to assess whether the *Strategic Capability Model* exhibits reasonable statistical behavior under ordinary least squares estimation and to evaluate whether the estimated relationships reflect systematic cost patterns rather than artifacts of model misspecification. Residual analysis indicated no significant departure from normality, and formal testing (Shapiro-Wilk, $p = 0.11$) supported this conclusion. Constant error variance was confirmed through the Breusch-Pagan test ($p = 0.06$), and robust standard errors yielded consistent results, providing additional assurance against heteroskedasticity.

We also examined visual diagnostics, including residual-fitted and scatter plots to assess scale behavior and functional form. These plots indicate stable variance patterns and consistent relationships across the range of fitted values, supporting the use of logarithmic specifications and reinforcing the robustness of the *Strategic Capability Model* for early-stage cost estimation.

We assess multicollinearity using VIFs, all of which fell well below commonly accepted thresholds, suggesting no collinearity concerns. Examination of residual-versus-fitted plots confirmed that the applied log-log transformations captured the underlying functional form without systematic specification error. Finally, Cook's distance values indicated no observations exerted undue influence on the regression results.

Together, these diagnostics provide confidence that the model is well specified, statistically valid, and appropriate for applied cost estimating contexts.

Implications for the Field & Additional Research Opportunities

The family of parsimonious models developed in this research provides a practical and adaptable approach to early-stage cost estimation. Because they rely on only two capability-based metrics, analysts can apply these models to a wide range of programs and airframes, enabling credible cost predictions in today's competitive global environment. The final specification also lends itself to clear and interpretable communication of results for decision-makers. Both stealth capability and weight capability are design attributes available early in AoA and trade space studies, reinforcing the model's utility for supporting affordability assessments and requirements-setting decisions at formative stages of program development. These models are designed specifically to support rapid, defensible decision-making up through Milestone A reviews, whereas later acquisition phases require higher-fidelity models which account for mature system designs. Beyond any standalone value these relationships may provide, they are equally useful as components within comprehensive parametric cost estimating models. The transparent, repeatable, and logically grounded selection of cost drivers in this research allows practitioners to incorporate these CERs into broader parametric frameworks, supporting model structuring, benchmarking, and cross-check validation regardless of program maturity or acquisition milestone.

Importantly, this work highlights that one should not evaluate CER usefulness on R^2 alone. While prior research, including Rojo et al. (2023), demonstrates that strong predictive performance can be achieved by incorporating a rich set of performance, mission, and categorical variables, the present study demonstrates a substantial share of variation in recurring flyaway cost can be captured using a more parsimonious specification composed of variables observable much earlier in the acquisition lifecycle. To reduce model complexity, our approach emphasizes early availability and interpretability, facilitating consistent benchmarking across platforms and enabling more meaningful comparisons that avoid noise introduced by less interpretable or less accessible metrics.

In this way, our research aligns with acquisition reform initiatives such as WSARA and BBP, which call for improved cost control and more informed decision-making with affordability constraints in mind (U.S. Department of Defense, 2013). More recent Department of War acquisition reform guidance further reinforces these priorities by emphasizing the importance of credible, early cost estimates to support trade-offs and reduce downstream cost growth (Office of the Secretary of War, 2025). The inclusion of physical dimensions as cost drivers provides a complementary pathway for early-stage estimation. By grounding these physical indicators in a statistically rigorous framework, this research equips cost estimators with tools applicable as early as the Materiel Solution Analysis (MSA) phase, supporting Rough Order of Magnitude (ROM) estimates. In doing so, it strengthens the analytical foundation for affordability assessments and design trade-offs that shape the nation's warfighting investments.

Future research can extend this effort in two primary directions. First, the models provided in this research can be applied to next-generation platforms to evaluate their validity against emerging airframes. Second, our models can serve as a baseline for iterative refinement, beginning with the simple two-variable structure and incorporating additional predictors as data availability, methodological advances, or research priorities evolve. These extensions would strengthen both the robustness of modern CERs and the advancement of best practices in defense acquisition analysis.

References:

Cisneros, C. J., White, E. D., Lucas, B. M., Ritschel, J. D., Fass, R. D., & Valentine, S. M. (2024). Production Box Cost Estimating Relationships for DoD Avionics. Defense Acquisition Research Journal, 158-179.

Congressional Research Service. (2024). Great Power Competition: Implications for Defense - Issues for Congress. Washington D.C.: Congressional Research Service.

Department of Defense. (2024). Department of Defense Instruction 5000.73. Washington D.C.: U.S. Department of Defense.

Hess, R. W., & Romanoff, H. P. (1987). Aircraft Airframe Cost Estimating Relationships. Santa Monica: RAND.

Hildebrandt, G. G., & Sze, M.B. (1986). Accounting for The Cost of Tactical Aircraft. Santa Monica: RAND.

Hu, S. P., & Smith, A. (2021). Using Dummy Variables in CER Development . Journal of Cost Analysis and Parametrics, 76-90.

International Society of Parametric Analysts. (2008). Parametric Estimating Handbook. Vienna: ISPA.

Jones, G., White, E. D., Ryan, E. V., & Ritschel, J. D. (2014). Investigation into the ratio of operating and support costs to life-cycle costs for DoD weapon systems. Defense Acquisition Research Journal, 21(1), 442–464.

Large, J. P., Campbell, H. G., & Cates, D. (1976). Parametric Equations for Estimating Aircraft Aifrframe Costs. Santa Monica: RAND.

Levenson, G., Boren, H., Tihansky, D., & Timson, F. (1972). Cost-Estimating Relationships for Aircraft Airframes. Santa Monica: RAND.

Light, T., Timson, F., & Younossi, O. (2024). Parametric Flyaway Cost Estimating Relationships for Manned and Remotely Piloted Military Aircraft. Santa Monica: RAND.

Lipovetsky, S., & Conklin, M. (2001). Analysis of Regression in Game Theory Approach. Applied Stochastic Models in Business and Industry, 324.

Office of the Secretary of Defense. (2022). 2022 National Defense Strategy of the United States of America. Washington, D.C.: U.S. Department of Defense.

Office of the Secretary of Defense. (2022). Analysis of Alternatives Cost Estimating Handbook. Washington D.C.: Office of the Secretary of Defense.

Office of The Secretary of Defense. (2022). DOD Cost Estimating Guide. Washington D.C.: Office of The Secretary of Defense.

Office of the Secretary of Defense. (2025). Operating and Support Cost-Estimating Guide. Washington D.C.: Office of the Secretary of Defense.

Office of the Secretary of War. (2025). Acquisition Transformation Strategy. Washington D.C.: U.S. Department of War.

Reese et al. (2021). Integrating Cost as a Decision Variable in Wargames. Air and Space Power Journal, 35-46.

Rogers, C. (1992). Advanced Composite Materials: The Air Force's Role in technology Development. Santa Monica: RAND.

Rojo, K. M., White, E. D., Lucas, B. M., Ritschel, J. D., Valentine, S. M., & . (2023). Cost Estimating Relationships for T100 Flyaway Costs. Journal of Cost Analysis and Parametrics, 24-36.

U.S. Department of Defense. (2013). Better Buying Power 2.0 Fact Sheet. Washington: Office of the Secretary of Defense.

U.S. Government Accountability Office. (2020). Cost Estimating and Assessment Guide. Washington: GAO.

Weisberg, S. (2013). Applied Linear Regression. John Wiley & Sons, Incorporated.

Wooldridge, J. M. (2018). Introductory Econometrics: A Modern Approach. Boston: Cengage.

Younossi, O., Kennedy, M., & Graser, J. C. (2001). Military Aiframe Costs: The Effects of Advanced Materials and Manufacturing Processes. Sanat Monica: RAND.

First Lieutenant Austin J. Romanov, USAF, is a Cost Analyst at the Air Force Life Cycle Management Center, at Hanscom AFB, Ma. He holds a BBA in Management from Florida Atlantic University and an MS in Cost Analysis from the Air Force Institute of Technology.

Dr. Robert D. Fass is an Assistant Professor in the Department of Systems Engineering and Management at AFIT. He holds a BA in Economics, an MBA from the University of New Mexico, and a PhD in Business Administration and Management from New Mexico State University. Dr. Fass's research interests include cost analysis, decision analysis, risk analysis, behavioral economics, organizational behavior, organizational change, and government acquisition policy.

Dr. Jonathan D. Ritschel is a Professor of Cost Analysis in the Department of Systems Engineering and Management at AFIT. He received his BBA in Accountancy from the University of Notre Dame, MS in Cost Analysis from AFIT, and PhD in Economics from George Mason University. His research interests include public choice, defense cost economics, cost analysis and economic institutional analysis.

Major Michael J. Brown is an Assistant Professor of Cost Analysis in the Department of Systems Engineering and Management at AFIT. He holds a BS in Economics from the United States Air Force Academy, a MS in Cost Analysis from AFIT, and a MA and PhD in Economics from the University of Tennessee. Major Brown's research interests include recruitment and retention, econometric analysis, and governmental acquisitions.

Shawn Valentine is an Operations Research Analyst at Air Force Life Cycle Management Center/Financial Management at Wright-Patterson AFB, OH. He graduated from Ohio University with a BS in Actuarial Science and an MS in Financial Economics.

Using Bayes' Theorem to Develop CERs – Extending the Gaussian Model

Dr. Christian B. Smart

Bayes' Theorem is a mathematical method for combining prior experience with new data. It is extremely important in leveraging limited information, which is often the case in cost estimating. This paper is an extension of a previous ICEAA paper that dealt with the application of Bayes' Theorem to cost estimating. In this update we show how the Bayesian approach for linear models can be extended for different and more realistic assumptions.

Introduction

Historically, parametric methods in cost estimating have used frequentist techniques. Frequentist statistics is the classical method that uses a sample of data as inputs. If you have taken Statistics 101 in college, most if not all the class was oriented towards this approach. For example, traditional linear and nonlinear regression analysis is a frequentist approach. The challenge with this method is that it requires a large amount of data. Statisticians have conducted numerous studies using random data and have concluded that you need 50 data points for a regression analysis with 10 additional data points for every independent variable you want to include. For example, if you want to include three independent variables in your analysis, you need 80 data points. The number of highly specialized systems used in the Department of Defense and NASA means that most data sets typically have nowhere near that much. For example, the Missile Defense Agency has only developed a handful of different kill vehicles, and NASA has only developed a few crewed launch vehicles. When looking at truly applicable data, the sample size shrinks even further – when considering launch vehicles, the primary systems that NASA has completed have been those for the Apollo and Shuttle programs. The Apollo program began in the 1960s, and the Shuttle program began in the 1970s. Thus, there are no directly applicable historical data points within the last 40 years. Considering the changes that have taken place in the realm of technology since then, there really is no applicable historical data at all for these systems.

For small data sets like these, Bayesian methods can help provide more accurate estimates. Bayesian methods leverage all your experience, making them less subject to being overwhelmed by noise. This prior experience can be subjective or objective. The objective data could involve the use of similar data that is not directly applicable.

This approach has proven to be successful in a multitude of applications. Bayesian techniques were used in World War II to help crack the Enigma code used by the Germans, thus helping to shorten the war. John Nash's equilibrium for games with incomplete or imperfect information is a form of Bayesian analysis (John Nash's life was portrayed in the film *A Beautiful Mind*). Actuaries have used Bayesian methods for over 100 years to set property and casualty insurance premiums. Bayesian voice recognition researchers applied their skills as leaders of the portfolio and technical trading team for the Medallion Fund, a $5 billion hedge fund which has averaged annual returns of 35% after fees since 1989.

As you can see, the Bayesian method has been used in several applications in which there is money on the line. I place a high degree of confidence in methods that have proven themselves under such circumstances. They are not mere "ivory tower" exercises without practical application. The practitioners who have used these methods have risked significant sums of money, literally billions of dollars, in using these methods. Thus, many of the users of Bayes' Theorem have "skin in the game" (Taleb 2018).

A previous paper on this subject by the author dealt with the normal/Gaussian model (Smart 2014). This paper focused on the context of log-transformed ordinary least squares regression and assumed that both the log of the prior information and that the log of the new information is normally distributed. In addition, this paper assumed that the variance is a known, fixed quantity. The prior experience is typically based on a large amount of information, either objective or subjective, and thus the assumption of normality for the prior information is not suspect. However, the variance is typically not known – indeed it is estimated with the square of the standard error of the prior model in the case of objective prior data. Even worse, the "known" variance is treated by the modeling process as the square of the standard error of the estimate of the regression equation based on a small sample, which is not an accurate estimate of the true variance. Also, the reason why Bayes' Theorem is used is that there is a small amount of data. In this case, when the variance is unknown and there is a small amount of data, the new information should be assumed to follow a Student's t-distribution because it accounts for the extra uncertainty that arises from estimating the population variance from a small sample.

When the sample size is small, the estimate of the standard deviation is not very reliable, so using a normal distribution would underestimate the true variability. The t-distribution corrects for this by having heavier tails, which gives more probability to extreme values and leads to wider, more cautious confidence intervals.

In the previous paper on Bayesian regression, it was shown that the resulting estimate of a Bayesian approach to regression is a weighted average of the regression coefficients for the two data sets. The weights are determined by the uncertainty about the coefficients. The impact of using the normal model with known variance is that it will typically assign too much weight to the parameters of the smaller data set – the small data set will provide a better fit than the true underlying population, so the standard error of the estimate will be significantly lower than the overall population.

The next section of this paper provides an overview/ review of Bayes' Theorem. This is followed by the application of Bayes' Theorem to a sanitized version of real-life regression problem encountered by the author. The simple approach that was discussed in the author's previous paper (Smart 2014) is applied to this example. However, there are two assumptions in this simple approach that need to be revisited. Methods for applying Bayesian regression with more realistic assumption are discussed in great detail. These techniques are applied to the example, and it is shown that using more realistic assumption results in greater predictive accuracy for the example. The results obtained, including uncertainty analyses, are compared.

Bayes' Theorem

In this section a review of Bayes' theorem is provided. The distribution of the model given values for the parameters is called the *model distribution*. *Prior* probabilities, which represent initial beliefs or assumptions about a parameter or hypothesis before observing any new data, are assigned to the model parameters. After observing data, a new distribution, called the *posterior* distribution, is developed for the parameters, using Bayes' Theorem.

The conditional probability of event A given event B is denoted by Pr(A/B).

In its discrete form, Bayes' Theorem states that

$$\boldsymbol{Pr(A|B) = \frac{Pr(A)Pr(B|A)}{Pr\ (B)}}$$

For example, consider testing for illegal drug use. Many people have had to take such a test as a condition of employment with the federal government or with a government contractor. What is the probability that someone who fails a drug test is not a user of illegal drugs? Bayes' Theorem can be used to answer such questions.

Suppose that 95% of the population does not use illegal drugs. Also suppose that the drug test is highly accurate. If someone is a drug user, it returns a positive result 99% of the time. If someone is not a drug user, the test returns a false positive only 2% of the time.

In this case: A is the event that someone is not a user of illegal drugs, and B is the event that someone test positive for illegal drugs. The complement of A, denoted A', is the event that someone is a user of illegal drugs.

From the law of total probability,

$$\mathbf{Pr(B) = Pr(\boldsymbol{B|A})\,Pr(\boldsymbol{A}) + Pr(\boldsymbol{B|A'})\,Pr(\boldsymbol{A'})}$$

Thus, Bayes' Theorem in this case is equivalent to:

$$\mathbf{Pr}(\boldsymbol{A}|\boldsymbol{B}) = \frac{\mathbf{Pr}(\boldsymbol{B}|\boldsymbol{A})\,\mathbf{Pr}(\boldsymbol{A})}{\mathbf{Pr}(\boldsymbol{B}|\boldsymbol{A})\,\mathbf{Pr}(\boldsymbol{A}) + \mathbf{Pr}(\boldsymbol{B}|\boldsymbol{A}')\,\mathbf{Pr}(\boldsymbol{A}')}$$

Plugging in the appropriate values

$$\mathbf{Pr}(\boldsymbol{A}|\boldsymbol{B}) = \frac{\mathbf{0.02(0.95)}}{\mathbf{0.02(0.95) + 0.99(0.05)}} \approx \mathbf{27.7\%}$$

Thus, even with accurate drug tests it is easy to obtain false positives. This is a case of inverse probability, a kind of statistical detective work where an attempt is made to determine whether someone is innocent or guilty based on revealed evidence. More typical of the kind of problem that cost estimators will want to solve is the following: There is some prior evidence or opinion about a subject, and there is also some direct empirical evidence available. How should the prior evidence be leveraged and how can it be combined with the current evidence to form an accurate estimate of a future event?

It's simply a matter of interpreting Bayes' Theorem. Pr(A) is the probability that is assigned to an event before seeing the data. This is called the *prior* probability. Pr(A/B) is the probability after seeing the data. This is called the *posterior* probability. Pr (B/A)/Pr(B) is the probability of seeing these data given the hypothesis. This is the *likelihood*.

Bayes' Theorem can be re-stated as

Posterior ∝ *Prior * Likelihood*

An example of this application of Bayes' Theorem can be found in the Monty Hall Problem. This is based on the television show *Let's Make a Deal*, whose original host was Monty Hall. In this version of the problem, there are three doors. Behind one door is a car. Behind each of the other two doors is a goat. You pick a door. Monty, who knows what is behind the doors, then opens one of the other doors that has a goat behind it. Suppose you pick door #1. He then opens door #3, showing you the goat behind it, and asks you if you want to pick door #2 instead. See Figure 1. Is it to your advantage to switch your choice?

To solve this problem, let A_1 denote the event that the car is behind door #1, A_2 the event that the car is behind door #2,

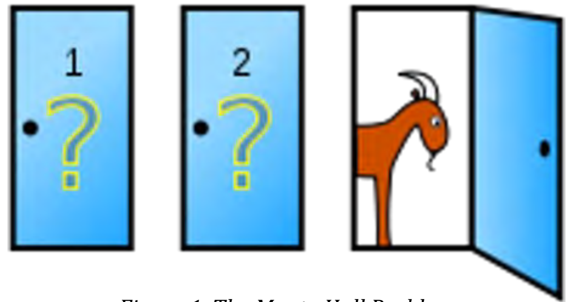


Figure 1. The Monty Hall Problem

and A_3 the event that the car is behind door #3. Your original hypothesis is that there was an equally likely chance that the car was behind any one of the three doors. Thus, the prior probability, before the third door is opened, that the car was behind door #1, which is denoted by $\Pr(A_1)$, is 1/3. Also, $\Pr(A_2)$ and $\Pr(A_3)$ are also equal to 1/3.

Once you picked door #1, you were given additional information. You were shown that a goat is behind door #3. Let B denote the event that you are shown that a goat is behind door #3. The probability that there is a goat behind door #3 is best calculated by considering three conditional probabilities.

The probability that you are shown the goat is behind door #3 is an impossible event since the car is behind door #3. Thus $\Pr(B|A_3) = 0$. Since you picked door #1, Monty will open either door #2 or door #3, but not door #1. Thus, if the car is actually behind door #2, it is a certainty that Monty will open door #3 and show you a goat. Thus

$\Pr(B|A_2)=1$. If you have picked correctly and have chosen the right door, then there are goats behind both door #2 and door #3. In this case, there is a 50% chance that Monty will open door #2 and a 50% chance that he will open door #3. Thus $\Pr(B|A_1)=1/2$.

By Bayes' Theorem,

$$\boldsymbol{Pr(A_1|B)} = \frac{\boldsymbol{Pr(A_1)\,Pr(B|A_1)}}{\boldsymbol{Pr(A_1)\,Pr(B|A_1) + Pr(A_2)\,Pr(B|A_2) + Pr(A_3)\,Pr(B|A_3)}}$$

Plugging in the probabilities that derived yields

$$\boldsymbol{Pr(A_1|B)} = \frac{\mathbf{(1/3)(1/2)}}{\mathbf{(1/3)(1/2) + (1/3)(1) + (1/3)(0)}} = \frac{\mathbf{1/6}}{\mathbf{1/6 + 1/3}} = \mathbf{1/3}$$

Also,

$$\boldsymbol{Pr(A_2|B)} = \frac{\mathbf{(1/3)(1)}}{\mathbf{(1/3)(1/2) + (1/3)(1) + (1/3)(0)}} = \frac{\mathbf{1/3}}{\mathbf{1/6 + 1/3}} = \mathbf{2/3}$$

And since you already know that the car is not

behind door #3, $\Pr(A_3|B)=0$.

