3-2007

Predicting the Effects of Longitudinal Variables on Cost Schedule Performance

James D. Foreman

PREDICTING THE EFFECT OF LONGITUDINAL VARIABLES

ON COST AND SCHEDULE PERFORMANCE

THESIS

James D. Foreman, Major, USAF

AFIT/GIR/ENC/07M-01

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GIR/ENC/07M-01

AFIT/GIR/ENC/07M-01

PREDICTING THE EFFECT OF LONGITUDINAL VARIABLES ON COST AND SCHEDULE PERFORMANCE

THESIS

Presented to the Faculty

Department of Mathematics and Statistics

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Information Resource Management

James D. Foreman, BS, MA

Major, USAF

March 2007

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AFIT/GIR/ENC/07M-01

PREDICTING THE EFFECT OF LONGITUDINAL VARIABLES

ON COST AND SCHEDULE PERFORMANCE

James D. Foreman, BS, MA

Major, USAF

Approved:

// signed //
Dr. Edward D. White (Chairman) Date

// signed //
Dr. Kevin L. Elder (Member) Date

// signed //
Mr. John V. Bielecki (Member) Date

Abstract

Determining accurate cost and schedule is a crucial step to planning acquisition expenditures but history has shown that estimates are routinely low. Several researchers have attempted to forecast cost and schedule growth; we pick up this stream of research with a new approach. Our data collection and analysis focused on bringing in new data sources and added longitudinal variables to account for changes that took place over time. We assessed cost and schedule parameters for 37 major acquisition programs between Milestones II and III, resulting in 172 input variables and 5 regression models, 2 for schedule slippage and 3 for cost growth.

All five models passed statistical scrutiny and exhibited an Adjusted r^2 in excess of 0.80. The primary discriminator was the inclusion of strictly qualitative variables, taken from Selected Acquisition Report narratives and change justifications. We called these "soft" variables and coded them on a scale of 1 to 5 in the categories of funding problems, political problems, technical challenges, and contractor cost growth. Models with and without soft variables are presented to demonstrate their relative benefit. Finally, implications and implementation examples provide users a path to what-if analysis and decision-making.

Acknowledgements

My heartfelt gratitude goes to everyone who supported this effort. First, to the good Lord, to whom I owe everything. Second, to my wife and children who suffered through my absence, but gave their steadfast encouragement, while I toiled away. Lastly, I thank my advisor and committee members who motivated this research and provided valuable guidance and feedback, keeping me on the straight path.

Table of Contents

List of Figures

List of Tables

PREDICTING THE EFFECT OF LONGITUDINAL VARIABLES

ON COST AND SCHEDULE PERFORMANCE

I. Introduction

Overview

Weapon systems procurement is a long-standing hot issue within the Department of Defense (DoD) due to a reputation of cost and schedule overruns and the resulting congressional scrutiny. Acquisition reform is reducing the level of contract ambiguity and allowing milestone decision authorities and project managers enhanced abilities to administer their procurement programs, but there remains a need to accurately predict future program costs and the opportunity cost of major program changes. Recent research has focused on finding and refining variables that describe the dynamics of cost and schedule growth with the intent of mathematically predicting their impact. These variables come principally from the Selected Acquisition Reports (SARs), the primary documents submitted by the DoD to Congress regarding the status of Major Defense Acquisition Programs (MDAPs) (Jarvaise et al., 1996:3). Over time, the SARs have undergone significant evolutionary changes at the hands of Congress and other organizations such as the Government Accountability Office (GAO) (Cross, 2006:23). This instability results in a data set that is less than ideal for making statistical conclusions but recent research has found some predictive capabilities.

When attempting to balance program cost, procurement schedule, and product quality, cost and schedule garner the most emphasis and quality is generally taken for granted. The acquisition system incorporates a rigorous requirements validation process that drives the product's minimum acceptable quality, or capability, and with this dimension held constant, cost and schedule must absorb program fluctuations. Therefore, this research focuses on quantifying internal and external change effects on cost growth and schedule slippage.

As with any government or commercial endeavor, accurately estimating per-item cost is an important first step in determining whether to make a purchase. In contrast, the DoD defines capabilities needed to overcome a potential threat and attempts to purchase that capability, at almost any cost. These philosophies clash when new and unproven technologies come into play, making it very difficult to estimate total cost. Since the needed capability is often still on the drawing board, technological challenges escalate cost and create an unpredictable schedule. Despite reform initiatives and laws requiring technological maturity, the problem has remained relatively constant over several decades. The problem finds recognition in several studies, including a 1993 RAND study stating that of the Acquisition Category (ACAT) I programs, approximately 20 percent will experience cost growth from initial estimates (Drezner, 1993:xiii).

While cost growth catches the congressional scrutiny, perhaps more apparent to the end user is schedule growth. As with cost, immature technologies, poor contractor performance, and funding changes create schedule delays and weapon systems are slow to field. As recently as April of 2006, the GAO reported that even with recent reforms, there are still cost and schedule problems (GAO-06-368, 2006: Introduction). History continues to demonstrate that cost and schedule growth will frequently occur and identifying growth triggers will save time and money.

Specific Issue

A review of several studies covering different aspects of cost and schedule growth showed that although promising indicators are available, there is little consistency among resulting models. Previous approaches incorporated static variables generated from the most recent SAR or contract data (Singleton, 1991; Wandland, 1993; Sipple, 2002; Bielecki, 2003; Moore, 2003; Genest, 2004; Lucas, 2004; McDaniel, 2004; Rossetti, 2004; Monaco, 2005). Those that strayed from this philosophy endeavored to demonstrate the effect of some specific historical change, such as acquisition reform, on cost growth (Abate, 2004; Phillips, 2004). These studies found their best available predictor variables but little consistency or consensus as to what variables might apply across programs or time periods.

With these traditional snapshot approaches nearly exhausted, we focus on the recommendation to view variables in a longitudinal manner (Cross, 2006:100). We also see the need to look outside the SAR confines to any other likely source, including the political climate, economic conditions, and the threat of enemy aggression. Supporting our stated purpose, we concentrate on finding readily available longitudinal variables, in the SAR and elsewhere, that predict total acquisition cost early enough in the process to affect change.

Scope and Limitations

After an extensive search for data, Cross determined that the SAR is the most reliable source and that others proved virtually useless (Cross, 2006:94). However, other researchers have pointed out that the SAR is less useful for cost calculations (Gordon, 1996:11). One challenge with using SAR data is that over time, the acquisition process has changed and along with it, terminology. This creates a mismatch between programs that challenges the analyst to determine valid comparisons across acquisition reform initiatives. For example, the Milestone III event had clear meaning until 2000, when the Full-Rate Production (FRP) decision review took its place in the acquisition vernacular. As a rule, we consider these equivalent. Keeping this in mind, we focus on SAR data and the most universally accepted acquisition events that can be determined regardless of acquisition process changes. Chapter III presents a detailed review of key events and outlines our assumptions of equivalency across major acquisition reforms.

Previous research applied both logistic and multiple regression techniques to build a predictive model but with mixed success. Depending upon the variables selected, missing data points resulted in such a small sample that logistic regression proved inadequate (Cross, 2006:65). However, multiple regression and least squares analysis have been successful and provide a good starting point. Since we employ a new longitudinal variable concept, we do not artificially limit our analysis to any specific technique but rather, we conduct an exploratory analysis using techniques appropriate to the resulting data.

Research Objectives

Specifically, this research establishes a relevant model by 1) determining the significance of historical data in light of external influences and acquisition reform initiatives, 2) building a longitudinal database of pertinent historical data, and 3) identifying non-traditional variables and confounders that influence the resulting model. The end goal is to produce an easy-to-use model that predicts cost and schedule growth, from readily available information, in time for the program manager and milestone decision authority to take action. A good model is able to answer the question, "if I initiate a program with the given characteristics of magnitude, quantity, difficulty, and external environment, how much cost growth and schedule slippage will occur?"

Thesis Overview

Chapter II reviews the current literature on the subjects of acquisition reporting and the SAR along with a detailed review of previous work in this research stream. After a thorough review to set the groundwork, Chapter III presents a detailed research strategy, discusses data gathering, states preliminary assumptions, and frames the analytical methodology. Once the data and methods are determined, we build and validate our model and discuss the results in Chapter IV. Finally, Chapter V presents conclusions, lessons learned, and ideas for follow-on research.

II. Literature Review

This chapter reviews previous research in the area of statistical cost and schedule growth for major DoD acquisition programs and summarizes the achievements made in this research stream. While several organizations such as the RAND Corporation and the GAO have conducted similar research, students from the Air Force Institute of Technology have extensively utilized SAR data in their statistical analyses. This review does not completely recapitulate the body of previous work but rather establishes a footing from which to take the next step by first outlining major contributions and second, reviewing the current state of the acquisition process. From these building blocks, a methodology will be constructed for the current effort.

AFIT Research

At least 23 Air Force Institute of Technology (AFIT) theses have been written addressing cost and schedule growth since 1986. Of these, approximately one third have focused on building a comprehensive SAR database to support statistical modeling with the intent of giving project managers a tool to predict cost and/or schedule growth. Figure 1 shows the magnitude of this research stream.

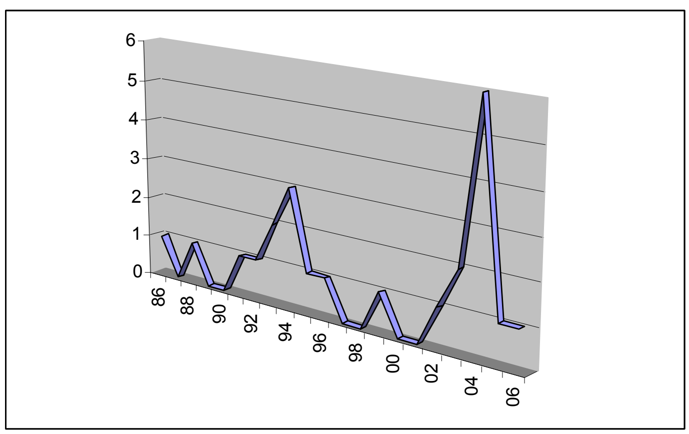


Figure 1 - AFIT cost and schedule growth theses by year, 1986 - 2006

Singleton, 1991

Singleton focused her research on making accurate program cost estimates from a grass roots approach of engineering methods and work breakdown (1991:39). The result was a "most probable cost (MPC) estimate" that could be used early in the source selection process. Subject data was derived from 16 Aeronautical Systems Division (ASD) programs between 1980 and 1988. Problem areas were identified by phase configuration in the development phase and schedule in the production phase. These represent two of the three factors listed by the ASD Research and Cost Division as creating challenges for all acquisition programs: technical risk, configuration stability, and schedule risk (Singleton, 1991:vii).

During her research, Singleton assembled a panel of industry experts who identified controllable and contributing factors. Controllable cost factors included unrealistic inflation estimates, lack of competition, high-risk design, poor management, specification changes, unrealistic schedules, concurrent production and development efforts, and technical advances (Singleton, 1991:39). Contributing factors were contractor experience, contractor familiarity with government business, technical risk, degree of Engineering and Manufacturing, Development (EMD), and production overlap, comparability to historical data, requirement stability, data availability for comparable systems, and schedule slippage (Singleton, 1991:50).

Singleton also listed three approaches to estimating costs. The first approach was parametric, which dictates correlating current design parameters to historical costs. The second approach was estimating by analogy. In this approach, the current program is compared to similar programs with differences accounted for via adjustments to technical definitions. The final and perhaps most ambiguous was the expert opinion approach. Expert opinion is subjective but may be the only option for new, beyond state-of-the-art, products.

Singleton proposed using a range rather than a point estimate to overcome the overlapping of different estimating techniques in use (1991:24). A single-value point estimate clouds decisions when competing alternatives are close (i.e. no statistical difference). Singleton derived unique cost growth range tables for different process phases (Table 1 shows the developmental phase) from which the decision-maker could predict a range of growth factors given their assessment of technical risk, configuration stability, and schedule.

Table 1 - Cost Growth Range Table Example (Singleton, 1991:70)

Development Potential Cost Growth Range					
Tech Risk	Config Stability	Schedule Impact	Upper CF	Med CF	Lower CF
High	High	Low		1.18	
High	High	High	1.10	1.06	1.02
Low	High	Low		1.31	
Low	High	High	1.57	1.02	0.83
High	Low	High	2.25	1.60	1.28
Low	Low	Low	1.07	1.05	1.03
Low	Low	High	1.70	1.51	1.31
					CF = Cost Factor

Gordon, 1996

Gordon provides a rich source of information gathered from the 1986 to 1996 period. His research showed that there were several inconsistencies in the causal variables offered. For example, Nystrom (1995) found influences due to stage of completion but Elkington and Gondeck (1994) come to the opposite conclusion.

Wandland (1993) says that contract type is not a factor and Buchfeller and Kehl (1994) conclude that there was no significant difference in cost growth due to contract category but others disagree (Nystrom, 1995; Terry and Vanderburgh, 1993; Blacken, 1986). Although these researchers focused on different outcomes, the fact that specific variables have a contradictory impact among studies demonstrates the need for more dependable indicators. Researchers have obviously found it difficult to define a parsimonious cost growth model.

The central question of Gordon's research was whether contract cost performance is sensitive to contract baseline volatility (1996:12). Gordon focused his efforts quantifying the effect of baseline changes and cites the fact that weak requirements lead to contracts not being fully defined even when awarded, making later modifications necessary (1996:9). This created a source of instability commonly called the "rubber baseline" (Gilbraeth, 1986:139) as demonstrated by the apparent differences in estimated cost growth. Aggregate cost growth based on a review of 197 programs is about 20 percent (Drezner et al., 1993:49). However, because cost growth is the difference between a baseline estimate and the *latest* prediction (baseline) of total cost (Hough, 1992:10), and the observation that the average cost overrun is only 8 percent, the 12 percent difference must be attributable to contract modifications resulting in contract baseline volatility.

Although the SARs attempt to identify the source of reported cost growth attributable to six categories: Economic, Quantity, Schedule, Engineering, Estimating, and Other Changes, they lack resolution and detail when working with cost data. Hough found that the practices employed in preparing the SAR could mask, delay, or exclude significant areas of cost growth. Because the SAR is an estimation report rather than a measurement tool, it is subject to manipulation by the program managers preparing it (Hough, 1992). In contrast, the Defense Acquisition Executive Summary (DAES) reports performance measurement data as well as cost, schedule and technical estimates. This increased awareness of the details allows the analyst to gauge validity of the estimates (Gordon, 1996:11).

Drezner found that estimates are biased lower than final cost (Gordon, 1996:12). Using the SAR database, correcting for quantity and inflation effects, Drezner showed that planning and development estimates are on average 20 percent below the final cost, including the cost of changes as well as cost overruns. Furthermore, he showed that these results were sensitive to program size, maturity, modification programs versus new starts, and program duration. Interestingly, prototyping was also influential but inversely related. Programs that used prototypes actually had poorer estimates and therefore, greater cost growth.

Gordon pointed out in his review of Terry and Vanderburgh (1993) that a combined measure of cost and schedule performance, the Schedule Cost Index (SCI), is the best predictor of final cost at completion. Of importance is the fact that this study addressed cost at completion rather than cost growth. By doing so, it was one of the first studies to divorce contract performance (i.e. overruns) from program baseline changes or contract modifications (Gordon, 1996:42).

Gordon continued his review with three 1994 theses. Buchfeller and Kehl (1994) found no significant differences between cost variances between contracts categorized by military service, program phase, contract type, or stage of completion. Their sensitivity analysis also failed to address possible differences between stable and unstable contracts. Elkinton and Gondeck (1994) attempted to quantify cost growth using a “Budget at Completion Adjustment Factor” derived from historical data and found that this measure of instability did not improve cost estimates over techniques based solely on unadjusted program performance. Finally, Pletcher and Young (1994) discovered that baseline stability was a predictor of contracts that improve cost performance over time.

Even though there was significant anecdotal evidence that baseline instability should cause cost growth, statistical analysis failed to show that the hypothesized relationship existed (Gordon, 1996:44). These findings run counter to the cited research. Earlier researchers (Hough, 1992; Pletcher and Young, 1994; Terry and Vanderburgh, 1993) were concerned with the instability of contracts and speculated that contract performance is sensitive to baseline changes. Gordon’s findings indicate that this sensitivity cannot be demonstrated, leaving the vast majority of the variance in contract performance to be attributable to other variables (Gordon, 1996:50).

Romasz, 1999

Romasz focused on base support function contracts. While these contracts are most often of less magnitude than major weapon systems, they may provide valuable insight into what factors cause cost growth. Indeed, these smaller programs have many of the same issues as their larger counterparts. The GAO found that “inadequately crafted statements of work have necessitated changes to contracts, which have often resulted in cost increases” and that “increases in federally established wage rates . . . are a source of increased contract costs” (GAO, 1997:5). Romasz could not determine if cost growth was occurring during the period from 1986 through 1994 (1999:64).

Complicating factors cited were a lack of data and limited usability, leading to a loss of statistical degrees of freedom (Romasz, 1999:66).

Sipple, 2002

Sipple's major contribution was a two-step statistical analysis. He determined that much of the data was centered in a point mass, effectively watering down the potential impact of other, more indicative variables. The solution was to first use logistic regression to determine if cost growth would occur and then use multiple regression to determine the magnitude. Using a SAR database covering the 1990 to 2000 timeframe, 78 variables were extracted for analysis of engineering cost growth during the EMD phase.

Up to this time, most work had concentrated on cost in terms of dollars. Sipple quotes the need of visibility on "cost of delay" as well (Westgate, 2000:16). By example, he states that making quantity or schedule changes is often the largest cost driver (Sipple, 2002:10). However, sacrificing schedule is usually easier than sacrificing cost. The idea of optimizing program schedules instead of subjecting them to budget constraints faces great resistance by program managers under the current politics of the acquisition-funding environment (Westgate, 2000:17). When reviewing past research, Sipple stated that "it was more descriptive than inferential" and that "more realistic estimates" were needed (2002:9).

The Office of the Secretary of Defense (OSD) Cost Analysis Improvement Group (CAIG) gives guidelines for documenting cost and estimating uncertainty for DoD system acquisition programs. First, they mandate that "areas of cost estimating uncertainty will be identified and quantified". Second, the CAIG prescribes "the use of probability distributions or ranges of cost” to quantify uncertainty. Third, they ask that the uncertainty be “attributable to estimating errors” (Department of Defense, 1992:22).

Similar to the ASD Research and Cost Division parameters referred to by Singleton, Sipple cited the Air Force Materiel Command (AFMC) Financial Management Handbook that recognizes three risk parameters: technical, schedule, and cost risk (AFMC, 2001:11-12). Cost growth occurs due to urgency of the program, technical difficulties, amount of concurrency, and the degree of testing (Tyson, 1994:S-5).

Sipple also pointed out a difference between program categories. Missile programs tend to experience more variability than aircraft programs. Closer management scrutiny and “protection from schedule stretch” were possible reasons for the more consistent cost growth in aircraft programs (Tyson, 1994:S-2).

Like Gordon, Sipple was concerned with Drezner’s conclusion that prototyping seemed to have an inverse effect on cost growth. “We compared the cost outcomes of prototyping and non-prototyping programs, expecting to find that a prototype development strategy contributes to cost control through reduction of uncertainty…it may also be true that prototyping was [only] conducted for programs with higher degrees of technical uncertainty” (Drezner, 1993:51). Interestingly, programs that included prototyping had a relatively higher cost growth. This result may be due in part to the timing of the prototype phase within the context of the overall program schedule, since earlier prototyping makes data available earlier, thus potentially affecting the baseline cost estimate at the time of EMD start (Sipple, 2002:35).

Sipple follow-ons

Bielecki (2003), Moore (2003), Genest (2004), Lucas (2004), McDaniel (2004), and Rossetti (2004) furthered Sipple's work by looking at different portions of the SAR database through the two-step regression approach. In general, they all found positive results but predictor variables were seldom the same, pointing to a possible underlying inconsistency or fallacy in using SAR data. In addition, these studies limited themselves to only the most recent SAR for each program.

Bielecki presented a worthwhile look at the acquisition environment. Since the fall of the Berlin Wall, the DoD budget has been under ever increasing downward pressure. Doing more with less is the daily mantra, particularly within a major weapons system program office. Moreover, weapons programs with exorbitant cost growth during this period of reduced funding, garnered harsh congressional and Presidential attention (Bielecki, 2003:10). Bielecki offered us a key turning point in history with mention of the A-12 program's cancellation in 1991. Then Secretary of Defense Cheney cancelled the program after costs inexplicably skyrocketed and "no one could tell him the program's final cost" (Christensen, 2004:105).

Like Gordon's rubber baseline, Bielecki used Hough's discussion to describe the problem of inconsistency. The analyst must recognize that the "selected" baseline may not be consistent over time. This inconsistency stems from two types of events: rebaselining and evolutionary changes. Rebaselining occurs when the program office develops a new baseline estimate in the middle of an acquisition phase. The new program estimate replaces the old estimate; yet, it retains the original estimate's designation (PE, DE, or PdE[1]). Evolutionary model changes occur when modifications are made to a program such that the "current model only remotely resemble what was originally estimated" (Hough, 1992:12-14). Detecting either a rebaselined or evolutionary changed program from a non-changed program is difficult at best and extremely hard to normalize out of SAR data (Hough, 1992:12-14; as referenced by Bielecki, 2003:29).

Moore contributed his insight with discussions of buffering and new variables. Buffering occurs when a program manager overstates the budget so that as cost growth occurs, it can be absorbed (Moore, 2003:2). This number padding is very tempting since it relieves the program manager from having to lobby for increased funding as growth occurs. The perception is that limiting cost overruns lessens the chance of program cancellation. Moore also identified the First Unit Equipped (FUE)[2] variable. He found it to be significant but cited a scarcity of data points as a potential problem (2003:26). FUE-based variables were not available for a majority of programs, limiting the results (2003:54).

Genest addressed the political aspects of acquisition by reviewing legislation intended to curtail cost overruns. One such law is the Nunn-McCurdy Act, which brings more visibility and scrutiny to programs that incur large cost increases (2004:1). Genest also compared the results and similarities between Sipple, Bielecki, and Genest models. Each model was reasonably predictive but it is difficult to find common predictor variables; maturity and prototyping being the only two that occur in three out of the four

[1] Depending on the phase of the acquisition cycle, the baseline values are represented by the Planning Estimate (PE), the Development Estimate (DE), or the Production Estimate (PdE).

[2] "First unit equipped" is discussed in more detail in Chapter 3.

logistic regression models and maturity alone in three of the multiple regression models. Genest put it this way: "we do not find any common variables between the four models nor do we expose any trend to shed light on future cost growth research…comparison of these models, predictor variables, and validation results reveals no considerable advantage realized from one model to the next" (2004:52).

Monaco, 2005

Monaco also applied the two-step logistic and multiple regression approach but he added the aspect of predicting schedule. Monaco referenced a 1990 RAND study stating that the average schedule slip of a major weapons system program is 33% (Drezner and Smith, 1990:44). RAND also reports that most programs choose an extended schedule to avoid [cost] overruns (Drezner and Smith, 1990:iii). Monaco's research uncovered a comprehensive list of potential schedule drivers that served as a useful addition to our work. Going further, and adding to Singleton (1991) and Sipple (2002), Monaco quoted Drezner and Smith's factors of unstable funding, technical difficulty, external guidance, and external events (Drezner and Smith, 1990:33).

One reason for continued schedule slippage in the procurement of major weapons systems is the low level of technical maturity of the system when it enters the EMD phase. Once the development phase begins, the government incurs a large fixed investment in the form of human capital, facilities, and materials. Any significant changes will have a large rippling effect on schedule and cost (Rodrigues, 2000:2). Furthermore, once in the development environment, external pressure to keep the program moving becomes dominant. Preserving cost and schedule estimates becomes paramount to securing budget approval. If a program manager decides that an additional year is needed to reach the desired level of technical maturity, they run the risk of reduced funding, which could lead to program cancellation (Rodrigues, 2000:6). Managers are more likely to accept a lower level of technology than risk losing the program. Unfortunately, low levels of maturity lead to increased risk, which in turn leads to the likelihood of schedule delays, increased costs, and quantity reduction (Monaco, 2005:11).

Monaco took the path set by Nelson and Trageser (1987:2-17) of separating programs by mission type: cargo, tanker, attack, and fighter aircraft. Separating programs in this way allowed comparison by technical difficulties and perceived urgency of warfighter need. A positive correlation existed between the mission type and schedule duration as indicated by larger increases for longer duration fighter aircraft compared to shorter duration cargo aircraft (Nelson and Trageser, 1987:2-17).

Via his results, Monaco showed that yet another set of predictor variables indicated likelihood of a schedule slip. He also pointed out that while the Milestone III (MSIII) occurring before Initial Operational Capability (IOC) is predictive, it is most likely acting as a proxy for total quantity planned (Monaco, 2005:109). Other research did not specifically bring out this concern of imposter or proxy variables but this could be a reason for inconsistency among what are otherwise equivalent models.

Table 2 shows the magnitude of the missing data problem mentioned by several researchers. The impact is that any programs missing a data point that is being used for regression analysis will not be considered, effectively reducing the entire dataset. For example, if FUE was to assessed, at best, only 19.4 percent of the data could be used. This creates a problem for drawing robust conclusions from an already limited database.

Table 2 - Data availability, percentage of programs with recorded dates (Monaco, 2005:110)

Schedule Date	% of Programs with Recorded Schedule Date
First Unit Equipped	19.4%
Preliminary Design Review	23.9%
Production Contract Award	29.9%
Critical Design Review	37.3%
EMD Contract Award	59.7%
Initial Operational Capability	77.6%

Finally, Monaco emphasized usability. In line with our stated objective, for a predictor variable to be of value it is important for the independent variables to be both understandable and available when the program office accomplishes the development estimate (Monaco, 2005:33). A confusing or hard-to-derive variable would be of little use. A model that uses prominent data has utility and is easily defendable.

Cross, 2006

Cross took Monaco's analysis one step farther by adding a variable to capture the effect of rebaselining.[3] Several researchers expressed a concern with the potential volatility driven by rebaselining a program but stopped short of trying to determine its true effect. Cross used the number of times a program has been rebaselined to predict both schedule and cost growth and in the process, determined that such a longitudinal variable does not work well with logistic regression. Sipple's two-step process would not work in this case.

Cross's major contribution was in discovering the importance of a longitudinal approach, looking at changes over time such as the number of rebaselines. Previous research all but exhausted the two-step method and although predictors were found,

[3] For further discussion of what constitutes a baseline and how it may change over time, see the SAR baseline discussion in Appendix B.

inconsistency from one model to the next revealed a weakness. Future research will need to uncover new ground to make significant progress. This is another reason why future research should focus longitudinally since we cannot find or recreate missing variables like FUE (Cross, 2006:100). Additionally, Cross pointed out that we would be remiss not to address 2005 GAO recommendations (2006:99). These recommendations included looking at cost estimates over the life of a program by comparing the first full estimate (usually at MS B) with the current Approved Program Baseline (APB).

Abate, 2004; Phillips, 2004

Abate and Phillips conducted similar research but from a more formal cost analysis background. Their major effort was in developing a hybrid Adjusted Cost Growth (ACG) model, which looked at cost growth throughout the life cycle of an acquisition program. Abate limited his research to missile systems, from 1991 – 2001, while Phillips conducted the same analysis for aircraft. Since they went in to the research looking for changes over the long term, they theorized that 1996's major acquisition reform might change the amount of cost growth. Abate presented a good review of acquisition reform (2004:3). For missile systems, this hypothesis held but for aircraft, it did not. Surprisingly, annual cost growth of the post-reform period (i.e. after 1996) was significantly higher (Abate, 2004:iv).

Like Cross, Abate and Phillips considered rebaselining in their analysis. Abate, however, took steps to neutralize its effects rather than use it as a predictor. His cost normalization process attempted to remove external effects and focus on purely programmatic issues. The result was an Adjusted Cost Growth Factor (ACGF) for each SAR year (Abate, 2004:10). An ACGF greater than 1.0 represents a program that incurred cost growth, while an ACGF less than 1.0 identifies favorable cost performance within a program (Abate, 2004:50). A plot of ACGF by SAR year could reveal cost growth trends.

Abate again reported the weaknesses of using SARs. The analysis revealed several complicating factors involved in performing cost growth calculations. Initially, the data included in cost growth calculations are somewhat subjective, as one must carefully interpret the SAR's qualitative and quantitative sections. Proper data extraction from the SAR is perhaps best classified as an art rather than a science, as numerous organizations have developed different cost data from the same source documents (Abate, 2004:72).

Phillips brought out the idea of the learning curve presented by McCrillis (see Figure 2). In short, lower quantities create a non-linear increase in per-unit cost since there are fewer units over which to spread fixed costs such as facilities and tooling. Normalizing using the learning curve slope affects the data by either increasing or decreasing the amount of a program's cost variance. A weapon system's baseline cost "is established assuming a specific quantity of units. As the number of units increases, the unit cost will go down even though the program cumulative total cost increases. As the number of units decreases, the unit cost increases even though the program cumulative total decreases" (McCrillis, 2003).