Thus, you have a 1/3 of picking the car if you stick with your initial choice of door #1, but a 2/3 chance of picking the car if you switch doors. It is in your interest to switch doors.

Did you think that there was no advantage to switching doors? You're not alone. Marilyn Vos Savant, famous as having the world's highest IQ at 228, wrote a column for *Parade* magazine for many years. In 1990 a reader posed the Monty Hall problem to her, and she provided the correct answer. But many people, including people with Ph.Ds., including some mathematicians, derided Marilyn for being wrong. Even the famous mathematician Paul Erdos found the problem to be counterintuitive (Hofmann, 1998). But the correct answer is that once door #3 is opened and revealed to have a goat behind it, there is a two-thirds chance that the car is behind door #2. If you're still not convinced, conduct a Monte Carlo simulation to see that this is the correct answer.

For the application of Bayes' Theorem to cost estimating, the continuous form of Bayes' Theorem is required. If the prior distribution is continuous, Bayes' Theorem is written as

$$\pi(\theta|x_1,\dots,x_n) = \frac{\pi(\theta)f(x_1,\dots,x_n|\theta)}{f(x_1,\dots,x_n)} = \frac{\pi(\theta)f(x_1,\dots,x_n|\theta)}{\int \pi(\theta)f(x_1,\dots,x_n|\theta)d\theta}$$

where:

$\boldsymbol{\pi(\theta)}$ is the *prior density*, the initial density function for the parameters that varies in the model. It is possible to define an *improper prior density*, one which is nonnegative, but whose integral is infinite;

$\boldsymbol{f(x|\theta)}$ is the conditional probability density function of the model. It defines the model's probability given the parameter θ;

$\boldsymbol{f(x_1, \dots, x_n|\theta)}$ is the conditional joint probability density function of the data given θ.

Typically, the observations are assumed to be independent given θ, and in this case,

$$f(x_1,\dots,x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

and $\boldsymbol{f(x_1, \dots, x_n)}$ is the unconditional joint density function of the data $\boldsymbol{x_1, \dots, x_n}$. It is calculated from the conditional joint density function by integrating over the prior density function of θ:

$$f(x_1,\dots,x_n) = \int \pi(\theta)f(x_1,\dots,x_n|\theta)d\theta$$

$\boldsymbol{\pi(\theta|x_1, \dots, x_n)}$ is the *posterior density function*, the revised density function for the parameter θ based on the observations $\boldsymbol{x_1, \dots, x_n}$. (Unless expressly noted, all integrals are over the entire range of the variable over which the function is being integrated.)

$\boldsymbol{f(x_{n+1}|x_1, \dots, x_n)}$ is the *predictive density function*, the revised unconditional density based on the sample data. It is calculated by integrating the conditional probability density function over the posterior density of θ:

$$f(x_{n+1}|x_1,\dots,x_n) = \int f(x_{n+1}|\theta)\pi(\theta|x_1,\dots,x_n)d\theta$$

Bayesian Regression

This section considers ordinary least squares CERs of the form

$$Y = a + bX + \varepsilon$$

This involves prior distributions about a and b, as well as e. The parameters a and b are the coefficients of the linear equation being fitted – a is the intercept and b is the slope. The parameter e is the residual, which is the difference between the actual and predicted value.

For the application of Bayesian regression, this is written this in mean deviation form:

$$Y = \alpha_{\bar{x}} + \beta(X - \bar{X}) + \varepsilon$$

This form makes it easier to establish prior inputs, since it is easier to think of an average value for prior cost than it is for the intercept of the least-squares equation.

For nonlinear regression log-transformed versions of the power equation $Y = aX^b$ will be considered. In linear (transformed) space the equation is

$$lnY = lna + bLnX$$

While weight is not a true cost driver, it is an excellent proxy for program scope and for most systems is highly correlated with cost. Weight-based CERs will be used in the examples in the paper.

Bayesian regression will be applied to an example of sanitized cost and weight data for commercial-like earth-orbiting satellites. The claim has been made that these satellites are much cheaper than the average satellite purchased using traditional government acquisition practices. While only a handful of satellites have been built using the streamlined approach, they are on average much cheaper than traditional programs. The difference is remarkable – $4,000 per pound vs. $18,000 per pound, a reduction of more than 75%.

What is the best way to model these more economical satellites? From my source data set of 72 Earth-orbiting robotic satellites, if used alone, the cost estimate will likely be overestimated!

An alternative would be to develop a CER for just the new missions alone. A power equation can be fit to this data set. See Figure 2. The R^2 coefficient is 63%, which is not bad. So why not just use that? Because the number of data points is an order of magnitude smaller than recommended for the minimum number of data points in a classical regression.

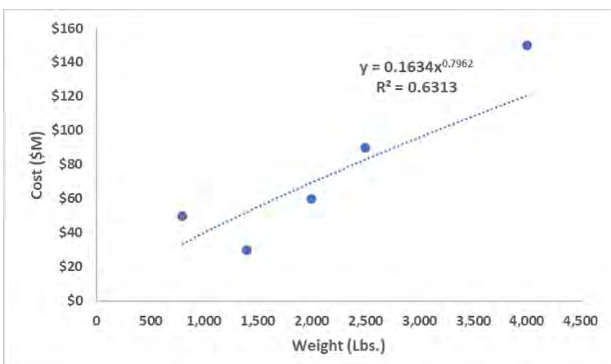


Figure 2. Regression of cost vs. weight for set of five new mission satellites data points.

In comparison to the larger set of 72 data points that do not include the five data points in the smaller set, another decent fit to the historical data is obtained, with an R^2 equal to 61%. See Figure 3.

The Bayesian approach allows us to combine the earth-orbiting spacecraft data with the smaller data set. A specific type of hierarchical approach like the one used in credibility theory in insurance for setting premiums for property and casualty is used. The earth-orbiting spacecraft data is treated as prior information. The prior information is updated with

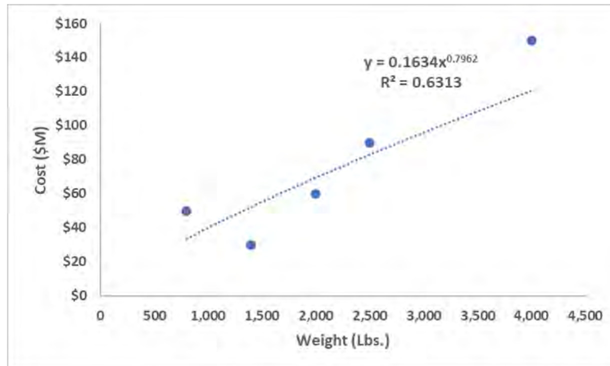


Figure 3. Regression of cost vs. weight for 72 earth-orbiting robotic satellites data points.

the set of five data points. This specific type of hierarchical approach involves two sets. One a general set, and the other a more specific subset or subtype. In this case, the commercial-like satellites are earth-orbiting satellites, and they are procured by government agencies. This is the general set. The commercial-like satellites are a proper subset of the earth-orbiting satellites. This is the specific subset.

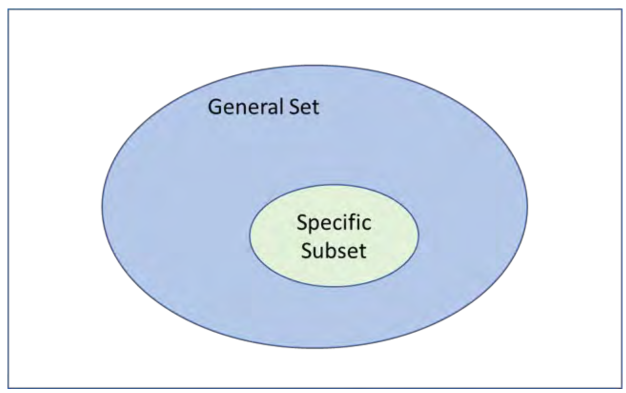


Figure 4. Conceptual illustration of the data used in a hierarchical approach.

This particular hierarchical approach has been used for over a century in setting property and casualty insurance premiums, particularly in smaller markets and for less common types of insurance.

There is a significant amount of probability theory in the derivation of Bayesian methods for regression. But I will avoid this here and refer the reader interested in the mathematical details to my earlier paper on this subject (Smart 2014).

The bottom line is that to implement the method you only need to understand the concept of standard error of the coefficients. To combine the two regression equations using Bayes' Theorem, each coefficient, the intercept and the slope, separately. Bayes' Theorem combines the coefficients using the weighted averages of each individual coefficient. These weights are based on the amount of uncertainty relative to one another. The greater the uncertainty in the coefficient relative to the other coefficient, the smaller its weight. The smaller the uncertainty in the coefficient relative to the other coefficient, the greater its weight. This makes intuitive sense – greater confidence should be placed in the coefficient with the less variation and assigned a greater weight in taking the average of the two.

The formula for combining the two coefficients is thus very simple. It takes the inverse of each of the variances, which is called the *precision*, and forms a weighted average of the precisions of each coefficient.

Given two independent, unbiased coefficients $\tilde{\theta}_1$ and $\tilde{\theta}_2$ with precisions $\rho_1 = 1/Var(\tilde{\theta}_1)$ and $\rho_2 = 1/Var(\tilde{\theta}_2)$ respectively, the minimum variance estimate of the weighted average is given by

$$\frac{\rho_1}{\rho_1+\rho_2}\tilde{\theta}_1 + \frac{\rho_2}{\rho_1+\rho_2}\tilde{\theta}_2$$

The variance is the square of the standard error. The standard error of each coefficient is provided by statistical packages, including the data analysis add-in for Excel and the "LINEST" function in Excel.

Next it will be shown how the weighted average is calculated in Excel using the LINEST function. LINEST is a matrix function. To enter it you select a range of cells for the output, type in "=LINEST(y values, x values, true, true)." The first "true" is to include the intercept, and the second is to include all the statistics. The output for a single independent variable appears as in Table 1.

b-value	**a-value**
Std err (b)	**Std err (a)**
R^2	**Std Err of Est**
F Statistic	**Degrees of Freedom**
SS Reg	**SS Resid**

Table 1. "LINEST" Excel function output format.

The first row are the coefficients. The "a-value" is the intercept, and the "b-value" is the slope. The second row is the standard errors of each coefficient. These are the only two sets of values needed from each equation to apply Bayes' Theorem.

Recall that equation used has the form

$$Y = \alpha_{\bar{x}} + \beta(X - \bar{X}) + \varepsilon$$

It is being applied to the log transformations of the data.

When the LINEST function is applied to the larger data set the results in Table 2 are obtained.

0.8858	**4.6087**
0.0809	**0.0956**
0.6311	**0.8111**
119.7361	**70**
78.7810	**46.0569**

Table 2. LINEST results for larger data set.

The a-value coefficient is equal to *4.6087* and its standard error is *0.0956*. The b-value coefficient is *0.8858* and its standard error is *0.0809*.

See Table 3 for the LINEST results for the smaller data set.

0.8144	**4.1359**
0.2588	**0.1416**
0.7675	**0.3167**
9.9033	**3**
0.9930	**0.3008**

Table 3. LINEST results for the small data set.

The a-value coefficient for the smaller data set is equal to *4.1359* and its standard error is *0.1416*. The b-value coefficient is *0.8144* and its standard error is *0.2588*. Note that the standard error calculation involves division by the degrees of freedom, so there is a penalty applied for having less data.

To combine the a-value coefficients requires calculating the variances of each as the square of the standard error. Thus, the variance of the a-value for the larger data set is equal to $0.956^2 \approx 0.0091$ and the variance of the a-value for the smaller data set is equal to $0.146^2 \approx 0.0201$. The weight for the larger data set's a-value coefficient is calculated as

$$\frac{\frac{1}{0.0091}}{\frac{1}{0.0091}+\frac{1}{0.0201}} \approx 68.7\%$$

and the coefficient weight for the a-value of the smaller data set is *1-0.687 = 31.3%.*

For the b-values similar calculations are used to determine the coefficient weight for the larger data set as

$$\frac{\frac{1}{0.0065}}{\frac{1}{0.0065}+\frac{1}{0.0670}} \approx 91.1\%$$

and the coefficient weight for the b-value of the smaller data set is *1-0.911 = 8.9%.*

The Bayesian estimate of the a-value coefficient is thus *4.6087*0.687+4.1359*.313 = 4.4607*, and the b -value coefficient is *0.8858*0.911+0.8144*.089 = 0.8794*.

There are two different data sets, each of which has its own average weight, but since the smaller data set is the "sample" and the larger data set is "prior information", the mean of the logs of the weights for the smaller data set is used, which yields

$$ln(Cost) = 4.4607 + 0.8794(ln(Weight) - 7.5161) + \varepsilon$$

Rearranging terms results in

$$ln(Cost) = -2.1491 + 0.8794ln(Weight) + \varepsilon$$

Exponentiating both sides yields the power equation

$$Cost = 0.1166Weight^{0.8794} \cdot \varepsilon$$

In all equations the error term denoted by the Greek letter epsilon ε simply denotes the difference between the estimated cost and the actual.

Note that because of the equation form we use to average the two data sets using Bayes' Theorem, the a-value coefficient in this case happens to be smaller than either coefficient of the two regressions for the individual data sets. Comparing the estimates provided by these equations for another commercial-like acquisition whose actual weight was 3,280 lbs. and whose cost was $180 million, we find that the equation based on the smaller data set alone predicts cost at $100 million, while the equation based on the larger data set alone predicts cost at $368 million. The Bayesian regression equation provides an estimate that is closer than either of these to the actual, at $144 million.

This particular approach to Bayesian estimating with a normal prior, normal likelihood, and known variance has also been used in software cost estimating (Boehm et al., 2018). In that particular application the prior is established via expert judgment.

Assumptions

There are two simplifying assumptions made that are dubious in this type of analysis.

One is the assumption that the variance of the estimating equation is known. That is, not only is it a fixed quantity, but the exact value is known. However, the value that is used is estimated from the likelihood, which is the smaller data set. Thus, there is a high degree of uncertainty in the variance, which means this assumption is not valid.

A second assumption is that the smaller data set has residuals that are normally distributed. The assumption that the residuals are normally distributed in the larger data set is not an issue, however when you have a small data set with an unknown variance, the residuals are better modeled with a Student's *t*-distribution. The use of the Student's t-distribution to model the log-space residuals of log-transformed ordinary least squares was first advocated several years ago at an ISPA-SCEA conference (Druker et al. 2009) and has since been implemented in the ACE-IT cost estimating platform.

The distribution of the variance is easy to estimate, and that will be dealt with next. Incorporating the change from known variance and the use of the Student's t-distribution to model the residuals is more challenging, and that will be dealt with second.

Distribution of the unknown variance

By Cochran's Theorem (Cochran 1934), when there are *p* parameters in a linear regression,

$$\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sigma^2} \sim \chi^2(n-p)$$

where $\chi^2(n - p)$ is a chi-square distribution with *n-p* degrees of freedom and ^ denotes a sample/ population estimated parameter).

Also, the estimate for the square of the standard error is

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-p}$$

and thus

$$\frac{(n-p)\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-p)$$

What is of interest is the distribution of σ^2, which follows a scaled-inverse chi-square distribution with parameters *n-p* and $\widehat{\sigma^2}$.

The scaled-inverse chi-square distribution has probability density function given by

$$\frac{\left(\widehat{\sigma^2}\frac{n-p}{2}\right)^{\frac{n-p}{2}}}{\Gamma\left(\frac{n-p}{2}\right)}\frac{e^{-\frac{n-p\widehat{\sigma^2}}{2\ \sigma^2}}}{\sigma^{2(1+\frac{n-p}{2})}}$$

The scaled-inverse chi-square is only defined over positive values and has positive skew. See Figure 1 for a graph of a scaled-inverse chi-square with parameters *n-p = 70* and $\widehat{\sigma^2}$ *=66%.*

Also, the variance follows an inverse gamma with parameters $\frac{n-p}{2}$ and $\frac{n-p}{2}\widehat{\sigma^2}$, and the inverse of the variance follows a gamma distribution with the same parameters. The inverse gamma and gamma distributions are more commonly encountered than the scaled-inverse chi-square and are more commonly available in statistical software.

Conjugate priors

One of the nice features of using the normal prior and normal likelihood model with known variance is that the posterior is also a normal distribution. That is, the prior and the posterior have the same distribution. When the prior and posterior have the same distribution using conjugate priors makes calculating the posterior relatively easy and analytically tractable. Departure from conjugate priors means losing this analytical tractability. When the assumption that the variance is known is removed, the relatively straightforward normal conjugate prior model for regression is lost. Recall that the essence of Bayes' Theorem is that the posterior is proportional to the product of the likelihood and the prior distribution; that is, for uncertain parameters *y* and *θ*.

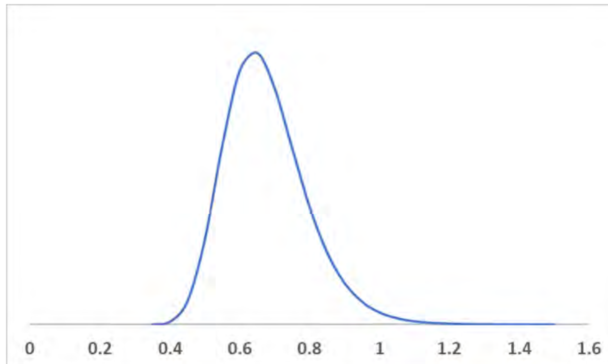


Figure 5. Probability density function of a scaled-inverse chi-square with parameters 70 and 0.66.

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

When the regression is centered about zero (as described in the previous section), the variance is known and the prior and likelihood are both normal, then the likelihood is (note that in the remainder of the paper bold font will be used to denote matrices and vectors).

$$p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\mu},\sigma^2) = \boldsymbol{N}(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2) \propto e^{-\frac{1}{2\sigma^2}||y-\bar{y}\cdot\overrightarrow{1_n}-X\beta||_2^2}$$

Where $\bar{y}$ is the sample average of the dependent variable; the joint prior of the coefficients is

$$p(\boldsymbol{\beta}) = N(\boldsymbol{\beta}|\boldsymbol{\beta_0},\boldsymbol{V_0})$$

Then by Bayes' Theorem, the posterior is

$$p(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{y},\sigma^2) \propto N(\boldsymbol{\beta}|\boldsymbol{\beta_0},\boldsymbol{V_0}) \cdot N(\boldsymbol{y}|\boldsymbol{X\beta},\sigma^2\boldsymbol{I_n}) \sim N(\boldsymbol{\beta}|\boldsymbol{\beta_N},\boldsymbol{V_N})$$

where

$$\boldsymbol{\beta_N} = \boldsymbol{V_N}\boldsymbol{V_0^{-1}}\boldsymbol{\beta_0} + \frac{1}{\sigma^2}\boldsymbol{V_N}\boldsymbol{X^T}\boldsymbol{y}$$

$$\boldsymbol{V_N^{-1}} = \boldsymbol{V_0^{-1}} + \frac{1}{\sigma^2}\boldsymbol{X^T}\boldsymbol{X}$$

$$\boldsymbol{V_N} = \sigma^2\left(\sigma^2\boldsymbol{V_0^{-1}} + \boldsymbol{X^T}\boldsymbol{X}\right)^{-1}$$

See Appendix 1 for the derivation.

The posterior predictive distribution is also normal, since

$$p(\boldsymbol{y}|\boldsymbol{x},\mathcal{D},\sigma^2) = \int N(\boldsymbol{y}|\boldsymbol{x}^T\boldsymbol{\beta},\sigma^2)N(\boldsymbol{\beta}|\boldsymbol{\beta}_N,\boldsymbol{V}_N)d\boldsymbol{\beta} = N(\boldsymbol{y}|\boldsymbol{x}^T\boldsymbol{\beta}_N,\sigma^2+\boldsymbol{x}^T\boldsymbol{V}_N\boldsymbol{x})$$

Note that it is not necessary to calculate the integral to derive the predictive posterior, it just needs to be noted that since $y = \boldsymbol{x}^T\boldsymbol{\beta}+\boldsymbol{\varepsilon_1}$ where $\varepsilon_1 \sim N(\boldsymbol{0},\boldsymbol{\sigma^2})$ and $\boldsymbol{\beta} = \boldsymbol{x}^T\boldsymbol{\beta}_N+\boldsymbol{\varepsilon_2}$ where $\boldsymbol{\varepsilon_2} \sim N(0,\boldsymbol{V}_N)$, then .

$$y = \boldsymbol{x}^T(\boldsymbol{\beta}_N+\varepsilon_2)+\varepsilon_1 = \boldsymbol{x}^T\boldsymbol{\beta}_N+\boldsymbol{x}^T\varepsilon_2+\varepsilon_1 \sim N(\boldsymbol{x}^T\boldsymbol{\beta}_N,\sigma^2+\boldsymbol{x}^T\boldsymbol{V}_N\boldsymbol{x})$$

The mean of the predictive distribution is thus the mean of the posterior, and the variance is the sum of the observed variance and another term that depends on the variance of the parameters of the posterior distribution. This latter value depends on the degree to which the input value for the prediction is close to the training data.

In the case of unknown variance, it turns out that if the inverse gamma distribution is used to model the variance, as discussed in the previous section, then with normal prior on the coefficients conditional on the variance, and normal likelihood, it is the case that

$$p(\boldsymbol{\beta},\sigma^2) = p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2)$$

In this case $p(\boldsymbol{\beta}|\sigma^2)$ follows a normal distribution, and $p(\sigma^2)$ follows an inverse gamma distribution. This conditional product is called a normal-inverse gamma (IG) distribution, that is,

$$p(\boldsymbol{\beta},\sigma^2) = p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2) \sim N(\boldsymbol{\beta}|\boldsymbol{\beta_0},\sigma^2\boldsymbol{V_0})IG(\sigma^2|a_0,b_0)$$

$$= \frac{b_0^{a_0}}{(2\pi)^{\frac{D}{2}}|\boldsymbol{V_0}|^{0.5}\Gamma(a_0)}(\sigma^2)^{-\left(a_0+\frac{D}{2}+1\right)} \cdot e^{-\frac{1}{2\sigma^2}[(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\boldsymbol{V_0^{-1}}(\boldsymbol{\beta}-\boldsymbol{\beta_0})+2b_0]}$$

Where D is the number of variables in the regression equation. With the normal-inverse gamma prior and a normal likelihood, the posterior joint distribution of the coefficients and the variance is also a normal-inverse gamma with

$$p(\boldsymbol{\beta},\sigma^2|\boldsymbol{X},\boldsymbol{y}) = N(\boldsymbol{\beta}|\boldsymbol{\beta}_N,\sigma^2\boldsymbol{V}_N)IG(\sigma^2|a_N,b_N)$$

where

$$\boldsymbol{\beta}_N = \boldsymbol{V}_N(\boldsymbol{V}_0^{-1}\boldsymbol{\beta}_0+\boldsymbol{X}^T\boldsymbol{y})$$

$$\boldsymbol{V}_N = (\boldsymbol{V}_0^{-1}+\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

$$a_N = a_0 + n/2$$

$$b_N = b_0 + \frac{1}{2}(\boldsymbol{\beta}_0^T\boldsymbol{V}_0^{-1}\boldsymbol{\beta}_0+\boldsymbol{y}^T\boldsymbol{y}-\boldsymbol{\beta}_N^T\boldsymbol{V}_N^{-1}\boldsymbol{\beta}_N)$$

See Appendix 2 for the derivation of this last expression.

The posterior predictive distribution in this case is not a normal distribution. Rather it is a Student's t-distribution $\mathcal{T}\,(mean, variance, d.f.)$, i.e.,

$$p(\boldsymbol{y}|\boldsymbol{x},\mathcal{D}) = \mathcal{T}\left(\boldsymbol{y}|\boldsymbol{x}\boldsymbol{\beta}_{N,}\frac{b_N}{a_N}(1+\boldsymbol{x}^T\boldsymbol{V}_N\boldsymbol{x}),2a_N\right)$$

With mean = $\chi\beta_N$, variance = $\frac{b_N}{a_N}(1+\boldsymbol{X}\boldsymbol{V}_N\boldsymbol{X}^T)$, and degrees of freedom = $2a_N$.

See Appendix 3 for the derivation.

For our particular application, the number of degrees of freedom for the *t*-distribution is roughly the sum of the number of data points in the prior and the number of data points in the likelihood. As long as the total of these two is at least 30, the Student's *t*-distribution will be approximately normal.

The expression for the coefficients and the variance matrix is like the case when the variance is known – the difference is that the variance disappears from the two expressions once the assumption that it is known is removed.

Markov Chain Monte Carlo

When the assumption of known variance is removed, the result is still analytically tractable, if the likelihood is normal. But what happens if a Student's t-distribution is used for the likelihood instead? In this case – a normal prior and a Student's t likelihood – the resulting posterior is not a conjugate prior so the result cannot be derived analytically. In this case a simulation method must be used. Most cost analysts are familiar with Monte Carlo simulation. Markov chain Monte Carlo simulation is a specific type of Monte Carlo simulation that is different than what is typically used in cost risk analysis. To begin, Markov chain Monte Carlo (MCMC) simulation picks a random parameter value to consider. The simulation will continue to generate random values subject to some rule for determining what makes a good parameter value. Given the prior distribution, if the simulation generated parameter value is better than the previous at explaining the data, it is added to the chain of parameter values with a probability determined by how much better it is than the previous.

The basic idea in Markov chain Monte Carlo is to sample each variable in turn, conditioned on the values of all the other variables in the distribution. Given a joint sample $x(n)$, generate a new sample $x(n+1)$ by sampling each component in turn, based on the values of $x(n)$. If there are two variables, x_1 and x_2, calculate $x_1(n+1) \sim p(x_1|x_2(n))$ and $x_2(n+1) \sim p(x_2|x_1(n+1))$.

As mentioned earlier, start the simulation from an arbitrary state. This means choosing an initial value for the parameters without requiring it to already follow the target distribution. The key idea is that the Markov chain is designed so that, over time, it "forgets" this starting point and converges toward the desired posterior distribution.

Because of the conditional simulations, it takes some time for the chain to converge to its stationary distribution. The samples collected from the period before the chain converges are discarded. These discarded samples are generated during what is called the burn-in phase. Because the simulation starts from an arbitrary state, it is common to use multiple initial values, which generate multiple chains, one for each set of initial values for the parameters.

Two tools are discussed that can be used to do the Markov chain Monte Carlo simulations. One is R – the built-in statistical capabilities in base R to conduct the simulation. Another is WinBUGS, one of several simulation tools that are designed specifically for doing a particular type of Markov chain Monte Carlo called Gibbs sampling. WinBUGS is free Windows software for doing Bayesian inference Using Gibbs Sampling. WinBUGS requires some simple programming. The syntax is similar to R, but because it is specifically designed to solve this particular kind of problem, less code is required than in base R. WinBUGS also has numerous diagnostic tools and graphs. However, if you want to do analysis, you need to export the simulations results back to R. One option for working around this is to use the R2OpenBUGS package for R to run WinBugs from R, which then produces the results in R for additional analysis. R code is provided below for the case of the Student's t likelihood with unknown variance.

Note that in R, comment lines begin with the hashtag symbol #. In the R code, the number of trials is set equal to 10,000, with a burn-in period equal to the first 1,000 trials.

In the Markov chain Monte Carlo simulation, the likelihood of the posterior for the candidate values is calculated. If the difference between the log likelihood of the current values is larger than a random value, this value is accepted. Otherwise, it is rejected. Some of the values with higher likelihood are rejected and some value with lower likelihood are accepted, in order to more fully explore the parameter space. Accepting too many candidate samples leads to highly correlated draws over time. Thus, a relatively large rejection rate, on the order of 40-80%, is desired.

The R code is simple. There is no need to actually code the simulation. What is needed is to specify the likelihood, the priors, the data, and the initial values. The *likelihood* describes how probable the observed data is given a particular value of a parameter or hypothesis and is based on a pdf chosen to represent the uncertainty about the residuals of the CER. The *priors* are the initial beliefs about the parameters, the *data* is the evidence you have to regress over, and the *initial values* are arbitrary starting points that could be based on using standard regression on the data. Note that in the code below the x values are the log transformations of the weight of each data point and the y values are the log transformation of the cost of each data point.

R code:

```
install.packages("R2OpenBUGS")

library('R2OpenBUGS')

#Example

  x = c(8.2938,7.750,7.2235,6.6598,7.6533,8.0956)
```

```
y = c(4.9182,4.5276,3.5965,3.68437,3.9526,NA)
n <- length(x)
data <- list("x","y","n")
inits <- function(){
  list(beta=c(5,0.7),tau=1,gamma=c(0,0,0,0,0,0))
  list(beta=c(4,0.9),tau=0.1,gamma=c
  (0,0,0,0,0,0))
}
Case3 = bugs(data,inits,
        model.file="model.txt",
        parameters=c("beta","tau","gamma"),
        n.chains=1,n.iter=20000,n.burnin=500
        0,n.thin=1,
        codaPkg=FALSE, debug = TRUE)
summary(Case3)
```

Three cases are considered and the results of Bayesian regression for them are compared. In the first case, which is the Bayesian regression example calculated earlier using only Excel formulas, the prior is a normal distribution, the likelihood is a normal distribution, and the variance is known. As discussed, the assumption of known variance is not realistic, so the second case relaxes this assumption but does not change the assumptions regarding the likelihood or the prior; this is still analytically tractable even though the details are a little more complicated. In the third case, recognizing that since the assumption of normality is suspect because the data set is small, the likelihood is instead assumed to follow a Student's t-distribution.

The results of the three cases are in Table 4. Case 1 is the normal prior and normal likelihood with known variance. Case 2 is the normal prior and normal likelihood with unknown variance. Case 3 is the normal prior and Student's t likelihood with unknown variance.

	Case 1	Case 2	Case 3
log-space intercept	4.4607	4.5800	4.5780
linear intercept	0.8794	0.8846	0.8880
slope	0.1166	0.1263	0.1229
prediction ($ millions)	144	163	163

Table 4. Comparison of the results of the three cases.

The mean results of cases 2 and 3 are virtually identical. Changing the assumption of the variance causes less weight to be placed on the likelihood, and more on the prior. Recalling that the actual cost for the mission (as discussed in the application of Bayesian regression when the prior and likelihood are normally distribution and the variance is known) is equal to $180, removing the assumption of known variance results in a more accurate prediction.

What about the variance? The variance is needed to model the uncertainty and risk for the estimate. For case 1, the predictive variance at the new estimate in log space is *0.1085*, only slightly higher than the variance of the likelihood. For case 2, when the assumption that the variance is known is removed, the predictive distribution is a Student's t distribution with *75* degrees of freedom and variance equal to *0.9061*, much higher than for case 1, but also much more realistic. For case 3, there is no analytical predictive distribution so it must be calculated from the results of the Markov chain Monte Carlo simulation. The R code for this is (considered in the context of the previous code):

```
Ynew<-numeric(T)
munew<-numeric(T)
for(j in B+1:T){
  munew[j]<-alph[j]+beta[j]*(8.0956-7.5161)
  Ynew[j]<-rt(1,3)*(sigma2[j])^0.5+munew[j]
}
```

In this code, *8.0956* is the new weight in log space, and *7.5161* is the average log space weight for the sample.

The sample variance is *2.1508* in log space. The histogram is very wide, and ranges from *-31* to *84*. A truncated histogram from *0* to *15* is displayed in Figure 6. Fitting a normal distribution using the MASS (Modern Applied Statistics with S) package in R, a normal and a Student's t-distribution is fitted. The normal distribution has variance equal to the sample variance, while the Student's t has variance equal to *2.15084* and three degrees of freedom.

Note that a normal distribution in log space is a lognormal distribution in unit space, by definition. A

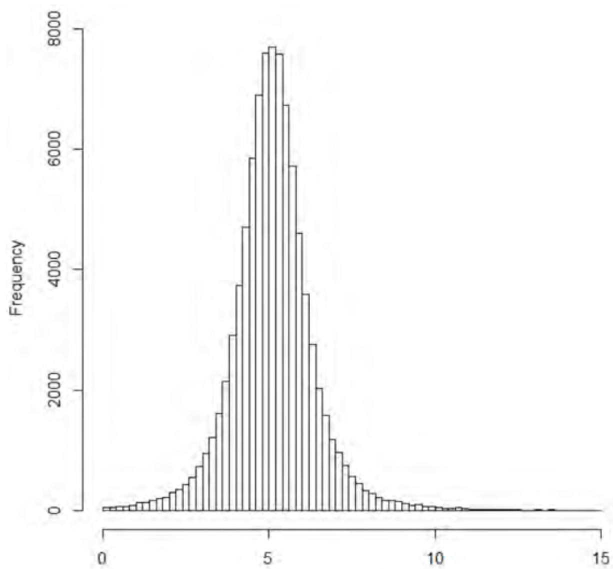


Figure 6. Truncated histogram for Student's t likelihood and unknown variance.

Student's t-distribution in log space is a log-t distribution in unit space.

A lognormal distribution has finite mean and variance, and you can parameterize a lognormal by these two parameters. A log t-distribution has neither a finite mean nor a finite variance. As a result, it has an extremely heavy right tail. See Table 5 for a comparison of the posterior predictive S-curves generated by the four different assumptions. The first assumption, normal likelihood of the estimate (in log space) with known variance results in the narrowest S-curve, with a *5th-95th* percentile range equal to $*164* million. Relaxing the variance assumptions leads to a wider S-curve, with a *5th-95th* percentile range equal to $*762* million. Changing the distribution of the likelihood to a student's t in log space results in two different S-curves, depending on whether a normal distribution or a t-distribution is used to model the predictive distribution in log space, which is a lognormal or log-t distribution in unit space. A lognormal distribution has a *5th-95th* percentile range equal to $*1.8* billion. The log-t distribution has a narrower range from the *5th* to the *95th*, equal to $*1.1* billion. However, the tail of the log-t is very heavy which causes the range to be narrower for the log-t since more of the weight is in the far right tail, which is beyond the *95th* percentile.

Although the log-t range is narrower, once you go to the *99th* percentile and above, the log-t distribution begins to blow up, going from $*7* billion at the *99th* percentile to $*843* billion at the *99.9th* percentile. That is, there is a *1* in *1,000* chance that the cost will be at or above $*843* billion for an estimate whose *50th* percentile is $*163* million. To get a more succinct sense of how these S-curves compare, look at the coefficients of variation (CV), which is the ratio of the standard deviation of the estimate to its mean. A variety of cost growth studies indicate that a reasonable value for CV is *50%* at the beginning of development (Smart 2018). The CV for the base case with normal likelihood and known variance is *34%*, lower than the rule of thumb. The CV for normal likelihood and unknown variance is *94%*, much higher and definitely in the reasonable range. When the likelihood is changed to a Student's t-distribution, the CV when a normal is fit in log space (lognormal in unit space) to the predictive equation is *275%*; when a Student's t-distribution is fit to the predictive equation, this is a log-t in unit space, which since it does not have a finite mean or variance, the CV does not exist.

Confidence level	Normal likelihood with known variance	Normal likelihood with unknown variance	Student t likelihood, unknown variance, normal error	Student t likelihood, unknown variance, Student's t error
5%	$84	$33	$15	$23
10%	$94	$48	$25	$41
20%	$109	$73	$47	$72
30%	$121	$99	$76	$100
40%	$132	$128	$112	$129
50%	$144	$163	$163	$163
60%	$156	$207	$236	$205
70%	$171	$269	$352	$266
80%	$190	$365	$560	$370
90%	$220	$558	$1,068	$642
95%	$247	$795	$1,819	$1,169
99%	$310	$1,565	$4,942	$7,295
99.5%	$336	$2,016	$7,125	$21,662
99.9%	$398	$3,434	$15,151	$842,898

Table 5. S-curve comparison for the different assumptions. Costs are in millions of dollars.

Even though the assumption of a t likelihood is more correct in the case of small samples, some common sense in establishing S-curves is needed. A risk range that includes values orders of magnitudes above the median is not reasonable. On the other extreme, the assumption of known variance is unrealistic, so a compromise that assumes unknown variance but normal likelihood seems reasonable. Since the analysis is leveraging experience from a larger data set this makes sense.

Summary

This paper has extended the linear model by changing two key assumptions. One of these is the assumption of known variance, and the second is the assumption of a normal distribution about the likelihood of the log-transformed ordinary least squares equation.

The assumption of known variance is not reasonable. The variance of a sample is not known with certainty. What is worse is that the "known" variance is actually estimated from the likelihood. The likelihood in the problem set up has a small sample size, so even the estimate of this variance has significant uncertainty. It was shown that the variance follows a specific distribution, namely a scaled inverse chi-squared (or equivalently an inverse gamma), and the parameters for the prior can be calculated from the number of degrees of freedom and the estimated variance of the estimating equation in the regression analysis used to establish the prior distribution.

The assumption of a normal likelihood is suspect. The problem set up is designed to work with likelihoods for which there are a small number of data points, otherwise a traditional frequentist regression method would have been appropriate. In this case, as has been pointed out by others (Druker et al., 2009), in such situations a Student's t-distribution is more appropriate.

An analysis was provided by relaxing the assumptions. First, the assumption of known variance was removed, which leads to a tractable result. The inclusion of uncertainty about the variance of the likelihood in the form of a prior distribution for the estimating variance leads to greater weight on the prior and less weight on the likelihood, which changes the mean. The posterior predictive equation follows a Student's t-distribution with high degrees of freedom as it benefits from the large sample used to establish the prior. For this number of degrees of freedom, the normal and Student's t-distribution are indistinguishable so the predictive equation can be modeled with a normal distribution.

Second, continuing with unknown variance, but now changing the assumption that the likelihood is normally distribution to a Student's t-distribution. This is not analytically tractable since without the assumption of a normal likelihood, the prior is not conjugate with the likelihood. In this case, a specific type of Monte Carlo simulation called Markov chain Monte Carlo needs to be used. It was shown how this can be accomplished using both a statistical programming platform called R, and a platform specifically designed for this type of Bayesian analysis called WinBUGS.

The assumption of known variance drives the mean. Changing to unknown variance significantly increases the mean and noticeably increases the variance. Changing this in turn to a Student's t likelihood does not change the mean but drives the variance much higher.

In conclusion, using the Bayesian approach can benefit a cost estimator when they are faced with an all-too common situation of limited data. The simple linear model has been extended to include more realistic assumptions. However, when departing from the assumption that the likelihood is normally distributed, common sense should be applied when applying an extremely heavy-tailed distribution. The approaches used in this model also apply when the prior is specified using expert judgment.

Next Steps

This paper deals with parametric models, ones which use specific distributions, such as the normal, lognormal, or Student's t-distribution. It does not cover regression methods that do not use a distribution to model the residuals, such as the Minimum Unbiased Percentage Error (MUPE) method or the Zero-bias Minimum Percent Error

(ZMPE) method. Markov Chain Monte Carlo simulation can be applied to histograms for both the prior and likelihood, and a future paper will deal with the application of MCMC techniques to ZMPE and MUPE regression methods.

APPENDIX 1

In both this appendix and in Appendix 2, the matrix identity

$$\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u} - 2\boldsymbol{\alpha}^T\boldsymbol{u} + \boldsymbol{\alpha}^T\boldsymbol{A}^{-1}\boldsymbol{\alpha} = (\boldsymbol{u} - \boldsymbol{A}^{-1}\boldsymbol{\alpha})^T\boldsymbol{A}(\boldsymbol{u} - \boldsymbol{A}^{-1}\boldsymbol{\alpha})$$

is used, where ***u*** and ***a*** are vectors, and ***A*** is a matrix.

The two exponents need to be combined, and have it result in something that is quadratic in ***b***. Thus, the ***u*** term in the matrix identity needs to represent ***b***.

Set $\boldsymbol{\alpha} = \boldsymbol{V}_0^{-1}\boldsymbol{\beta}_0 + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{y}$ and $\boldsymbol{A} = \boldsymbol{V}_0^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}$.