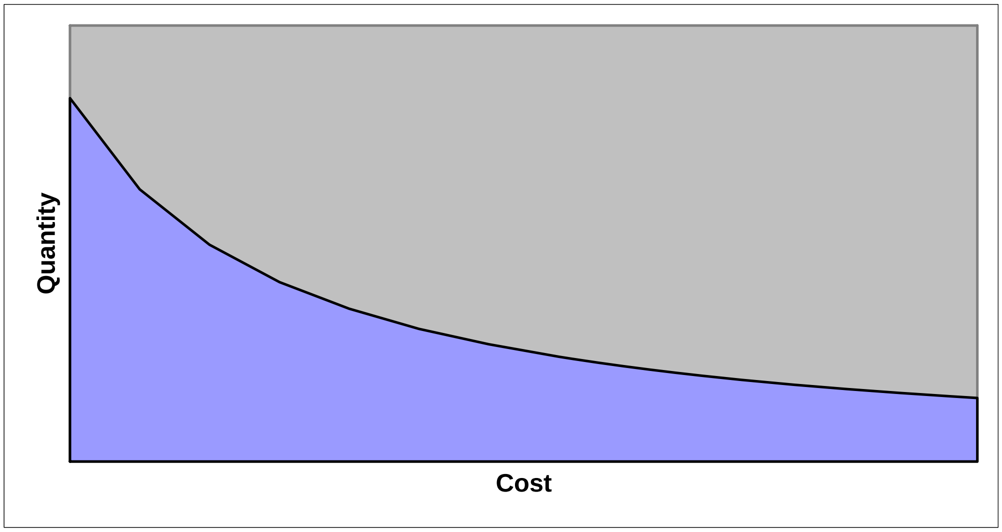


Figure 2 - Learning curve slope (McCrillis, 2003; as referenced by Phillips, 2004:44)

DoD Acquisition Performance Research

Many RAND studies and GAO reports document acquisition program performance and provide a source of lessons learned. The often-cited 1993 Drezner study attempted to identify the extent of a historical cost growth problem in DoD acquisition by focusing on two primary research objectives: quantifying the magnitude of cost growth in weapon systems and identifying factors affecting cost growth. Utilizing SARs dated through December 1990, Drezner compiled a database of 197 major weapon systems for cost growth analysis. Two significant findings resulted from this study. First, there has been “no substantial improvement in average cost growth (approximately 20 percent) over the last 30 years, despite the implementation of several initiatives intended to mitigate the effects of cost risk and the associated cost growth (Drezner et al., 1993:xiv). Second, researchers could not definitively account for observed cost growth patterns. Thus, no ‘silver bullet’ policy option is available for mitigating cost growth

(Drezner et al., 1993:xi). Two factors, program size and maturity, did stand out among the rest as having the greatest effect on total program cost (Drezner et al., 1993:xii).

A 1996 study in which Drezner worked with Jarvaise and Norton analyzed data in the Defense Systems Cost Performance Database (DSCPD) constructed and maintained by RAND. Their general conclusion was that "though the issue has been studied extensively over the last several decades, the results of these studies appear not to have translated into policy changes that have had a measurable impact on cost growth" (Jarvaise, et al., 1996:xi). The authors pointed out the weaknesses in their database so that decision-makers might understand its limitations as well as its usefulness. One key issue to remember is that SARs are generated only for the largest acquisition programs, representing only 45 to 55 percent of total procurement (Jarvaise, et al., 1996:6). If smaller programs have differing growth patterns, conclusions made with sole reference to the SARs may be misleading.

A 2006 RAND study conducted by Arena et al. looked at historical data to find cost growth of completed weapon system programs. They observed the following:

> • Average adjusted total cost growth for the completed program is 46 percent from MS II and 16 percent from MS III.
>
> • This analysis shows about a 20 percent higher growth than the previous RAND SAR study. We attribute this increase to using only completed programs in the current analysis. As we demonstrate, cost growth continues for both development and production well past MS III—likely due to requirements changes and system upgrades. Another contributing factor may be the sample selection (e.g., excluding ship programs).
>
> • Cost growth bias does not disappear until three-quarters of the way through system design, development, and production. At this point, the system is well understood and a solid estimating basis is available. (Arena et al. 2006:39).

As with prior research, Arena et al. observed very few correlations with cost growth but in general, programs with longer duration had greater cost growth. As an aside, they found that electronics programs tended to have lower cost growth. They also considered possible differences between the services but found none.

In the same vein as Abate and Phillips, Arena et al. explored the possibility of cost growth improvements over time. They found it difficult to pinpoint any specific period of improvement or significant change due to reform initiatives. Addressing trends that did appear, they stated: "…the data do show an improving trend with time. However, our data for recent programs are biased toward ones with shorter duration, and programs that take less time to complete tend to have lower cost growth. Therefore, we cannot say whether the trend is due to improvement or sample selection" (Arena et al., 2006:39). They also noted a trend toward reducing quantities and that quantity growth seems to be less of an issue.

In a 1999 study, Christensen added further support for the 20 percent average annual cost growth identified in the 1993 Drezner report, finding similar results with the DAES database as Drezner found with the SAR database (Christensen et al., 1999:251). More specifically, this study analyzed an eight-year window around the implementation of the Packard Commission's recommendations to determine if cost growth improved because of these reform efforts. Christensen's research identified that the Packard Commission's recommendations "did not reduce the average overrun percent experienced on 269 completed defense acquisition contracts over an eight year period (1988 through 1995). In fact, the cost performance experienced on development contracts and on contracts managed by the Air Force worsened significantly (Christensen et al., 1999:251). Failure of the Packard Commission's recommendations to control cost growth as designed reveals the need for continued monitoring of newly implemented acquisition reform efforts. (Abate, 2004:32)

Christensen advocated in a 2004 article that the 1991 cancellation of the Navy's A-12 program was a powerful catalyst for acquisition change. He cited numerous studies that confirm that program managers chronically understate the final projected cost of their programs – the Estimated Acquisition Cost (EAC). In its generic form, EAC is calculated as:

EAC = Cumulative Actual Cost + (BAC – Earned Value) / Performance Index

where BAC is the Budget at Completion and the performance index is a factor used to adjust the budget upward to account for typical understatement (Christensen, 2004:3). When calculated at different stages, the EAC gives what is essentially a lower bound to the final cost range (Christensen, 2004:6). The utility being that at any particular stage, a program manager could forecast how much similar programs have overrun their best estimates. Table 3 shows the generic results and how far below final cost the estimates were at different points of contract completion.

Table 3 - EAC percent completion / percent below final cost

Percent contract completion	EAC percent below final cost
20	18.1
50	8.2
70	2.1

For the last five years, the GAO has reviewed the status of several major weapon systems acquisitions. The latest report, GAO-06-391, presents their assessment of 52 systems chosen for their high dollar value, stage in acquisition, and congressional interest (GAO-06-391, 2006:2). They found that the DoD often exceeds cost estimates by 30 to 40 percent and that programs experience cuts in planned quantities, missed deadlines, and performance shortfalls. They proposed managing programs based on levels of knowledge versus traditional milestones. One such area of knowledge is technical maturity. They stated that programs that start with immature technologies average research and development cost growth of 34.9 percent while those that begin with mature technologies experience only 4.8 percent (GAO-06-391, 2006:2).

The report also pointed out that a significant portion of the recognized total development cost increases took place after programs were approximately half way into their product development cycle. This suggests that cost growth due to immature technology occurs even after design approval. The GAO stated that programs experienced a cumulative increase in development costs of 28.3 percent throughout their product development and that approximately 8.5 percent of the total development cost growth occurred up until the time of the average critical design review. The remaining 19.7 percent occurred after the average critical design review. “If past is prologue, the decisions to continue to move programs through development without the requisite knowledge will continue to result in programs that are not delivered on time nor with the quantities and capabilities promised” (GAO-06-391, 2006:13).

Summary

The defense acquisition system has suffered many improvement attempts over the last 50 years but cost growth and schedule slippage continue. Efforts to determine what might predict growth have turned up would-be indicators but the variety of contributing factors coupled with inconsistency from model to model indicates that there are causal factors still hidden. However, several researchers have developed novel ways of analyzing the available data, and present us with a platform from which to start our analysis.

First, we pursued longitudinal variables to uncover time-based effects. Considering the exhaustive research into internal factors and program parameters, a comparison against external factors such as political climate, adversary positioning, and the economy was deemed an appropriate addition. Next, cost and schedule factors were normalized to level the playing field when comparing multiple programs and time periods. Third, while the SAR database is arguably the most consistent data source, we assessed any valid database that might have yielded a key piece of information. Finally, since cost, schedule, and requirements are intertwined, we compared them in unison. Chapter III pulls these concepts together into a plan for building our database and conducting our analysis. Chapters IV and V present our analysis, discussion, and conclusions.

III. Methodology

Introduction

This chapter outlines the data collection process and describes analysis techniques employed in Chapter IV. Since the primary goal was to track program changes over time, the principle effort of this research was in producing a longitudinal database from which we could extract a pool of predictor variables. In building the database, we determined from where to collect the data, what programs to include, what data to collect, and how to address missing or dissimilar data. Included is a discussion of assumptions as well as strengths and weaknesses we uncovered. Finally, we review the statistical techniques used to cull variables and build regression models during the analysis phase.

Data Collection and Assessment

Building the database started with determining appropriate sources. Previous research pointed almost exclusively to the SARs but in addition, some researchers used the DAES database to source more specific cost information, the advantage being a higher reporting resolution. SARs are submitted on an annual basis with the requirement for additional reports if a significant event occurs such as moving from a development baseline to a production baseline. Disadvantages include the cumbersome size of the database and lack of information from early programs. Since our work focused on program changes over time, we needed consistent reporting across all programs, from at least MSII and through MSIII.

The SAR database[4] has many advantages including: strict reporting format which improves consistency of the data, annual SAR training for those submitting SAR reports which also improves consistency of the data, and increased scrutiny of data since SARs are presented to Congress (Bielecki, 2003: 31). As a result, we determined that availability and consistency of the SARs presented the best source for both schedule and cost data. In addition, readily available information about defense spending, inflation rates, and Consumer Price Index (CPI) was pulled from the Office of Management and Budget (OMB) and the U.S. Department of Labor (DoL).

Next, we established criteria for what programs to include (see Figure 3). First, since SARs are required only for MDAPs, all programs had to fall into that category. Considering the fact that conditions change over time, and the older the data becomes the less indicative it is of current conditions, we chose to limit programs to those that had not yet achieved MSIII at the end of 1996. With acquisition reforms and change initiatives in the early nineties, this presented a logical place to start. Next, in order to maintain consistency across programs, and to provide a stable basis for comparison, we needed hard dates at the beginning and end of the comparison timeframe.

1. The program is an MDAP (ACAT 1C or 1D).
2. The program had not reached MSIII by the end of 1996.
3. The program has reached Milestone III (or C).
4. The program has a Milestone II.
5. Milestones II and III are not the same.
6. Development estimates are available.
7. The program has subjective relevance.

Figure 3 - Program Selection Criteria

[4] SARs are maintained by the Under Secretary of Defense, Acquisitions, Training, and Logistics, in the Defense Acquisition Management Information Retrieval (DAMIR) database.

The easiest delineation became the development phase, typically defined by the time between MSII and MSIII. Therefore, we required that all programs considered be at or past MSIII. After an initial look at some of the possible programs, we discovered that not all became MDAP programs before MSII was established or, as in the case with commercial derivatives, a program may have started at MSIII (e.g. the C-130J). Furthermore, programs initiated under an acquisition streamlining effort may not have traditional milestones. The Stryker, for example, started production before it officially met MSIII. These programs were necessarily excluded by the requirements that a program must have a SAR when MSII occurred, and that MSII and MSIII were not the same. Finally, we performed an initial quality cut by subjectively eliminating programs based on their relevance to this research. For example, we excluded a nuclear aircraft carrier program because of its excessive procurement cycle and large single-unit cost.

The SARs provide several kinds of information: schedule, cost, quantity, performance, and narrative. Critical to this research was the change in cost and schedule so we considered any data reflecting these two factors. For schedule, we recorded MSI, MSII, MSIII, Low-rate Initial Production (LRIP), and Initial Operational Capability (IOC) for each SAR, paying close attention to SAR date, APB, and whether the value was a planning, development, production, or current estimate (CE).

Under the cost category, we recorded only changes in the Program Acquisition Unit Cost (PAUC). This simplification allowed us to focus on overall program cost and avoid inconsistencies among programs and over time. PAUC proved to be an accurate and meaningful variable throughout the research but in the future, more effort could be spent breaking out costs for individual areas such as research and development, or account categories such as military construction. To calculate total cost, we multiplied PAUC by the estimated quantity, so we also recorded quantity changes for each SAR.

The SARs proved problematic for performance data. First, performance characteristics were often classified, making tracking their changes cumbersome. More importantly, there was very little commonality among programs and therefore no solid basis for comparison. Without a common quantitative measure for requirements, we relied upon manually rating the SAR narratives for this type of information.

Narratives include any textual explanation of what happened during the SAR period. We placed emphasis on the executive summary but we also gleaned important information from cost and schedule change explanations. Each SAR's narratives were rated in three categories: technical problems, funding problems, and political changes. We recorded both presence (1 if the condition was present, 0 if not) and magnitude (1 to 5, see Figure 4) for each.

1 – no program delay or impact
2 – created a delay
3 – created a delay or challenge significant enough to cause a rebaseline
4 – caused a work stoppage
5 – resulted in program cancellation

Figure 4 - Magnitude ratings

For each program, we calculated a number of occurrences, an average number of occurrences per SAR, and an average magnitude per SAR. The narratives turned out to be a rich source of data but comments in early SARs were very brief and seldom exhibited attributable characteristics such as technological or political challenges. However, we noticed a subtle change in the quality of narrative reporting after approximately 1990 when more detailed change explanations became common.

We addressed missing and dissimilar data by bracketing the missing value with known good data or looking for common ground upon which to make a comparison. For example, if a SAR did not report LRIP or if it was to-be-determined for a given year, but the years before and after (the bracket) presented the same date, we made the assumption that no substantive change had taken place and used the bracket value. When a value was missing altogether, we searched for a logical equivalent. For example, some programs reported a date for Required Assets Available (RAA) instead of IOC. One program manager even argued for the exclusion of IOC since it was determined by the major command employing the system and that RAA was a more accurate acquisition-based term. In this case, we used RAA in the place of IOC, assuming that it would behave in the same manner statistically.

Another instance of missing variables surfaced when programs were not initially at the MDAP level and therefore began submitting SARs at MSII. Planning estimates, and in some cases MSI, were not reported under these circumstances. We handled these variables in two different ways, depending upon the analysis technique. First, if the analysis considered only the presence of a value, zero was entered for the missing variables because without any value assigned, the analysis software would ignore all data for the program in question, reducing the already small dataset. Second, if we considered the variable in isolation, we removed the zero and left the field empty so as not to bias the field to an arbitrary zero value, the result being that programs missing data were not used in the analysis.

Complications

Some older programs went through a transitional period wherein milestone titles and meanings changed. For example, the Longbow Apache listed two MSII decision points in the December 1989 SAR. The first was for an internal Army board, the Army Systems Acquisition Review Council, and the second was for the Defense Acquisition Board (DAB). All programs began reporting DAB baselines with the 1988 annual SAR but some carried duplicate baselines for a while. As the transition become complete, later programs listed principally the DAB baseline and milestones.

Milestone III presents another challenge to continuity across SARs and programs. Initially, MSIII marked the transition from development to production but by the late 1980s, common practice was to list Milestone IIIa as the LRIP decision point and Milestone IIIb as the Full-rate Production (FRP) decision point. In 1992, DoD Instruction 5000.2, "Operation of the Defense Acquisition System," officially changed Milestone IIIa to LRIP and IIIb to FRP. The Joint Surveillance and Target Attack Radar System (JSTARS) SARs submitted in June of 1991 shows this by making a clear transition from IIIa to LRIP and IIIb to FRP but even up to 1997, the Longbow Hellfire program listed "Milestone III (LRIP)" and "Milestone III (FRP)," carrying over the older terminology. We held these terms as logical equivalents during data collection.

Likewise, terminology changed again in 2000 when Milestones I, II, and III became Milestones A, B, and C. While MSC and MSIII are virtually equivalent, this is not the case with MSI and MSA. However, the impact to this research was minimal since we focused primarily on MSII and III. Rarely, a SAR reported Milestone C but no LRIP. In this case, we assumed them to be equivalent.

Cross pointed out confusion over the predominantly Army-used term FUE (2006: 46). We also noticed the term's use along with the pseudo-equivalents IOC and RAA. Consider the Joint Air-to-Surface Standoff Missile program that was developed simultaneously for both the F-16 and B-52. Each aircraft had its own definition of IOC, based upon user requirements, not on the physical system development or the acquisition program. They also reported RAA for the weapon itself which was different from and independent of the multiple IOCs.

The IOC requirements can also fluctuate throughout a program's life cycle. For example, the CH-47F program reported in its 2001 annual SAR that the IOC definition changed from 16 aircraft to 14. Adding to the confusion, the Abrams Upgrade program made a clear distinction between IOC and FUE, indicating that IOC was linked to operational capability at a training location while FUE indicates that the first combat-ready unit is fully equipped. As a general rule, FUE was preferred over IOC for programs with both so that a reasonable comparison could be made to programs listing only IOC but having a definition more in line with the "ready for combat" concept than simply "ready for training." For programs with both an RAA and either an IOC or FUE, the RAA date was used.

We made other assumptions and notes during collection to allow inclusion of as many programs as possible. The following list presents the balance:

1. All of a month's activities were reflected on the 1st. We assumed that on a scale of years, plus or minus 30 days was inconsequential but the simplification allowed program events to be seen as simultaneous or equivalent. This more accurately represents the fact that activities surrounding a December 20th decision were also present for the December 31st SAR reporting date.

2. Upgrade programs have the advantage of starting with a proven weapon system and their development time is generally shorter so it was important to

differentiate whether a program was an upgrade or not. When determining if the program was an upgrade, we asked the question "can the product stand alone?" The F-18E/F program was an upgrade to the C/D program – the modifications could not stand as a weapon system in themselves.

3. Unavailability of test aircraft or DoD test personnel was counted as a policy issue as opposed to a contractor delay.

4. To make the best use of IOC dates, we used the estimated IOC date for five programs (219, 278, 330, 341, and 354).[5] Actual dates were not yet available but since these IOC dates were to occur in the near future, we assumed that they would not change substantially.

5. During analysis, we arbitrarily set MSIII as 90 percent program completion, calculated by time. The measure allowed comparison of programs by percent completion but very short programs, those with only two or three SARs between MSII and MSIII, were easily skewed.

6. It was not clear in some programs what constitutes a prototype; so we assumed that if the program did not specifically mention a prototyping effort as part of the development phase, then it did not have one. This presented a weakness in this variable.

7. Programs combine and split during their lifecycle. The B-1B Conventional Mission Upgrade Program, for example, included three components, two of which were eventually recombined, and two separate timelines. The Longbow Apache split milestone tracking between the fire control radar and rockets. In this case we assumed that the milestones between the two components were consistent but we assessed each situation individually, looking for a consistent measure.

8. Baseline amendments for older (pre-1987) SARs were attributed as rebaselines (e.g. C-17 December 1987 SAR).

9. When multiple contracts were listed or awarded for the same milestone, the earliest was recorded in the database (e.g. MH-60 December 2001 SAR).

10. Cross (2006) did not include 12 programs (200, 219, 240, 260, 278, 294, 330, 341, 354, 367, 537, 551) used in this research so their variables were either brought up to date, substituted for, or removed.

11. Annual SAR submissions were cancelled for 2000. This did not have a direct impact on our analysis but could have skewed data an unknown amount.

[5] Program numbers and their equivalent titles can be found in Appendix C.

12. Programs can be specified as an MDAP but have vastly different characteristics. A high quantity, low cost program (rocket) behaves quite differently than a low quantity, high cost program (ship).

13. Adjusting baselines was often used as a management tool to "resynchronize" a program but also had the possibility of hiding cost and schedule problems from the casual observer. For 2006, Nunn-McCurdy breech reporting changed, removing the ability to hide overages by rebaselining and therefore, there may be fewer rebaselines in the future.

14. We recorded PAUC in base year dollars, which removed the complication of escalation. However, the base year changes as major transitions (e.g. from development to production) occur so particular attention was paid to costs reported in the same SAR as MSII or MSIII achievement to ensure uniformity.

Prior work in this research stream addressed differing issues and vulnerabilities when making predictions based on uncontrolled historical data (Gordon, 1996:38). This research was conducted ex-post facto, with no attempt to predetermine design. Therefore, we were limited to the data available as extracted from historical records so no effort to control for extraneous variables was possible. Interaction with a dynamic and often unpredictable environment was anticipated to be a major intervening variable (Gordon, 1996: 38).

While we attempted to account for confounding variables, several threats to internal validity made the establishment of a causal relationship problematic. First, the art of program management has changed over time, the body of knowledge growing from experience. Second, as demonstrated by the changes in terminology, an instrumentation effect was also possible. Third, program selection may have systematically biased the dataset. It is possible that the sample was the most or least likely grouping available to demonstrate cost and schedule changes. To our advantage, the possibility of data manipulation by program managers was controlled through the use of a certification procedure for performance data management, an audit function, and independent reporting (Gordon, 1996: 38). Therefore, we expect the data reported to be free of excessive manipulation.

In addition to internal validity, threats to external validity limit this study's generalizability. Along with other selection criteria, we limited the study to MDAPs and therefore, the results cannot be reliably applied to smaller programs. Generalizability rests then on the assumption that current and future programs will not differ substantially from historical ones.

Analysis Process and Statistical Techniques

We endeavored to find new and different ways to approach the problems of cost and schedule growth by looking to new data sources and assessing changes to unique variables over time. The foundation of this analysis was laid in statistical tests and linear regression modeling. In practice, the approach was to collect as much data as possible and push it through statistical analysis until only the significant variables remained, satisfying the necessary assumptions along the way. We assumed a significance level of $\alpha = 0.05$ throughout.

The goal of regression is to develop a formula comprised of fixed amounts of the input variables that will accurately predict the response variable. However, data rarely behaves well enough to be fit perfectly, leaving an amount of error – the residuals. Furthermore, if the formula, or model, adequately explains the data and there are no missing input variables, or pieces of the puzzle, the residuals will show no pattern; they will simply be noise. Statistically, this means that they will be independent, normally distributed around the residual mean, and will have constant variance. If a pattern is present, there is an unexplained but significant piece missing from the model such as the presence of mixed data types. Throughout the discussion of the regression models, we addressed normality and constant variance, along with outliers and other significant points of interest but independence presented a challenge.

Since we worked ex-post facto, there was no opportunity to address independence while collecting data and there was no way to guarantee independence actually exists. However, we assumed that all programs were executed in sufficient isolation to not violate the assumption that independence exists. One could argue that a program (e.g. new aircraft) could not proceed until another program (e.g. new radar system for multiple aircraft) met a certain milestone, but after reviewing the dataset, this interaction appeared to be minimal. Statistical tests that demonstrate independence could not be performed without specific ordering in the data, which we did not have.

Model validation was the final consideration. Once we built the regression models, we validated them using Tukey's jackknife approach (1987: 30). Jackknifing determines how a model is influenced by subsets of observations and by using this technique, we could determine presence of weakness due to data variability. More discussion on the mechanics of this procedure is offered in the analysis chapter.

Chapter Summary

This chapter addressed data sources, program selection, variable selection, and missing data. Next, we reviewed assumptions and notes we made during data collection. Finally, we previewed the statistical regression techniques and associated assumptions. Chapter IV presents the detailed analysis we conducted and Chapter V provides conclusions and recommendations.

IV. Analysis

This chapter outlines the analysis conducted from determining the appropriate response variables to building the regression models. First, we chose response variables that reasonably answered the questions of cost growth and schedule slippage, keeping in mind the goals of usefulness and equivalence among programs. Next, we classified each variable into one of seven categories: absolute dates, program characteristics, number of occurrences, qualitative variables, year-referenced era variables, percent completion-referenced variables, and dummy variables that isolated significant program groupings. Finally, we culled the variables, constructed regression models for cost growth and schedule slippage, and discussed possible application and usefulness.

Response variables

The stated goal of this research was to quantify internal and external change effects on cost growth and schedule slippage. However, we first needed to define what these terms meant to the potential user and what variables would best fit the desired model output. A key component to comparing the wide range of dissimilar programs was finding a common ground. The SARs became that ground but reporting procedures have changed several times in the last 30 years and it was often difficult to extract the same type of data from different SARs in the same program, or across programs. The only consistently accurate timeframe over which we could collect data was from MSII to MSIII so these terms bounded our response variables.

Since programs change over time, adopting new baselines due to quantity or schedule changes, for example, we could not simply use the most recent estimate compared to the final cost as a response variable. The true difference lay between the initial estimate of the total acquisition cost and the most recent or final estimate, adjusted for inflation and quantity changes (Jarvaise, et al., 1996:1). Similarly, Hough defined cost growth as "the difference between the most recent or final estimate of the total acquisition cost for a program and the initial estimate" (Hough, 1992:10). With the Development Estimate (DE) set at MSII and the end of our target phase at MSIII, the most logical cost response variable became:

"Cost_delta_MSII_MSIII_2005_percent_of_MSIII_cost"

This variable is a construction of the change in PAUC from the SAR reporting MSII achievement to the SAR reporting MSIII achievement, converted to 2005 dollars through the standard DoL Consumer Price Index (CPI) method, and expressed as a percentage of PAUC at MSIII. Using PAUC instead of total cost allowed us to separate the effect of quantity and control for it separately. Converting to 2005 dollars allowed comparison across time periods and mitigated inflation and other escalation effects. Finally, percentage growth provided the means to compare the programs side-by-side by mitigating PAUC differences (i.e. a ten percent change might mean $10B for one program and $100M for another.)

Addressing schedule, we considered several variables that would indicate program delays and found that the most significant schedule effects occurred at the end of the target phase – MSIII. Our schedule representative then was:

"Perc_MSIII_growth"

Percent MSIII growth is the difference between the initial MSIII estimate (the DE) and the actual MSIII, expressed as a percentage. As with cost, using a percentage minimized the effect of comparing very long and very short programs.

Since cost and schedule are separate but logically dependent, we took a quick look at how these variables related to each other. Figure 5 shows a multivariate scatterplot produced with JMP® 6 (SAS Institute, 2005). Correlation between the two was small (*r* = 0.2106), indicating that there were probably different factors affecting their outcomes. In addition, several programs presented themselves as potentially influential points. Programs 2, 3, 9, 24, and 35 (in the green circles) stood out and proved to be contentious throughout the analysis. While it was premature to exclude any programs at this point, we now had two specific groupings (3, 9 and 2, 24, 35) to watch.

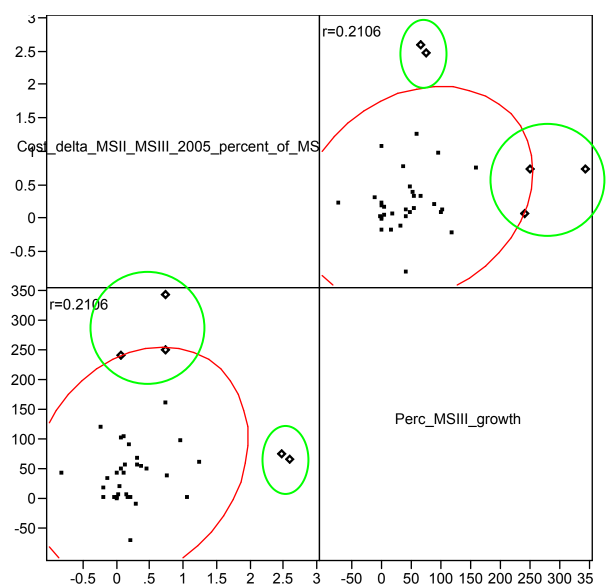


Figure 5 - Cost and Schedule comparison scatterplot

Predictor variables

From the literature review, we determined the importance of longitudinal variables and external factors, that cost and schedule factors should be normalized to level the playing field when comparing multiple programs and time periods, and that the SAR database is the most consistent data source but any valid database might yield a key piece of information. We derived a list of 172 program characteristics and variables about which we could either collect or calculate data. Of the variables extracted from the SARs, many were static, fixed program characteristics (e.g. the first SAR date). We collected data for other SAR variables multiple times for each program (e.g. the latest MSIII estimate), and they were used to calculate longitudinal variables such as the number of MSIII estimates occurring between MSII and MSIII. The remaining variables fell into the external data category and included things such as inflation rates, spending appropriations, calendar years elapsed, and controlling political party.

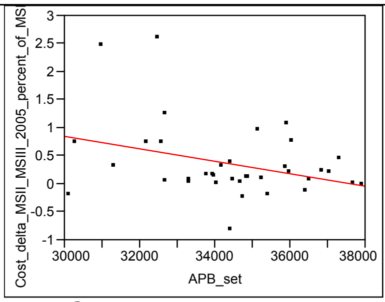
The following discussion addresses variables that in and of themselves are significant at the 0.05 level in predicting either cost or schedule, and others worth mentioning. We define each predictor variable and provide a linear fit for both cost and schedule responses in the following tables. The diagrams indicate two aspects worth mentioning. First, the dots represent the actual response of each program and show how scattered or different the responses were. Second, the line shows the response that the variable predicted. A perfect fit would have all the dots on the line so looking at how tightly the dots grouped together and their relationship to the line gives an indication of how powerful the predictor variable was. The line's slope, either negative (higher on the left) or positive (higher on the right) shows how the predictor impacts the response. For example, the first variable in Table 4, "APB_set" shows a negative slope for both cost and schedule. One could interpret this as "programs that received program approval later in time had less cost and schedule growth." To demonstrate how a variable might influence one response but not the other, we placed both cost and schedule responses side by side for each. Take care not to generalize based on these simple comparisons and keep in mind that each of these variables was considered here in isolation. When combined, their cumulative effects might be quite different.

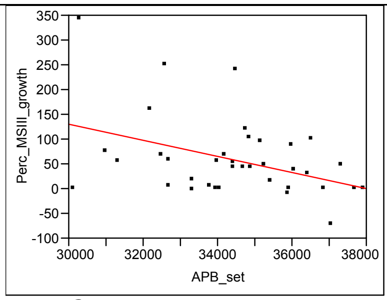
Each plot shows the individual p-value and Adjusted r^2. Adjusted r^2 estimates the proportion of the response attributable to the model (a single variable in this case) rather than error. It gives us a convenient indication of how strongly the variable and the response were linearly related. Since the correlation coefficient, r, ranges from -1 to 1, we used the square of r to remove the minus sign and convert the range to 0 to 1. A value of zero would indicate that the model was no more able to predict the response than the sample mean and a one would indicate a perfect fit. As a further measure, an algorithm adjusts r^2 downward to diminish the benefit of adding more and more input variables, which reduces the degrees of freedom (JMP® 6, SAS Institute, 2005). Therefore, it is possible to end up with a very low or even negative Adjusted r^2, either of which indicates the variable had negligible predictive strength.

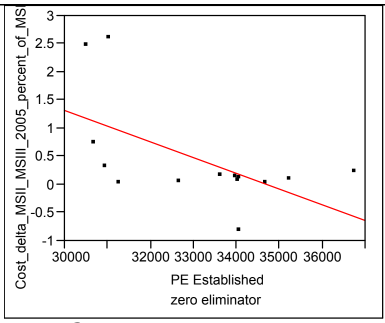
Predictor variables – Absolute dates

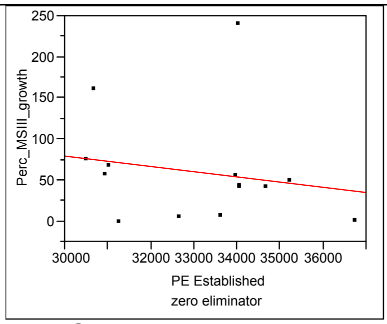
Absolute date variables were referenced to the Microsoft© Excel (2003) default day count index of January 1, 1900. For example, February 12, 2005 is equivalent to 39395. This allowed comparison within and across programs referenced to an absolute baseline.

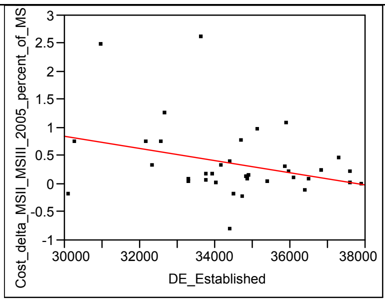
Table 4 - Predictor variables - Absolute dates

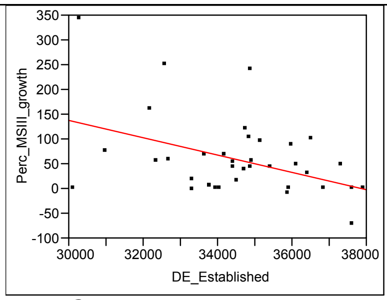
<table>
<tr><th>Variable</th><th>Cost response</th><th>Schedule response</th></tr>
<tr><td>APB_set – the date that the DE baseline was established.</td><td>

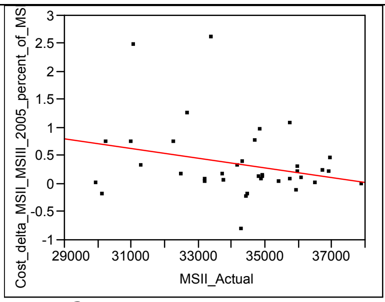
Adj r^2 = 0.08, p-value = 0.04</td><td>

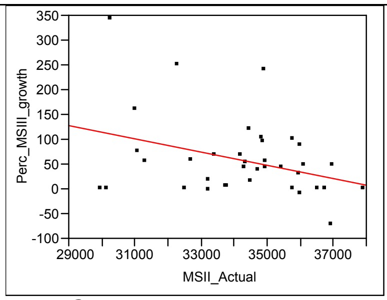
Adj r^2 = 0.14, p-value = 0.01</td></tr>
<tr><td>PE_Established – the date that the PE was established, if there was one. If a program has a PE, it is more likely to have less cost growth if the PE is established later (i.e. newer programs).</td><td>

Adj r^2 = 0.28, p-value = 0.03</td><td>

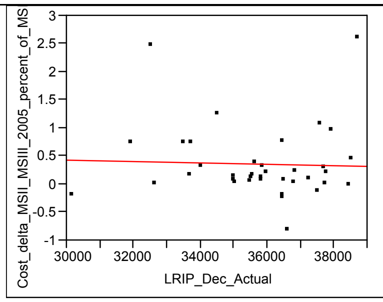
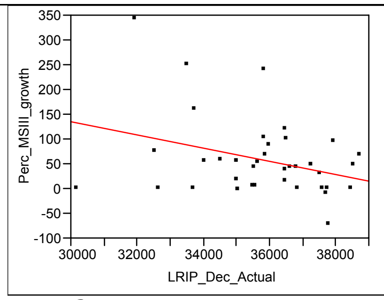
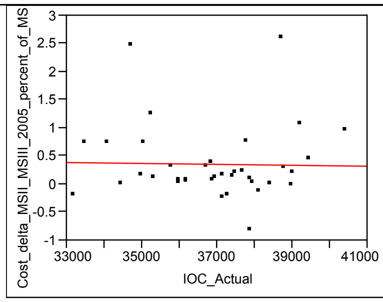
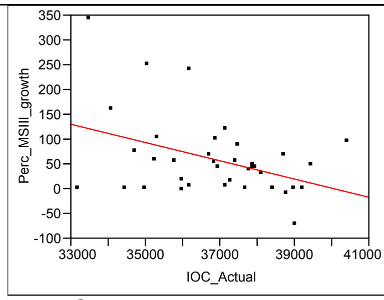
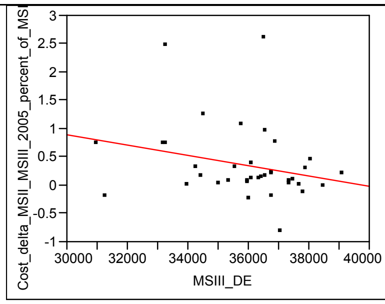
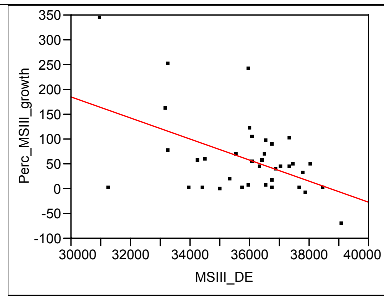
Adj r^2 = -0.04, p-value = 0.51</td></tr>
<tr><td>DE_Established – the date that the DE was established.</td><td>

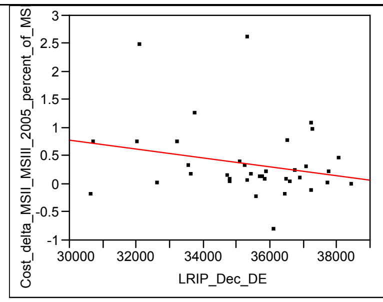
Adj r^2 = 0.08, p-value = 0.05</td><td>

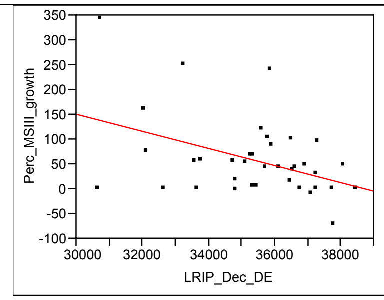
Adj r^2 = 0.15, p-value = 0.01</td></tr>
<tr><td>MSII_Actual – the actual MSII date.</td><td>

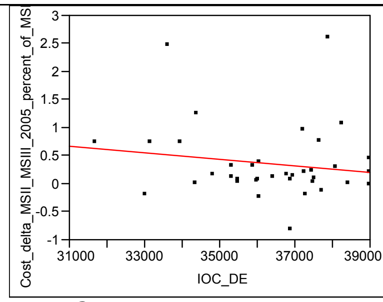
Adj r^2 = 0.04, p-value = 0.12</td><td>

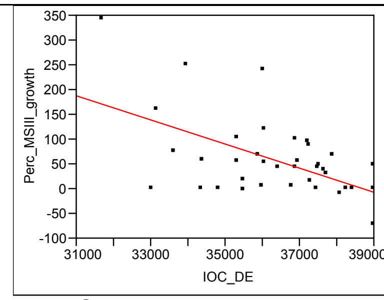
Adj r^2 = 0.09, p-value = 0.04</td></tr>
</table>

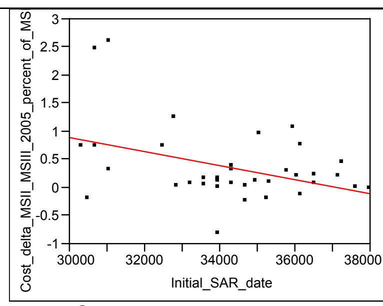
Variable	Cost response	Schedule response
LRIP_Dec_Actual – the actual LRIP decision date. This variable is fairly consistent across programs but as SARs changed over time, terminology and reporting requirements also changed. In some instances, the LRIP decision date was inferred from other information.	 Adj r^2 = -0.03, p-value = 0.83	 Adj r^2 = 0.08, p-value = 0.05
IOC_Actual – the actual IOC date. For 6 programs that had reached MSIII but not IOC, the estimated IOC date was used. Since these dates were in the near future, we assumed that they would not have changed significantly.	 Adj r^2 = -0.03, p-value = 0.90	 Adj r^2 = 0.14, p-value = 0.01
MSIII_DE – first MSIII DE, set at the initial APB.	 Adj r^2 = 0.04, p-value = 0.12	 Adj r^2 = 0.22, p-value = 0.002

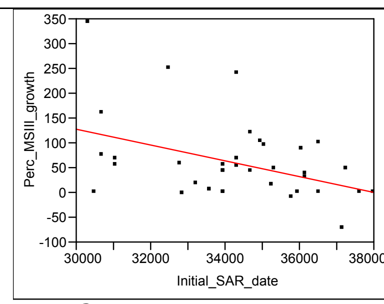
<table>
<tr><th>Variable</th><th>Cost response</th><th>Schedule response</th></tr>
<tr><td>LRIP_Dec_DE – first LRIP decision DE, set at the initial APB.</td><td>

Adj r^2 = 0.03, p-value = 0.15</td><td>

Adj r^2 = 0.16, p-value = 0.009</td></tr>
<tr><td>IOC_DE – first IOC DE, set at the initial APB.</td><td>

Adj r^2 = -0.003, p-value = 0.35</td><td>

Adj r^2 = 0.27, p-value = 0.006</td></tr>
<tr><td>Initial_SAR_date – date the first SAR was submitted. This date is the submission due date. This variable provides a good indicator of how long ago a program started.</td><td>

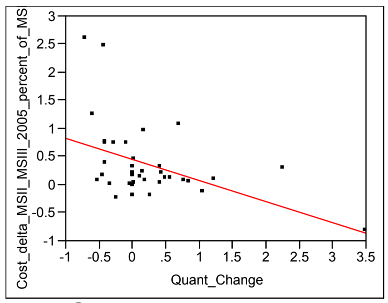
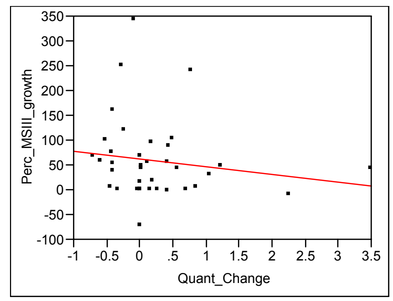
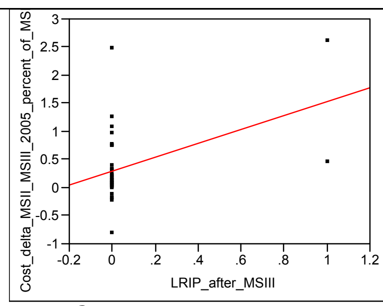
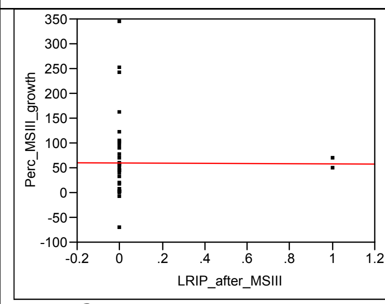
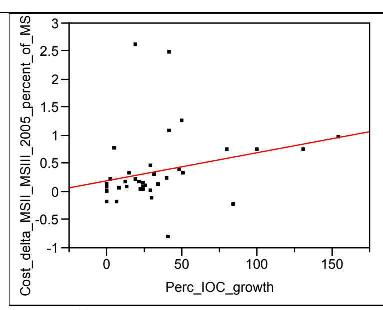
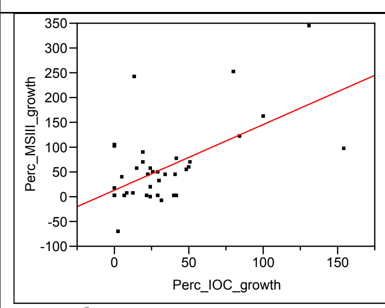
Adj r^2 = 0.13, p-value = 0.02</td><td>

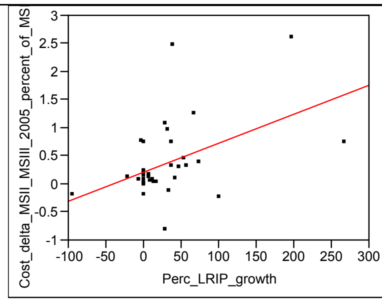
Adj r^2 = 0.15, p-value = 0.01</td></tr>
</table>

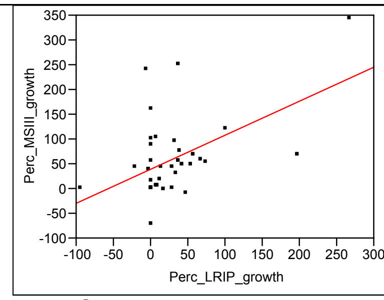
Predictor variables – Program characteristics

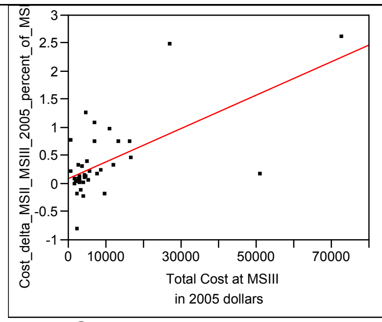
Program characteristic variables describe observable features.

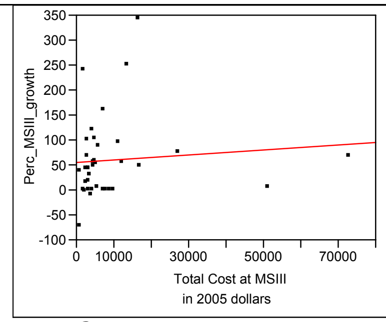
Table 5 - Predictor variables - Program characteristics

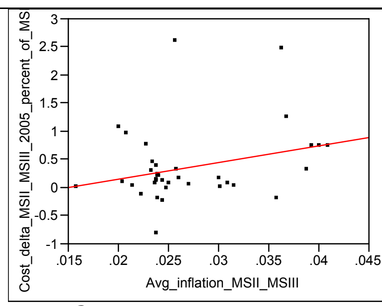
Variable	Cost response	Schedule response
Quant_Change – percentage quantity change from the estimate at MSII to the estimate at MSIII. It is interesting to note that neither the initial or final quantity were predictive (Quant_Change * 100 = percent).	 Adj r^2 = 0.18, p-value = 0.005	 Adj r^2 = -0.04, p-value = 0.37
LRIP_after_MSIII – value = 1 if LRIP occurred more than three months after MSIII.	 Adj r^2 = 0.16, p-value = 0.008	 Adj r^2 = -0.03, p-value = 0.98
Perc_IOC_growth – the difference between IOC DE and actual IOC, expressed as a percentage. This variable naturally tracks with our chosen schedule variable, Percent MSIII growth.	 Adj r^2 = 0.05, p-value = 0.11	 Adj r^2 = 0.32, p-value = 0.0002

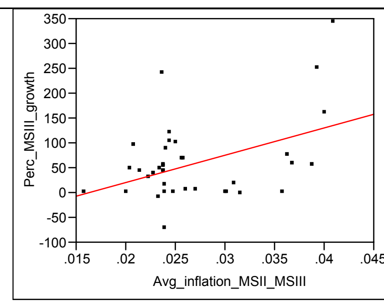
<table>
<tr><th>Variable</th><th>Cost response</th><th>Schedule response</th></tr>
<tr><td>Perc_LRIP_growth – the difference between LRIP decision DE and actual LRIP decision, expressed as a percentage. Like IOC, this variable tracks with our chosen schedule variable.</td><td>

Adj r^2 = 0.19, p-value = 0.004</td><td>

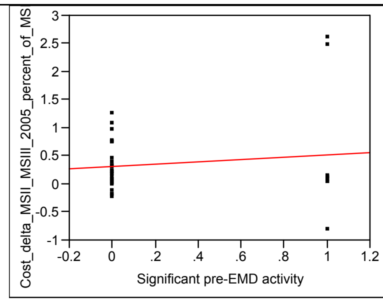
Adj r^2 = 0.24, p-value = 0.001</td></tr>
<tr><td>Total Cost at MSIII in 2005 dollars – quantity at MSIII * PAUC at MSIII, converted to 2005 dollars by the CPI. There was significant predictive ability between total cost and cost growth but an insignificant effect on schedule.</td><td>

Adj r^2 = 0.38, p-value = <0.0001</td><td>

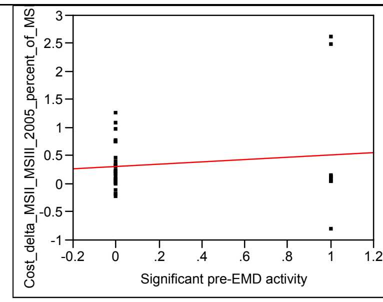
Adj r^2 = -0.02, p-value = 0.59</td></tr>
<tr><td>Avg_inflation_MSII _MSIII – the average annual inflation that occurred between MSII and MSIII.</td><td>

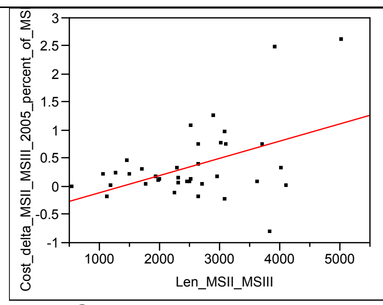
Adj r^2 = 0.06, p-value = 0.09</td><td>

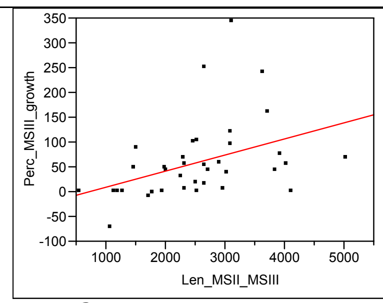
Adj r^2 = 0.16, p-value = 0.007</td></tr>
</table>

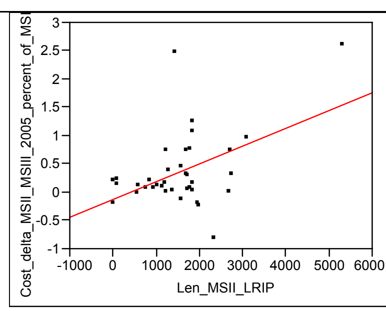
<table>
<tr><th>Variable</th><th>Cost response</th><th>Schedule response</th></tr>
<tr><td>Significant pre-EMD activity – a measure of how long the program was reported in the SARs before MSII. Value = 1 if initial SAR date was less than the actual MSII date by more than 360 days. Since most programs are not reported in the SARs until MSII, this effectively separates larger programs that were known to be MDAPs early.</td><td>

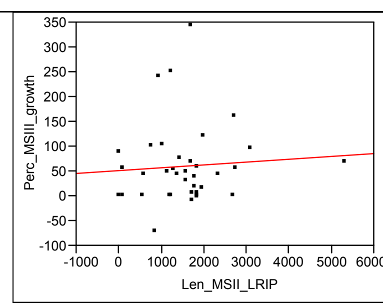
Adj r^2 = -0.007, p-value = 0.39

While this variable does not appear significant here, it is useful when combined with others in the final regression models.</td><td>

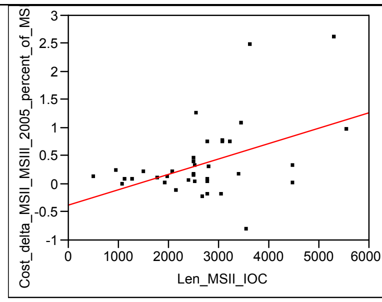
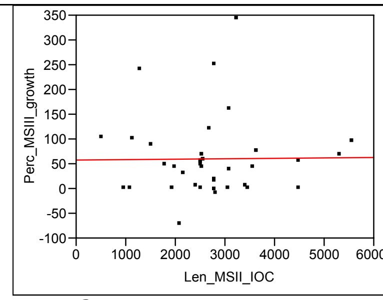
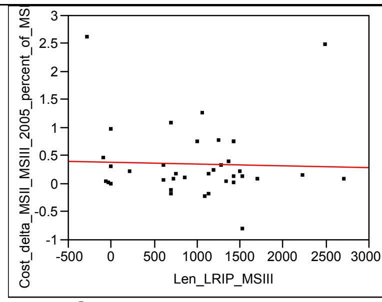
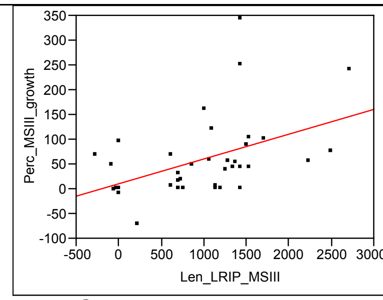
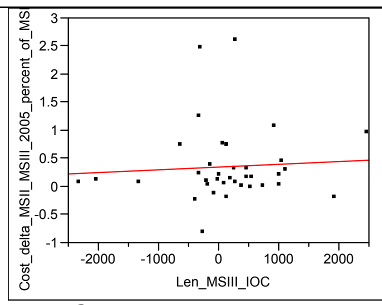
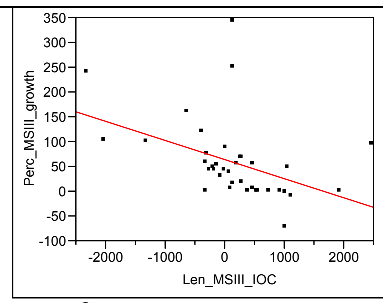
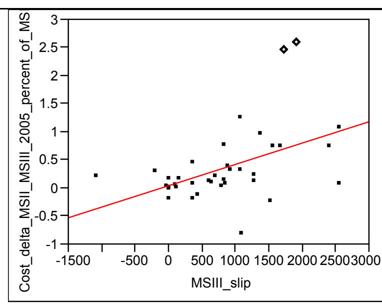
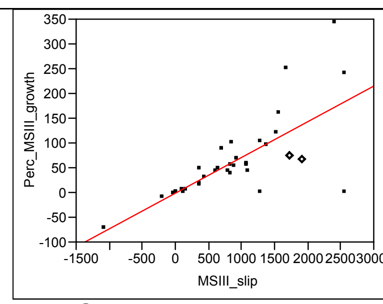
Adj r^2 = -0.002, p-value = 0.71</td></tr>
<tr><td>Len_MSII_MSIII – the length of time in days between MSII and MSIII. When programs 3 and 9 were excluded from the analysis, the cost p-value increase to 0.46, losing all significance, while the schedule p-value decreases to 0.008, further demonstrating these programs as particularly influential.</td><td>

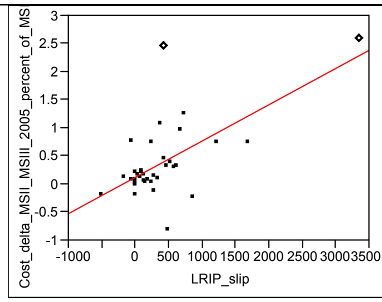
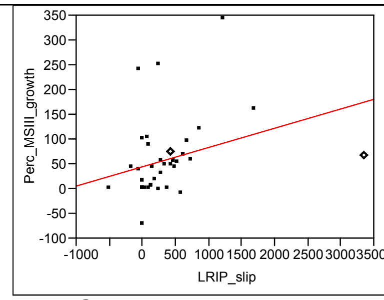
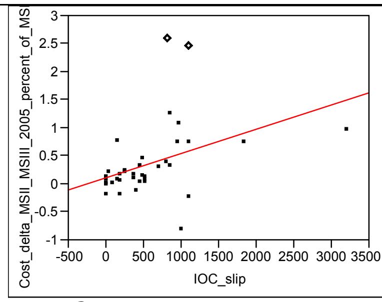
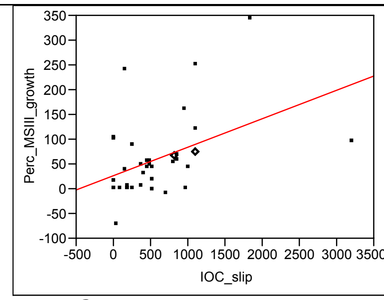
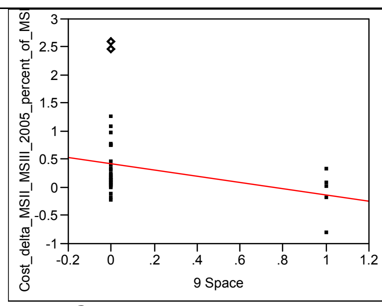
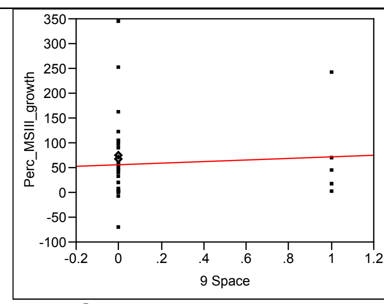
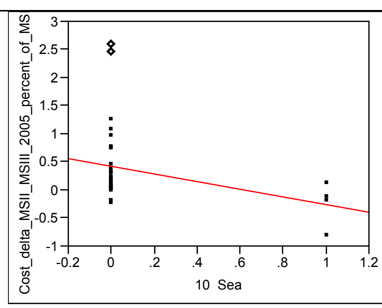
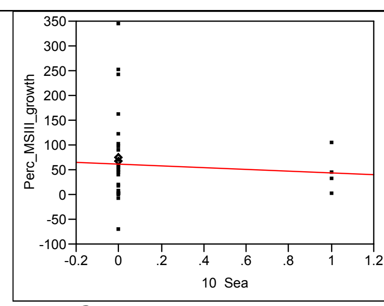
Adj r^2 = 0.18, p-value = 0.005</td><td>

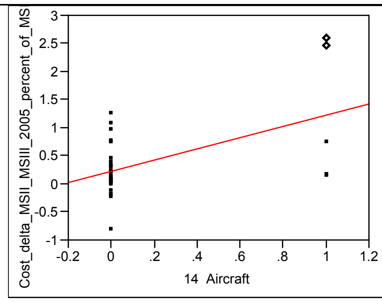
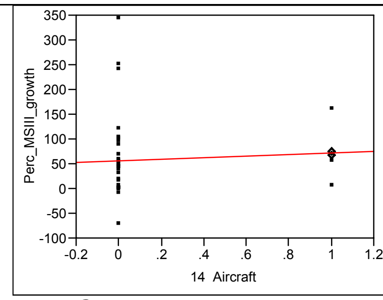
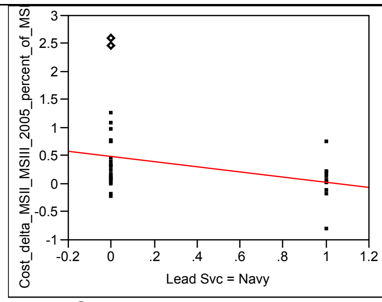
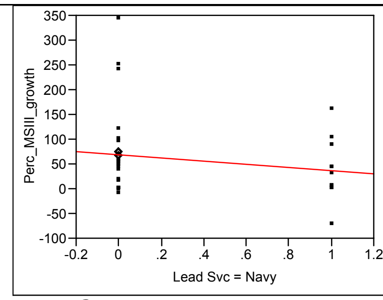
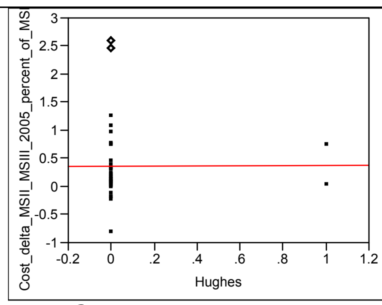
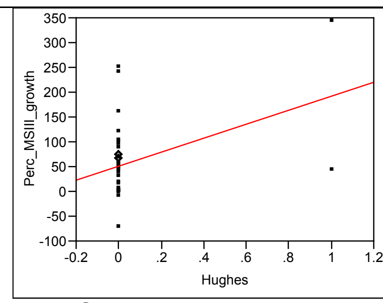
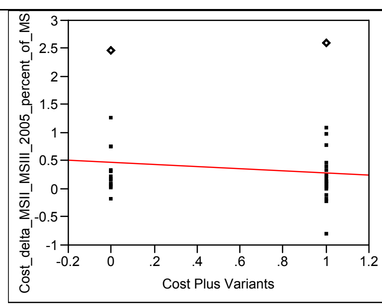
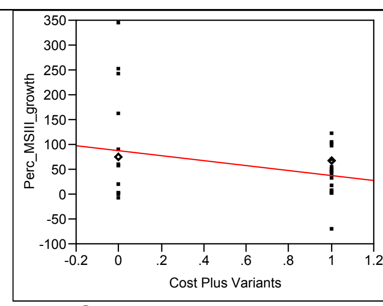
Adj r^2 = 0.14, p-value = 0.01</td></tr>
<tr><td>Len_MSII_LRIP – the length of time in days between MSII and LRIP decision.</td><td>

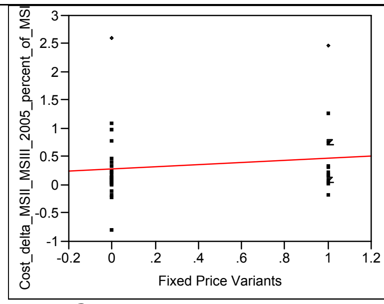
Adj r^2 = 0.20, p-value = 0.003</td><td>

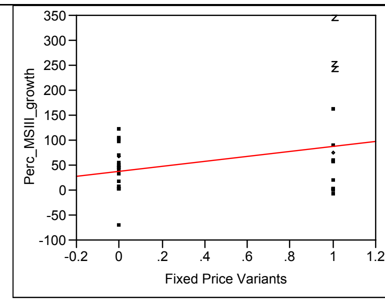
Adj r^2 = -0.02, p-value = 0.67</td></tr>
</table>