Begin with the expression .

$$(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\boldsymbol{V}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \frac{1}{\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

Expanding this -

$$\boldsymbol{\beta}^T\left(\boldsymbol{V}_0^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{\beta} - 2\left(\boldsymbol{V}_0^{-1}\boldsymbol{\beta}_0 + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{y}\right)^T\boldsymbol{\beta} + \boldsymbol{\beta}_0^T\boldsymbol{V}_0^{-1}\boldsymbol{\beta}_0 + \frac{1}{\sigma^2}\boldsymbol{y}^T\boldsymbol{y}$$

Rearranging terms yields

$$\boldsymbol{\beta}^T\boldsymbol{V}_0^{-1}\boldsymbol{\beta} - 2(\boldsymbol{V}_0^{-1}\boldsymbol{\beta}_0)^T\boldsymbol{\beta} + \boldsymbol{\beta}_0^T\boldsymbol{V}_0^{-1}\boldsymbol{\beta}_0 + \frac{1}{\sigma^2}\boldsymbol{y}^T\boldsymbol{y} - \frac{2}{\sigma^2}(\boldsymbol{X}^T\boldsymbol{y})^T\boldsymbol{\beta} + \frac{1}{\sigma^2}(\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{X}\boldsymbol{\beta}$$

Ignore the last two terms as they are constants. Then addition and subtraction of the constant term $\boldsymbol{\alpha}^T\boldsymbol{A}^{-1}\boldsymbol{\alpha}$ completes the square.

APPENDIX 2

See Appendix 1 for the key matrix identity used. Two terms will be combined and will result in a single term that is quadratic in ***b***. Thus, the ***u*** term in the matrix identity will represent ***b***, as in Appendix 1. Set

$\boldsymbol{\alpha} = \boldsymbol{V}_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}^T\boldsymbol{y}$ and $\boldsymbol{A} = \boldsymbol{V}_0^{-1} + \boldsymbol{X}^T\boldsymbol{X}$

Begin with the expression $(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\boldsymbol{V}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$

Expanding the expression, it is found that $\boldsymbol{\beta}^T\boldsymbol{V}_0^{-1}\boldsymbol{\beta} - 2(\boldsymbol{V}_0^{-1}\boldsymbol{\beta}_0)^T\boldsymbol{\beta} + \boldsymbol{\beta}_0^T\boldsymbol{V}_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{y}^T\boldsymbol{y} - 2(\boldsymbol{X}^T\boldsymbol{y})^T\boldsymbol{\beta} + (\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{X}\boldsymbol{\beta}$

Rearranging terms yields $\boldsymbol{\beta}^T(\boldsymbol{V}_0^{-1} + \boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta} - 2(\boldsymbol{V}_0^{-1}\boldsymbol{\beta}_0 + X^Ty)^T\beta + \beta_0^TV_0^{-1}\beta_0 + y^Ty$

The first two terms are $u^T A u - 2\alpha^T u$ so $\alpha^T A^{-1}\alpha$ are added and subtracted in order to complete the square. After applying the identity it is found that $(\beta - A^{-1}\alpha)^T A(\beta - A^{-1}\alpha) + \beta_0^T V_0^{-1}\beta_0 + y^T y - \alpha^T A^{-1}\alpha$ using *A* and *α* to keep the expression simple.

Noting that $V_N = A^{-1}$ and $\beta_N = A^{-1}\alpha$ yields $(\beta - \beta_N)^T V_N^{-1}(\beta - \beta_N) + \beta_0^T V_0^{-1}\beta_0 + y^T y - \beta_N^T V_N^{-1}\beta_N$ which is what needs to be proved. Note that that in this case constants are not ignored, unlike in the known variance case. The difference in this case is that there is an inverse gamma parameter in the exponent, so the additional terms are not constants as in the known variance case.

APPENDIX 3

To see that $p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{D}) = \mathcal{T}\left(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}_{N,} \frac{b_N}{a_N}(1 + \boldsymbol{X}\boldsymbol{V}_N\boldsymbol{X}^T), 2a_N\right)$, note that

$$p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{D}) = \int_0^\infty \int_{-\infty}^\infty p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) d\boldsymbol{\beta} d\sigma^2$$

$$= \int_0^\infty \int_{-\infty}^\infty N(\boldsymbol{X}^T\boldsymbol{\beta}, \sigma^2) N\left(\boldsymbol{\beta}|\boldsymbol{\beta}_{N,}\sigma^2\boldsymbol{V}_N\right) IG(\sigma^2|a_N, b_N) d\boldsymbol{\beta} d\sigma^2$$

$$\int_0^\infty \left[\int_{-\infty}^\infty N(\boldsymbol{X}^T\boldsymbol{\beta}, \sigma^2) N\left(\boldsymbol{\beta}|\boldsymbol{\beta}_{N,}\sigma^2\boldsymbol{V}_N\right) d\boldsymbol{\beta}\right] IG(\sigma^2|a_N, b_N) d\sigma^2$$

It has already been shown that for the known variance case that

$$\int_{-\infty}^\infty N(\boldsymbol{X}^T\boldsymbol{\beta}, \sigma^2) N\left(\boldsymbol{\beta}|\boldsymbol{\beta}_{N,}\sigma^2\boldsymbol{V}_N\right) d\boldsymbol{\beta} = N\left(\boldsymbol{X}^T\boldsymbol{\beta}_N, \sigma^2(1 + \boldsymbol{X}^T\boldsymbol{V}_N\boldsymbol{X})\right)$$

Substituting yields

$$\int_0^\infty N\left(\boldsymbol{X}^T\boldsymbol{\beta}_N, \sigma^2(1 + \boldsymbol{X}^T\boldsymbol{V}_N\boldsymbol{X})\right) IG(\sigma^2|a_N, b_N) d\sigma^2$$

To keep the notation simple, let $\boldsymbol{V}^* = 1 + \boldsymbol{X}^T\boldsymbol{V}_N\boldsymbol{X}$ and $\phi = (\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\beta}_N)^T \boldsymbol{V}^{*-1}(\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\beta}_N)$ Then the above expression is $\frac{{b_N}^{a_N}}{(2\pi)^{\frac{D}{2}}|\boldsymbol{V}^*|^{\frac{1}{2}}\Gamma(a_N)} \int_0^\infty (\tau)^{a_N + \frac{D}{2} + 1} e^{-\tau\left(b_N + \frac{\phi}{2}\right)} \frac{1}{\tau^2} d\tau$

Let $\tau = \frac{1}{\sigma^2}$. Then τ has the same range and $d\sigma^2 = -\frac{d\tau}{\tau^2}$.

$$\frac{{b_N}^{a_N}}{(2\pi)^{\frac{D}{2}}|\boldsymbol{V}^*|^{\frac{1}{2}}\Gamma(a_N)}\int_0^\infty (\tau)^{a_N+\frac{D}{2}+1}e^{-\tau\left(b_N+\frac{\phi}{2}\right)}\frac{1}{\tau^2}d\tau$$

$$=\frac{{b_N}^{a_N}}{(2\pi)^{\frac{D}{2}}|\boldsymbol{V}^*|^{\frac{1}{2}}\Gamma(a_N)}\int_0^\infty (\tau)^{a_N+\frac{D}{2}-1}e^{-\tau\left(b_N+\frac{\phi}{2}\right)}d\tau$$

The integrand has the form of a gamma distribution, the probability density function for which is defined as

$$p(x)=\frac{1}{\Gamma(k)\theta^k}x^{k-1}e^{-\frac{x}{\theta}}$$

Letting $k=a_N+\frac{D}{2}$ and $\theta=\left(b_N+\frac{\phi}{2}\right)^{-1}$, write

$$\frac{{b_N}^{a_N}\Gamma(a_N+\frac{D}{2})\left(b_N+\frac{\phi}{2}\right)^{-\left(a_N+\frac{D}{2}\right)}}{(2\pi)^{\frac{D}{2}}|\boldsymbol{V}^*|^{\frac{1}{2}}\Gamma(a_N)}\int_0^\infty \frac{1}{\Gamma(a_N+\frac{D}{2})\left(b_N+\frac{\phi}{2}\right)^{-\left(a_N+\frac{D}{2}\right)}}(\tau)^{a_N+\frac{D}{2}-1}e^{-\tau\left(b_N+\frac{\phi}{2}\right)}d\tau$$

The range of the gamma distribution is the nonnegative real numbers, so the integrand evaluates to *1*, leaving

$$\frac{{b_N}^{a_N}\Gamma(a_N+\frac{D}{2})\left(b_N+\frac{\phi}{2}\right)^{-\left(a_N+\frac{D}{2}\right)}}{(2\pi)^{\frac{D}{2}}|\boldsymbol{V}^*|^{\frac{1}{2}}\Gamma(a_N)}$$

This expression is proportional to

$$\frac{\Gamma(\frac{2a_N+D}{2})\left(1+\frac{(y-\boldsymbol{X}^T\boldsymbol{\beta}_N)^T(\frac{b_N}{a_N}(1+\boldsymbol{X}^T\boldsymbol{V}_N\boldsymbol{X}))^{-1}(y-\boldsymbol{X}^T\boldsymbol{\beta}_N)}{2a_N}\right)^{-\left(a_N+\frac{D}{2}\right)}}{\pi^{\frac{D}{2}}\left|2a_N\cdot\frac{b_N}{a_N}\boldsymbol{V}^*\right|^{\frac{1}{2}}\Gamma(a_N)}$$

which is a multivariate Student's t distribution with mean $\boldsymbol{X}^T\boldsymbol{\beta}$, variance $\frac{b_N}{a_N}(1+\boldsymbol{X}^T\boldsymbol{V}_N\boldsymbol{X})$, and $2a_N$ degrees of freedom.

References

Boehm, B., A. Hira, K. Qi, and E. Venson, "Calibrating Use Case Points Using Bayesian Analysis," presented at the 2018 International Cost Estimating and Analysis Association Professional Development & Training Workshop.

Cochran, W. G., 1934,"The distribution of quadratic forms in a normal system, with applications to the analysis of covariance," Mathematical Proceedings of the Cambridge Philosophical Society. 30 (2).

Congdon, P., 2006, *Bayesian Statistical Modelling*, 2nd Edition, Wiley, West Sussex.

Druker, E.R., R.L. Coleman, and P.J. Braxton, "Don't Let the Financial Crisis Happen to You: Why Estimates Using Power CERs are Likely to Experience Cost Growth," presented at the 2009 ISPA-SCEA Annual Conference.

Foussier, P.M.M., "The Benefits of the Bayesian Approach vs. the Frequentist Approach when Dealing with Low Data Sample," presented at the 2010 ISPA-SCEA conference.

Hoffman, P., The Man Who Loved Only Numbers: The Story of Paul Erdos and the Search for Mathematical Truth, Hyperion, 1998, New York, New York.

Smart, C.B., "Enhancing Risk Calibration Methods," presented at the 2018 International Cost Estimating and Analysis Association Professional Development & Training Workshop.

Smart, C.B., "Covered with Oil: Incorporating Realism in Cost Risk Analysis," Journal of Cost Analysis and Parametrics, 2015.

Smart, C.B., 2014, "Bayesian Parametrics: How to Develop a CER with Limited Data and Even Without Data," presented at the 2014 International Cost Estimating and Analysis Association Professional Development & Training Workshop.

Taleb, N.N., *Skin in the Game*: *Hidden Asymmetries in Daily Life*, Random House, New York, 2018.

Christian B. Smart, Ph.D., CCEA, is a leading expert in cost estimating, project risk analysis, and statistical modeling for complex aerospace and defense programs. With more than twenty years supporting NASA, the Department of Defense, and major government contractors, he has helped transform how organizations understand uncertainty and make high-stakes decisions. Dr. Smart's work blends rigorous mathematics with practical insight, and his research on cost and schedule growth has become influential across the federal acquisition community. He is the author of Solving for Project Risk Management. Dr. Smart holds a Ph.D. in Applied Mathematics and is known for his clear, engaging explanations of complex analytical concepts.

Management: Using Time Series Forecasts and Regression Models

Anna B. Peters

Mark W. Hodgins, Ph.D.

Abstract

The recent digitization of contractor earned value management (EVM) data affords cost analysts a newfound ability to execute robust statistical and data science techniques that better predict total project cost and schedule realism. Time series analysis, a well-established method in private sector finance, is one such method. Auto regressive integrated moving average (ARIMA) models may capture the persistence and patterns in EVM data, as measured by CPI, SPI, and schedule execution metrics (SEMs). As a second option, macroeconomic regression models can measure the relationship between contract performance and external considerations, like unemployment and inflation, over time. Both techniques, moreover, may forecast future changes in EVM variables of interest, like IEAC. This presentation will discuss how these types of time series models and forecasts are employed on real acquisition programs and their associated integrated program management data and reports (IPMDAR) data using Python based tools to raise program analysts' alertness to emergent acquisition risks and opportunities.

EVM is a best practice that Department of Defense (DoD) acquisition programs implement to monitor cost and schedule requirements and performance. However, the implementation of these analytic principles is not a panacea for effective cost and schedule performance monitoring as large-scale DoD acquisition programs routinely experience cost overruns and schedule delays (Meier, 2010). The Government Accountability Office (GAO) completed several studies comparing the frequency and severity of cost and schedule overruns in large acquisition programs in 2000 vs 2008 and found a higher rate of cost and schedule overruns in 2008 than 2000 (Meier, 2010). Specifically, Meier (2010) determined that, based on several GAO study results in this vein, there was "in absolute terms, a 702% cost growth increase" and a "31% increase in schedule delays" over the 7-year period considered.

Fortunately, several new data sources, techniques, and metrics have created an opportunity to apply EVM analysis using advanced data science techniques to assist in the prevention of cost and schedule overruns. As an example, the relatively new requirement for large DoD acquisition programs to electronically publish an IPMDAR each reporting period provides the analyst with much higher fidelity data relative to non-digital Integrated Program Management Data and Analysis Reports (IPMRs) (IPMDAR Contract Performance, Schedule Performance, Implementation, 2020). Moreover, the evolution of data science platforms provides the cost and schedule community with low-to-no cost tools and techniques to examine cost and schedule information. Finally, some newer advanced schedule metrics created by the National Reconnaissance Office (NRO), named the SEMs, are becoming another robust option to track a program's schedule performance.

Given these opportunities, this study presents how relatively advanced statistical techniques are applied to acquisition data in order to improve the cost and schedule performance forecasting. Prior to discussing the specific methodological implementations, Section 2 examines the recent data source, technique, and metric opportunities in greater detail. Section 3 then describes the implementation of time series analysis forecasting on acquisition data sets. The study team executes the ARIMA technique to forecast future values of standard EVM, like cost performance index (CPI), and schedule performance index (SPI), and future measures of the NRO SEMs. External factors may also affect DoD acquisition performance. Section 4 executes an analytic process using linear regression with lags to explore how macroeconomic indicators, such as unemployment and inflation rates, may influence acquisition health. Section 5 and 6 explore future research and conclude the study, respectively.

Analytic Enablers

The application of new statistical and data science approaches to forecast acquisition performance is enabled in this effort by emergent metrics, data sources, and techniques. Specifically, the NRO SEMs provide additional descriptive metrics to conduct time series analysis, IPMDARs represent a new data set with more detailed contract performance information, and data science tools and techniques enable robust analysis. Each of these opportunities is discussed in detail throughout the remainder off this section

New data metrics are one key enabler. The NRO created eight advanced schedule analysis techniques, referred to as SEMs, to provide objective and improved measures of schedule performance (NRO, 2022). Specifically, NRO leadership wanted the metrics to serve as advanced predictions of inadequate schedule accomplishment before threshold levels are violated and costly re-programming is required (Schultheis, 2021). The NRO study team used Naval Air Systems Command (NAVAIR) data to derive their methods and test metric reliability on a portfolio of completed projects (Schultheis, 2021).

This analysis focused on three of the eight SEMs to support acquisition performance forecasting:

1. Current Baseline Realism Index (BRI) – Percentage of planned events that actually finished in the planning period; indicator of how well the contractor is following the plan in the period. NRO recommends BRI is applied using a six-reporting period cumulative window. (NRO, 2021)

$$Current\ BRI = \frac{\sum of\ tasks\ baselined\ to\ finish\ in\ RP\ \&\ completed\ in\ RP}{\sum tasks\ baselined\ to\ finish\ in\ RP}$$

- BRI ≥ 0.80 considered favorable and “on plan”. There are no major milestone delays or baseline resets expected in next 6-12 months.
- BRI ≤ 0.20 considered unfavorable and “way off plan”.

2. Current Baseline Progress Index (BPI) – Percentage of planned events that actually finished in or before the planning period; indicator of how many of the planned events in the period have actually been accomplished. (NRO, 2021)

$$Current\ BPI = \frac{\sum of\ tasks\ baselined\ to\ finish\ in\ RP\ \&\ completed\ in\ or\ before\ RP}{\sum tasks\ baselined\ to\ finish\ in\ RP}$$

- BPI ≤ 0.35 considered unfavorable and “way off plan”.

3. Current Baseline Execution Index (BEI) – Percentage of total events that actually finished in the current period; pace of work. (NRO, 2021)

$$Current\ BEI = \frac{\sum of\ tasks\ finished\ in\ RP\ even\ if\ not\ originally\ planned\ to\ finish\ in\ RP}{\sum tasks\ baselined\ to\ finish\ in\ RP}$$

The favorable and unfavorable thresholds for BRI and BPI were estimated from a statistical analysis of historical data from completed government acquisition programs (NRO, 2021). Programs actual acquisition performance was used to classify the favorable and unfavorable levels based on realized issues and positive impacts. Due to the data driven methods used to derive the metrics and the associated classifiers, data science and statistical techniques employing the descriptive SEMs in a predictive manner may provide new and improved analytic insights. In fact, JHU APL and the NRO are collaborating to further advance predictive analytics to improve acquisition program decision making and forecasting.

IPMDARs are a second enabler. Data Item Description (DID) DI-MGMT-81861C specifies the attributes of IPMDARs and contractually requires their use on significant DoD acquisition efforts (e.g., typically in excess of $20 million) (DID for IPMDAR, 2021). IPMDARs include three primary data components, which are the contract performance dataset (CPD), the schedule performance dataset (SPD), and the performance narrative report (IPMDAR Contract Performance, Schedule Performance, Implementation, 2020). The CPD includes most of the earned value information and the program baseline data. The SPD contains the contractor’s integrated master schedule (IMS), and the narrative reports provide supporting documentation to accompany the CPD and SPD.

Critically, IPMDARs possess instructions to ensure formatting rigor and maturity. File format specifications (FFSs) and data exchange instructions (DEIs) provide precise digitization instructions for transparency, replicability, and auditability for the CPD and SPD. In addition, the submittals are

Item Type	Description
Database (DB)	SQL (relational) database
DB management system	PostgreSQL
SQL toolkit and object-relational mapper	SQLAlchemy
PostgreSQL DB adapter	Psycopg2 (2.9.5)
Software (Programming Language)	Anaconda (Python)

Table 1. Outlines the major elements of the data environment

required on a monthly basis. The digitized data, reliable formatting, and frequency of reporting afford the ability to execute time series analysis on major development contracts. In turn, the newly available IPMDAR data for current and emergent acquisition programs provides an opportunity to improve upon the standard EVM analysis limitations.

The data environment is a third and final enabler for advanced statistical analysis. Low-to-no cost data science environments are now common and a burgeoning supply of data scientists with the requisite human capital execute complex analytics within these environments. Python is one such environment, which the study team selected for executing analytic techniques across the data pipeline. Python is highly used and a top ranked data science programming language, with significant community support. Python programming is relatively accessible because this open-source language possesses highly readable and writable syntax/code. Finally, Python supports a variety of emergent, predictive analytics on disparate sized relational data sets, which aligns with the study team's requirements.

Overall, the environment is multi-tiered (e.g., n-tier), where the first tier contains the database and the schemas, storing the structured data. A benefit of this architecture is the ability to process the information in any desired manner (e.g. data normalization or transformations). The second tier is the business logic, which is the analytics tier in this case. The business logic has access to the database through a customized python library via SQL queries and is supported by the psycopg2 Python library. The business logic tier is where the time series and macroeconomic regression analyses are conducted. The third tier is the presentation layer, which presents the analytics' visualizations using Jupyter notebooks and a plotly-dash dashboard (Plotly API, 2023). Table 1 outlines the major elements of the data environment. Table 2 lists the major Python libraries used across all analyses discussed in this report.

Python Library	Purpose
ARIMA, from statsmodels.tsa.arima.model	ARIMA time series modeling
auto_arima, from pmdarima.arima	Time series analysis, specifically auto_arima
dateutil, from relativedelta	Managing date-time formatted data
express, from plotly	Data visualization
graph_objects, from plotly	Data visualization
matplotlib.pyplot	Creating object-oriented plots
numpy	Scientific computing
pandas	Building and manipulating data structures
product, from itertools	Creating iterators for efficient looping
seaborn	Creating statistical data visualizations

Table 2. Major Python libraries used in this analysis

Time Series Analysis
Analytical Background

Time series modeling serves as a focal point of this effort due to the enabling data, metrics, and tools discussed in the previous section. In general, time series analysis is a statistical method used to determine potential trends in a data set over time. A statistical background is briefly discussed here before turning to how this analysis can be applied to EVM.

In time series modeling, time is an independent variable that is used to evaluate the potential for statistically significant relationships with a dependent variable (Cryer, et. al., 2008). The time period or frequency can vary with each time series model, allowing for different relationships to be investigated. Autoregressive integrated moving average (ARIMA) is one of the most common forms of time series analysis. ARIMA has been used extensively in the financial industry to forecast stock prices over time. ARIMA is also used frequently in epidemiology to predict the spread of disease over time.

The ARIMA model is a combination of two other time series model formats: the auto regressive (AR) model and the moving average (MA) model (Cryer, et. al., 2008). The AR model is constructed so that the variable of interest, also referred to as Y, depends only on its own lags (Cryer, et. al., 2008). The MA model is constructed so that Y depends on the lagged forecast errors (Cryer, et. al., 2008). The integration of the two model types is done once the stationarity assumption has been validated. The stationarity assumption can be validated in a variety of ways. In this analysis the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) unit-root test was applied (Kwiatkowski, et. al., 1992). ARIMA modeling can be applied using seasonal or non-seasonal trends. This analysis focuses on non-seasonal ARIMA modeling, due to acquisition programs and the IPMDAR data typically lacking seasonal tendencies. The theoretical ARIMA equation is below, Equation 1, where $Y_{(t-p)}$ and $\phi_q \epsilon_{(t-q)}$ are the Autoregressive (AR) and Moving Average (MA) components of the model, respectively (Cryer, et. al., 2008):

The more commonly used and abbreviated notation for a non-seasonal ARIMA model is: ARIMA(p, d, q) (Cryer, et. al., 2008). The ARIMA(p, d, q) model is comprised of three terms: p, d, and q. P is the order of the AR term (the number of lags of Y to be used as predictors), d is the minimum number of differencing needed to make the time series stationary (if the series is already stationary then d=0), and q is the order of the MA term (number of lagged forecast errors that should go into the ARIMA model) (Cryer, et. al., 2008).

Critically, ARIMA models can also be used to forecast future trends based on historical data to predict what the future values of the dependent variable could be (Cryer, et. al., 2008). In this analysis, one of the IPMDAR metrics (e.g., CPI, SPI, and SEMs) is used as the dependent variable to build an ARIMA model. That model is then used to forecast future CPI, SPI, IEAC, BRI, BPI, and BEI values (Montgomery, 2015).

When building the framework to forecast future metric values this analysis followed the best practice of separating the dataset into test and validation subsets. The test data makes up 80% of the original dataset and is used to create models to explain the data available. The remaining 20% of the data is put into a validation subset, which is used to verify the models chosen using the test data set. Time series analysis requires a continuous time series to use as the independent variable, so the data was separated based on the first 80% in time as the test dataset and the last 20% in time as the validation dataset. After validating the models adequately explain the data, they are applied to the entire dataset.

Confidence intervals were added to each metric's forecast window to account for the level of uncertainty inherent in any modeling forecast. The confidence intervals display what range the actual metric results could have based on the forecasted point estimate. The higher the confidence interval

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \cdots + \phi_q \epsilon_{t-q}$$

Equation 1. Theoretical ARIMA

percentage, the wider the range of data considered as part of the uncertainty bounds for each forecast point estimate. This analysis displays a 90%, 80%, and 50% confidence interval band for all forecast results. The varying levels are employed to reflect varying risk preferences among decision makers.

The specific ARIMA(p, d, q) values for each model were chosen using a python function called ‘auto_arima’ to optimize the process of determining the best model. The ‘auto_arima’ function automatically checks the stationarity assumption using the KPSS test and then iterates through possible combinations for the best ARIMA model. The final model is chosen based on which iteration has the lowest value of the Akaike information criterion (AIC) (Keaton, 2011). The final ARIMA models are then used to forecast the dependent variable values up to six-reporting periods into the future. Forecasts are calculated six reporting periods into the future to assist in the next rolling wave of EVM planning that acquisition programs typically complete every six months. The forecasted values are based on the historical dependent variable values that are contained within the model.

ARIMA using NRO SEMs

The study team used the ARIMA method to build time series models for the BRI, BPI, and BEI SEMs. Time was used as the independent variable in each model. The dependent variable for each model was either the BRI, BPI, and BEI values. These relationships were also used to forecast or extrapolate significant trends to future time periods. In this analysis, ARIMA can provide insight to address how the SEMs are predicted to change in the near future and if the forecasts are on track with previously defined SEM values.

In order to model the SEM BRI, BPI, and BEI metrics effectively, independent non-seasonal ARIMA models were used. The dependent variable metrics used in the ARIMA models are calculated in several different ways. The study team calculated several variants for each dependent variable based on time durations. While the BRI calculation is limited to a 6-reporting period cumulative view, BPI and BEI were estimated for each reporting period using 1-reporting period of data (current time and non-cumulative), 3-reporting periods of cumulative data, and 6-reporting periods of cumulative data. Creating three models for BPI, and BEI allows for a more comprehensive look at the BPI, and BEI values because each model reacts differently to the historical data. Typically, an ARIMA model using a SEM calculated with 1-reporting period of non-cumulative data is the most reactive to changes in the data, an ARIMA model using a SEM calculated with 3-reporting periods of cumulative data is a bit smoother and less reactive than the one reporting period model, and finally an ARIMA model using a SEM calculated with 6-reporting periods of cumulative data tends to be the least reactive and most conservative of the three modeling views.

The three model views of BEI below demonstrate the difference in the model and forecasts chosen with each view. Finally, the 6-reporting period cumulative views of BRI, BPI, and SPI were also compared to demonstrate the difference between the new SEMs and traditional EVM in identifying program schedule risk. The reader must note that all analytic visualizations presented are performed on notional data and intended to serve as illustrations of how to execute the analysis and interpret the results. However, the general results are reflective of trends observed when applying the analysis to several acquisition programs.

The three BEI models, which are the 1-reporting period (Figure 1), 3-reporting period cumulative (Figure 2), and 6-reporting period cumulative (Figure 3), reflect the differences in the responsiveness of the metric based on how much historical data is used to build the dependent variable. As predicted, the non-cumulative view is the most reactive and the 6-reporting period cumulative view is the most conservative with the smoothest trendline fit. Overall, all three BEI models show a downward trend indicating that there are likely schedule risks happening within the program. The BEI forecasts for each model mostly follow this general downward trend as well.

Figure 4 compares the 6-reporting period cumulative view for BRI, and BPI against a traditional EVM metric, SPI. BPI predicts that the program will be “way off plan” starting in reporting period 22. BRI predicts that the program will be “way off plan” in reporting period 35. The results of the BRI model and forecasts indicate that each month the

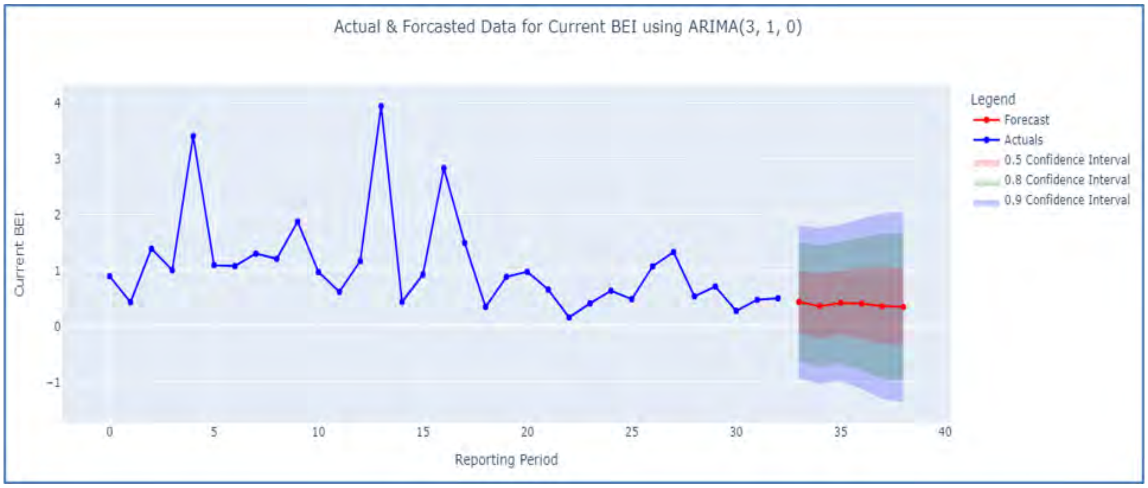


Figure 1. Current BEI model, for non-cumulative view

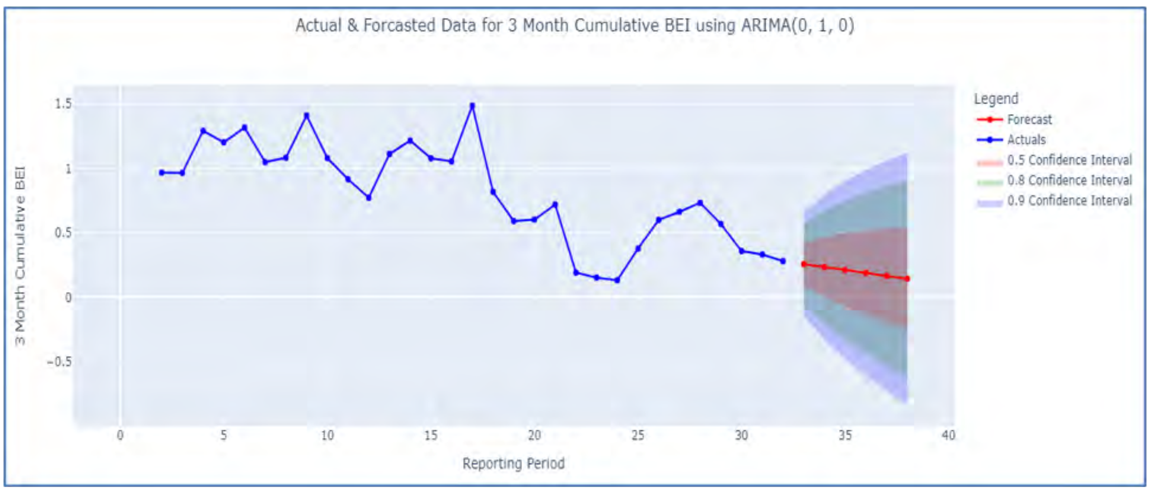


Figure 2. Current BEI model, 3 reporting period cumulative view

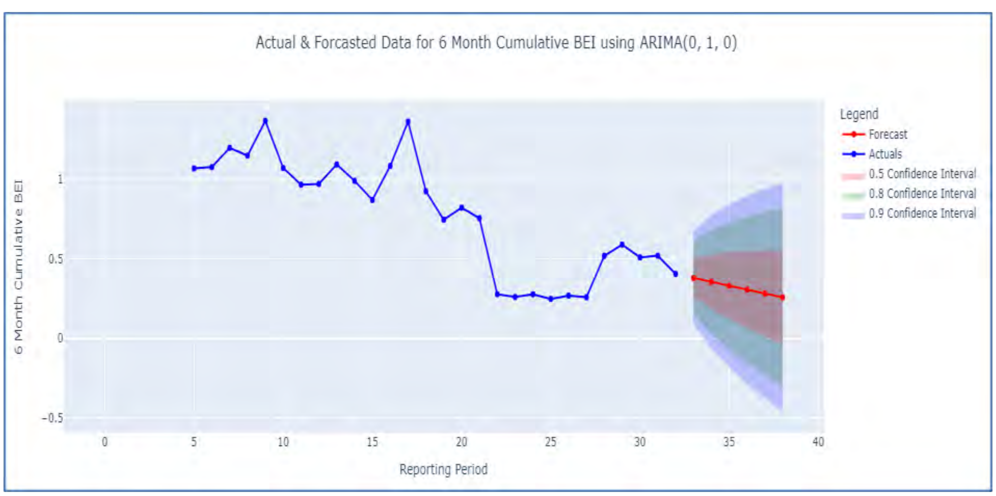


Figure 3. Current BEI model, 6 reporting period cumulative view

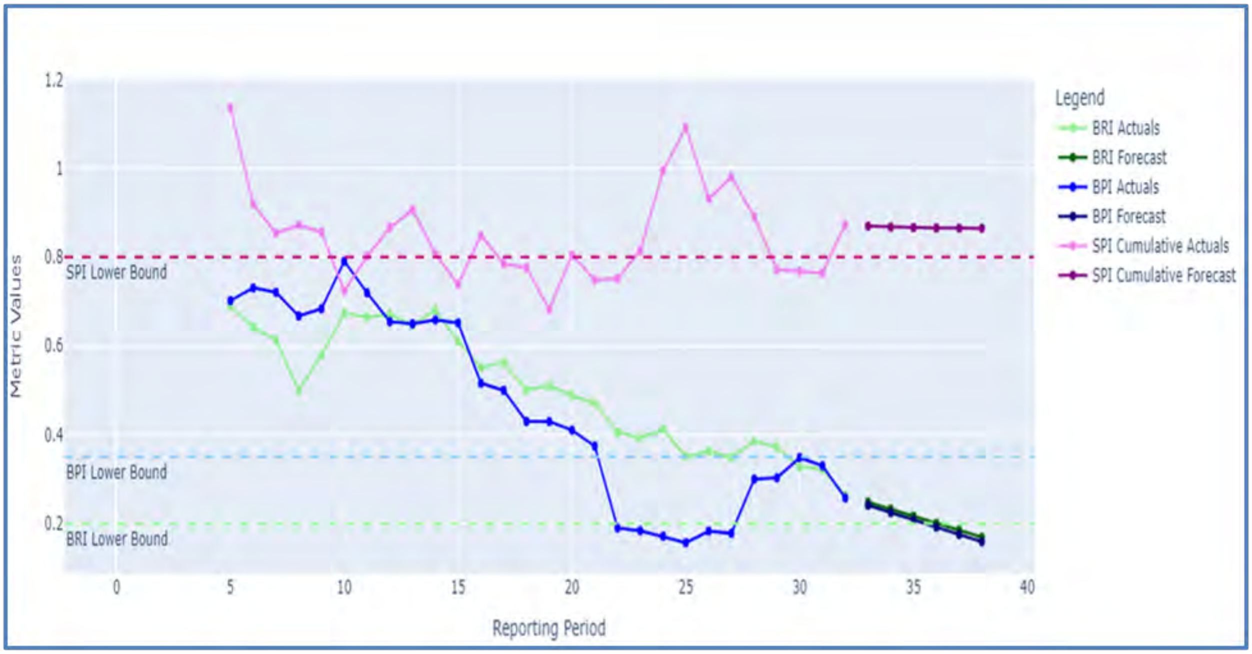


Figure 4. Comparing the 6 Reporting Period Cumulative Views of BRI, BPI, and SPI

percentage of completed tasks that were baselined to be executed has reduced; the schedule issues are forecasted to persist despite the recent improvement. The results of the BPI model and forecasts indicated that, similarly to BRI, the percentage of completed tasks that were baselined to be executed leading up to the current reporting period reduces; the schedule issues are forecasted to persist. On the contrary, the SPI values are not indicative of significant program schedule risk. The SPI values briefly drop below the 0.8 mark but quickly return to a reasonable SPI range, indicating there is not a lingering program schedule risk. As a result, the SEMs are more responsive to schedule issues. In this example, BRI forecast the program will be “way off plan” starting in reporting period 35 and BPI detects the program to be “way off plan” starting in reporting period 22 whereas SPI is expected to trend right above the 0.8 threshold and not raise alarm of a major schedule issue. Implementation of the SEMs may allow program personnel more time to address these issues before it potentially became a critical issue.

ARIMA using CPI and SPI

The study team also executed ARIMA modeling to forecast EVM metrics, like cumulative SPI and cumulative CPI, at varying levels of the work breakdown structure (WBS). In this case, the ARIMA models use time as the independent variable and the CPI or SPI calculations, using the IPMDAR dataset, as the dependent variable. These metrics are fundamental, descriptive EVM metrics. CPI is used to keep track of how the project cost is performing (Keaton, 2011). SPI is used to keep track of how the project schedule is performing (Keaton, 2011). Both metrics set a threshold of greater than one to be considered favorable performance and less than one to be considered unfavorable performance (Keaton, 2011). These relationships were also used to forecast or extrapolate significant trends to future time periods.

In this analysis, ARIMA can provide insights on how CPI and SPI are estimated to change in the near future, if predicted CPI and SPI rates are on track with recent history, and how abnormal changes to earned value may affect future cost and schedule performance. Moreover, the capability to drill up and down through the WBS levels allows account

$$Cumulative\ CPI_t = \sum_{i=1}^{j} \frac{BCWP_i}{ACWP_i} \qquad Cumulative\ SPI_t = \sum_{i=1}^{j} \frac{BCWP_i}{BCWS_i}$$
$$where\ j = latest\ reporting\ period$$

$$Threshold: \begin{cases} Favorable, & > 1 \\ Unfavorable, & < 1 \end{cases}$$

Equation 2.

managers to conduct root cause analysis with models that are updated and refined each reporting period.

An independent estimate at complete (IEAC) prediction is feasible as well. An IEAC prediction may improve upon the standard IEAC through the incorporation of cost and schedule forecasts from the ARIMA CPI and SPI analysis. For reference, the standard IEAC is used as a test for "reasonableness" for the contractor provided estimate at complete (EAC) (Keaton, 2011). IEAC can also be used as an external approximation of the budget at complete (BAC) (Keaton, 2011). BAC is a stable value that is a baseline budget planned for the entire program effort based on the amount of authorized funding. The BAC can change when a program re-baseline occurs. This analysis will focus on the comparison of IEAC to BAC. Tracking the IEAC can help balance work priorities, re-plan remaining tasks, and adjust the technical approach to complete the project with the remaining resources (Keaton, 2011).

This notional analysis uses the MIL-STD-881F Appendix G Ground Vehicle Systems WBS structure (DoD, 2022). The study focuses on the analytic implementation and results for two example WBS elements. In order to emphasize the multidimensional capabilities of this model, one example is executed at the third level of the WBS structure, while the second example is presented at the fourth level of the WBS structure. The example third level WBS element is 1.5.1 Development Test and Evaluation (DoD, 2022). The example fourth level WBS element is 1.5.1.1 Cybersecurity Test and Evaluation (DoD, 2022).

The dependent variables chosen for these ARIMA models were the IPMDAR CPI and SPI metrics. The study team forecasted CPI and SPI at varying WBS levels; however, the calculation of these indices at higher fidelity, subsystem and component levels, introduced data quality challenges. Accounting revisions and missing data, for example, led to normalization and statistical issues. In order to limit noise in the IPMDAR data caused by accounting revisions and missing data, the ARIMA models use rolling cumulative CPI and cumulative SPI measurements as the dependent variables (Equation 2).

Independent non-seasonal ARIMA models were used to model the IPMDAR cumulative CPI and cumulative SPI metrics effectively. The ARIMA models were used to forecast 6 future reporting periods of CPI or SPI values. The forecasted cumulative CPI and cumulative SPI are incorporated into a novel, predictive estimate of the IEAC to compare against BAC. There are several ways to solve for IEAC but Keaton (2011) provides a general equation (See Equation 3).