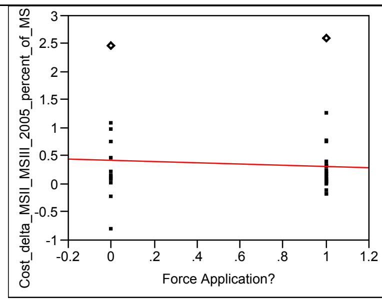
Variable	Cost response	Schedule response
Len_MSII_IOC – the length of time in days between MSII and IOC.	 Adj r^2 = 0.19, p-value = 0.005	 Adj r^2 = -0.03, p-value = 0.95
Len_LRIP_MSIII – the length of time in days between LRIP decision and MSIII.	 Adj r^2 = -0.03, p-value = 0.85	 Adj r^2 = 0.18, p-value = 0.005
Len_MSIII_IOC – the length of time in days between MSIII and IOC.	 Adj r^2 = -0.02, p-value = 0.69	 Adj r^2 = 0.16, p-value = 0.009
MSIII_slip – the difference in days between MSIII DE and MSIII actual. This variable is closely correlated with our chosen schedule variable.	 Adj r^2 = 0.20, p-value = 0.003	 Adj r^2 = 0.52, p-value = <0.0001

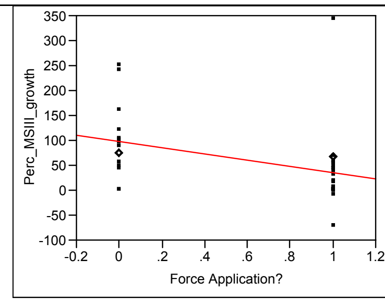
Variable	Cost response	Schedule response
LRIP_slip – the difference in days between LRIP DE and LRIP actual.	 Adj r^2 = 0.38, p-value = <0.0001	 Adj r^2 = 0.07, p-value = 0.06
IOC_slip – the difference in days between IOC DE and IOC actual.	 Adj r^2 = 0.14, p-value = 0.01	 Adj r^2 = 0.17, p-value = 0.007
9 Space – value = 1 for space-based programs such as satellites.	 Adj r^2 = 0.06, p-value = 0.08	 Adj r^2 = -0.02, p-value = 0.68
10 Sea – value = 1 for sea-based programs such as ships.	 Adj r^2 = 0.08, p-value = 0.05	 Adj r^2 = -0.02, p-value = 0.70

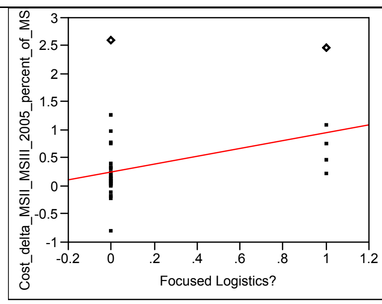
Variable	Cost response	Schedule response
14 Aircraft – value = 1 for aircraft programs.	 Adj r^2 = 0.25, p-value = 0.0009	 Adj r^2 = -0.02, p-value = 0.69
Lead Svc = Navy – value = 1 for programs assigned to the Navy for management.	 Adj r^2 = 0.08, p-value = 0.05	 Adj r^2 = 0.006, p-value = 0.27
Hughes – value = 1 if Hughes received the first contract award. The variable is suspect due to low number of Hughes contracts and their wide variance.	 Adj r^2 = -0.03, p-value = 0.96	 Adj r^2 = 0.14, p-value = 0.01
Cost Plus Variants – value = 1 if the first contract awarded was a variant of a cost-plus contract. Each program may have many different type contracts issued over time but this analysis only considers the first contract.	 Adj r^2 = -0.001, p-value = 0.40	 Adj r^2 = 0.07, p-value = 0.06

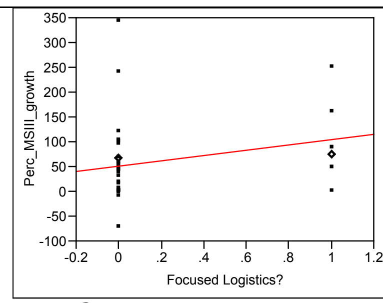
<table>
<tr><th>Variable</th><th>Cost response</th><th>Schedule response</th></tr>
<tr><td>Fixed Price Variant – value = 1 if the first contract awarded was a variant of a fixed price contract; this analysis considers only the first contract.</td><td>

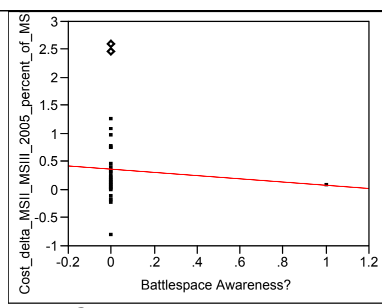
Adj r^2 = -0.008, p-value = 0.40</td><td>

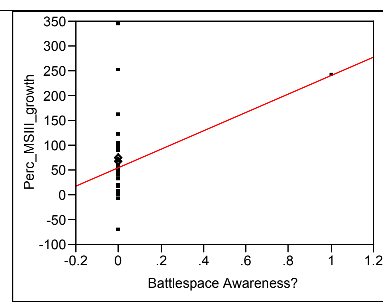
Adj r^2 = 0.07, p-value = 0.06</td></tr>
<tr><td>Force Application? – value = 1 if the program falls under the force application Functional Capability Area (FCA).</td><td>

Adj r^2 = -0.02, p-value = 0.62</td><td>

Adj r^2 = 0.13, p-value = 0.02</td></tr>
<tr><td>Focused Logistics? – value = 1 if the program falls under the focused logistics FCA.</td><td>

Adj r^2 = 0.13, p-value = 0.015</td><td>

Adj r^2 = 0.03, p-value = 0.14</td></tr>
<tr><td>Battlespace Awareness? – value = 1 if the program falls under the battlespace awareness FCA. This variable is weak since only one program fell into this category.</td><td>

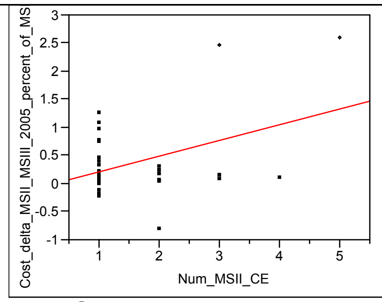
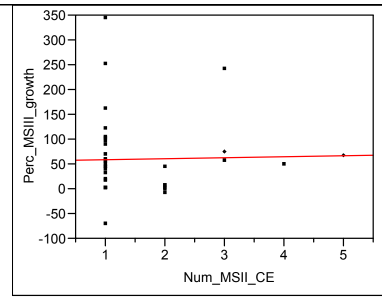
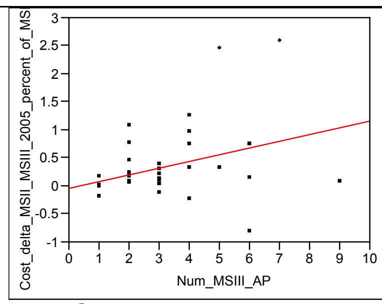
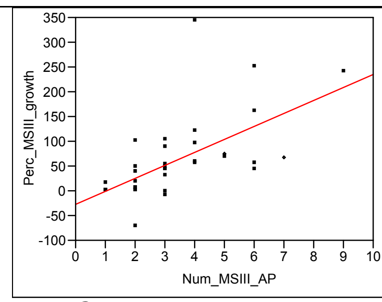
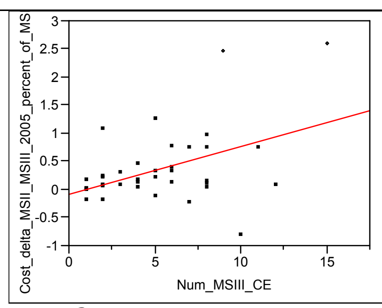
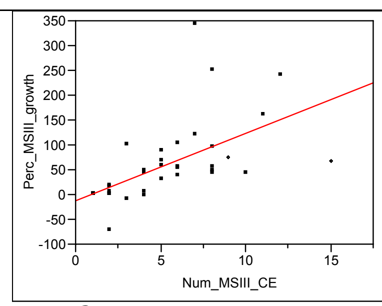
Adj r^2 = -0.02, p-value = 0.68</td><td>

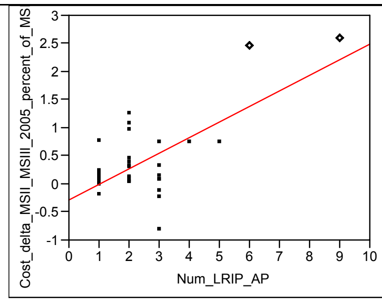
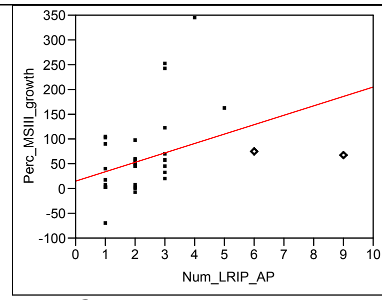
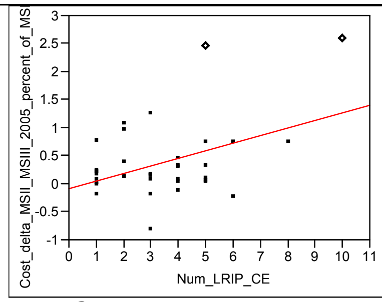
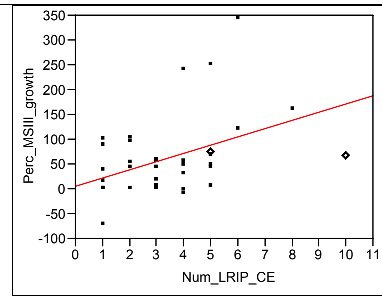
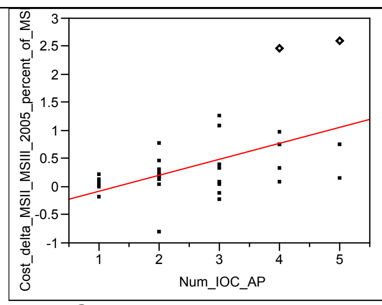
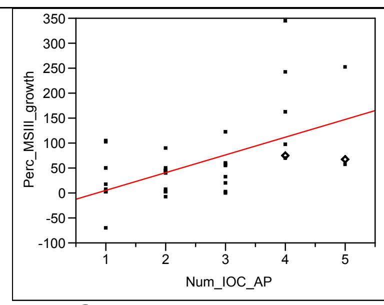
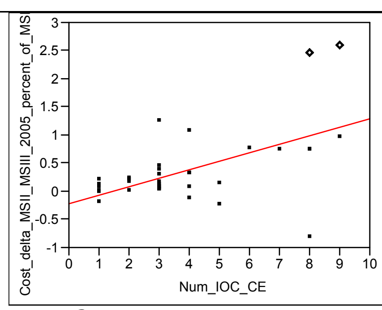
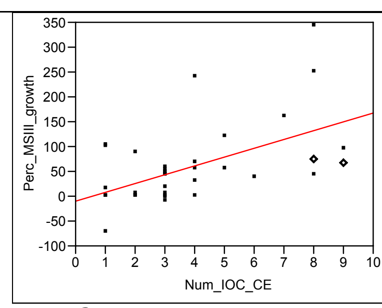
Adj r^2 = 0.12, p-value = 0.02</td></tr>
</table>

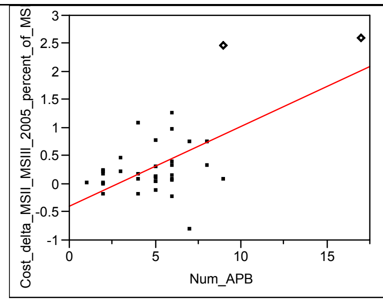
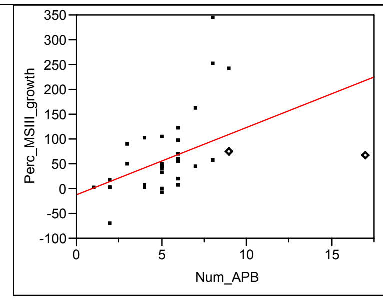
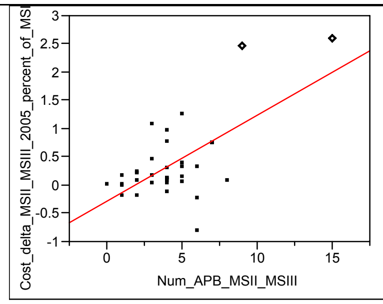
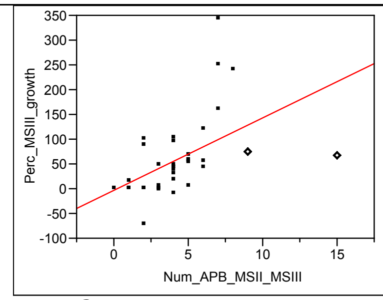
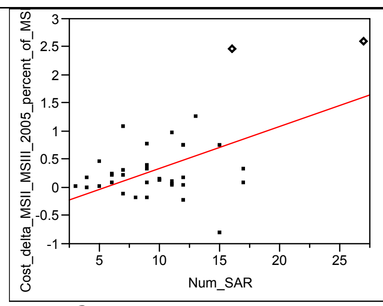
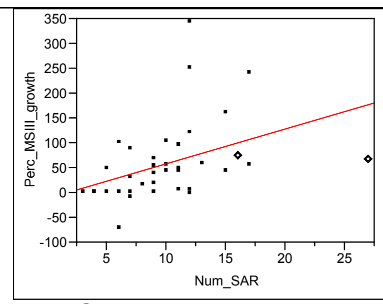
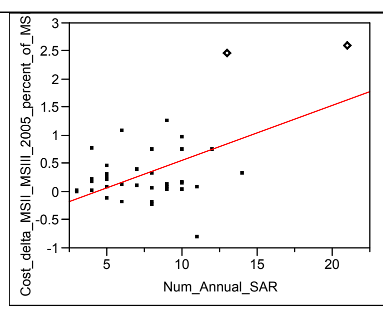
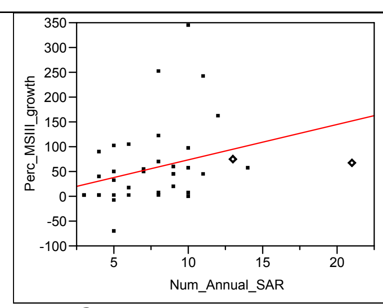
Predictor variables – Number of occurrences

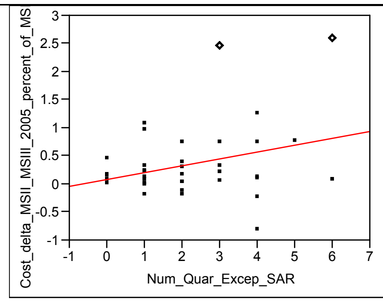
These predictor variables count the number of times something occurred, as reported in the SARs.

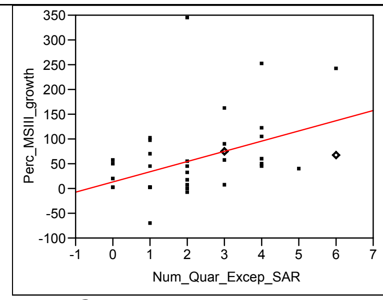
Table 6 - Predictor variables - Number of occurrences

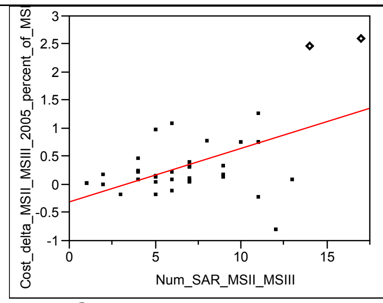
Variable	Cost response	Schedule response
Num_MSII_CE – the number of times the MSII CE changed.	 Adj r^2 = 0.14, p-value = 0.01	 Adj r^2 = -0.03, p-value = 0.88
Num_MSIII_AP – the number of times an approved program change for MSIII occurred.	 Adj r^2 = 0.10, p-value = 0.03	 Adj r^2 = 0.37, p-value = <0.0001
Num_MSIII_CE – the number of different current estimates for MSIII.	 Adj r^2 = 0.18, p-value = 0.005	 Adj r^2 = 0.32, p-value = 0.0002

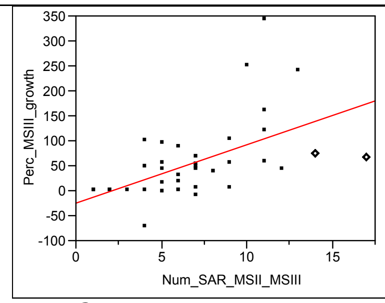
Variable	Cost response	Schedule response
Num_LRIP_AP – the number of times an approved program change for LRIP occurred.	 Adj r^2 = 0.45, p-value = <0.0001	 Adj r^2 = 0.13, p-value = 0.017
Num_LRIP_CE – the number of different current estimates for LRIP.	 Adj r^2 = 0.17, p-value = 0.006	 Adj r^2 = 0.17, p-value = 0.006
Num_IOC_AP – the number of times an approved program change for IOC occurred.	 Adj r^2 = 0.29, p-value = 0.0004	 Adj r^2 = 0.30, p-value = 0.0003
Num_IOC_CE – the number of different current estimates for IOC.	 Adj r^2 = 0.29, p-value = 0.0004	 Adj r^2 = 0.28, p-value = 0.0005

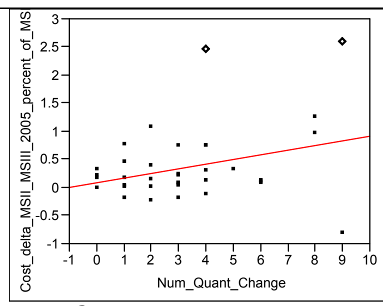
Variable	Cost response	Schedule response
Num_APB – the number of approved program baselines. More baselines tend to occur in longer, more complicated, and more volatile programs.	 Adj r^2 = 0.36, p-value = <0.0001	 Adj r^2 = 0.21, p-value = 0.002
Num_APB_MSII_MSIII – the number of approved program baselines reported from the actual MSII through the baseline reported at the same time as the actual MSIII.	 Adj r^2 = 0.40, p-value = <0.0001	 Adj r^2 = 0.24, p-value = 0.001
Num_SAR – the number of SARs submitted through the latter of MSIII, IOC.	 Adj r^2 = 0.25, p-value = 0.0009	 Adj r^2 = 0.14, p-value = 0.014
Num_Annual_SAR – the number of annual SARs (as opposed to quarterly exception SARs) submitted.	 Adj r^2 = 0.26, p-value = 0.0007	 Adj r^2 = 0.08, p-value = 0.05

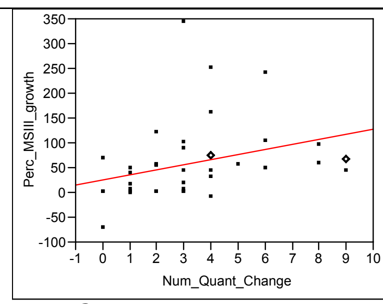
<table>
<tr><th>Variable</th><th>Cost response</th><th>Schedule response</th></tr>
<tr><td>Num_Quar_Excep_SAR – the number of quarterly exception SARs submitted. This variable is weak because an exceptional event could occur in the quarter an annual report was due, masking its importance.</td><td>

Adj r^2 = 0.06, p-value = 0.07</td><td>

Adj r^2 = 0.15, p-value = 0.009</td></tr>
<tr><td>Num_SAR_MSII_MSIII - the number of SARs submitted from the actual MSII through the submission that included the actual MSIII.</td><td>

Adj r^2 = 0.27, p-value = 0.0007</td><td>

Adj r^2 = 0.27, p-value = 0.0006</td></tr>
<tr><td>Num_Quant_Change – the number of times estimated quantity changed. This includes small changes due to a buy-to-budget philosophy.</td><td>

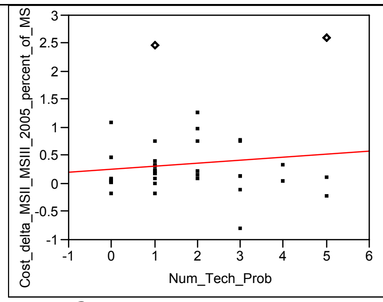
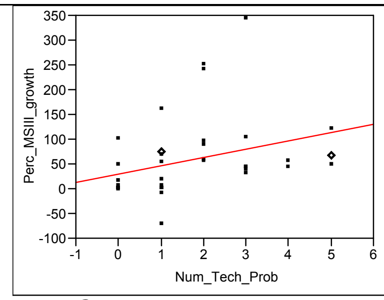
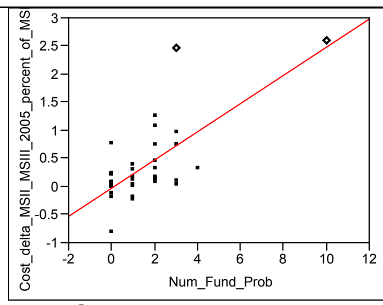
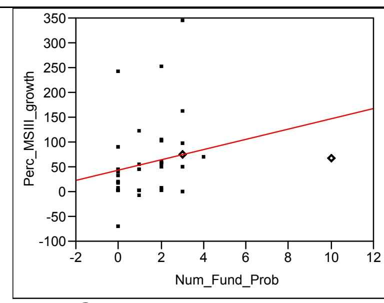
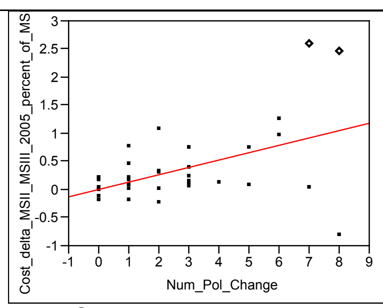
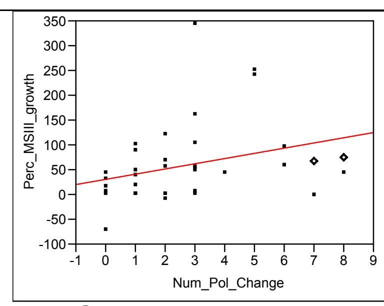
Adj r^2 = 0.07, p-value = 0.06</td><td>

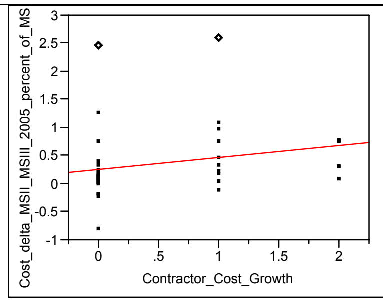
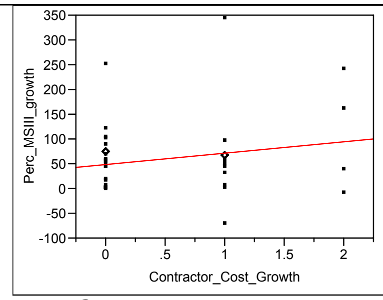
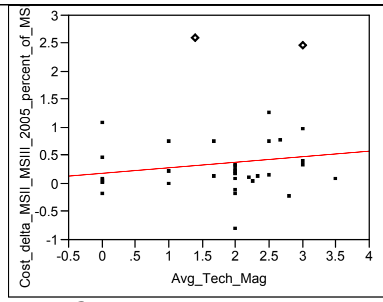
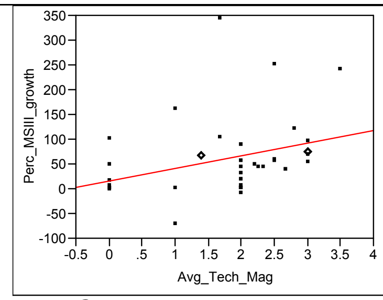
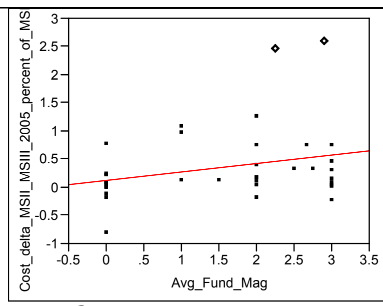
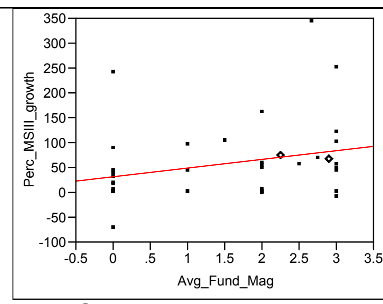
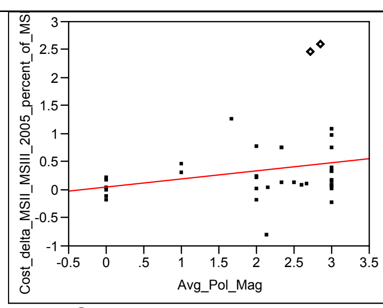
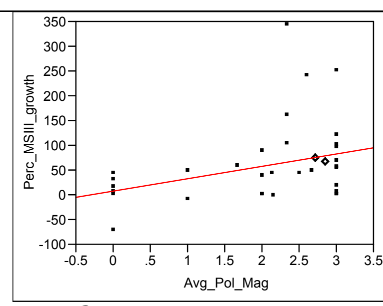
Adj r^2 = 0.07, p-value = 0.06</td></tr>
</table>

Predictor variables – Qualitative variables

The SAR narratives, including the executive summary and change explanations, provided a rich source of qualitative information. During data collection, all programs received the same treatment and only one rater completed the assessment, removing the question of inter-rater reliability. We also conducted a short dry run data collection to mitigate warm up bias at the beginning. However, there may still be undiscovered human error or system variances. We call these "soft" variables as opposed to "hard" or quantitative variables.

Table 7 - Predictor variables - Qualitative variables

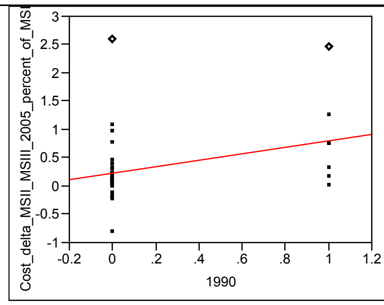
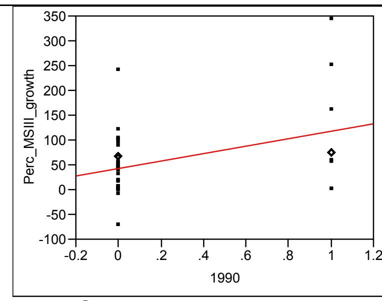
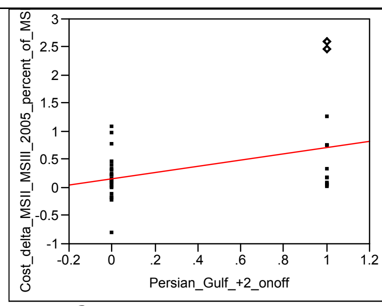
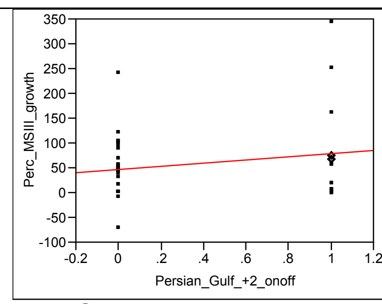
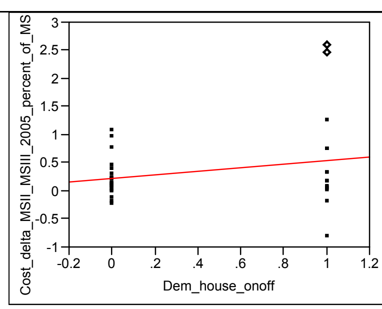
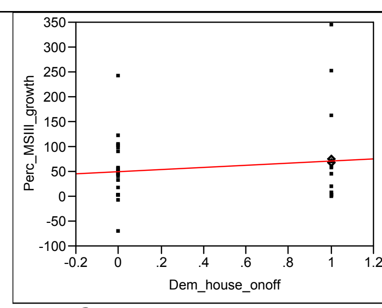
Variable	Cost response	Schedule response
Num_Tech_Prob – the number of SARs reporting unexpected technical challenges.	 Adj r^2 = -0.01, p-value = 0.48	 Adj r^2 = 0.08, p-value = 0.055
Num_Fund_Prob – the number of SARs reporting funding problems.	 Adj r^2 = 0.47, p-value = <0.0001	 Adj r^2 = 0.03, p-value = 0.15
Num_Pol_Change – the number of SARs reporting politically driven complications such as quantity changes due to new strategic direction or times of war.	 Adj r^2 = 0.20, p-value = 0.003	 Adj r^2 = 0.07, p-value = 0.06

Variable	Cost response	Schedule response
Contractor_Cost _Growth – the number of SARs reporting contractor-specific issues that generated cost growth.	 Adj r^2 = 0.02, p-value = 0.18	 Adj r^2 = 0.01, p-value = 0.22
Avg_Tech_Mag – the average magnitude of reported technical challenges. Magnitude ranges from 1 to 5 with 1 being no impact and 5 being program termination.	 Adj r^2 = -0.004, p-value = 0.36	 Adj r^2 = 0.09, p-value = 0.04
Avg_Fund_Mag – the average magnitude of reported funding problems.	 Adj r^2 = 0.05, p-value = 0.09	 Adj r^2 = 0.05, p-value = 0.10
Avg_Pol_Mag – the average magnitude of reported political changes.	 Adj r^2 = 0.03, p-value = 0.15	 Adj r^2 = 0.09, p-value = 0.04

Predictor variables – Year-referenced era variables

We assessed variables representing individual years between 1980 and 2005 along with year groupings or eras surrounding significant military and political activity. Most of these variables proved ineffective.

Table 8 - Predictor variables - Year-referenced era variables

Variable	Cost response	Schedule response
1990 – the time period between MSII and MSIII included the year 1990.	 Adj r^2 = 0.10, p-value = 0.03	 Adj r^2 = 0.13, p-value = 0.02
Persian_Gulf_+2 _onoff – the time period between MSII and MSIII included at least one of the years 1990 through 1991 plus two years (i.e. 1990 through 1993).	 Adj r^2 = 0.14, p-value = 0.01	 Adj r^2 = 0.01, p-value = 0.25
Dem_house_onoff – the time between MSII and MSIII included years when the House of Representatives was controlled by the Democratic party.	 Adj r^2 = 0.03, p-value = 0.14	 Adj r^2 = -0.01, p-value = 0.41

Predictor variables – Percent completion-referenced variables

We assessed two variables as they changed during the time between MSII and MSIII. As a reference point, MSIII represented an arbitrary 90 percent program completion. Calculations were made in 10 percent increments from 10 to 90 and the significance p-values (smaller is better, <0.05 is considered significant) were plotted in Figure 6.