In this analysis, the traditional descriptive IEAC metric was adapted to derive an IEAC metric incorporating the results of the cumulative CPI and cumulative SPI ARIMA forecasts. The descriptive IEAC (Equation 4) is modified into a comprehensive valuation that allows for the insertion of the cumulative CPI and cumulative SPI 6 reporting period forecast (Keaton, 2011). The forecasted IEAC (Equation 5) replaces part of the comprehensive IEAC calculation with 6 reporting period forecasts from the cumulative CPI, and cumulative SPI ARIMA models to allow for a more robust predictor when compared to the program's BAC. This methodology created a novel IEAC equation (Equation 5) with three distinct parts. The left most part of the equation contains all costs performed to date, as represented by the ACWP. The second, middle, part contains the forecasted CPI and SPI metrics for 6 future reporting periods. Finally, the

$$IEAC = ACWP + \frac{(BAC - BCWP)}{CPI_{cum} * SPI_{cum}}$$
$$where\ CPI_{cum}\ or\ SPI_{cum}\ is\ the\ cumulative\ value\ of\ CPI\ or\ SPI$$

Equation 3.

third part contains the costs planned by the program until the end of the contract, as represented by a calculation of the estimate to complete.

$$IEAC_{comp} = ACWP_{cum} + \frac{BAC_t - BCWP_{cum}}{CPI_{cum} * SPI_{cum}}$$
$$where\ t = time,\ ACWP_{cum} = \left(\sum_{i=1}^{t-1} acwp_i\right), BCWP_{cum} = \left(\sum_{i=1}^{t-1} bcwp_i\right),$$
$$CPI_{cum} = \left(\sum_{i=1}^{t-1} \frac{bcwp_i}{acwp_i}\right),\ SPI_{cum} = \left(\sum_{i=1}^{t-1} \frac{bcwp_i}{bcws_i}\right)$$

Equation 4. IEAC, Comprehensive

$$IEAC_t = ACWP_{cum} + \sum_{j=t+1}^{t+6} \left(\frac{\widetilde{BCWS_j} * \widehat{SPI_j}}{\widehat{CPI_j}}\right) + \frac{BAC_t - BCWP_{cum} - \sum_{j=t+1}^{t+6} \widehat{BCWP_j}}{CPI_{cum} * SPI_{cum}}$$
$$where\ t = time,\ ACWP_{cum} = \left(\sum_{i=1}^{t-1} acwp_i\right), BCWP_{cum} = \left(\sum_{i=1}^{t-1} bcwp_i\right),$$
$$CPI_{cum} = \left(\sum_{i=1}^{t-1} \frac{bcwp_i}{acwp_i}\right),\ SPI_{cum} = \left(\sum_{i=1}^{t-1} \frac{bcwp_i}{bcws_i}\right)$$
$$\widetilde{BCWS_j} = BCWS\ to\ complete\ within\ forecast\ window\ of\ ARIMA\ models,$$
$$\widehat{SPI_j} = forecasted\ value\ of\ SPI\ based\ on\ ARIMA\ models,$$
$$\widehat{CPI_j} = forecasted\ value\ of\ CPI\ based\ on\ ARIMA\ models$$

Equation 5.

The cumulative CPI and cumulative SPI models for WBS element 1.5.1 Development Test and Evaluation (Figure 5 and Figure 6) both show an overall negative trend in the program's cost and schedule performance, based on the notional data (DoD, 2022). The forecasts for both models predict this trend is going to continue for the next six reporting periods. Despite the overall negative trend, the cumulative CPI model never dropped below 0.8 indicating there is not a high-cost risk for the program based, on the current notional data. However, cumulative SPI floats right around the 0.8 threshold starting in reporting period 7 indicating there is some schedule risk for the program based on current notional data. Nonetheless, the 80% and 90% confidence intervals for both models forecast potential cost and schedule risk increases, which may warrant increased monitoring and enforcement activities by the program office.

Similar to the previous two models, the cumulative CPI and cumulative SPI models for WBS element 1.5.1.1 Cybersecurity Test and Evaluation (Figure 7 and Figure 8) both show an overall negative trend in the program's cost and schedule performance, again based on the notional data (DoD, 2022). The forecasts for both models predict this trend is going to continue for the next six reporting periods. In both the cumulative CPI and cumulative SPI figures a drastic drop in cost and schedule performance,

Figure 5. Cumulative CPI model, for WBS Element 1.5.1

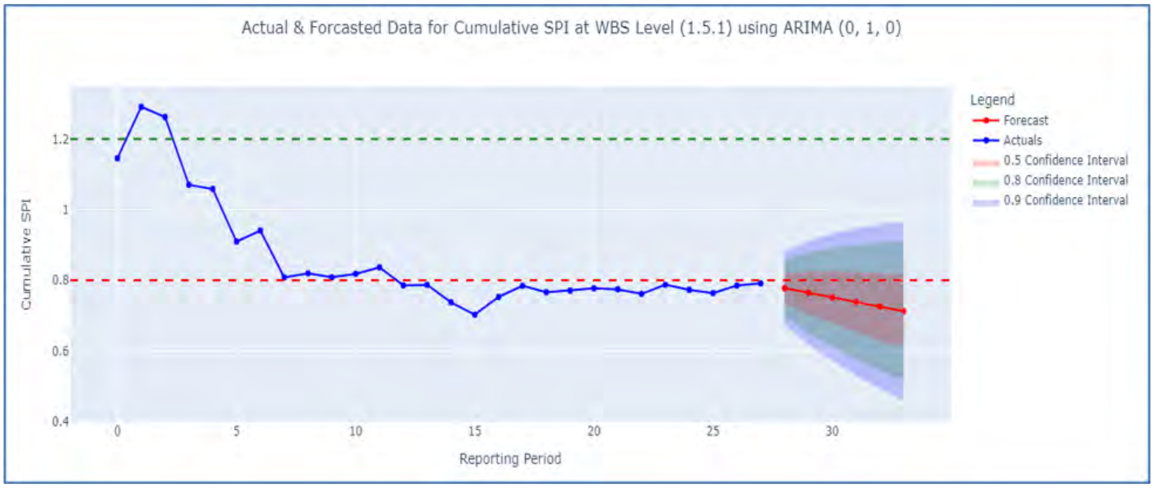


Figure 6. Cumulative SPI model, for WBS Element 1.5.1

Figure 7. Cumulative CPI model, for WBS Element 1.5.1.1

Figure 8. Cumulative SPI model, for WBS Element 1.5.1.1

Third Level WBS Elements				
WBS Element	**Name**	**IEAC**	**BAC**	**% Difference**
1.5.1	Development Test & Evaluation	$170,183	$240,910	-29%
1.5.5	Test & Evaluation Support	$968,086	$497,392	95%
1.9.1	Test & Measurement Equipment	$170,801	$86,010	99%
1.9.2	Support & Handling Equipment	$106,129	$88,494	20%
1.10.2	Contractor Technical Support	$245,399	$230,504	6%

Table 3. Comparing Descriptive IEAC and BAC for Third Level WBS Elements

below 0.8, is shown starting in reporting period 22. The trend for both metrics then slightly increases in the remaining reporting periods; but neither cumulative CPI or cumulative SPI are forecasted to return to a low risk level in the remaining reporting periods shown. These results indicate preventive actions, like a root cause analysis, may be necessary by the program office.

The notional results of the cumulative CPI and cumulative SPI model forecasts, for each WBS element, were also used to calculate the IEAC values of each WBS element shown in Table 3 and Table 4. Ideally the IEAC and BAC should be close to the same value. Significantly larger IEAC values compared to the BAC typically indicate contract cost or schedule growth. A large, positive percent difference would indicate this growth occurred in the results. As shown in our notional example, the comparative values are highly variable, which may better reflect reality. At the third level of the WBS, the test and measurement equipment element, in particular, possesses an IEAC much greater than the planned BAC. Using this predictive IEAC metric may help acquisition analysts to periodically check the potential effect cumulative CPI and cumulative SPI forecasts have on the program's overall budget.

Macroeconomic Regression

DoD acquisition program issues are not limited to activities and events internal to the program. External events may also drive cost and schedule deviations from a planned baseline. In discussing the root causes of Nunn-McCurdy acquisition breaches, Blickstein et al. (2012) find that exogenous events, such as industrial base changes, can lead to critical cost and schedule growth. As an example, evolution in the commercial satellite industry is cited as a root cause for a breach in the Wideband Global Satellite (WGS) program (Blickstein et al., 2012). Arena et al. (2014) assert that external considerations can

Fourth Level WBS Elements				
WBS Element	**Name**	**IEAC**	**BAC**	**% Difference**
1.1.1.2	Hull/Frame/Body/Cab	$64,172	$27,094	137%
1.1.1.6	Vehicle Electronics	$227,919	$180,301	26%
1.1.1.13	Special Equipment	$513,538	$318,294	61%
1.8.1.3	Test & Measurement Equipment (Electronics/ Avionics)	$200,637	$105,109	91%
1.9.2.1	Support & Handling Equipment (Airframe/Hull/ Vehicle)	$92,070	$16,010	475%

Table 4. Comparing Descriptive IEAC and BAC for Fourth Level WBS Elements

impact program stability, citing overall demand for production function inputs and technology improvement as specific factors. The authors conclude that exogenous events necessitate greater monitoring in order to mitigate or avoid significant acquisition issues (Arena et al., 2014).

Economic indicators were selected as independent variables to evaluate against acquisition performance. Inflation and unemployment are two predominant measures of interest because strong anecdotal evidence exists for their influence on firm production in recent years. As one example, recent levels of inflation led to acquisition instability that manifested as issues in talent retention for government contractors, especially small firms, that ultimately threatened the defense industrial base (Overman 2022). Moreover, the U.S. economy has experienced exceptionally tight labor markets since 2020, which has led to labor recruitment challenges (Waddell and Macaluso, 2022). DoD system contractors attempting to hire staff for a new development or production contract may encounter staffing shortfalls as a result.

Empirical evidence also exists for how inflation and unemployment are measures of economic performance (Figure 9). Okun's law, for instance, highlights the importance of unemployment to measuring overall economic performance, like gross output. The law measures the negative linear association between cyclical unemployment and the output gap (Equation 6), where cyclical unemployment is equal to the actual rate of unemployment minus the natural rate of unemployment; and the output gap is equal to the actual output minus the potential output. Okun's Law implies that changes in unemployment will behave countercyclically with economic output growth, such that a decline in cyclical unemployment increases the output gap. In terms of inflation, higher inflation levels can have significant costs, such as random redistribution of wealth, relative price variations, and more uncertainty about economic planning that leads to fewer long-run

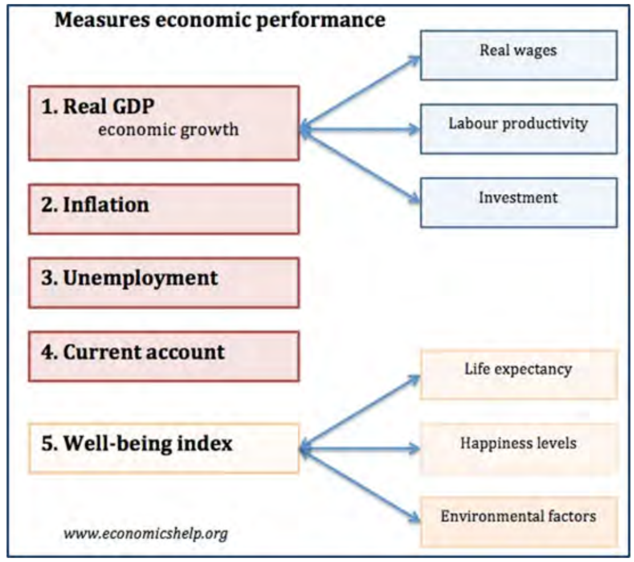


Figure 9. Measures of Economic Performance

contracts (Ball and Cecchetti, 1990). DoD contractors may find that certain portions of the supply chain become relatively more expensive, which can impact contract performance. In addition, the greater uncertainty in contracting from higher inflation can lead to a lower probability that the integrated baseline accurately reflects emergent economic realities.

Therefore, the study team statistically examined whether these factors influence acquisition program health in terms of cost and schedule performance. The following research question was posed to guide the analysis. Do economic indicators that have a significant relationship with gross output, possess comparable relationships with DoD acquisition performance?

Economic data was collected from the Bureau of Labor Statistics (BLS). Specifically, the producer price index (PPI), which measures the change in domestic producers selling prices for their output, was used as a measure for inflation. Likewise, the study team collected local area unemployment statistics (LAUS) to serve as measurements for inflation. Several normalization procedures were executed to integrate the macroeconomic data with the IPMDAR data sets and to conduct regression-

$$\Delta output\ gap = \alpha + \beta(\Delta[actual\ unemp\ rate\ - natural\ unemp\ rate]) + \epsilon$$

Equation 6. Okun's Law

$$CPI_{cumulative} = \beta_{Intercept} + \beta_1 Unemployment\_rate_{lag=t} + \varepsilon$$

Equation 7. Unemployment Regression

based analysis at varying levels, to include a producer's industry and production locations.

The IPMDAR file format specification includes fields for contractor information by WBS element. The study team collated the contractor data, such as contractor and sub-contractor name and ID. Next, the team used the Defense Logistics Agency (DLA) to identify commercial and government entity (CAGE) codes and unique entity identifiers (UEIs) for each of the contractors listed in the IPMDARs. A contractor table is subsequently produced, which includes contractor location information at the county level (e.g., Federal Information Processing Standards [FIPS] codes) based on the CAGE codes and UEIs and the firm's primary North American Industry Classification System (NAICS) codes. The NAICS code lists the specific industry a firm is associated. Ultimately, the completed contractor table includes foreign keys, such as FIPS and NAICS code, which allows the analyst to interface with both the BLS macroeconomic data and the IPMDAR cost and schedule performance tables.

Once the normalization procedures are completed, regressions are executed with CPI and SPI as the variables of interest and unemployment and inflation as the independent variables. Inflation regressions may be executed nationally and by NAICS. Unemployment regressions are executed by FIPS codes at the national, state, or county level or by NAICS codes. Both unemployment and inflation regressions are run at varying lagged rates, to include zero lag, under the assumption that some amount of time may be necessary for the macroeconomic parameters to influence contract performance. Measures of contract performance, like cumulative CPI and cumulative SPI, are calculated after rolling up the raw contract and schedule data by the level of interest (e.g., work location or industry). An example for unemployment as a single independent regressor with an intercept and error term is provided in Equation 7.

Results have been executed for several programs to date. Due to the sensitivity of the program data, notional results are displayed in Figure 10, showing a positive and statistically significant relationship between 12 month lagged unemployment at the county level for Dallas County, Texas and the cumulative contract CPI. One potential interpretation is that a "tight" labor market, exacerbated by limited labor supply with the requisite human capital to manufacture DoD systems, leads to worse acquisition performance. A tight labor market occurs when the actual unemployment rate is less than the natural rate of 4.75%, leading to a negative cyclical rate. The tight labor market portion of this graph, where unemployment is less than 4.75% aligns with Okun's law. On the contrary, the positive relationship of contract performance and higher levels of unemployment provides an inverse relationship to Okun's law. Reason exists for the

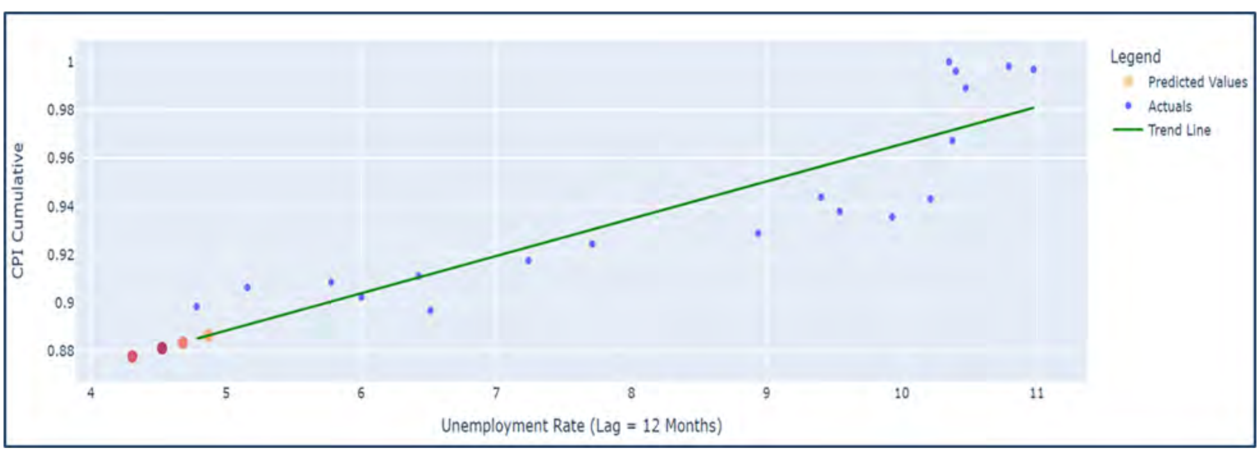


Figure 10. Unemployment and Cumulative CPI for Dallas County, TX

results to diverge, as macroeconomic performance and contract performance are unique measures. The continuous positive relationship may occur because higher levels of unemployment entail lower competition for labor supply, which enables defense contractors to hire and appropriately staff their contracts.

Overall, the specification for IPMDAR affords analysts an opportunity to explore how exogenous program factors may impact contract performance at relatively high fidelity. Given inflation projections provided by entities like the Congressional Budget Office, analysts can even forecast future CPI or SPI performance with a prediction interval based on a regression's results. Contractor information provided by WBS elements and macroeconomic information obtained from organizations -- Bureau of Labor Statistics, Bureau of Economic Analysis, and Census Bureau -- is mature, stable, and very detailed. Thus, the analyst may examine how external program factors impact their programs of interest broadly, at key geographic areas of design and manufacturing, and for specific industrial sectors. The analytic value is potentially substantial, given the current defense industrial base and supply chain considerations for DoD acquisition.

Path Forward

The study team explored the application of predictive acquisition analysis using a variety of time series analyses and regressions with lagged economic performance indicators. However, many additional techniques are available to test, in large part due to proliferation of applicable data, metrics, and data science tools. The continued expansion of the portfolio of IPMDAR data through direct support of MDAPs and coordination with OSD would provide additional examples and use cases upon which to build and refine the existing methodological approaches.

The integration and exploration of additional economic factors is a second advantageous course of action. The Census Bureau, Federal Reserve, and National Bureau of Economic Analysis are three additional data sources offering cross-sectional and longitudinal statistics on a variety of economic factors may exogenously influence DoD acquisition performance. Future efforts may test and train the time series and macroeconomic models on both a broader set of economic data and a portfolio of IPMDARs. This exploration of economic factors would help identify and verify best fitting analytics, and establish classifiers (i.e., objective and threshold values) to help inform decision makers when opportunities or risks are imminent so that preventive action can be taken.

Another analysis path might involve evaluating additional data science and statistical techniques such as panel regressions or clustering and classification techniques. Panel regressions, which combine cross-sectional and time-series data to control for unobserved dependencies and endogeneity, could be used for a variety of research purposes. In regards to panel regressions, future research efforts may address whether work performance in a specific region of the U.S. is impacting one, or many, projects over time. Moreover, this type of analysis may compare acquisition performance by service, command, program size, prime system integrator, etc. A second analysis path is the application of clustering and classification techniques such as K-nearest neighbors (K-nn) or logistic regression. K-nn is a supervised classifier that minimizes distance of items from a centroid of measurements, which can be used to classify completion risk of activities over time, based on variables of interest (e.g. budgeted work remaining, active months). A final analysis path is logistic regression: a probabilistic model of an event (e.g. issue or opportunity) taking place which can be used to correlate cost and schedule variables of interest (e.g. average CPI or task mid-point SPI) with total cost outcomes (e.g. over budget or under budget) at varying WBS levels.

Conclusions

The study was constrained by several limitations, the most notable was data accessibility, as the analysis was executed on a single, on-going program. Using data from an on-going program rather than using data from completed projects limits the ability to assess the overall level of improvement provided by time series analysis. Nonetheless, forecasts can be timely in alerting the program of directional trends in cost and schedule health relative to standard EVM. In addition, these techniques have successfully transferred to multiple acquisition programs since the time this analysis was originally constructed. Adequate sample sizes are another limitation. In order to conduct ARIMA analysis, multiple years of cost and schedule data (without re-baselines) are necessary to create a properly sized consecutive time series dataset.