The first variable considered was percent cost growth calculated as the change in PAUC. This variable converges on the cost response variable (maroon line) as expected but it is interesting to note that it is fairly predictive by 40 percent program completion. What this means is that given the change in PAUC at a point in the program, it will be more indicative of the final growth as you get further along. In this case, once you pass 40 percent of program completion, you can accurately predict your final cost growth just from the growth you have experienced so far. However, knowing your cost growth was not predictive of your schedule slippage, as shown by the cyan line that does not converge to a low p-value.

The second variable viewed in this way was the number of Approved Program Baselines (APBs). The number of APBs became predictive for cost very early on, at around 20 percent program completion (orange line). The number of APBs grew to be predictive for schedule at about 40 percent program completion (purple line). While these variables alone did not accurately predict final cost or schedule growth at the beginning of the development phase, they did show that early change indicators could accurately predict final outcomes at partial program completion. Future research could expand on this concept and help make better mid-program estimates.

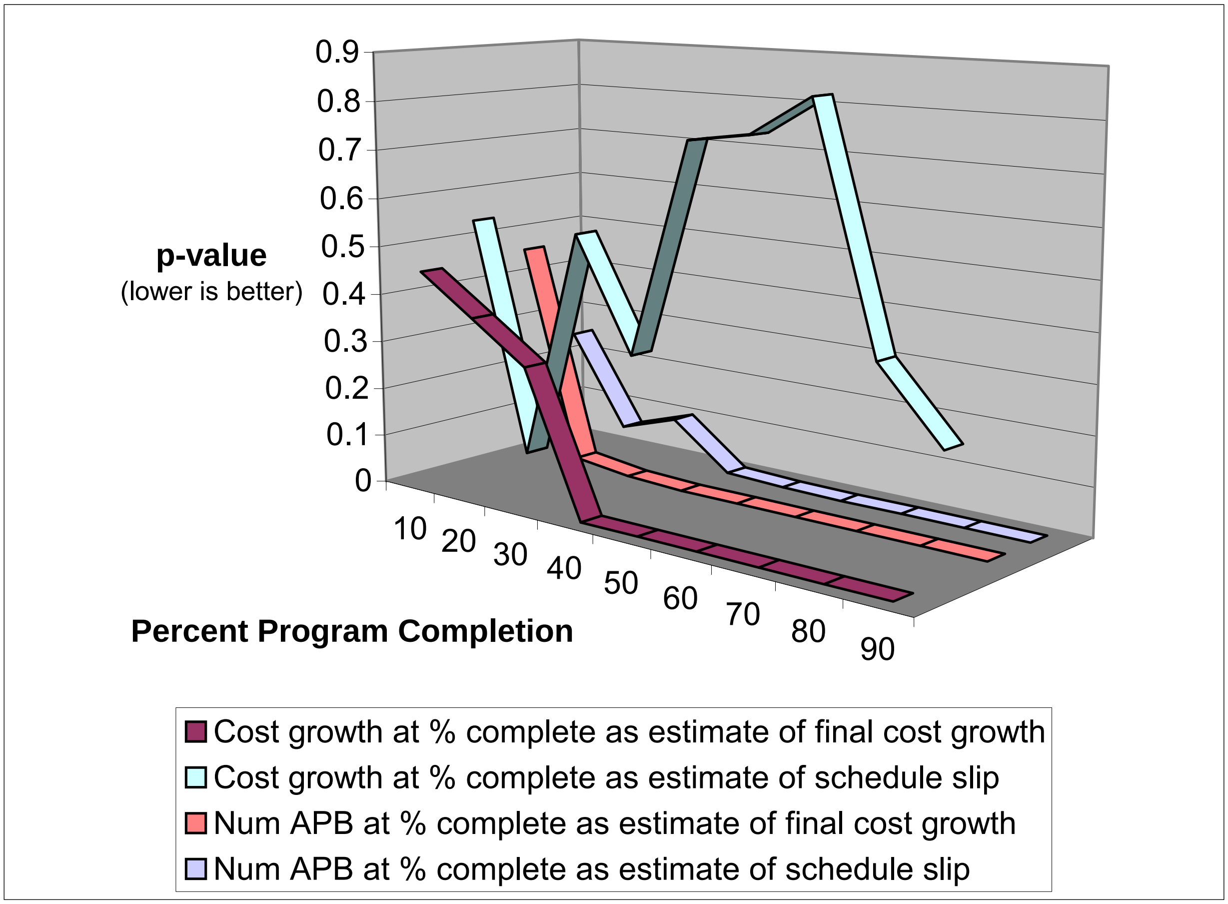
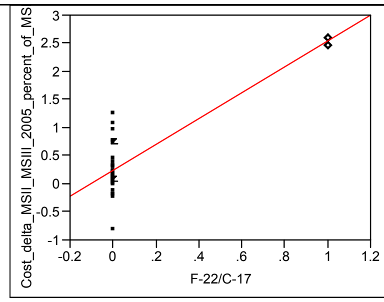
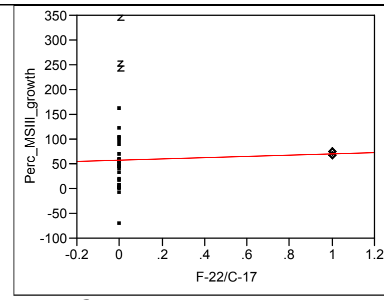
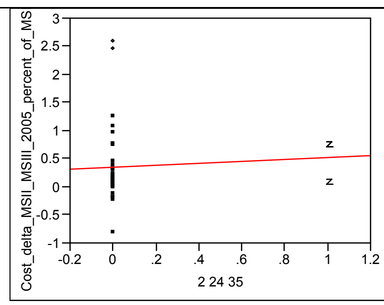
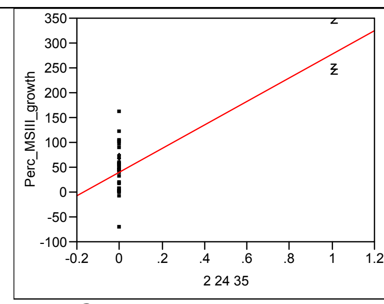
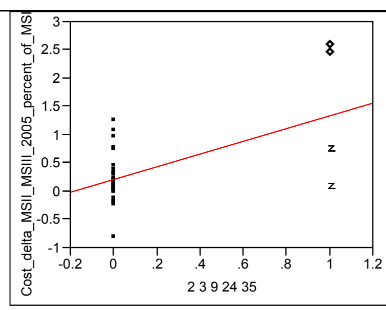
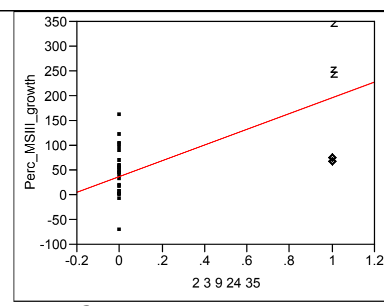


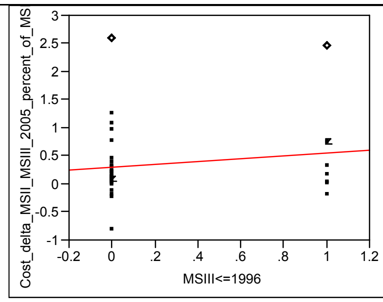
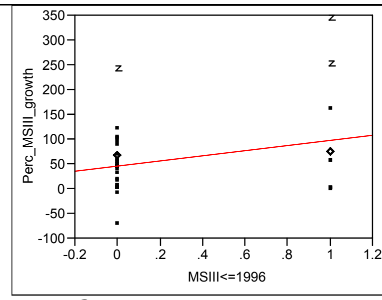
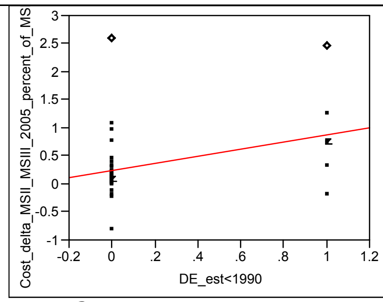
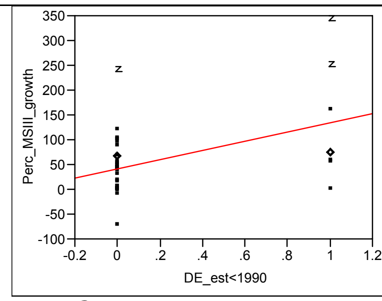
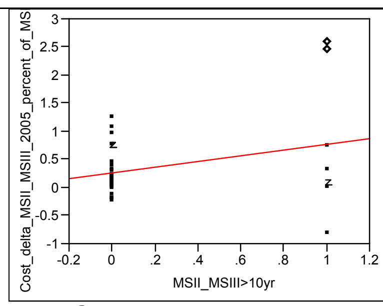
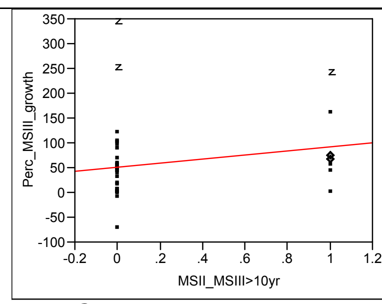
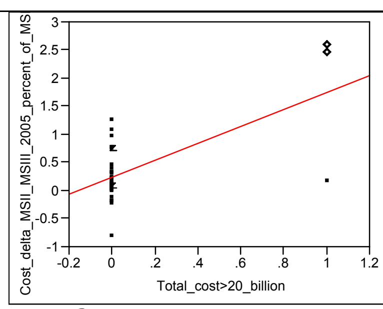
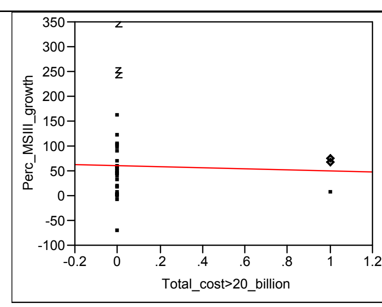
Figure 6 - P-value change for cost growth and number of APBs by percent program completion

Predictor variables – Dummy variables isolating significant program groupings

When groups of programs stand out, they can be isolated using dummy variables to assess their impact as a single entity. If the representative dummy variable proves to be predictive, the researcher can look for commonalities that might explain how the programs in this subset are similar. For example, the dummy variable "F-22/C-17" (programs 3 and 9) was very predictive in our cost models. When we assessed these programs, they stood out in program length and number of rebaselines so separating them from the rest was a logical step. The regression models contain a detailed discussion of dummy variables used.

Table 9 - Predictor variables - Dummy variables

Variable	Cost response	Schedule response
F-22/C-17 – value = 1 if program is F-22 or C-17. Used to isolate influences of these two programs. (the C-17 is program 3 and the F-22 is program 9).	 Adj r^2 = 0.63, p-value = <0.0001	 Adj r^2 = -0.02, p-value = 0.83
2 24 35 – value = 1 if program is number 2, 24, or 35. These programs were unique in that they had significant development delays and very long and/or complicated procurement cycles.	 Adj r^2 = -0.02, p-value = 0.66	 Adj r^2 = 0.66, p-value = <0.0001
2 3 9 24 35 – value = 1 if program was number 2, 3, 9, 24, or 35.	 Adj r^2 = 0.33, p-value = 0.0001	 Adj r^2 = 0.45, p-value = <0.0001

Variable	Cost response	Schedule response
MSIII<=1996 – MSIII occurred before or during the year 1996.	 Adj r^2 = 0.002, p-value = 0.31	 Adj r^2 = 0.05, p-value = 0.09
DE_est<1990 – value = 1 if the DE was established before 1990.	 Adj r^2 = 0.12, p-value = 0.02	 Adj r^2 = 0.19, p-value = 0.004
MSII_MSIII>10yr – value = 1 if the time between MSII and MSIII exceeded 3600 days.	 Adj r^2 = 0.07, p-value = 0.06	 Adj r^2 = 0.01, p-value = 0.24
Total_cost>20_billion – value = 1 if MSIII PAUC * quantity was greater than $20 billion.	 Adj r^2 = 0.38, p-value = <0.0001	 Adj r^2 = -0.03, p-value = 0.83

Regression models – Schedule

We present two models that predict schedule, the first with only three variables and no soft, qualitative variables and the second with five variables, including soft variables. There are two reasons for the distinction. First, the relatively small sample size of 37 would normally require use of a small number of predictor variables to consider a model worthwhile. Since this was an observational study of events that occurred in the past, we were not able to manipulate conditions, design how data was produced, or dictate quantity. When designing an experiment, a researcher determines sample size based upon the power and accuracy desired, with an idea of how many predictor variables will be used. This ratio of data points to explanatory variables often comes out to approximately 10 to 1 but to avoid overfitting the model at least a 5 to 1 ratio should be used (Bartlett, Kotrlik, and Higgins, 2001:46). In our case, the 10 to 1 goal meant three variables. However, during the analysis, we determined that there were often more than three unique and significant predictors. Therefore, we also offer a five-variable model that maintains a 7 to 1 ratio.

In practice, adding more variables can have the affect of decreasing the Adjusted r^2 and can artificially give credence to variables that could not otherwise stand alone. In addition, adding more variables does not always produce a significant increase in predictive capability. Variables interact with each other in a model and using more variables makes a model rigid; it becomes situation specific. In other words, the model might predict the response very well but if a program is added or removed, the complex model is more likely to fall apart. The final models are a balance between minimizing the number of variables and maximizing model effectiveness as measured by Adjusted r^2. A simple, easy to understand, and easy to use model was the goal.

The second difference between the models is the use of soft variables and there are two reasons for this. First, showing that models with and without soft variables have similar predictive capabilities provides a modicum of validation to the use of soft variables. Second, offering a model without soft variables provides the opportunity for future research to add other soft variables to a clean, quantitative model.

Schedule Model I

Schedule Model I (SM-I) used three variables to predict schedule growth with an Adjusted r^2 of 0.81 and a p-value of < 0.0001.[6] Figure 7 shows the quality of fit this model represents (the figure shows the standard r^2 as output by the analysis software but we used *Adjusted* r^2 to assess the models because it takes into account the number of explanatory variables). The three variables used in this model were “Significant pre-EMD activity,” “Num_MSIII_CE,” and “2 24 35.” The model response equation is given as:

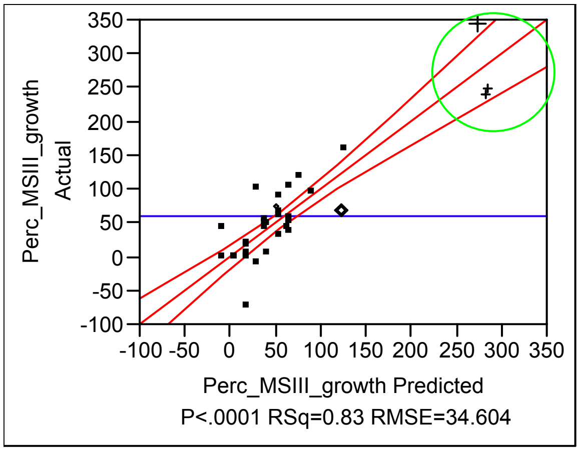


Figure 7 - SM-I Actual by predicted plot

$$\text{Percent MSIII growth} = -7.7 - 50.5 * \text{“Significant pre-EMD activity”}$$
$$+ 12.0 * \text{“Num\_MSIII\_CE”} + 193.9 * \text{“2 24 35”}.$$

[6] Statistical analysis of each model is presented in Appendix E.

As a means of comparing relative impact, we present the standardized beta coefficients:

“Significant pre-EMD activity”	–0.27
“Num_MSIII_CE”	+0.52
“2 24 35”	+0.67

We used JMP® to calculate the standardized coefficients, but the process is straightforward. To standardize, subtract the sample mean of a given variable then divide by its standard deviation. This effectively puts the input variables on a common scale that shows their relative significance by direct comparison. From the resulting coefficients, you could say that the dummy variable “2 24 35” had the most impact, being more than twice as strong as “Significant pre-EMD activity” (and in the opposite direction) but only slightly more than “Num_MSIII_CE”. Implementing the model, however, requires use of the non-standardized coefficients as given in the response equation.

Looking at the parameter estimates, we found “Significant pre-EMD activity” had a negative influence, meaning that programs submitting SARs before MSII exhibit less schedule growth between MSII and MSIII. The other two variables had a positive influence. “Num_MSIII_CE” is a measure of how many times the MSIII estimate changed and indicates program volatility and development length. The “2 24 35” dummy variable groups three programs together (green circle in Figure 7), selected by their affect on the model. Once this variable was introduced, the residuals behaved more appropriately and the model fit, as represented by the Adjusted r^2, increased by approximately 0.1.

To justify this grouping, we went back to the SARs and looked for commonalities that made these programs stand out from the others. The three programs, numbers 2 – Advanced Medium Range Air-to-Air Missile (AMRAAM), 24 – National Airspace System (NAS), and 35 – Family of Medium Tactical Vehicles (FMTV) did not, on the surface, appear similar. However, digging deeper, they demonstrated similar qualities. First, they were all complicated programs. The AMRAAM had an extended development cycle with a seven-year MSIII slip. Multiple changes and modernizing steps were added throughout the program. The NAS SARs tracked four different timelines and included multiple duplicate milestones. The FMTV actually represents a grouping of vehicles of different types: 2 ½ and 5 ton trucks, tractors, vans, wreckers, etc. and also experienced a seven-year MSIII slip. All three programs also had extended procurement cycles (33, 19, and 32 years respectively). Turning to our statistical analysis, these programs stood out as having a high number of IOC estimates, indicating program volatility, and they had a high number of SARs between MSII and MSIII, again indicating volatility as well as length of development. A program manager supervising a complicated program with an anticipated long EMD, many product variants, and lengthy production, should consider using this dummy variable when predicting schedule slip.

Implementing this model requires only three pieces of information about the program: 1) will significant pre-EMD activity occur (i.e. more than 360 days between the initial SAR report and MSII), 2) how many times will the MSIII estimate change as reported in the SARs, and 3) does this program look like the programs included in the dummy variable “2 24 35”? To demonstrate, we will answer these questions with yes, 6, and no.

With these answers, the predicted response would be:

Percent MSIII growth = –7.7 – 50.5 * 1 + 12.0 * 6 + 193.9 * 0 = 13.8%.

Before we could place confidence in this result, we had to assess the model for trustworthiness. From the model's Adjusted r^2 and p-value, it appeared to be accurate so we searched for problems that might have made us question the results. First, we checked for multicolinearity among the variables using the Variance Inflation Factor (VIF). While some research indicates a VIF of less than 10.0 is acceptable, we targeted a VIF of less than 2.0 to avoid having too much overlap in explanatory power between the variables (Neter, et al., 1996: 387). The highest VIF in this model was 1.7.

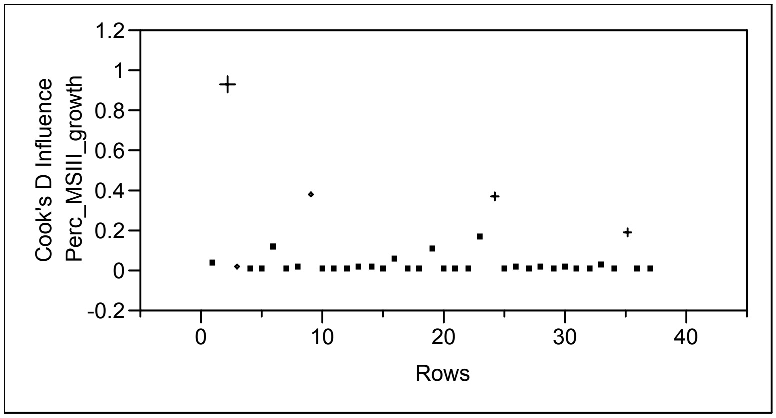


Figure 8 - SM-I Cook's Distance

Next, we assessed influential data points via Cook's Distances. Cook's Distance considers a single program's influence on the model. If the result was less than about 25 percent, the individual program did not have a significant impact on the model. If the value approached 50 percent, the program ~~had a significant effect and the model could be~~ substantially different without it (Neter, et al., 1996:381). Figure 8 shows three programs of concern with scores >0.25. To determine if these three programs made an unacceptable impact, we excluded them from the model and re-ran the analysis. The new p-value was still <0.0001, indicating that the programs did not unduly skew the results, and that their absence would not have changed our conclusions. The programs were therefore included for the remainder of the analysis.

Once the model passed these checks, we analyzed the statistical assumptions of independence, normality, and constant variance. As discussed in Chapter III, we assumed independence based upon inherent differences and separation between programs. However, we were able to conduct statistical tests for normality and constant variance. Using descriptive measures, we looked at the distribution of the studentized residuals (Figure 9). The distribution appeared normal so we conducted a Shapiro-Wilk (S-W) goodness-of-fit test to confirm (Neter, et al., 1996:111). The test revealed a p-value of 0.69, indicating normality (<0.05 would indicate that the hypothesis of normality failed). Next, we addressed constant variance with the Breusch-Pagan (B-P) test (Neter, et al., 1996:115). This test resulted in a p-value of 0.27. Like the S-W test, a p-value of <0.05 would have indicated failure. Finally, we moved on to model validation.

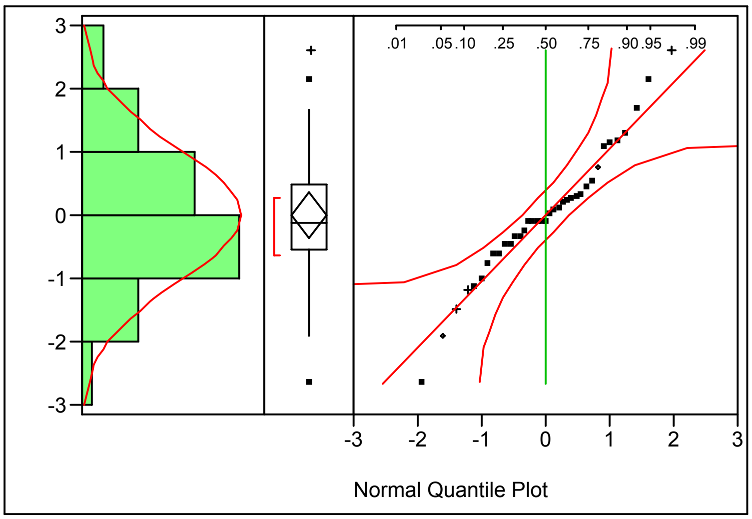


Figure 9 - SM-I Assumption of Normality

Due to the small sample size, we used the entire database to build the regression models, leaving us without the possibility of reserving a portion of the database against which to validate. Therefore, we adopted a variation of the jackknife technique pioneered by Tukey (1987:30). This technique determines if a subset of the data might act differently than the whole sample, giving us an idea of validity in the same way that reserving a portion of the data for comparison would.

Implementing the jackknife procedure, we used JMP® to calculate an individual Prediction Interval (PI) for each program. A "1" was then assigned to each program if its response variable was within the 95 percent PI, "0" if not. Next, we randomly ordered the programs and, using a portion size of eight (approximately 20 percent), computed an average of the number of "1's" that occurred in that portion. Then, the portion was incremented and a new average was calculated until all combinations were complete. Finally, a mean and standard deviation were calculated from the results of all possible portion averages and a Confidence Interval (CI) was established.

For SM-I, the CI was from 0.95 to 0.99. Therefore, we can say with 95 percent confidence that given any eight randomly selected programs, the model correctly predicts the amount schedule growth between 95 and 99 percent of the time. However, this outcome must be tempered with the fact that the model results were also based on a 95 percent PI, compounding the potential error. Regardless, with a p-value of <0.0001, an Adjusted r^2 of 0.81, and a 95+ percent confidence in the results, the model proved to be quite effective.

Schedule Model II

Schedule Model II offers five variables to explain schedule slippage. The result was a p-value of <0.0001 and Adjusted r^2 of 0.85 (see Figure 10). This model includes the same dummy variable "2 24 35" as SM-I (green circle) as well as "Significant pre-EMD activity." The remaining three variables were "MSIII before IOC?," "Num_Fund_Prob," and "Force Application?" The resulting model formula is:

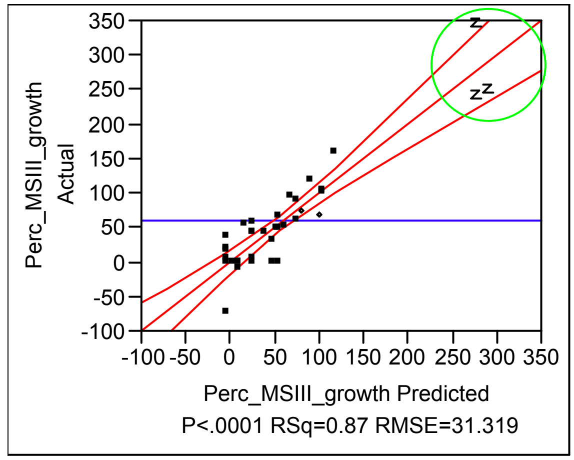


Figure 10 - SM-II Actual by predicted plot

Percent MSIII growth = 74.1 – 50.5 * "MSIII before IOC?"

– 36.4 * "Significant pre-EMD activity" + 14.3 * "Num_Fund_Prob"

– 28.9 * "Force Application?" + 235.0 * "2 24 35".

The standardized beta coefficients are:

"MSIII before IOC?"	–0.31
"Significant pre-EMD activity"	–0.20
"Num_Fund_Prob"	+0.33
"Force Application?"	–0.18
"2 24 35"	+0.81

We assessed this model in the same manner as before (highest VIF score = 1.3, Cook's Distances passed scrutiny, S-W p-value = 0.54, B-P p-value = 0.09) and validation was successful with a jackknife CI of 0.95 to 0.99.

In this grouping of predictors, “Num_MSIII_CE” proved to be less significant. In its place, we found “Num_Fund_Prob” and two new hard variables, “MSIII before IOC?” and “Force Application?” First, “Num_Fund_Prob” serves as a count of SARs reporting funding problems. A funding problem could have been a simple comment in the executive summary that the President’s budget cut program spending or a direct reference to cuts that caused quantity decreases. As a soft variable, however, we attempted to isolate its influence by creating quantitative variables that might embody the same information. For example, we used defense appropriations to link political changes. The second new variable we found significant was “MSIII before IOC?” Similar variables have presented themselves in previous research, lending added credibility (Monaco 2005:106). Finally, we found the FCA category “Force Application?” to be predictive.

All variables except for “Num_Fund_Prob” have a yes or no (1 or 0 in the model formula) response. For example, if you answer yes to the question of whether MSIII occurs before IOC, then the model will predict less schedule growth. The same is true if there is significant pre-EMD activity or if the FCA is Force Application. On the other hand, if the new program exhibits characteristics like programs represented by the dummy variable “2 24 35,” there will be significant schedule slippage (subjectively ~200 percent). “Num_Fund_Prob” is not a simple yes or no variable but rather, it provides a range of impact; more funding problems indicate more schedule problems.

Regression models – Cost

We started our analysis of the cost regression models in the same manner as the schedule models. The list of input variables and stated assumptions were the same but to tell the whole story, we have included three cost models. As with schedule, there is a soft variable model, Cost Model I (CM-I), and a hard variable model, CM-II, but this time each required four variables (a 9 to 1 ratio) to achieve similar predictive ability. The third model, CM-III, adds "Significant pre-EMD activity" to CM-I and demonstrates a common tie to the schedule models.

Cost Model I

Cost Model I uses four variables; two soft: "Num_Pol_Change," "Contractor_Cost_Growth," one hard: "Quant_Change," and one dummy: "F-22/C-17." The model predicts cost growth with a p-value of <0.0001 and Adjusted r^2 of 0.80. Assessment confirmed that the model was valid (highest VIF score = 1.6, Cook's Distance passed, S-W p-value = 0.21, B-P p-value = 0.44). Validation yielded a CI of 0.95 to 0.99. Figure 11 shows the model's fit and the formula is given as:

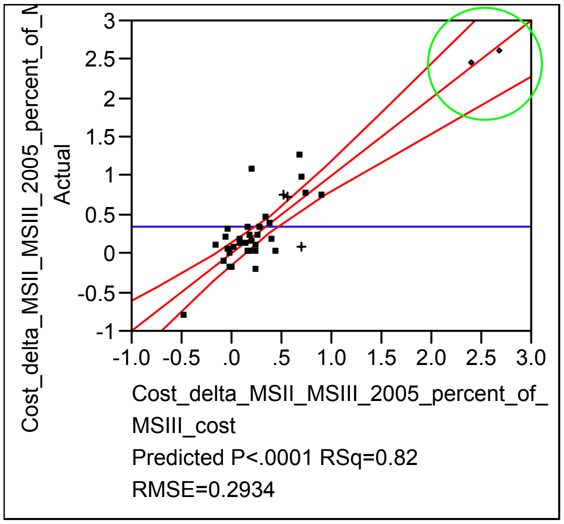


Figure 11 - CM-I Actual by predicted plot

Percentage cost growth = – 0.012 – 0.32 * "Quant_Change" + 0.08 * "Num_Pol_change" + 0.26 * "Contractor_Cost_Growth" + 1.61 * "F-22/C-17".

The standardized beta coefficients are:

"Quant_Change"	–0.39
"Num_Pol_change"	+0.29
"Contractor_Cost_Growth"	+0.27
"F-22/C-17"	+0.56

"Num_Pol_Change" is a count of the SARs reporting a specific politically driven change to the program. For example, policy changes that affect how we fight future wars might cut spending for out-of-date weapon systems that are still in development. The model showed that a higher number of changes correlated to higher cost growth. The variable "Contractor_Cost_Growth" counts how many SARs report cost growth directly attributable to the contractor. This variable, more than the other soft variables, depends upon accurate reporting by the program manager and is therefore less trustworthy. However, this model demonstrates its strength when compared to CM-II.

Quantity change did not play much of a role in schedule slippage between MSII and MSIII but it proved significant in predicting cost, showing up in all three models. Since our cost measure was per unit total cost, it was insulated from the changes in total program cost due simply to buying more items. Therefore, the relationship between "Quant_Change" and cost growth per unit reflects the overhead and manufacturing losses incurred by reducing the number of units and losing the efficiency of long production runs.

Finally, the dummy variable "F-22/C-17" groups these two programs similar to the "2 24 35" variable in the schedule models. However, we needed not look beyond statistical measures to justify this grouping. These two programs stood out, almost by themselves, in several areas including program age, length, cost, number of rebaselines, number of funding problems, number of political changes, quantity reductions, and total cost over $20B. Program managers of these types of programs could expect close to 250 percent cost growth by the time they reach MSIII. The green circle in Figure 11 shows their position relative to the other programs and the standardized coefficients indicate that this variable was almost twice as influential as the others.

Cost Model II

This model employs all hard variables (plus the F-22/C-17 dummy variable) and draws a firm correlation between changes in schedule and cost. The model was assessed for assumption compliance and validated (p-value <0.0001, Adjusted r^2 = 0.80, highest VIF score = 1.3, Cook's Distance passed, S-W p-value = 0.09, B-P p-value = 0.88). Validation yielded a CI of 0.92 to 0.97. Figure 12 shows model fit and the location of the F-22 and C-17 programs (green circle). The model formula is given as:

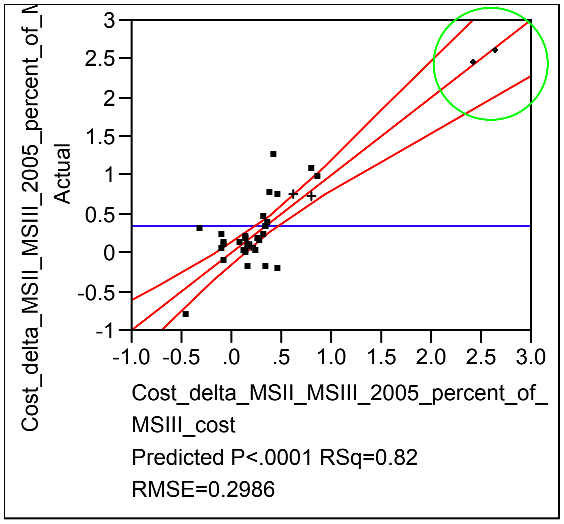


Figure 12 - CM-II Actual by predicted plot

Percentage cost growth = – 0.035 – 0.22 * "Quant_Change" + 1.84 * "F-22/C-17" + 0.00018 * "Len_MSIII_IOC" + 0.00029 * "MSIII_slip".

The standardized beta coefficients are:

"Quant_Change"	–0.27
"F-22/C-17"	+0.64
"Len_MSIII_IOC"	+0.25
"MSIII_slip"	+0.36

As in CM-I, we used four variables: "F-22/C-17," "Quant_Change," "Len_MSIII_IOC," and "MSIII_slip." We have already discussed the first two and their affect was similar here. Turning to the MSIII-related variables, "Len_MSIII_IOC" is a measure of time between the declared MSIII and IOC. This variable showed a weak but positive effect on cost growth. One could say that the longer a period of time between MSIII and IOC, the higher the chance of cost growth. "MSIII_slip" also demonstrated a positive correlation, which makes logical sense: if you spend longer trying to figure out how to make something, the chances are good you estimated the costs incorrectly (probably too low) when you started. These two variables imply that when schedule gets drawn out, program cost goes up and since the comparison is in base year dollars, growth over time is due to something other than escalation. Both the hard and soft model developed approximately the same predictive capability and increasing the number of variables in either model yielded minimal gains. However, combining hard and soft variables did improve the results.

Cost Model III

This model shows the advantage of combining hard and soft variables by adding "Significant pre-EMD activity" to model CM-I. The model performed well, resulting in a p-value = <0.0001 and Adjusted r^2 = 0.84 (see Figure 13, F-22 and C-17 programs in green circle). The assumptions passed with no complications (highest VIF score = 1.8, Cook's Distance passed, S-W p-value = 0.17, B-P p-value = 0.49) and the resulting jackknife CI was 0.95

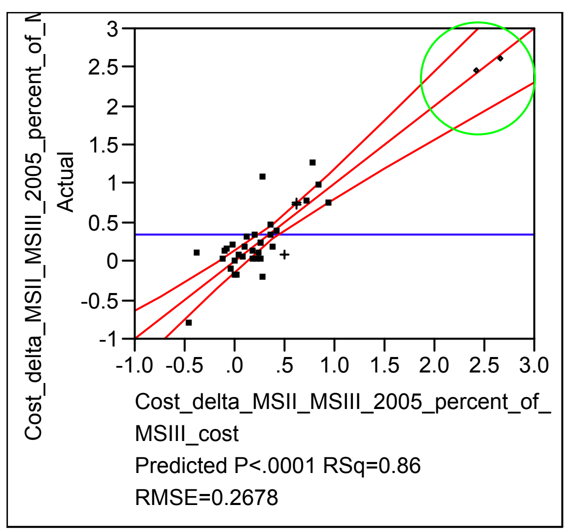


Figure 13 - CM-III Actual by predicted plot

to 0.99. Adjusted r^2 improved, but this increase required the addition of a fifth variable, creating a 7 to 1 ratio.

The model formula is:

Percentage cost growth = – 0.0099 – 0.27 * “Quant_Change”

- 0.36 * “Significant pre-EMD activity” + 0.11 * “Num_Pol_change”

+ 0.26 * “Contractor_Cost_Growth” + 1.83 * “F-22/C-17”.

The standardized beta coefficients are:

“Quant_Change”	–0.33
“Significant pre-EMD activity”	–0.24
“Num_Pol_change”	+0.38
“Contractor_Cost_Growth”	+0.27
“F-22/C-17”	+0.63

We have already discussed these variables in conjunction with the other models but it is important to note that “Significant pre-EMD activity” is now common to both schedule and cost models. It seems that pre-MDAP programs that spend more time before MSII have less schedule slippage and less cost growth. At the beginning of this chapter, we pointed out that there seemed to be different predictors for schedule and cost but this result shows at least some overlap.

Chapter Summary

This chapter addressed the detailed analysis of significant individual variables as well as regression models for both schedule and cost. Next, Chapter V addresses our conclusions, recommendations for implementation, and ideas for future research.

V. Conclusions and Recommendations

This thesis recounts our efforts to expand this stream of cost analysis research with longitudinal variables, addressing schedule slippage and cost growth of major acquisition programs. Prior research pointed the way to this longitudinal approach and methodology as demonstrated in the literature review. Our analysis addressed individual variables, one at a time, to explore their impact on the chosen schedule and cost response variables. Lastly, we used standard statistical techniques to derive regression models that correlated select input variables to our response variables and discussed model accuracy, validity, and meaning.

It is difficult to single out one answer to the question of "how much schedule slippage or cost growth will I have?" Statistical analysis can explain correlations but is less adept at showing causality. However, starting with a clean slate as in this research, we looked to any source of valid data for input variables. This was not an exhaustive effort but it was comprehensive and used both well-documented sources such as the SARs and other valid sources such as defense spending data and the consumer price index. This research also looked at static variables, those that did not change over time, and dynamic or longitudinal variables that did change throughout a program's execution. The result of this effort was a list of 172 variables for each of the 37 programs meeting entry criteria.

All five resulting models were effective, demonstrated by an Adjusted r^2 in excess of 0.80, and they all met the requisite assumptions and validation. However, some use fewer variables, or different types of variables, and may be easier to implement. Most input variables are easy to determine or estimate but all the models used a dummy variable to isolate the effects of influential data points. A subjective decision must be made as to whether or not to include any new program in that dummy category and therefore, a model's efficacy hinges on that determination. Table 10 compares the final models.

Table 10 - Regression model comparison

Model	Variable type	Number of Variables	Ratio	Adjusted r^2	Jackknife CI
SM-I	Hard	3	10 to 1	0.81	0.95 – 0.99
SM-II	Mix	5	7 to 1	0.85	0.95 – 0.99
CM-I	Mix	4	9 to 1	0.80	0.95 – 0.99
CM-II	Hard	4	9 to 1	0.80	0.92 – 0.97
CM-III	Mix	5	7 to 1	0.84	0.95 – 0.99

When choosing a model for use in estimating schedule and cost, the program manager must decide what types of information are available for input. How well do you know the political and economic environment? Can you predict the soft variables accurately? Are you far enough along in the program to determine the expected EMD length? These and many other questions need answers before any estimates will be reliable. However, the models can easily aid decision-making through what-if analysis. Try different values of each of the variables in a model and you will get an idea of how programs have behaved in the past. Table 11 shows percent estimated cost growth given different inputs to CM-III. These are only a few examples but the types of information available are evident.

Table 11 - Example model implementation

"Quant_Change" (%/100)	"Significant pre-Emd activity"	"Num_Pol_ Change"	"Contractor_ Cost_Growth"	“F-22/C-17”	Estimated cost growth
-0.5	0	0	0	0	13%
0.5	0	0	0	0	-14%
0	0	1	0	0	10%
0	1	1	0	0	-26%
0	1	4	1	0	33%
0	1	4	1	1	216%
0	0	4	1	1	252%

In addition to the formal analysis offered in Chapter IV, we noticed some trends worth reporting. First, it became apparent that there is a relationship between funding, schedule, and cost. The FA-18E/F 1995 SAR gives one example of how lack of funding caused a slip in test dates, which in turn delayed the development schedule. In several cases, DoD test personnel were not available, again slipping the development timeline. National and political issues also played a role. “Fact of life” changes such as the Global War on Terror created ripples throughout the acquisition system, reducing spending for some programs while increasing it for others. Functional capabilities became more important and getting equipment to the warfighter in the field received a new urgency. While we attempted to capture specific changes due to specific time periods, looking at programs by what calendar years they covered showed only weak correlation.

Our time period scrutiny revealed a potential weakness. The majority of our data came from younger programs due to the entrance requirement that MSIII occur after 1996. Figure 14 shows the number of programs that were between MSII and MSIII for the years 1980 to 2005. A noticeable mass of programs in the late 1990’s and early

2000's accentuates any national or political impact during this period. Future research could increase the resolution of these world events in the database and test for lag effects.

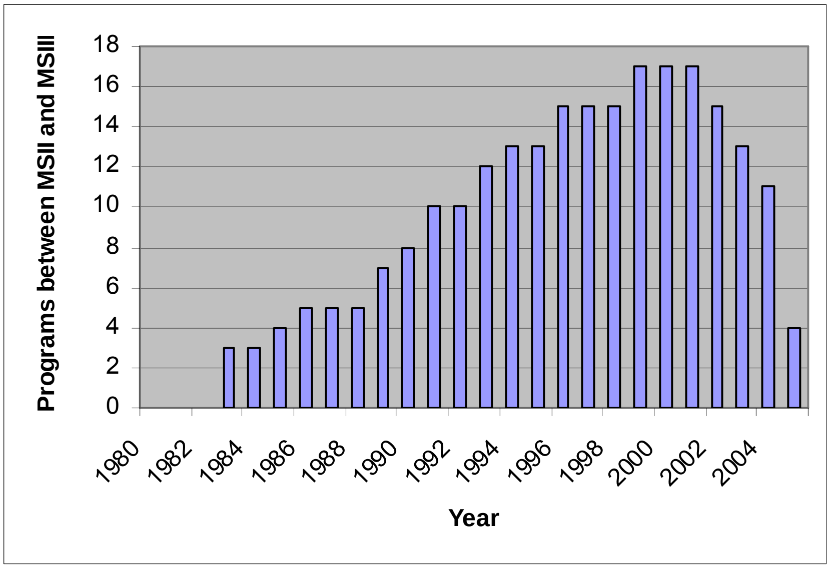


Figure 14 - Program location in history

Our final regression models provide the grounds for other possible recommendations. First, more research could highlight the effects of our chosen dummy variables. We grouped certain programs to enhance our analysis and justified their grouping from a historical perspective but more effort might uncover commonalities that could serve as new input variables and remove the need for a dummy. To review, we saw that complicated programs with many variants and long manufacturing runs had a significant impact on schedule response. Requirements drift was not directly measured but it was implied through the SAR narratives. When looking at cost response, the F-22 and C-17 programs stood out because of political and funding problems, which drove longer development, a high number of rebaselines, and significant quantity reductions.

Going beyond the dummy variables that isolated influential programs, we discovered more universal variables. Perhaps the most powerful was

"Significant pre-EMD activity." This variable showed up in three of the five models, proving valid for predicting both schedule and cost. The implication seems clear in that more pre-planning begets a smoother development phase. The weakness in this variable lies in how programs were reported. Only MDAPs were required to submit SARs so programs that did not reach that threshold until MSII, or programs that started at MSII (usually upgrades to existing systems), did not show "Significant pre-EMD activity." However, the indication could be that upgrade programs or those that *seem* simple, and therefore go straight to MSII, experience more problems and growth. We did not directly address technological maturity level but the willingness to begin programs at MSII indicates that some scrutiny of technological viability took place. Our soft variable count of the number of technical challenges did not prove to be predictive but future research could go deeper. The challenge will be finding consistent and reliable technical information from sources other than the SARs.

The analysis covered many other variables but it is pertinent to mention quantity change again. Since we were concerned primarily with development, quantity change did not significantly impact schedule because EMD quantities were mostly static – production quantities suffered the changes. However, costs were estimated based upon total production runs and manufacturers can recoup more of their development costs and increase production efficiencies with longer runs. The cost of reducing quantity becomes significant when the contractor can no longer absorb development costs and must increase unit cost to compensate, as predicted by the learning curve slope (Chapter II). This is not a new concept but this research confirms it once again. We must do our best to determine accurate quantity when the development baseline is set and resist the temptation to inflate the numbers to entice contractors or lower per-unit costs.

A final suggestion for future research would be to take our cost growth at percentage of program completion variables and expand them to include schedule, then develop a model for determining final schedule slippage or cost growth given a program's characteristics at a specific completion point.

Appendix A: Acronyms

ACAT – Acquisition Category
ACG – Adjusted Cost Growth
AFIT – Air Force Institute of Technology
AFMC – Air Force Materiel Command
APB – Acquisition Program Baseline
ASD – Aeronautical Systems Division
BAC – Budget at Completion
CAIG – Cost Analysis Improvement Group
CE – Current Estimate
CI – Confidence Interval
CM – Cost Model
CPI – Consumer Price Index
DAES – Defense Acquisition Executive Summary
DE – Development Estimate
DAB – Defense Acquisition Board
DoD – Department of Defense
DoL – Department of Labor
DSCPD - Defense Systems Cost Performance Database
EAC – Estimated Acquisition Cost
EMD – Engineering and Manufacturing Development
FCA – Functional Capability Area
FRP – Full-rate Production
FUE – First Unit Equipped
GAO – Government Accountability Office
GPO – Government Printing Office
IOC – Initial Operational Capability
LRIP – Low Rate Initial Production
MDAP – Major Defense Acquisition Program
MS – Milestone
OMB – Office of Management and Budget
OSD – Office of the Secretary of Defense
PAUC – Program Acquisition Unit Cost
PdE – Production Estimate.
PE – Planning Estimate
PI – Prediction Interval
PNO – Program Number
R&D – Research and Development
RAA – Required Assets Available
RAND – Research and Development Corporation
SAR – Selected Acquisition Report
SCI – Schedule Cost Index
SM – Schedule Model
VIF – Variance Inflation Factor

Appendix B: Selected Acquisition Reports

Selected Acquisition Reports (SARs) are submitted on an annual basis for MDAP programs. SARs summarize the latest estimates of cost, schedule, and performance status. These reports are prepared annually in conjunction with the President's budget. Subsequent quarterly exception reports are required only for those programs meeting the following criteria:

> 15% or more increase in the procurement estimate of the Program Acquistion Unit Cost (PAUC) compared to the PAUC in the currently approved Acquisition Program Baseline (APBA), or
>
> 15% or more increase in the current estimate of the Average Procurement Unit Cost (APUC) compared to the APUC in the currently approved APB, or
>
> Six-month of greater delay in the current estimate of any schedule milestone since the current estimated reported in the previous SAR, or
>
> Milestone B, Milestone C, or Full Rate Production Decision Review (Milestones II or III for grandfathered programs) and associated APB approval within 90 days prior to the quarter end date (DoD 5000.2-I).

The National Defense Authorization Act (NDAA) for FY 2006 made changes to the Nunn-McCurdy unit cost reporting statute for DoD major defense acquisition programs (10 USC§2433). The primary change was the addition of 30% and 50% unit cost thresholds against the original baseline estimate approved at System Development and Demonstration (Milestone B). The existing 15% and 25% unit cost thresholds were retained against the current baseline estimate.

Source: http://www.acq.osd.mil/ara/am/sar/2005-DEC-SARSUMTAB.pdf

SAR baseline discussion

The following discussion was excerpted from a 1996 RAND study that documented their in-house SAR database (Jarvaise, et al., 1996:5).

> Baseline Problems
>
> There are three types of baseline estimates (planning, development, and production) that are measured and tracked, each roughly corresponding to a decision point in the acquisition process. As a general rule, once a baseline has been established, the first estimate presented as that baseline should be used in calculating cost growth. However, at times, SAR baselines can be unstable. For instance, occasionally a second, more accurate estimate is substituted for the original estimate, generally improving cost performance as measured from this new baseline.
>
> Alternatively, changes that reflect an entirely different work scope from the original baseline may falsely portray poor cost performance. This information is generally classified and so is difficult to use in an unclassified environment. While earlier versions of DSCPD have made limited use of performance data, current versions have dropped this information because of data quality, measurement, and interpretation problems. Programs may even be canceled, then brought back with updated baselines, resulting in an apparent improvement in cost estimating performance. An example of this is the Precision Location Strike System (PLSS, Air Force). This program was canceled in 1981, resurrected in1983, and canceled again in 1986. The original DE for total system cost was $678.2 million (base-year 1977) for a quantity of three. The updated DE in the December 1983 SAR reported a total system cost of $635.5 million (base-year 1977) for a quantity of one. The new DE was significantly higher and would have resulted in a much lower cost growth factor had we used it as the baseline estimate. In some cases, using a new baseline may be justified if the program has significantly changed in scope, or the new system is different from the system for which the original DE was made. An example of this is the Bradley Fighting Vehicle System (Army), whose original DE was based on a predecessor vehicle, the Mechanized Infantry Combat Fighting Vehicle (MICV). The Bradley included a 25-mm gun and the tube-launched optically tracked wire-guided (TOW) missile system (the TOW system is a separate SAR program), while the MICV had only a 20-mm gun. Clearly, the original DE, when compared with the cost estimates for the Bradley, its 25-mm gun, and ammunition, would result in excessive cost growth. In this case, the original DE was not a fair basis for measuring cost growth; the current

DE (made after the cancellation of the MICV) was closer to a production baseline. We, therefore, added costs identified in the SAR as being associated with the new configuration to the PE and DE baselines to bring the estimates in line with the final design configuration of the vehicle.

Another baseline problem comes with combinations or separation of programs. Sometimes programs are reorganized and combined with other programs. Similarly, large programs consisting of several subsystems that were formerly contained in one program SAR are sometimes broken out into individual programs, each with its own SAR. These changes result in fairly severe distortions. Often, a large portion of the cost is lost or gained, while the baselines are unchanged, resulting in very large changes to the cost growth factors. The Submarine Combat System (SUBACS, Navy) is a good example of this. In December 1983, the SAR for SUBACS included a DE for two major subsystems, the AN-BSY 1 and the AN-BSY 2.

Subsequently, ANBSY 2 was removed from the SAR in December 1985, reestablished as a separate SAR program in December 1986, and was incorporated into the SSN-21 SAR in December 1990. While we would have liked to maintain consistency with the original DE and combine the two subsystems and treat them as one, the lack of detail reported for the AN-BSY 2 in the SSN-21 SAR made it impossible without making too many blind assumptions. In the end, the AN-BSY 2 costs were stripped from the SUBACS program and included in the SSN-21 program, thereby, changing both the AN-BSY 1 and SSN-21 baselines. If we had left the baselines as they were, we would have seen understated cost growth in the SUBACS program and greatly overstated cost growth in the SSN-21 program. Unfortunately, SARs do not provide enough information to separate models in a series. Thus, the costs of the F-15C/D or E versions cannot be separated from the original A/B version, even though the modifications were substantial. Thus, some observed development cost growth is due to development program costs for a major modification program added to the original development costs. Procurement costs may also increase because of the cost of performance enhancements not envisioned in the original SAR. In summary, changes to baselines have to be carefully scrutinized to preserve consistency over time within a program. If a large portion of the program has been dropped (or added), adjustments must be made to the baseline estimates to ensure that they reflect these changes. Failure to do so would result in large, unwarranted changes in cost growth factors. Often the SARs provide the necessary adjustment factors, but not always.

Appendix C – List of Programs

Program	PNO	Name
1	148	Patriot PAC-3
2	185	AMRAAM
3	200	C-17A
4	217	LHD 1
5	219	ATIRCMS/CMWS
6	240	T-45TS
7	248	Minuteman III PRP
8	260	GMLRS
9	265	F/A-22
10	274	JSTARS
11	278	CH-47F
12	280	Javelin
13	282	MH-60S
14	288	B1-B CMUP
15	289	Tactical Tomahawk
16	294	FBCB2
17	299	STRYKER (IAV)
18	302	Minuteman III GRP
19	330	AESA
20	341	Black Hawk Upgrade (UH-60M)
21	354	SDB
22	367	HIMARS
23	503	JDAM
24	537	NAS
25	541	Longbow Hellfire
26	549	F/A-18 E/F
27	551	NESP
28	554	MIDS-LVT
29	555	JASSM
30	560	JPATS
31	575	ABRAMS Upgrade
32	581	AIM-9X
33	582	CEC
34	601	BRADLEY Upgrade
35	746	FMTV
36	766	JSOW
37	831	LONGBOW Apache

Appendix D – List of Variables

The following is a complete listing of the 172 program characteristics and variables used in this research. The list is provided as a means for the reader to assess depth of study and uncover possible areas that could benefit from further research.

PNO
Cost_delta_MSII_MSIII_2005_percent_of_MSIII_cost
Perc_MSIII_growth
Prog_name
APB_set
PE_Established
PE_Established zero eliminator
DE_Established
MSI_Actual
MSI_Actual zero eliminator
MSII_Actual
MSIII_Actual
LRIP_Dec_Actual
IOC_Actual
MSIII_DE
LRIP_Dec_DE
IOC_DE
Initial_SAR_date
First_contract_award_date
Prototype
Upgrade?
Initial_Quant
Final_Quant
Quant_Change
MSIII before IOC?
LRIP_before_MSIII
MSIII_3mo_LRIP
LRIP_after_MSIII
Perc_IOC_growth
Perc_LRIP_growth
Total Cost at MSIII in 2005 dollars
Avg_inflation_MSII_MSIII
Average_approp_MSII_MSIII
has_MSI
has_MSI_or_upgrade
has_MSI_or_upgrade_or_foreman_prototype
Significant pre-EMD activity
PE ?
Len_MSII_MSIII
Len_MSII_LRIP
Len_MSII_IOC
Len_LRIP_IOC
Len_LRIP_MSIII
Len_MSIII_IOC
MSIII_slip
LRIP_slip
IOC_slip
Num_MSII_AP
Num_MSII_CE
Num_MSIII_AP
Num_MSIII_CE
Num_LRIP_AP
Num_LRIP_CE
Num_IOC_AP
Num_IOC_CE
Num_APB
Num_APB_MSII_MSIII
Num_SAR
Num_Annual_SAR
Num_Quar_Excep_SAR
Num_SAR_MSII_MSIII
Num_Quant_Change
Num_Tech_Prob
Num_Fund_Prob
Num_Pol_Change
Contractor_Cost_Growth
Avg_Tech_Mag
Avg_Fund_Mag
Avg_Pol_Mag
Avg_num_APB_MSII_MSIII
Avg_num_quant_change
Avg_num_tech_prog
Avg_num_polit_prob
Avg_num_fund_prob
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997

1998
1999
2000
2001
2002
2003
2004
2005
Persian_Gulf_90_91_ordinal
Persian_Gulf_90_91_onoff
Persian_Gulf_+2_ordinal
Persian_Gulf_+2_onoff
Bosnia_92_95_ordinal
Bosnia_92_95_onoff
Bosnia_+2_ordinal
Bosnia_+2_onoff
Afganistan_2002_ordinal
Afganistan_2002_onoff
Iraq_02_05_ordinal
Iraq_02_05_onoff
Dem_house_ordinal
Dem_house_onoff
Dem_senate_ordinal
Dem_senate_onoff
Dem_president_ordinal
Dem_president_onoff
7 Air
8 Land
9 Space
10 Sea
11 Electronic
12 Helo
13 Missile
14 Aircraft
15 Munition
17 Space (RAND)
18 Ship
21 Svs>1
Lead Svc = Navy
Lead Svc = AF
Lead Svc = Army
37 Lockheed-Martin
39 Boeing
40 Raytheon
41 General Dymics
McDonnell Douglas
Hughes
77 Class - C
76 Class - S
78 Class - U
Cost Plus Variants
Force Application?
Focused Logistics?
Force Protection?
Command and Control?
Battlespace Awareness?
Net Centric?
Joint Training?
Force Ap & Log
percent_cost_growth_10_percent_complete
percent_cost_growth_20_percent_complete
percent_cost_growth_30_percent_complete
percent_cost_growth_40_percent_complete
percent_cost_growth_50_percent_complete
percent_cost_growth_60_percent_complete
percent_cost_growth_70_percent_complete
percent_cost_growth_80_percent_complete
percent_cost_growth_90_percent_complete
Num_APB_by_10_percent_complete
Num_APB_by_20_percent_complete
Num_APB_by_30_percent_complete
Num_APB_by_40_percent_complete
Num_APB_by_50_percent_complete
Num_APB_by_60_percent_complete
Num_APB_by_70_percent_complete
Num_APB_by_80_percent_complete
Num_APB_by_90_percent_complete
F-22/C-17
2 3 9 24 35
MSIII<=1996
DE_est<1990
MSII_MSIII>10yr
Total_cost>20_billion

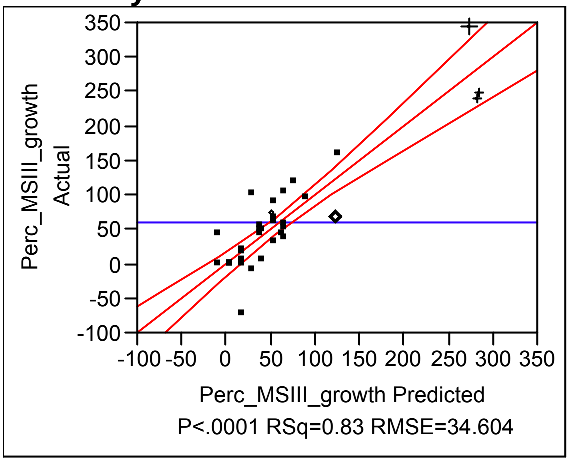
Appendix E – Regression Models

This appendix provides the complete analysis for each regression model for those who might want more information about the statistical output and compliance with assumptions. Some background in statistical analysis is required to understand fully this information but as a guide, we offer the following explanations:

- Actual by predicted plot – visual representation of how well the model fits the actual data. Points close to the line indicate a good fit and accurate model.
- Summary of fit – source of the Adjusted r^2 value discussed during analysis.
- Analysis of variance – source of the model's p-value ("Prob > F").
- Parameter estimate – source of each variable's p-value and VIF.
- Residual by predicted plot – visual representation of the residuals. A well disbursed plot with no visual trends indicates probable constant variance.
- Leverage plots – show each variable's predictive capability.
- Overlay plots showing Cook's Distance – indicate potential outliers (>0.25). Numbers below the plot indicate programs that exceeded the desired value.
- Overlay plot with studentized residuals – can reveal dependence or trends. A random but somewhat even magnitude across is good.
- Distributions with goodness of fit and S-W test – demonstrate normality in the residuals.
- Breusch-Pagan – test results for constant variance
- Jackknife confidence intervals – validation results

Schedule Model I

Response Perc_MSIII_growth
Whole Model
Actual by Predicted Plot



Summary of Fit

RSquare	0.829653
RSquare Adj	0.814167
Root Mean Square Error	34.60438
Mean of Response	58.74412
Observations (or Sum Wgts)	37

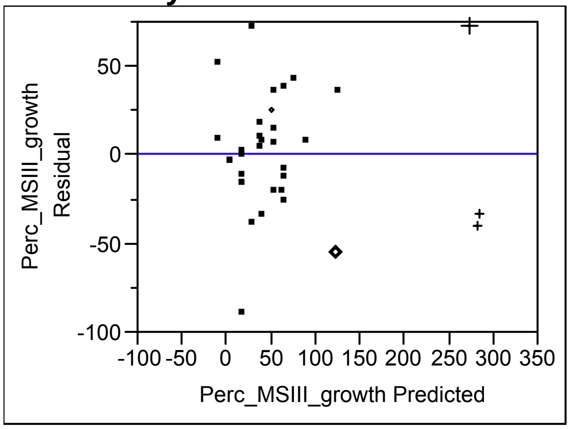
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	192458.74	64152.9	53.5740
Error	33	39516.29	1197.5	Prob > F
C. Total	36	231975.03		<.0001

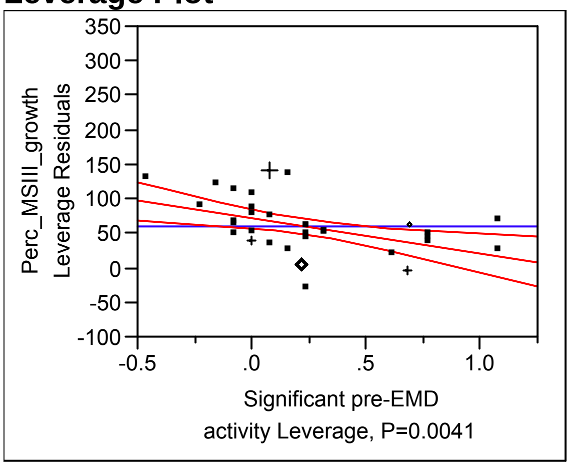
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-7.702895	10.815	-0.71	0.4813	.
Significant pre-EMD activity	-50.51628	16.37356	-3.09	0.0041	1.5248348
Num_MSIII_CE	12.018426	2.185669	5.50	<.0001	1.7027158
2 24 35	193.8704	22.36128	8.67	<.0001	1.1511438