Nevertheless, the methodological approaches discussed offer several advantages over current EVM practices. Time series forecasts on standard EVM data and the SEMs may be more responsive to trends in project performance relative to descriptive statistics, which may subsequently allow the analyst more time to identify and decompose potential cost and schedule risks, before they manifest into significant issues. While traditional time series analysis has historically been conducted on a portfolio of completed programs or projects, this framework is also set up to train, test, and execute analysis in real time on current projects. Moreover, conducting time series analysis in real time and at varying WBS levels may allow for improved root cause analysis. This application creates an opportunity to prevent potential cost and schedule issues from emerging. Finally, this analysis accounts for trends in external program factors, which can be a key reason for a cost or schedule breach, impacting program performance.

References

Alexander, Sean. “Earned Value Management Systems (EVMS) Basic Concepts”, Project Management Institute Washington, DC Chapter, 2002. https://apps.dtic.mil/sti/tr/pdf/ADA403490.pdf

Arena, Mark V., et al., Management Perspectives Pertaining to Root Cause Analyses of Nunn-McCurdy Breaches: Contractor Motivations and Anticipating Breaches, Vol. 6, Santa Monica, Calif.: RAND Corporation. MG-1171/6, 2014. As of January 27, 2019: https://www. rand. org/pubs/monographs/MG1171z6. html, 2014.

Anton, Philip S., William Shelton, James Ryseff, Samantha Cohen, Grant Johnson, Stephen B. Joplin, David Kravitz, Megan McKernan, and Alejandro Vigo Camargo. "Early Predictive Indicators of Contractor Performance: A Data-Analytic Approach." RAND Corporation, 15-42, 2022.

Ball, Laurence, Stephen G. Cecchetti, and Robert J. Gordon. "Inflation and uncertainty at short and long horizons." Brookings Papers on Economic Activity 1990.1 (1990): 215-254.

Blickstein, Irv, et al. "Methodologies in analyzing the root causes of Nunn-McCurdy breaches." (2012).

Cryer, Jonathan D., and Kung-Sik Chan. Time series regression models. Time series analysis: with applications in R: 249-276, 2008.

Data Item Description (DID) for Integrated Program Management Data and Analysis Report (IPMDAR), Number: DI-MGMT-81861C, Approval Date: 20210830.

Department of Defense, Distribution Statements on DoD Technical Information., DoD Instruction 5230.24, Office of the Under Secretary of Defense for Research and Engineering, January 10, 2023.

Department of Defense, Department of Defense Standard Practice Work Breakdown Structures for Defense Materiel Items, MIL-STD-881F, May 13, 2022.

Dowling, Austin W. Using Predictive Analytics to Detect Major Problems in Department of Defense Acquisition Programs, Department of the Air Force Air University, Air Force Institute of Technology, March 2012.

Integrated Program Management Data Analysis Report (IPMDAR) Implementation & Tailoring Guide, Department of Defense (DoD), 24 August 2021.

IPMDAR Contract Performance Dataset, Version 1.0, File Format Specification, March 12, 2020

IPMDAR Schedule Performance Dataset, Version 1.0, File Format Specification, March 12, 2020

Keaton, C. G., White, E. D., & Unger, E. J., “Using earned value data to detect potential problems in acquisition contracts”. Journal of Cost Analysis and Parametrics, 4(2), 148-159, 2011.

Kwiatkowski, Denis, et al. “Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?." Journal of econometrics 54.1-3 (1992): 159-178, 1992.

Meier, Steven R. "Causal inferences on the cost overruns and schedule delays of large-scale US federal defense and intelligence acquisition programs." Project Management Journal 41.1 (2010): 28-39.

Montgomery, Douglas C., Cheryl L. Jennings, and Murat Kulahci. Introduction to time series analysis and forecasting. John Wiley & Sons, 2015.

National Reconnaissance Office (NRO), “Advanced Schedule Analysis Techniques”, PowerPoint presentation, May 2021.

National Reconnaissance Office (NRO), “Predictive Analytics: Schedule Execution Metrics”, PowerPoint presentation, September 2022. https://ndiastorage.blob.core.usgovcloudapi.net/ndia/2022/ipmdFall/Bembers.pdf

Overman, T. Defense authorization act provides inflationary relief for government contractors. Reuters. December 20, 2022. https://www.reuters.com/legal/legalindustry/defense-authorization-act-provides-inflationary-relief-government-contractors-2022-12-20/

Python API reference for plotly. https://plotly.com/python-api-reference/. Accessed May 2023.

Schultheis, Candice, Ryan Carey, Robert Dunn. “An Analysis of Schedule Execution Metrics (SEMs) Developed by the National Reconnaissance Office (NRO)”, Naval Post Graduate School Thesis, March 2021.

Waddell, S. R., & Macaluso, C. (2022, September). Changing recruiting practices and methods in the tight labor market. Federal Reserve Bank of Richmond. https://www.richmondfed.org/publications/research/economic_brief/2022/eb_22-36

Anna B. Peters is a cost analyst and data scientist at Johns Hopkins University Applied Physics Lab, contributing to a team that specializes in affordability analysis and assessments for a variety of Department of War sponsors. Anna earned a M.S. in Applied Statistics from Bowling Green State University and a B.S. in Industrial Math and Statistics from West Virginia University. Anna holds a PCEA certification from ICEAA.

Mark W. Hodgins is a member of the Principal Professional Staff at JHU APL, where he leads investment decision analyses for the DOD. Mark specializes in defense economics, including cost imposition, economic statecraft, and comparative institutional assessments of military production. He earned a Ph.D. in Economics from George Mason University, a graduate certificate in Business Analytics from Indiana University Bloomington, and a M.A./B.A. in Economics from Johns Hopkins University. He holds certifications as a CCEA from ICEAA and a PMP from PMI.

The authors would like to acknowledge Adam J. Levin for his valuable assistance in the preparation of the figures and tables presented in this analysis.

Perception vs Reality: Determinants of AFIT Graduate Education Pursuit in the Financial Management Officer Corps

Capt Benjamin L. Ainsworth

Dr. Jonathan D. Ritschel

Dr. Robert D. Fass

Major Michael J. Brown

This study analyzes two surveys, one administered to Company Grade Officers (CGOs) and one to Field Grade Officers (FGOs), to examine factors shaping officers' intention to apply to the Graduate Cost Analysis (GCA) master's degree program at the Air Force Institute of Technology (AFIT). Structural Equation Modeling (SEM) suggests that leadership support is associated with intent to apply primarily through positive perceptions of career progression for AFIT graduates. No statistically significant barriers emerged in the survey data, though open-ended responses frequently referenced concerns about the Active Duty Service Commitment (ADSC). Survey evidence further shows stronger interest for AFIT attendance among early-career officers, suggesting that targeted recruiting should focus on second lieutenants.

Introduction

The Financial Management (FM) career path for Air Force FM officers continues to evolve, offering diverse opportunities in base-level and staff budget positions, acquisitions budgeting, and cost analysis. Some FM officers follow an Operation and Maintenance (O&M) budget-focused career path, gaining base-level budget and financial operations experience before advancing to Major Command (MAJCOM) staff or Air Staff roles to manage budgets at higher levels. This O&M budget-centric path aligns with the standard career roadmap outlined on the myFMHub website (United States Air Force, 2024). Other FM officers pursue a blended career path, transitioning from traditional budget roles to acquisitions and cost analysis positions. The Blended FM Career Path Roadmap, also available on myFMHub, details this alternative path.

To support the blended career path, formal education in cost analysis plays a foundational role. A significant portion of cost analysis officers begin their careers by completing the Cost Analysis master's degree at the Air Force Institute of Technology (AFIT). Established in 1982, the program has remained a cornerstone of FM officer development in cost analysts for over forty years. The program is fully accredited and provides a rigorous academic foundation in quantitative methods and military cost-related topics (Air Force Institute of Technology, 2024). Graduates go on to serve in key acquisition and cost-related roles across the Air Force, including positions within Systems Program Offices (SPOs), MAJCOM staffs, and the Air Force Cost Analysis Agency (AFCAA).

The importance of formally trained cost analysts has been well documented. A RAND Corporation study found that many organic cost estimators historically relied on on-the-job training, while AFIT graduates were recognized for their strong analytical capabilities and required less time to become proficient in cost-estimating roles (Vernez and Massey, 2009). More recently, acquisition reform efforts under the Warfighting Acquisition System (WAS) have emphasized the need for improved cost estimation, noting that reliable and credible estimates at program initiation are critical to improving acquisition outcomes (Secretary of War, 2025). Despite this renewed emphasis, and the opportunity for fully funded, full-time graduate education, officers must still decide whether pursuing AFIT aligns with their career goals.

Attending AFIT requires officers to forego operational assignments for an extended period, raising questions about potential impacts on career progression, including performance evaluations, promotion timing, and command opportunities. As a result, some officers perceive participation in AFIT as a potential risk to their career trajectory. Notably, the data does not support this perception, as a 2026 study found that career outcomes as measured by Development Team (DT) scores, command opportunities, and promotion timing are statistically equivalent for AFIT and non-AFIT FM officers (Ainsworth et al., 2026). Despite these results, the perception persists and may influence whether officers choose to pursue graduate education, regardless of actual career outcomes.

This study examines the factors shaping officers' decisions to apply to the AFIT Graduate Cost Analysis (GCA) master's program using survey data from Company Grade Officers (CGOs) and Field Grade Officers (FGOs). Specifically, this research examines the factors shaping officers' decisions to apply to AFIT, including perceptions of career impact, leadership influences, and other considerations associated with pursuing graduate education.

Theoretical Framework and Decision-Making in Graduate Education

Human capital theory provides a foundation for understanding decisions related to education. At its core, human capital refers to the knowledge, skills, and experience individuals accumulate that enhance their productivity and value in the workforce. Investments in education and training are therefore theorized to generate future returns, both for the individual through improved career outcomes and for organizations through increased capability and performance (Blundell et al, 1999). While this framework explains the economic value of education, it does not fully capture how individuals decide whether to pursue additional education, particularly in environments where perceived career risks and organizational influences may shape those decisions and the monetary incentive (e.g. pay, benefits, etc.) does not change.

Within defense acquisitions, the role of human capital has been examined in the context of program performance. Prior research has found that higher levels of education within acquisition teams are associated with improved outcomes, including fewer cost overruns in some cases (Feuring, 2007; Gray, 2009). However, more recent work suggests that team composition alone may not fully explain variation in acquisition outcomes, indicating that other factors may influence performance (Smith et al., 2025).

While human capital theory provides a foundation for understanding the value of education, it does not fully explain how individuals make decisions regarding whether to pursue additional education. To better understand this decision-making process, this study draws on Ajzen's (1991) Theory of Planned Behavior (TPB), a widely used framework in behavioral research. The TPB posits that an individual's intention to engage in a behavior is shaped by three key factors: attitudes toward the behavior, perceived social norms, and perceived behavioral control.

Applied to this study, attitudes toward behavior reflect officers' perceptions of how participation in AFIT may influence their career outcomes. Perceived social norms capture the influence of leadership and organizational expectations on an officer's decision to apply. Perceived behavioral control reflects the extent to which officers believe they are able to pursue graduate education given constraints such as service commitments or other barriers.

Ajzen's (1991) framework provides a structured approach for examining how these factors jointly influence behavioral intent. Because these constructs can be measured through survey instruments, the TPB is particularly well suited for analyzing officers' decisions to pursue graduate education. To evaluate the factors shaping FM officers' intention

Table 1. Constructs and Hypothesized Relationships to Intent to Apply to AFIT

Constructs	Hypothesized Relationship to DV	TPB - Attitude	TPB - Norms	TPB - Behavioral Control	Human Capital
Intent to Attend (Apply) AFIT	N/A				X
ADSC Requirement	-	X			
Supervisor Support	+		X		
Leadership Support	+		X		
Career Progression	+	X			X
GRE Requirement	-	X			
Intent to Stay (AF)	+	X			
Thesis Requirement	-	X			
PartTime vs FullTime GradEd	+	X			
Blended Career	-	X			
Location Perception	-	X			
AFIT Academic Student Quality	+			X	

Notes: "+" indicates a positive hypothesized relationship (variables move together), "-" indicates a negative hypothesized relationship (variables move in opposite directions), * Captures the respondent's confidence that he/she would be accepted to graduate school (based on grades and potential test scores)

to apply to the AFIT Cost Analysis master's program, this study utilized a survey instrument specifically designed around these established theoretical frameworks. Table 1 maps each resulting construct to its corresponding dimension within the TPB and Human Capital theories along with its hypothesized relationship to the dependent variable.

Data and Methods

The dataset used in this study is derived from two surveys developed by the research team: one administered to CGOs and one to FGOs. The CGO population includes officers in grades O-1 through O-3, while the FGO population includes officers in grades O-4 through O-6.

Prior to distribution, we obtained an exemption from the Institutional Review Board (IRB) process at AFIT, as no Personally Identifiable Information (PII) was collected. The surveys were distributed via the Qualtrics platform by the Office of the Deputy Assistant Secretary for Budget. The CGO survey was sent to 406 officers, and the FGO survey to 249 officers, with a response window from 19 May 2025 to 13 June 2025. A total of 101 usable CGO responses and 67 usable FGO responses were received, corresponding to response rates of approximately 25 percent and 27 percent, respectively.

Both surveys were designed to capture officer perceptions related to graduate education and career decision-making. Survey items assessed perceptions of leadership and supervisor support, expected career impacts of attending AFIT, and potential barriers to participation, including the Active Duty Service Commitment (ADSC), GRE requirement, thesis requirement, and geographic location. Responses were measured using a seven-point Likert scale. Each survey also included open-ended response sections and demographic questions. The primary dependent variable in this study is an officer's reported intention to apply to the AFIT Cost Analysis master's program. For FGOs, this measure reflects retrospective assessments of whether they would have pursued AFIT during their CGO years.

Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) is used to validate the measurement model underlying the survey constructs. CFA evaluates whether observed survey items load onto their intended latent factors, such as leadership support, supervisor support, perceived career impact, and barriers to pursuing AFIT. Standardized factor loadings are used to assess the strength of the relationship between observed items and their corresponding constructs. Consistent with prior research, loadings above 0.50 are considered acceptable indicators of construct validity, and items that do not meet this threshold are evaluated and removed if necessary to improve measurement quality (Hair et al, 1992; Cheung et al., 2024).

Cronbach's Alpha

Internal consistency of the survey constructs is assessed using Cronbach's alpha. This measure evaluates the extent to which items within a construct reliably capture the same underlying concept. Following established guidance, a coefficient of 0.70 or greater is considered indicative of acceptable reliability (Nunnally, 1978). Constructs falling below this threshold are reviewed to identify and address problematic items, including ensuring proper scoring of reverse -coded questions.

Correlation Matrix and Mediation Testing

After confirming acceptable factor loadings and internal consistency, the analysis evaluates the correlations between the constructs, the dependent variable (Intention to Apply to AFIT), and each other. These relationships form the foundation for subsequent Structural Equation Modeling (SEM) development. Constructs that significantly correlate at the $p < 0.05$ level with the dependent variable are retained for SEM testing, while constructs that significantly correlate with one another and the dependent variable are evaluated for potential full or partial mediation. This step identifies theoretically grounded pathways that guide the structural model.

Structural Equation Modeling

Structural Equation Modeling is used to estimate the relationships among the validated constructs. SEM simultaneously evaluates both the measurement and structural components of the model, enabling assessment of multivariate associations among leadership support, supervisor support, perceived career impact, and barriers to officers' intent to apply to AFIT. Model fit is evaluated using standard global fit indices, including the chi-square statistic (χ^2), Comparative Fit Index (CFI), Tucker–Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA). Prior research suggests that CFI and TLI values of approximately 0.90 indicate acceptable fit, while values of 0.95 or higher indicate good fit, and RMSEA values below 0.08 reflect reasonable model fit (Hu & Bentler, 1999; Hooper, Coughlan, & Mullen, 2008).

Sentiment Analysis and Latent Dirichlet Allocation Topic Modeling

Qualitative responses from the CGO and FGO surveys are analyzed using both sentiment analysis and Latent Dirichlet Allocation (LDA) topic modeling. Sentiment analysis is conducted using the VADER (Valence Aware Dictionary and sEntiment Reasoner) text-mining approach, which classifies responses as positive, neutral, or negative. LDA is then applied using the Python scikit-learn (sklearn) library to identify underlying themes in open-ended responses by grouping frequently co-occurring terms (Blei, Ng, & Jordan, 2003). Together, these methods provide additional insight into officer perceptions and complement the quantitative findings.

FGO Analysis

Due to the retrospective nature of the FGO survey, the FGO analysis focuses primarily on descriptive patterns and qualitative insights rather than full structural modeling. Sentiment analysis and topic modeling are also applied to FGO responses to identify recurring themes.

Results

Prior to model estimation, descriptive analysis was conducted to examine key patterns in the CGO survey data. Among the 101 usable CGO responses, defined as respondents who answered at least one Likert-scale item associated with the survey constructs, responses regarding the intention to apply to AFIT cluster near the midpoint of the scale (mean = 3.90) indicating mixed views toward pursuing the program. Second lieutenants report the highest intent to apply to AFIT (mean = 4.93), followed by first lieutenants (mean = 3.91), while captains report the lowest interest (mean = 3.63). The survey uses a 7-point Likert scale, where 4 represents a neutral response; values below 4 indicate lower intent, and values above 4 indicate greater intent to apply AFIT. These results suggest that interest in AFIT is strongest early in an officer's career and declines with time. This pattern indicates that FM leadership may benefit from prioritizing AFIT recruiting efforts toward second lieutenants within their first one to two years of commissioned service.

Confirmatory Factor Analysis-Results

The full modeling approach is applied to the CGO survey data. CFA is conducted to evaluate the measurement model and assess the alignment of survey items with their intended latent constructs. Initial results indicate that four items exhibit standardized loadings below the 0.50 threshold. Two of these items are associated with the Leadership Support construct, one with Career Progression, and one with the Blended Career Path construct. These items were reviewed for conceptual alignment and measurement validity, and all four were removed from the model. The removed items primarily reflect inconsistencies in question framing, such as reverse-coded wording or differences in evaluative direction relative to other items within the same construct. Following item removal, the CFA model was re-estimated. The revised measurement model demonstrated strong convergent validity and internal consistency across the remaining constructs, indicating that the retained items provide a reliable representation of the underlying latent factors.

Table 2. CGO Survey Cronbach's Alpha

Construct	Items	Alpha
Intention to Apply to AFIT	5	0.963
ADSC Requirement	4	0.822
Supervisor_Support	5	0.897
Leadership Support	4	0.909
Career Progression	5	0.915
GRE Requirement	2	0.844
Intent to Stay	5	0.919
Thesis Requirement	5	0.943
PartTime vs FullTime GradEd	5	0.935
Blended Career	4	0.922
Location Perception	5	0.927
AFIT Academic Student Quality	5	0.924

Cronbach's Alpha-Results

Following the CFA, the final Cronbach's alpha results presented in Table 2 indicate strong reliability across all constructs in the CGO survey dataset.

These results reflect the refined measurement model after removing the four survey items that demonstrated weak standardized loadings (< 0.50) during the CFA process to improve overall model fit. Every construct exceeds the 0.70 threshold for acceptable internal consistency, with most falling in the good (0.80–0.89) or excellent (≥ 0.90) range.

Correlation Matrix-Results

Following the internal consistency analysis, a final correlation matrix was generated that reports Cronbach's alpha values along the diagonal and inter-construct Pearson correlations in the lower triangle. See Table 3.