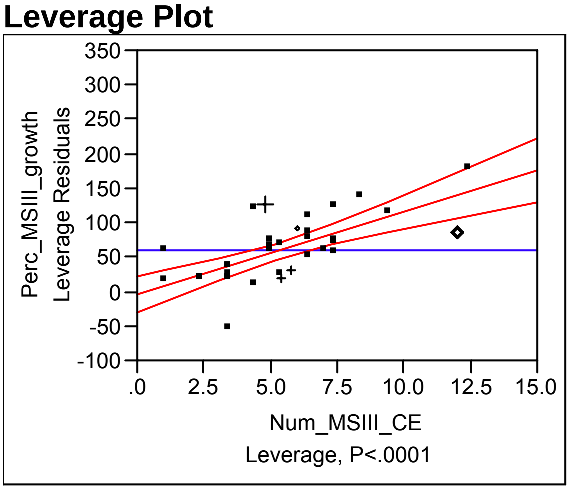
Residual by Predicted Plot



Significant pre-EMD activity
Leverage Plot

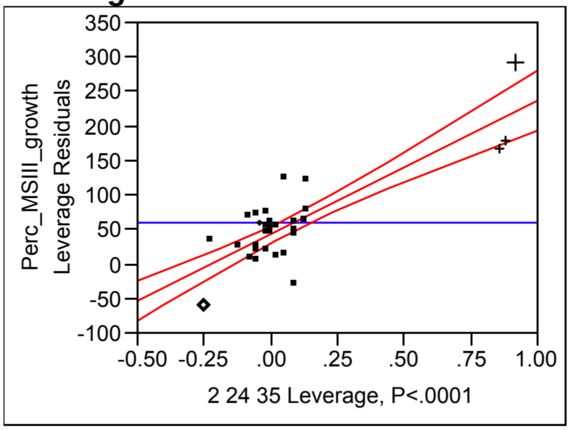


Num_MSIII_CE
Leverage Plot

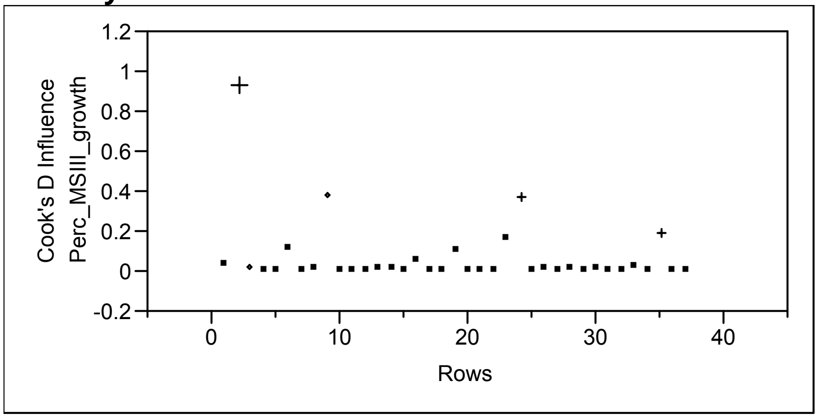


2 24 35

Leverage Plot

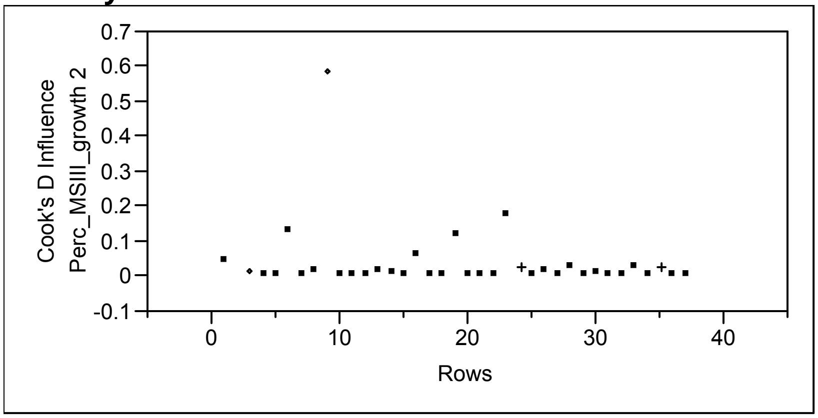


Overlay Plot



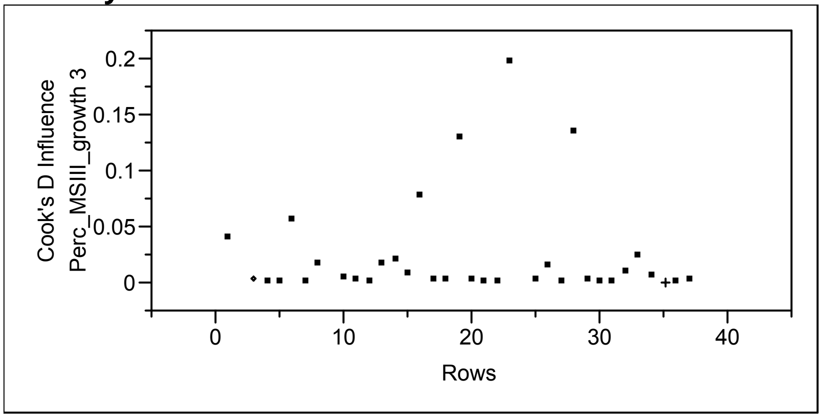
2, 9, 24

Overlay Plot



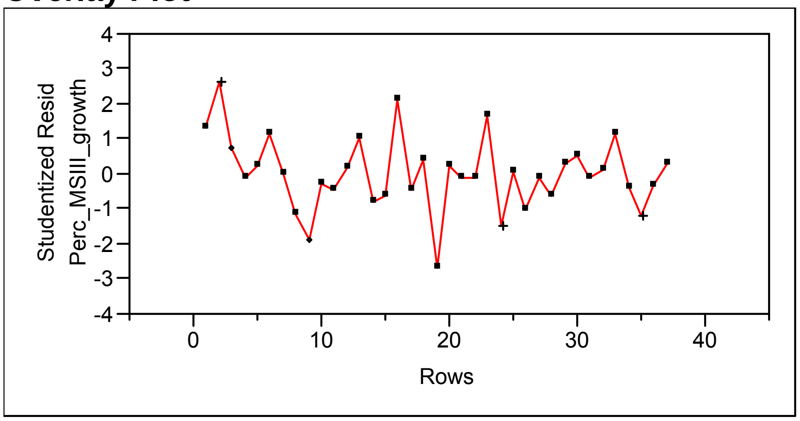
2 excluded, p-value = <.0001

Overlay Plot



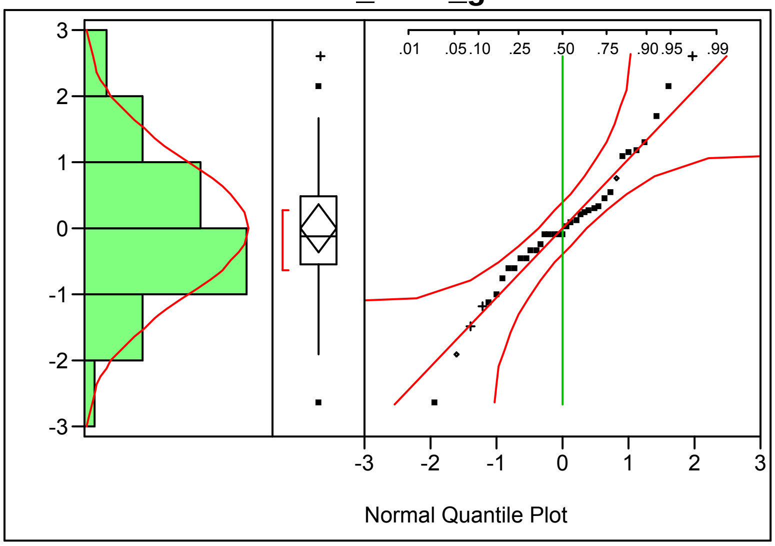
2, 9, 24 excluded, p-value = <.0001

Overlay Plot



Distributions

Studentized Resid Perc_MSIII_growth



Normal(-0.0002,1.05247)

Goodness-of-Fit Test

Shapiro-Wilk W Test

W	Prob<W
0.978865	0.6921

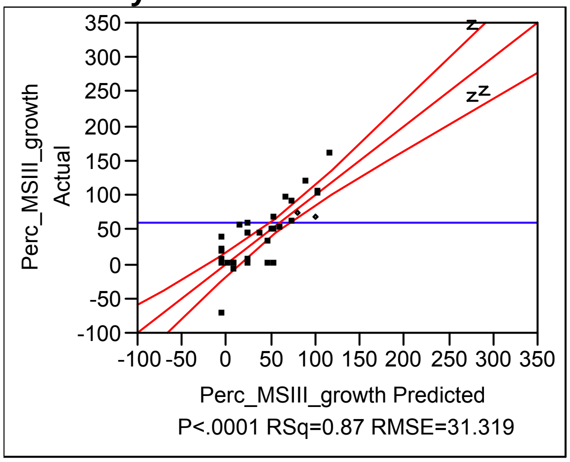
Note: Null Hypothesis = The data is from the Normal distribution. Small p-values reject the null.

Breusch-Pagan	
n	37
df Model	3
SSE	39516.29
SSM-r	8826470
TS	3.86908414
α	0.05
p-value	0.27595207

Jackknife Confidence Intervals	
MS3103	Schedule Response
std dev	0.052167724
mean	0.972972973
lower CI	0.953749984
upper CI	0.992195962

Schedule Model II

Response Perc_MSIII_growth
Whole Model
Actual by Predicted Plot



Summary of Fit

RSquare	0.868919
RSquare Adj	0.847777
Root Mean Square Error	31.31911
Mean of Response	58.74412
Observations (or Sum Wgts)	37

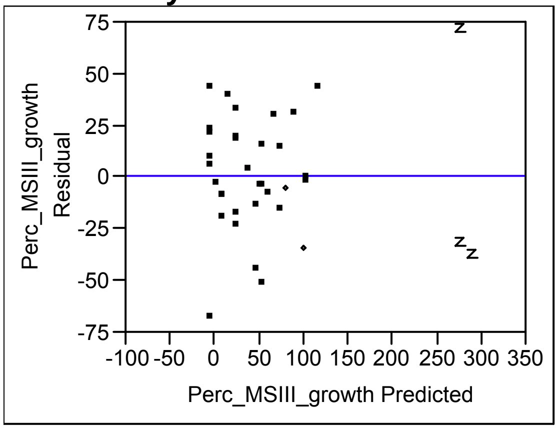
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	201567.54	40313.5	41.0990
Error	31	30407.49	980.9	Prob > F
C. Total	36	231975.03		<.0001

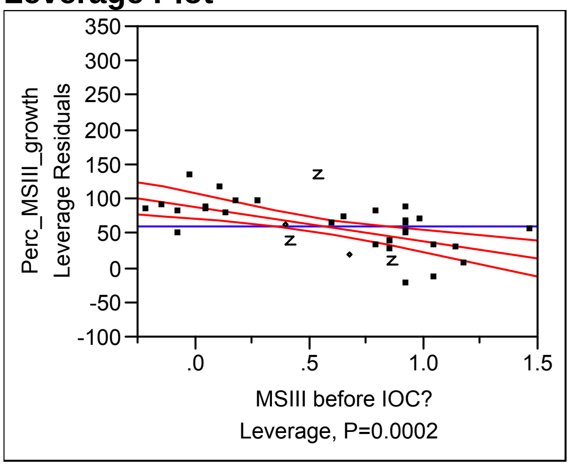
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	74.109499	11.19428	6.62	<.0001	.
MSIII before IOC?	-50.48311	11.84064	-4.26	0.0002	1.2748039
Significant pre-EMD activity	-36.40762	13.59431	-2.68	0.0117	1.2832006
Num_Fund_Prob	14.305909	3.069734	4.66	<.0001	1.1824257
Force Application?	-28.90747	11.2373	-2.57	0.0151	1.1203651
2 24 35	234.969	19.3236	12.16	<.0001	1.0494354

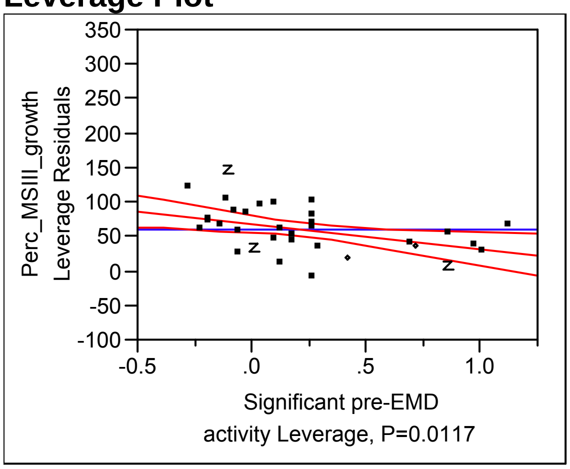
Residual by Predicted Plot



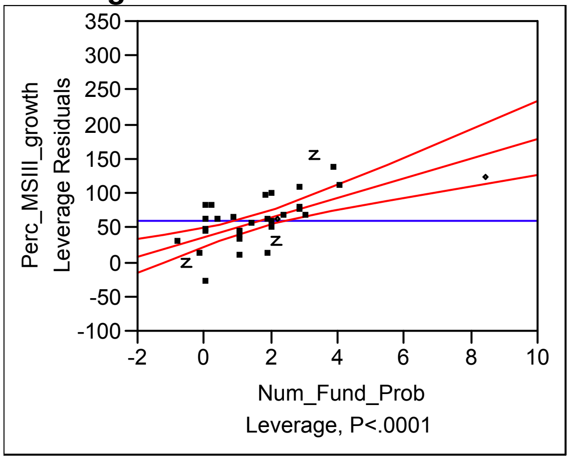
MSIII before IOC?
Leverage Plot



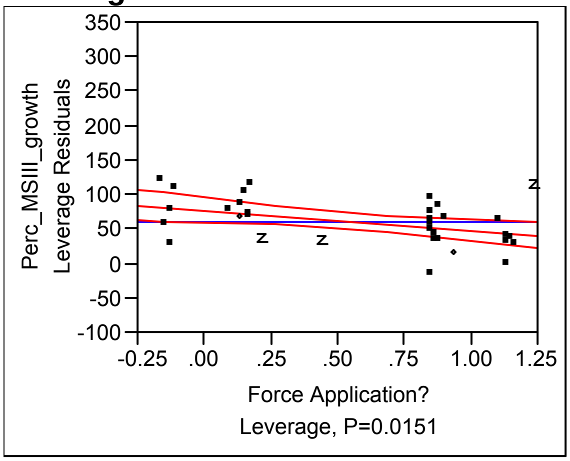
Significant pre-EMD activity
Leverage Plot



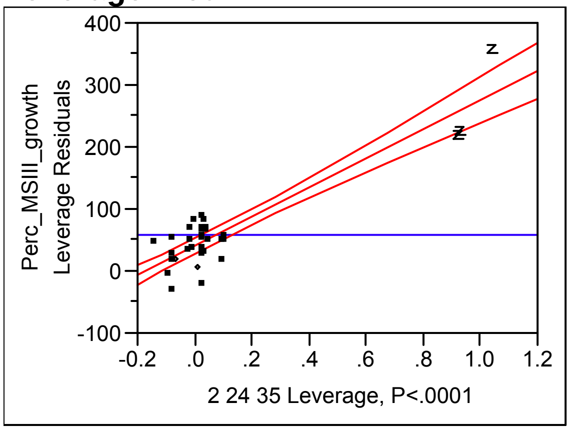
Num_Fund_Prob
Leverage Plot



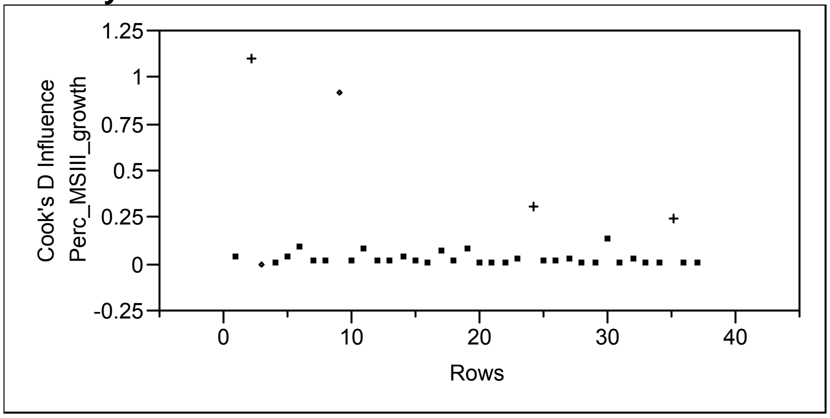
Force Application?
Leverage Plot



2 24 35
Leverage Plot

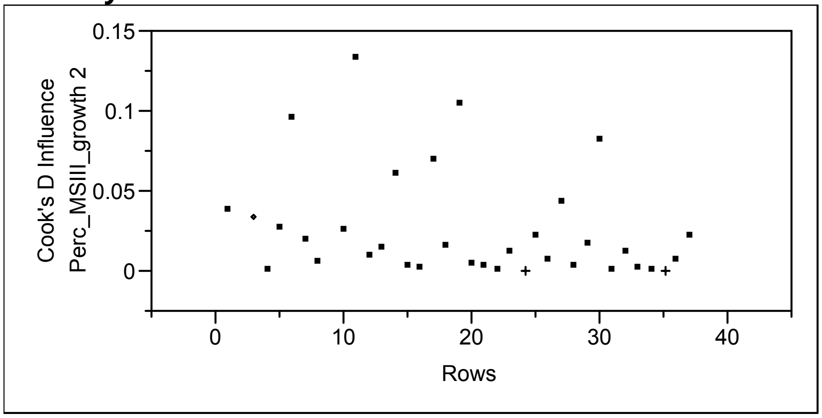


Overlay Plot



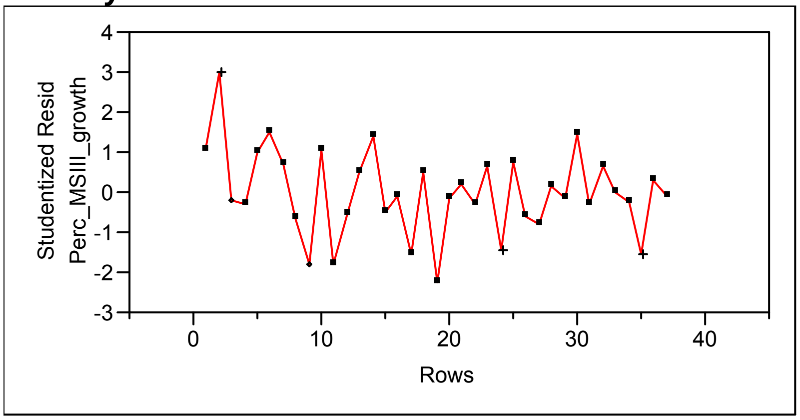
2, 9

Overlay Plot

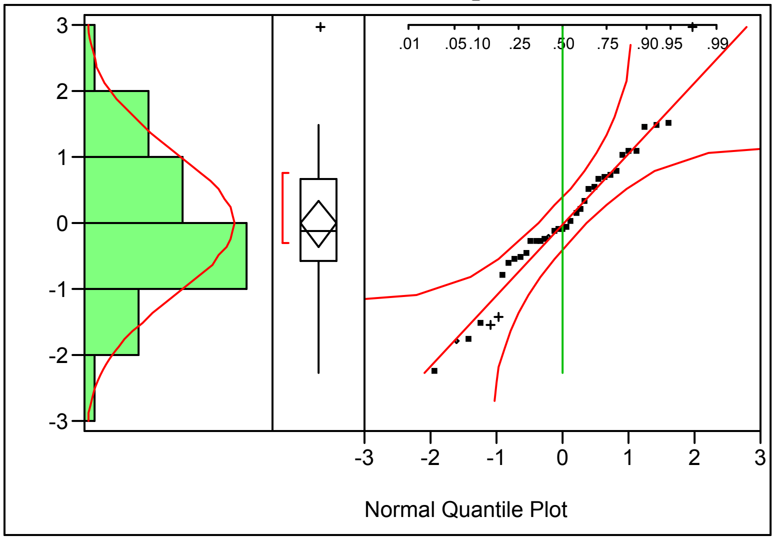


2, 9 excluded, p-value = <.0001

Overlay Plot



Distributions
Studentized Resid Perc_MSIII_growth



Normal(-0.0119,1.07819)

Goodness-of-Fit Test
Shapiro-Wilk W Test

W	Prob<W
0.974401	0.5400

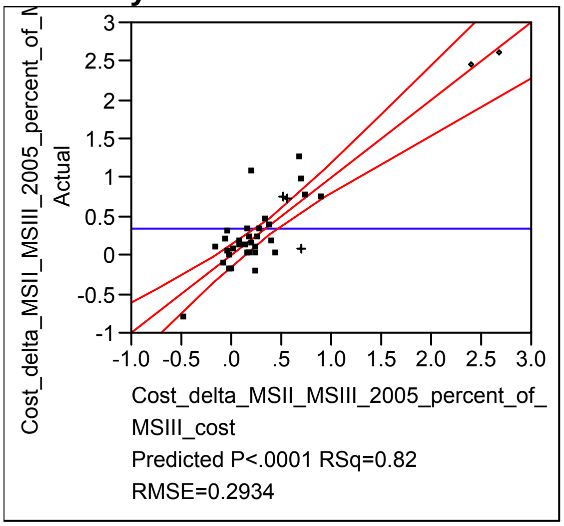
Note: Null Hypothesis = The data is from the Normal distribution. Small p-values reject the null.

Breusch-Pagan	
n	37
df Model	5
SSE	30407.49
SSM-r	12719984
TS	9.41670298
α	0.05
p-value	0.09355381

Jackknife Confidence Intervals	
MS5101	Schedule Response
std dev	0.052167724
mean	0.972972973
lower CI	0.953749984
upper CI	0.992195962

Cost Model I

Response Cost_delta_MSII_MSIII_2005_percent_of_MSIII_cost
Whole Model
Actual by Predicted Plot



Summary of Fit

RSquare	0.82478
RSquare Adj	0.802877
Root Mean Square Error	0.293368
Mean of Response	0.353836
Observations (or Sum Wgts)	37

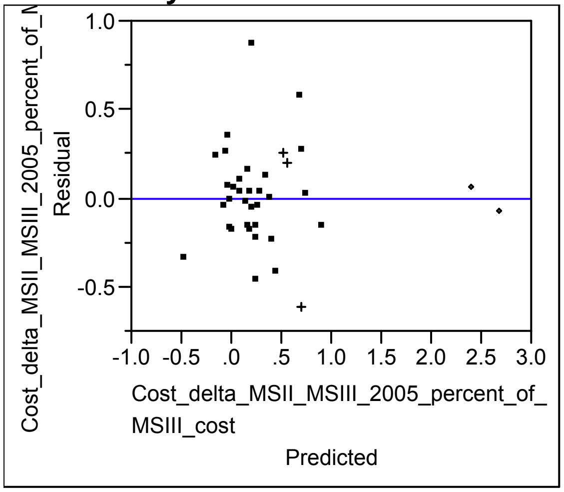
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	12.963690	3.24092	37.6568
Error	32	2.754072	0.08606	Prob > F
C. Total	36	15.717762		<.0001

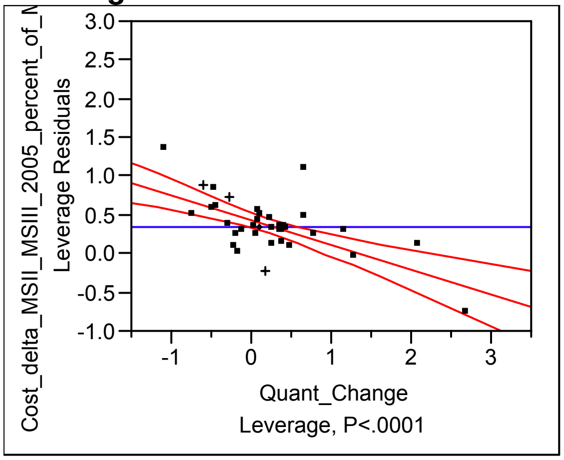
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-0.011822	0.087488	-0.14	0.8934	.
Quant_Change	-0.322547	0.068116	-4.74	<.0001	1.2431147
Num_Pol_Change	0.0824751	0.026103	3.16	0.0034	1.5653646
Contractor_Cost_Growth	0.2645406	0.071623	3.69	0.0008	1.027771
F-22/C-17	1.6121714	0.268548	6.00	<.0001	1.5853111

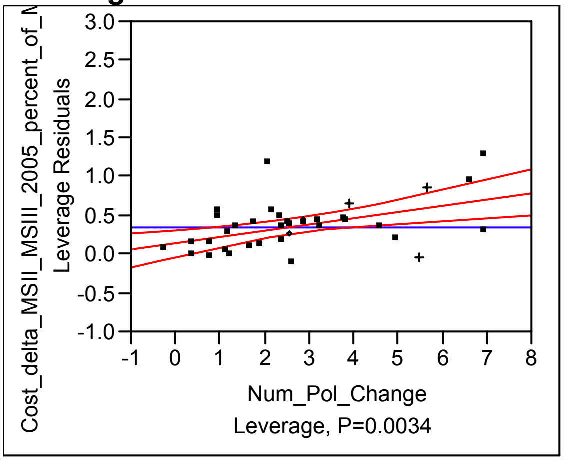
Residual by Predicted Plot



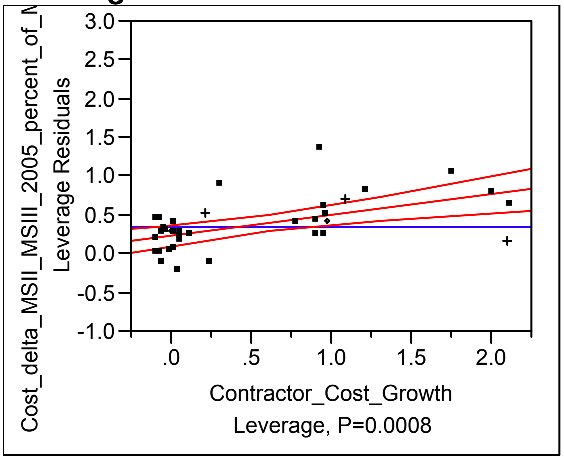
Quant_Change Leverage Plot



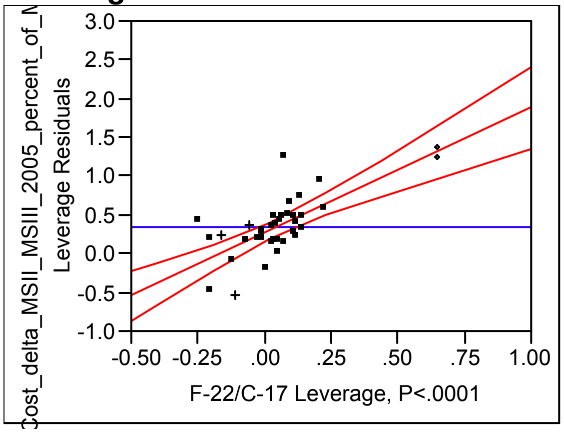
Num_Pol_Change Leverage Plot



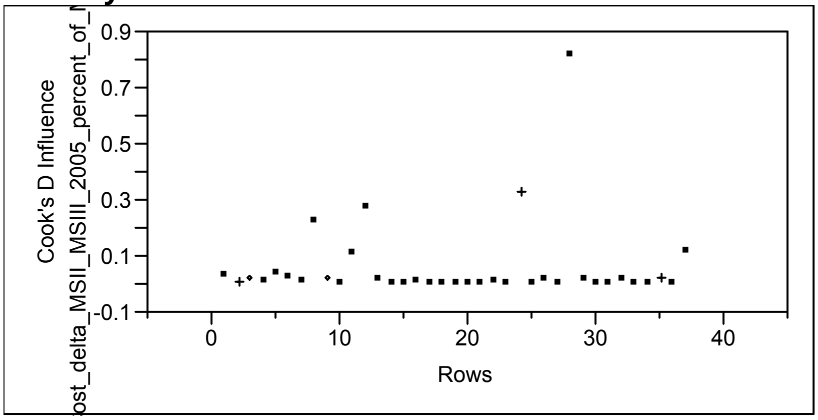
Contractor_Cost_Growth
Leverage Plot



F-22/C-17
Leverage Plot

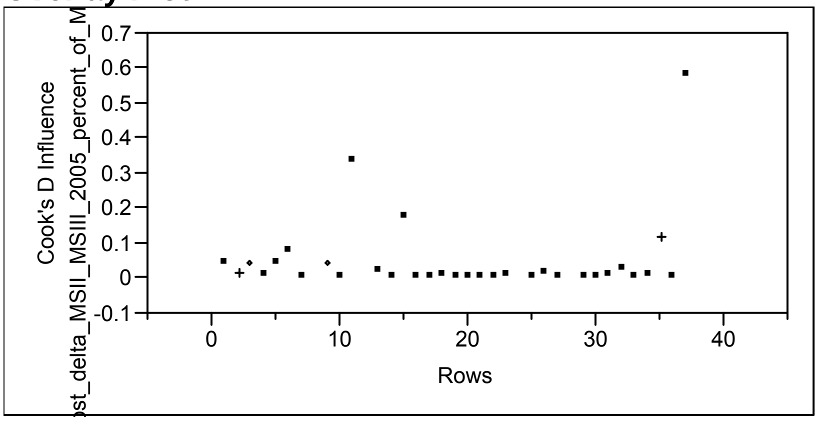


Overlay Plot



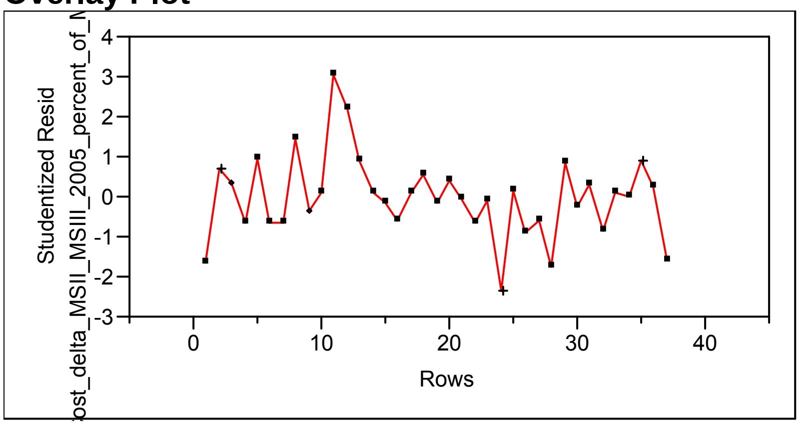
28, 24, 12, 8

Overlay Plot



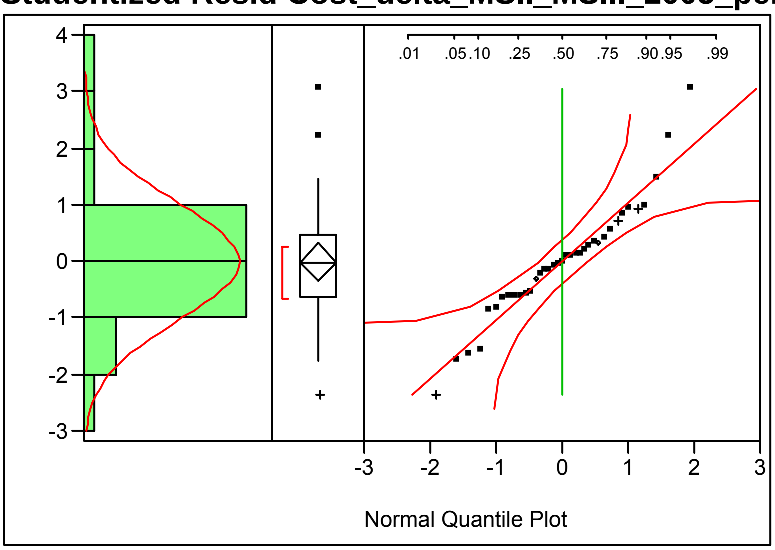
28, 24, 12, 8 excluded, p-value = <0.0001

Overlay Plot



Distributions

Studentized Resid Cost_delta_MSII_MSIII_2005_percent_of_MSIII_cost



Normal(-0.0127,1.03847)

Goodness-of-Fit Test

Shapiro-Wilk W Test

W	Prob<W
0.960196	0.2050

Note: Null Hypothesis = The data is from the Normal distribution. Small p-values reject the null.