Table 3. CGO Survey Correlation Matrix

Variables	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12
1. Likelihood of Attending AFIT	3.90	2.05	(0.96)											
2. ADSC Requirement	4.58	1.54	-0.079	(0.82)										
3. Supervisor Support	5.32	1.26	0.149	0.106	(0.90)									
4. Leadership Support	5.28	1.34	0.211 *	-0.072	0.449 ***	(0.91)								
5. Career Progression	3.99	1.51	0.601 ***	-0.063	0.180	0.537 ***	(0.92)							
6. GRE Requirement	3.59	1.70	0.124	-0.075	-0.164	-0.132	0.022	(0.84)						
7. Intent to Stay	5.27	1.57	0.084	-0.477 ***	0.141	0.187	0.115	0.147	(0.92)					
8. Thesis Requirement	3.72	1.71	-0.004	0.088	-0.226 *	-0.107	-0.038	0.439 ***	0.173	(0.94)				
9. PartTime vs FullTime GradEd	4.97	1.70	0.286 **	-0.071	-0.034	0.070	0.195	-0.119	0.002	-0.016	(0.94)			
10. Blended Career	3.27	1.41	-0.221 *	-0.079	-0.233 *	-0.493 ***	-0.372 ***	-0.047	-0.167	0.036	-0.044	(0.92)		
11. Location Perception	3.91	1.73	-0.211 *	0.196	-0.036	-0.326 ***	-0.257 *	0.145	-0.063	0.194	-0.038	0.088	(0.93)	
12. AFIT Academic Student Quality	5.98	1.08	-0.094	0.078	0.113	0.008	-0.136	-0.148	-0.017	-0.057	-0.308 **	0.048	0.014	(0.92)

Note: Chronbach's Alphas appear across the diagonal in parentheses.
* p-value < .05, ** p-value < .01, *** p-value <.001

Mediation Testing-Results

Only constructs that show statistically significant correlations with the dependent variable, Intention to Apply to AFIT, are retained for SEM consideration. These include Leadership Support, Career Progression, Part-Time vs. Full-Time Graduate Education, Blended Career, and Location Perception. From this set, we then assess which constructs are also significantly correlated with one another to identify theoretically and statistically plausible mediation pathways. This process yields five potential mediation chains:

1. Leadership Support → Career Progression → Intention to Apply to AFIT
2. Blended Career → Career Progression → Intention to Apply to AFIT
3. Location Perception → Career Progression → Intention to Apply to AFIT
4. Part-Time vs. Full-Time GradEd → Career Progression → Intention to Apply to AFIT
5. Leadership Support → Location Perception → Intention to Apply to AFIT

Of these, four mediation pathways show significant indirect effects, indicating meaningful mediation, while one lacks statistical support and is excluded from further consideration. This systematic, theory-driven, and statistically validated process ensures that only robust and interpretable mediation relationships carry forward to the SEM stage. Below are the four significant mediation pathways.

1. Leadership Support → Career Progression → Intention to Apply to AFIT (Full Mediation)
2. Blended Career → Career Progression → Intention to Apply to AFIT (Full Mediation)
3. Location Perception → Career Progression → Intention to Apply to AFIT (Full Mediation)
4. Part-Time vs. Full-Time GradEd → Career Progression → Intention to Apply to AFIT (Partial Mediation)

Structural Equation Modeling-Results

The final structural model includes one mediator, Career Progression, and four predictors that demonstrated statistically significant indirect effects during mediation testing. Three predictors, Leadership Support, Blended Career, and Location Perception, exhibit full mediation through Career Perception; therefore, their direct paths to Intention to Apply to AFIT are removed from the final model. Part-Time vs. Full-Time Graduate Education demonstrates partial mediation, so both its mediated pathway and its direct effect are retained. As a result, the only direct predictor of Intention to Apply to AFIT in the final model is Part-Time vs. Full-Time Graduate Education, consistent with its statistical significance in both the correlation matrix and the mediation analysis. This structure produces a theoretically grounded and parsimonious SEM that includes only empirically supported structural paths. Figure 1 displays the final SEM. [Note: The H1(+), H2(+) etc. notation indicates a hypothesis test and the direction, positive or negative, of the hypothesis test.]

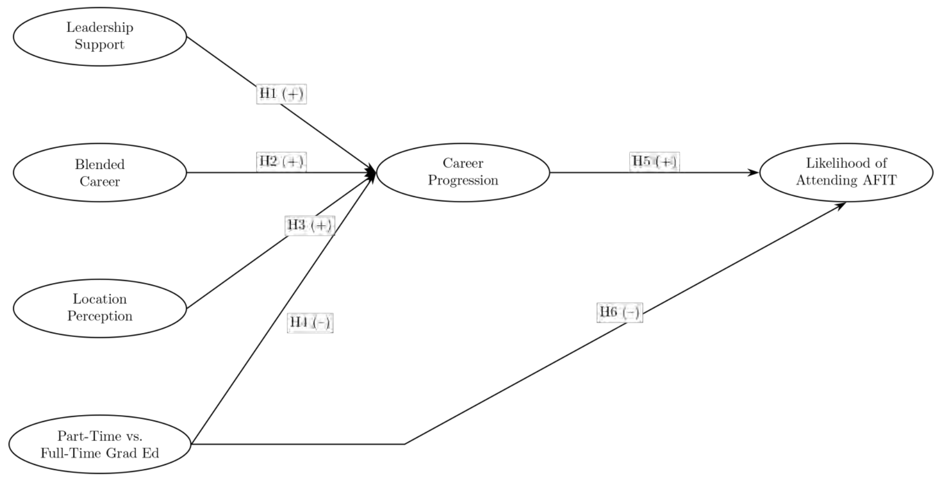


Figure 1. CGO SEM

Not displayed in Figure 1 is a binary control variable indicating whether CGOs already hold a master's degree. More than half of respondents, particularly among captains, report holding a graduate degree, which may influence responses to the dependent variable. This variable is therefore included in the SEM as a control to ensure that the observed relationships between the constructs and the intent of attending AFIT are not confounded by prior graduate education.

When interpreting results from the SEM analysis, it is important to note that respondents exhibiting extreme response patterns on the dependent variable, Intention to Apply to AFIT, are excluded. Specifically, 20 respondents selected "strongly disagree" across all items, while 11 selected "strongly agree" across all items, representing extreme responses with no variation on each end of the boundary. These groups represent individuals whose intent to attend is fixed regardless of leadership support, career perceptions, or perceived barriers. These individuals are not the object of interest in this study. Rather our interest is understanding what factors are associated with intent to apply in the marginal individual who has not yet made up his or her mind.

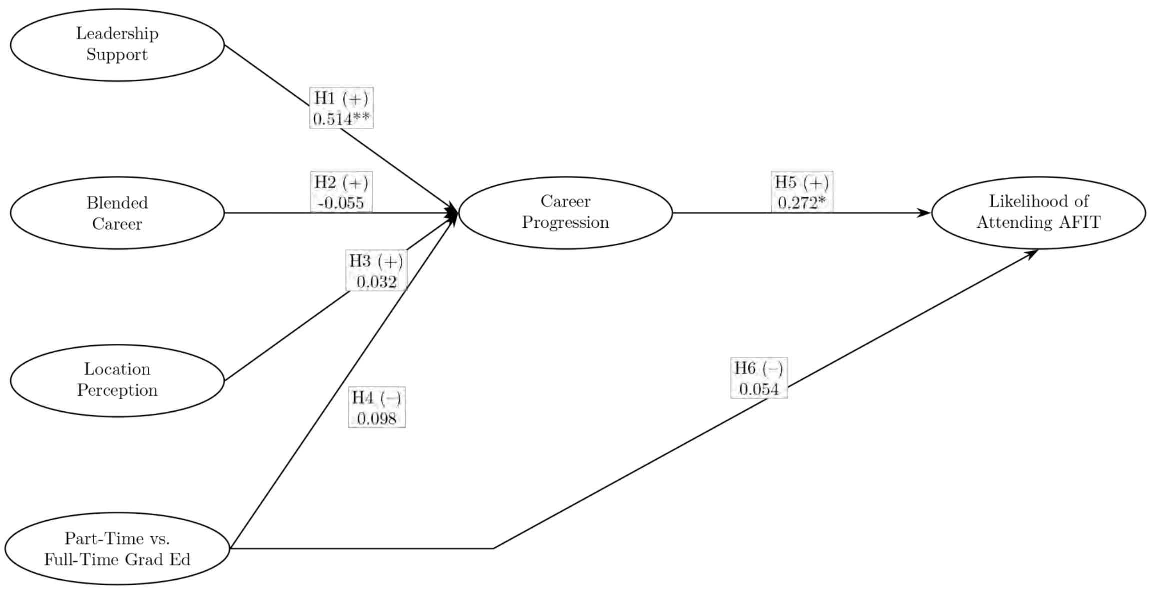
Additionally, 13 respondents exhibit substantial missingness across key constructs and are excluded from the SEM analysis. After applying these screening criteria, the final SEM sample consists of 57 CGO respondents, while all responses are retained for descriptive and measurement analyses.

The final SEM model shown in Figure 2 identifies two statistically significant relationships. The path from Leadership Support to Career Progression is statistically significant at the 0.05 level ($p = 0.02$), indicating that stronger perceived leadership support is associated with more favorable perceptions of career progression. The path from Career Progression to Intention to Apply to AFIT is significant at the 0.10 level ($p = 0.06$), suggesting that officers who perceive AFIT as beneficial to their career are more likely to express intent to apply. These results indicate that leadership support is associated with AFIT interest indirectly through its effect on career progression perceptions, rather than through a direct pathway.

Table 4 reports the global fit indices for the final SEM.

Table 4. CGO SEM Global Fit Indices

Global Fit Indices	
chi2	541.29
DoF	362
CFI	0.86
TLI	0.85
RMSEA	0.09
chi2/df	1.5



*Note: Standardized coefficients reported. * $p < .10$, ** $p < .05$.*

Figure 2. CGO Significant SEM Results

The CFI (0.86), TLI (0.85), and RMSEA (0.09) fall slightly outside conventional benchmarks. However, these results are consistent with the relatively small sample size (n = 57) and the number of estimated parameters driven by the 29 observed indicators in the model. SEM fit indices are known to exhibit a conservative bias under such conditions, particularly with samples below 100. Despite these limitations, the model demonstrates strong standardized factor loadings, high Cronbach's alpha values well above accepted thresholds, and a statistically coherent mediation pathway consistent with theory. Taken together, the measurement and structural evidence indicate that the model is stable and theoretically sound, supporting confidence in the findings despite the modest global fit statistics.

Qualitative Insights

Open-ended responses from the CGO survey provide additional context to the quantitative findings and reinforce key patterns identified in the SEM. Sentiment analysis indicates generally positive attitudes toward AFIT's educational value, suggesting that officers recognize the program as a meaningful professional development opportunity.

Topic modeling further reveals differences across career stages. Junior officers tend to view AFIT as a long-term investment in their career, whereas more senior captains, many of whom have already completed graduate education, express lower levels of interest. This pattern aligns with the descriptive findings showing higher intent to apply among second lieutenants relative to more senior CGOs.

While administrative factors such as the GRE and the ADSC requirement appear in the CGO open-ended responses, these factors do not emerge as statistically significant predictors of application intent in the SEM. This contrast suggests that, although barriers are salient in officer perceptions, they do not ultimately drive decision-making once broader career considerations are taken into account.

FGO responses provide additional context and largely reinforce the patterns observed in the CGO survey. Open-ended responses consistently emphasize the importance of timing when pursuing AFIT, with officers noting that early-career participation allows sufficient time to complete follow-on assignments and remain competitive for future leadership opportunities.

FGOs also highlight the role of leadership messaging, indicating that support from supervisors and senior leaders influences how officers perceive the value of AFIT. While most respondents view AFIT as beneficial for developing analytical and data-driven skills, some note that a lingering perception exists within portions of the FM community that participation in AFIT may hinder career progression.

Despite these concerns, respondents broadly recognize the long-term professional value of AFIT, particularly in developing technical expertise relevant to acquisition and cost analysis roles. These findings further support the conclusion that perceptions of career impact, shaped by leadership signaling, play a central role in officers' decisions to pursue graduate education.

Discussion and Conclusion

The importance of formally trained cost analysts has been well documented, and the AFIT Cost Analysis master's program provides officers with the analytical foundation necessary to contribute effectively to the defense acquisition mission. Despite these benefits, some officers have historically expressed reservations about pursuing AFIT, particularly due to concerns that stepping away from operational assignments may negatively impact career progression.

This study finds that CGO responses regarding the intention to apply to AFIT center near the midpoint of the scale (mean = 3.90), indicating mixed views toward pursuing the program. A closer examination shows that second lieutenants report the highest intention to apply to AFIT (mean = 4.93), while captains report the lowest interest (mean = 3.63). Qualitative findings reinforce this pattern. Sentiment analysis indicates generally positive attitudes toward AFIT's educational value, while topic modeling reveals that officers tend to view AFIT as a long-term investment in their careers, whereas more senior captains, many of whom have already completed graduate education, express lower levels of interest.

FGO responses provide additional context and largely reinforce these findings. Open-ended responses emphasize the importance of timing, with officers noting that early-career participation allows sufficient time to complete follow-on assignments and remain competitive for future leadership opportunities. While most respondents recognize the program's value in developing analytical and data-driven skills, some note that a lingering perception remains within portions of the FM community that AFIT participation may hinder career progression, despite empirical data to the contrary (Ainsworth et al., 2026).

Results from the SEM suggest that leadership support likely plays a central role in shaping officers' decisions to pursue AFIT. Specifically, leadership support is associated with perceptions of career progression, which in turn increases officers' intent of applying to AFIT. These findings suggest that leadership operates primarily through subjective norms, while perceptions of career progression reflect officers' attitudes toward the program, both of which are consistent with the TPB. In contrast, structural barriers such as the ADSC, GRE requirement, and thesis requirement are not associated with application intent.

Taken together, these results indicate that officers' decisions to pursue AFIT are associated with perceptions of career value rather than with structural constraints. While barriers are mentioned in open-ended responses, they do not appear to meaningfully influence decision-making once broader career considerations are taken into account. This distinction highlights a gap between perceived and actual drivers of behavior and underscores the importance of leadership signaling in shaping officer perceptions.

These findings carry several practical implications. First, leadership messaging represents the most influential lever for increasing AFIT participation. Consistent and visible endorsement from senior leaders may help shape perceptions of AFIT as a valued developmental pathway. Second, recruiting efforts may benefit from targeting officers early in their careers, particularly second lieutenants, where interest in AFIT is highest and program timing aligns most effectively with career development. Third, while structural barriers are not associated with application intent, they remain salient in officer perceptions and may still influence how the program is viewed within the broader FM community.

This study is subject to several limitations. First, the SEM analysis is based on a relatively small sample size (n = 57), which may affect model fit and generalizability. Second, sociological demographics such as gender were not surveyed, introducing potential endogeneity into the SEM regressors. Third, only one measure of behavioral control was included in the survey questions. This measure was not significant in the final SEM. Fourth, because this study relies on cross-sectional data, we cannot establish temporal precedence, demonstrate directional causality, or rule out equivalent alternative mediation models. Future research utilizing longitudinal designs is necessary to validate these directional pathways and mitigate potential common method variance. Lastly, the use of self-reported survey data introduces the potential for response bias, and the cross-sectional design limits the ability to make causal inferences over time.

Future research could expand beyond the active-duty population to include FM officers who have separated or retired. Examining retention and separation outcomes would provide additional insight into how AFIT participation and acquisition experience influence long-term career trajectories, including transitions to private-sector roles within the defense acquisition workforce.

Overall, this study finds that AFIT participation decisions are shaped more by how officers perceive the program's impact on their careers than by the structural features of the program itself, such as thesis requirements or acceptance criteria. Strengthening leadership messaging and aligning program timing with early-career development may therefore represent the most effective strategies for increasing participation in the AFIT Cost Analysis program.

References

Ainsworth, B.L., Ritschel, J.D., Brown, M.J., Fass, R.D., & White, E.D. (forthcoming 2026). Debunking the Myth: Acquisition Experience and Career Outcomes in the Air Force Financial Management Community, *Defense Acquisition Research Journal*

Air Force Institute of Technology, (2024). *Graduate School Academic Catalog 2024–2025*, "Cost Analysis (M.S.) Program Description," Department of Systems Engineering and Management, https://www.afit.edu/ENV

Ajzen, I. (1991). The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes*, 50 (2): 179–211.

Ajzen, I. (2001). Nature and Operation of Attitudes. *Annual Review of Psychology*, 52(1): 27–58, https://doi.org/10.1146/annurev.psych.52.1.27

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research,* 3(Jan): 993–1022, https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Blundell, R., Dearden, L., Meghir, C., & Sianesi, B. (1999). Human Capital Investment: The Returns from Education and Training to the Individual, the Firm and the Economy. *Fiscal Studies*, 20(1): 1–23.

Cheung, G.W., Cooper-Thomas, H.D., Lau, R.S., & Wang, L.C. (2024). Reporting Reliability, Convergent and Discriminant Validity with Structural Equation Modeling: A Review and Best-Practice Recommendations. *Asia Pacific Journal of Management,* 41(4): 745–783, https://doi.org/10.1007/s10490-023-09871-y

Feuring, J.C. (2007). *The Impact of Human Capital on the Cost of Air Force Acquisition Programs* (master's thesis, Air Force Institute of Technology, https://scholar.afit.edu/etd/3011.

Gray, S.E. (2009). *The Correlation of Human Capital on Costs of Air Force Acquisition Programs* (master's thesis, Air Force Institute of Technology), https://scholar.afit.edu/etd/2579

Hair, J.F., Anderson, R.E., Tatham, R.L., & Black, W.C. (1992). *Multivariate Data Analysis with Readings*, 3rd ed. New York: Macmillan.

Hooper, D., Coughlan, J., & Mullen, M.R. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. *Electronic Journal of Business Research Methods,* 6(1): 53–60

Hu, L., & Bentler, P.M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling: A Multidisciplinary Journal* 6(1): 1–55, https://doi.org/10.1080/10705519909540118

Nunnally, J.C. (1978). *Psychometric Theory*, 2nd ed. New York: McGraw-Hill.

Secretary of War (2025). *Transforming the Defense Acquisition System into the Warfighting Acquisition System to Accelerate Fielding of Urgently Needed Capabilities to Our Warriors*. Washington, DC: Department of War.

Smith, D.K., Ritschel, J.D., Lucas, B.M., White, E.D., & Valentine, S.M. (2025). Human Capital Impacts in Military Acquisitions. *Journal of Cost Analysis and Parametrics,* 12(1): 33–45.

United States Air Force, (2024). *Officer Force Development Roadmap*, Available at https://www.myfmhub.hq.af.mil/assets/admin/pages/media/works/Officer_FDR.pdf

Vernez G., & Massey, H.G. (2009). *The Acquisition Cost-Estimating Workforce: Census and Characteristics,* Santa Monica, CA: RAND Corporation.

*Capt **Benjamin L. Ainsworth**, USAF, is a recent graduate of the Air Force Institute of Technology (AFIT), where he earned a Master of Science in Cost Analysis. He is currently a cost analyst for the Air Force Cost Analysis Agency.*

*Dr. **Jonathan D. Ritschel** is a Professor of Cost Analysis in the Department of Systems Engineering and Management at AFIT. He received his BBA in Accountancy from the University of Notre Dame, MS in Cost Analysis from AFIT, and PhD in Economics from George Mason University. His research interests include public choice, defense cost economics, cost analysis and economic institutional analysis.*

*Dr. **Robert D. Fass** is an Assistant Professor in the Department of Systems Engineering and Management at AFIT. He holds a BA in Economics, an MBA from the University of New Mexico, and a PhD in Business Administration and Management from New Mexico State University. Dr. Fass's research interests include cost analysis, decision analysis, risk analysis, behavioral economics, organizational behavior, organizational change, and government acquisition policy.*

*Major **Michael J. Brown** is an Assistant Professor of Cost Analysis in the Department of Systems Engineering and Management at AFIT. He holds a BS in Economics from the United States Air Force Academy, a MS in Cost Analysis from AFIT, and a MA and PhD in Economics from the University of Tennessee. Major Brown's research interests include recruitment and retention, econometric analysis, and governmental acquisitions.*

The International Cost Estimating and Analysis Association is a 501(c)(6) international non-profit organization dedicated to advancing, encouraging, promoting and enhancing the profession of cost estimating and analysis, through the use of parametrics and other data-driven techniques.

www.iceaaonline.com

Submissions:

Prior to writing or sending your manuscripts to us, please reference the JCAP submission guidelines found at

www.iceaaonline.com/publications/jcap-submission

Kindly send your submissions and/or any correspondence to:
JCAP.editor@gmail.com

International Cost Estimating & Analysis Association

8280 Willow Oaks Corporate Drive, Suite 600
Fairfax, VA 22031

703-642-3090 | iceaa@iceaaonline.org