Breusch-Pagan	
n	37
df Model	4
SSE	2.754072
SSM-r	0.10737277
TS	9.68984946
α	0.05
p-value	0.04598909

11 chosen for exclusion based on residual plot

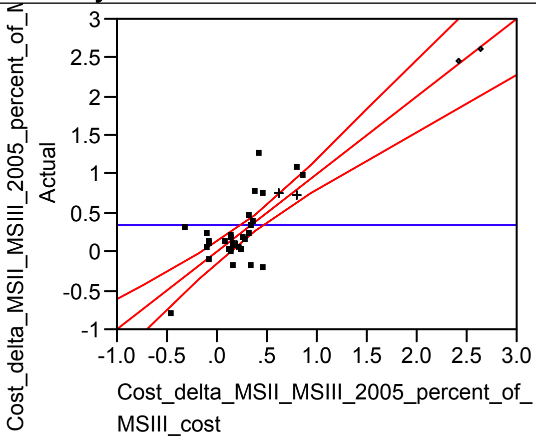
Breusch-Pagan	
n	36
df Model	4
SSE	1.862237
SSM-r	0.01973328
TS	3.68726526
α	0.05
p-value	0.44998084

11 excluded, p-value = <.0001

Jackknife Confidence Intervals	
MC4102	Cost Response
std dev	0.052167724
mean	0.972972973
lower CI	0.953749984
upper CI	0.992195962

Cost Model II

Response Cost_delta_MSII_MSIII_2005_percent_of_MSIII_cost
Whole Model
Actual by Predicted Plot



Cost_delta_MSII_MSIII_2005_percent_of_MSIII_cost
Predicted P<.0001 RSq=0.82
RMSE=0.2986

Summary of Fit

RSquare	0.818508
RSquare Adj	0.795822
Root Mean Square Error	0.298572
Mean of Response	0.353836
Observations (or Sum Wgts)	37

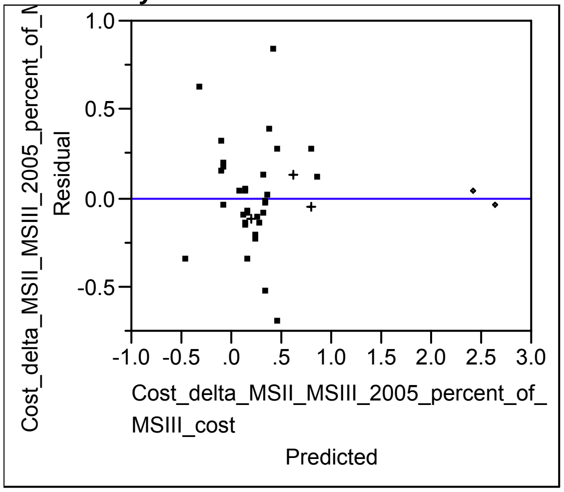
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	12.865118	3.21628	36.0791
Error	32	2.852644	0.08915	Prob > F
C. Total	36	15.717762		<.0001

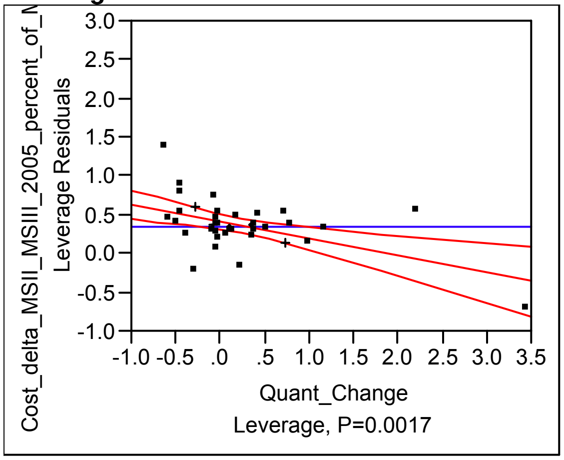
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	0.035459	0.080197	0.44	0.6614	.
Quant_Change	-0.220234	0.064105	-3.44	0.0017	1.0629955
F-22/C-17	1.8430427	0.234318	7.87	<.0001	1.1652286
Len_MSIII_IOC	0.0001831	6.181e-5	2.96	0.0057	1.2242032
MSIII_slip	0.0002925	0.000071	4.12	0.0002	1.3411277

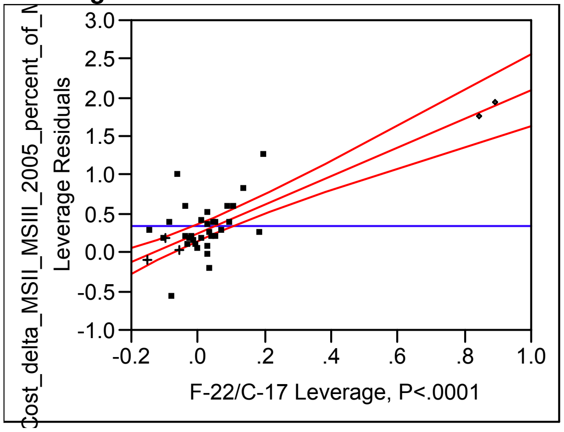
Residual by Predicted Plot



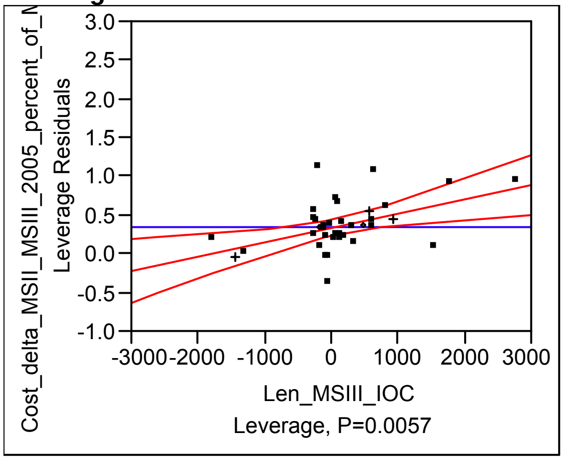
Quant_Change
Leverage Plot



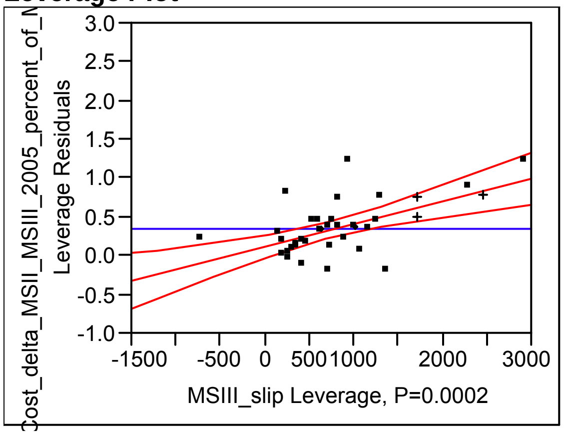
F-22/C-17
Leverage Plot



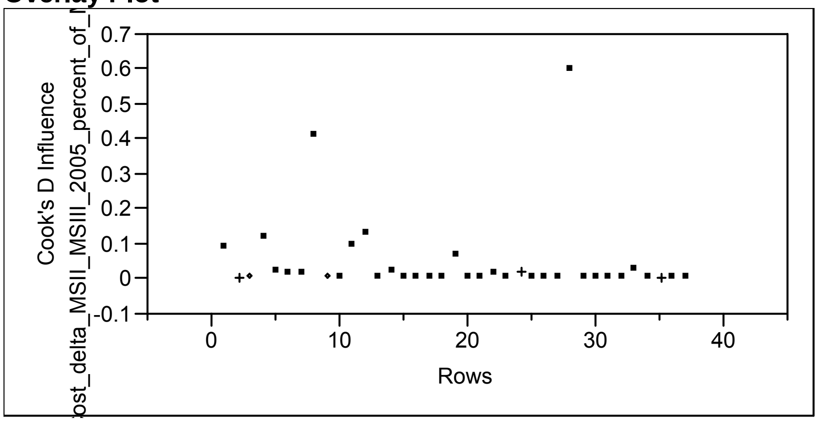
Len_MSIII_IOC
Leverage Plot



MSIII_slip
Leverage Plot

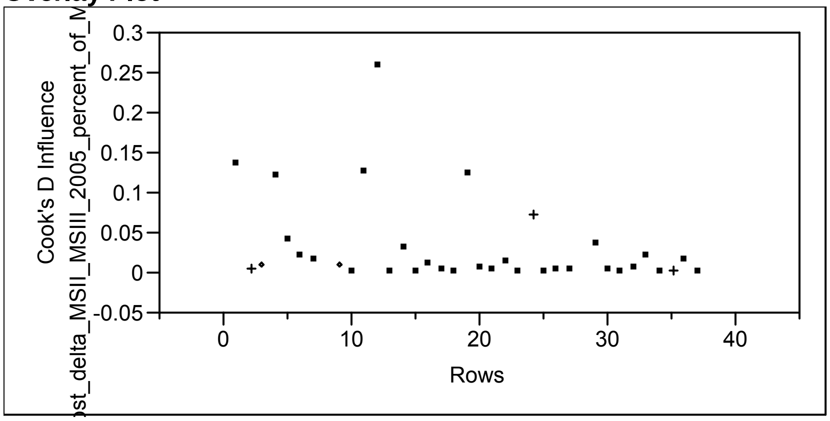


Overlay Plot



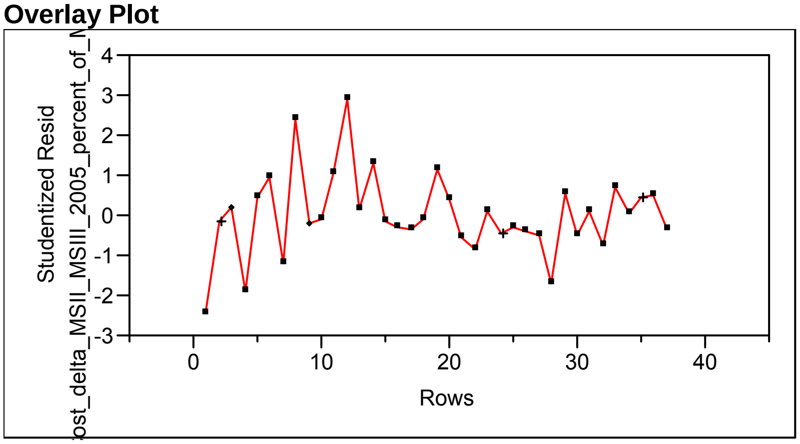
28, 8

Overlay Plot



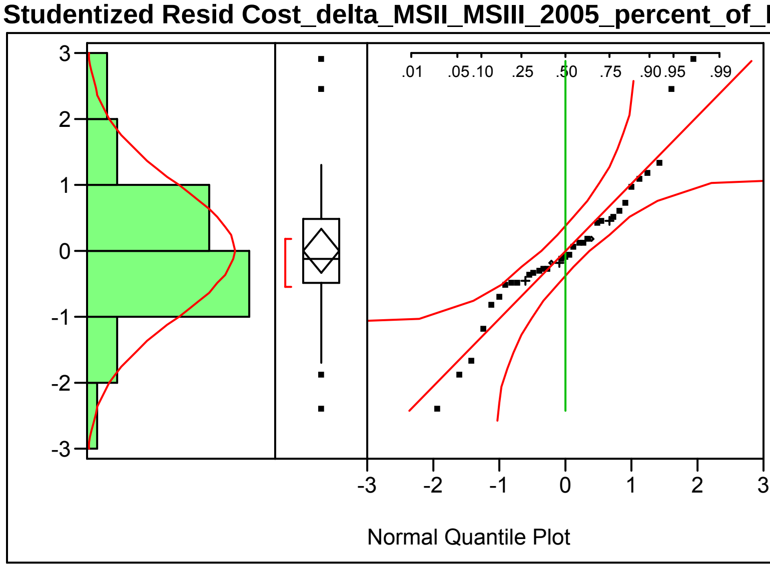
28, 8 excluded, p-value = <0.0001

Overlay Plot



Distributions

Studentized Resid Cost_delta_MSII_MSIII_2005_percent_of_MSIII_cost



Normal(0.00106,1.02485)

Goodness-of-Fit Test

Shapiro-Wilk W Test

W	Prob<W
0.948833	0.0885

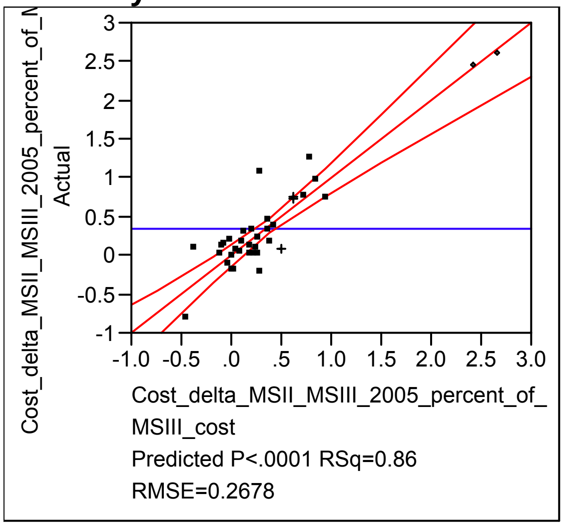
Note: Null Hypothesis = The data is from the Normal distribution. Small p-values reject the null.

Breusch-Pagan	
n	37
df Model	4
SSE	2.852644
SSM-r	0.01384519
TS	1.16460116
α	0.05
p-value	0.88389171

Jackknife Confidence Intervals	
MC4104	Cost Response
std dev	0.0627809
mean	0.945945946
lower CI	0.922812167
upper CI	0.969079724

Cost Model III

Response Cost_delta_MSII_MSIII_2005_percent_of_MSIII_cost
Whole Model
Actual by Predicted Plot



Summary of Fit

RSquare	0.858576
RSquare Adj	0.835765
Root Mean Square Error	0.267779
Mean of Response	0.353836
Observations (or Sum Wgts)	37

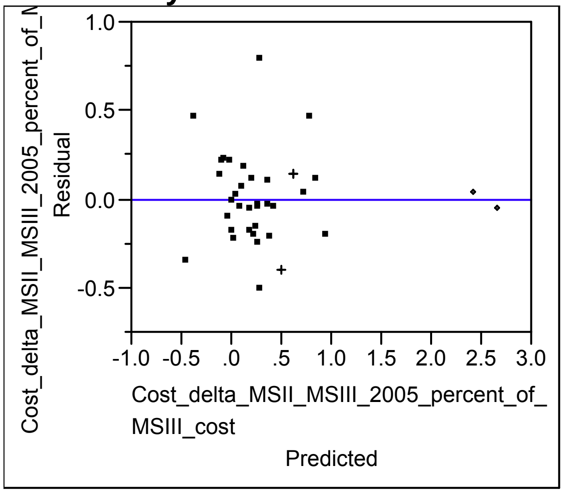
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	13.494890	2.69898	37.6397
Error	31	2.222873	0.07171	Prob > F
C. Total	36	15.717762		<.0001

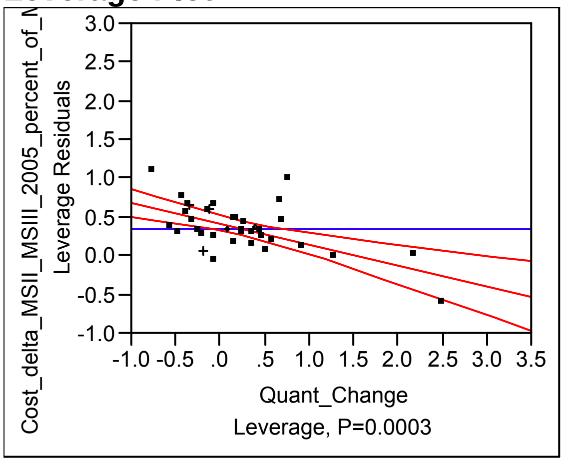
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-0.009872	0.079861	-0.12	0.9024	.
Quant_Change	-0.268342	0.065286	-4.11	0.0003	1.3706602
Significant pre-EMD activity	-0.362155	0.133058	-2.72	0.0106	1.6816302
Num_Pol_Change	0.1063309	0.025387	4.19	0.0002	1.7772011
Contractor_Cost_Growth	0.2599832	0.065397	3.98	0.0004	1.0284452
F-22/C-17	1.8267906	0.257495	7.09	<.0001	1.7493609

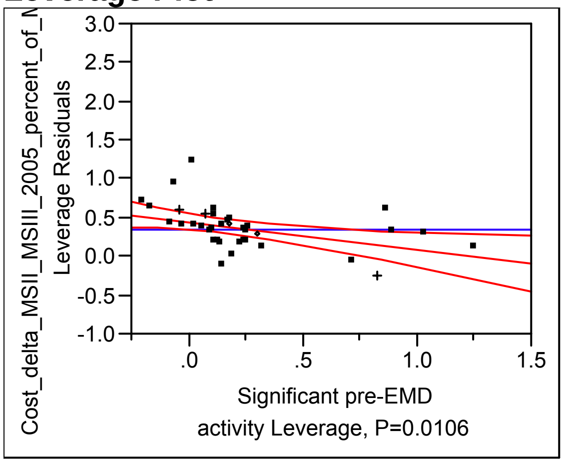
Residual by Predicted Plot



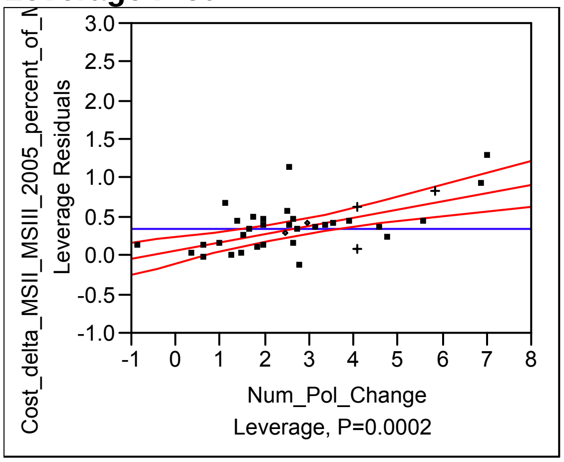
Quant_Change Leverage Plot



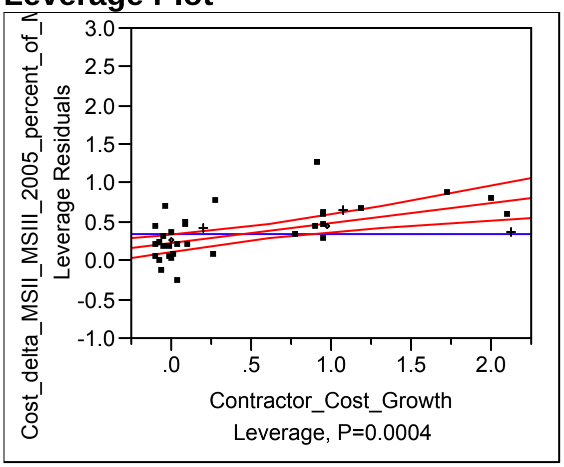
Significant pre-EMD activity Leverage Plot



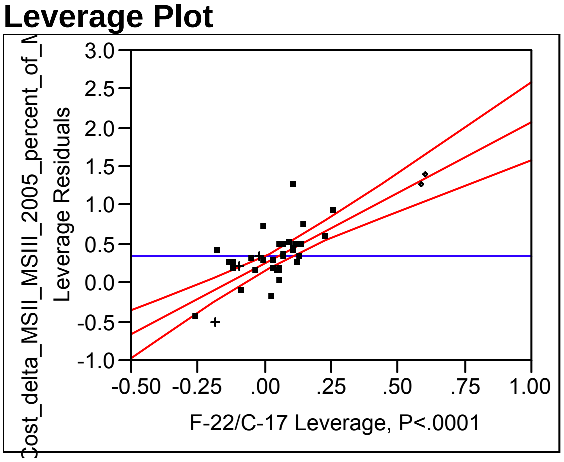
Num_Pol_Change
Leverage Plot



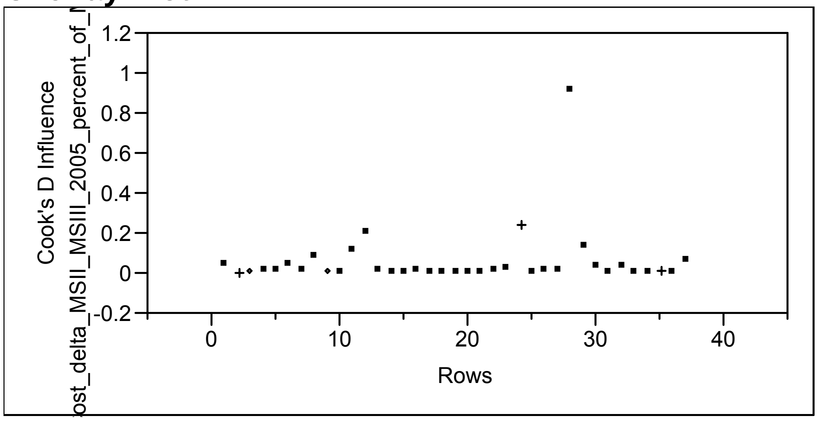
Contractor_Cost_Growth
Leverage Plot



F-22/C-17
Leverage Plot

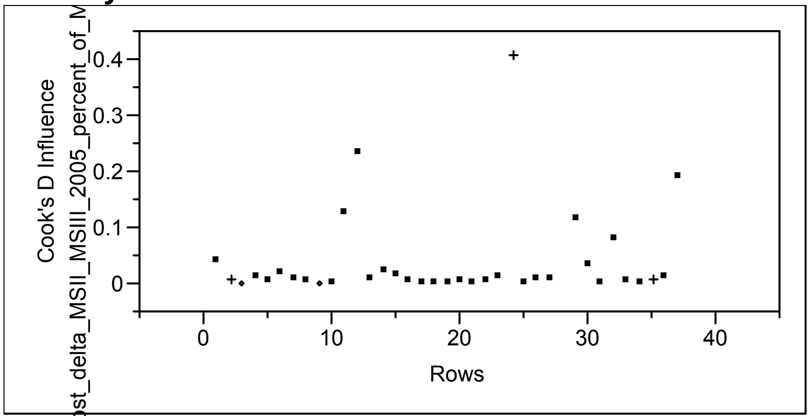


Overlay Plot



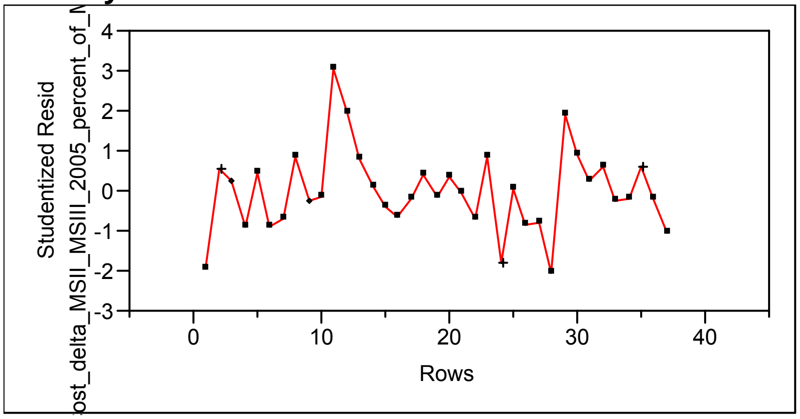
28

Overlay Plot

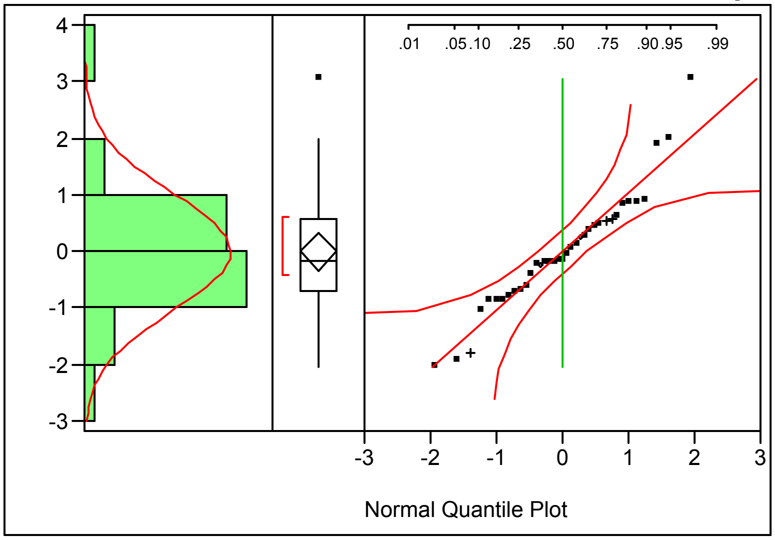


28 excluded, p-value = <.0001

Overlay Plot



Distributions
Studentized Resid Cost_delta_MSII_MSIII_2005_percent_of_MSIII_cost



Normal(-0.0099,1.03359)

Goodness-of-Fit Test
Shapiro-Wilk W Test

W	Prob<W
0.957662	0.1702

Note: Null Hypothesis = The data is from the Normal distribution. Small p-values reject the null.

Breusch-Pagan	
n	37
df Model	5
SSE	2.222873
SSM-r	0.03194845
TS	4.42582201
α	0.05
p-value	0.48986912

Jackknife Confidence Intervals	
MC5103	Cost Response
std dev	0.052167724
mean	0.972972973
lower CI	0.953749984
upper CI	0.992195962

Bibliography

Abate, Christopher C. A*n Analysis of Missile Systems Cost Growth and Implementation of Acquisition Reform Initiatives Using a Hybrid Adjusted Cost Growth Model.* Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2004 (ADA422659).

Air Force Material Command (AFMC). *AFMC Financial Management Handbook*. Wright-Patterson AFB OH, December 2001, http://www.deskbook.osd.mil.

Arena, Mark V., Robert S. Leonard, Sheila E. Murray, and Obaid Younossi. *Historical Cost Growth of Completed Weapon System Programs*. Santa Monica CA: RAND, 2006. http://www.rand.org/.

Bartlett, James E., II, Joe W. Kotrlik, and Chadwick C. Higgins. "Organizational Research: Determining Appropriate Sample Size in Survey Research," *Information Technology, Learning, and Performance Journal*: 43-50 (Spring 2001).

Bielecki, John V. *Estimating Engineering and Manufacturing Development Cost RiskUsing Logistic and Multiple Regression.* Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2003 (ADA413231).

Blacken, James E., Jr. *An Investigation of the Impact of Contract Changes on Cost Growth Within the LANTIRN System Program Office*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, September 1986 (ADA174395).

Buchfeller, Glenn W. and Donald E. Kehl. *Cost and Schedule Benchmarks for Defense Acquisition Contracts*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, September 1994 (ADA285001).

Christensen, David S., David A. Searle, and Caisse Vickery. "The Impact of the Packard Commission's Recommendation on Reducing Cost Overruns on Defense Acquisition Contracts," *Acquisition Review Quarterly*: 251-262 (Summer 1999).

Christensen, David S. "Is the cumulative SCI-based EAC an upper bound to the final cost of post-A12 defense contract?" *Journal of Cost Analysis and Management*: 1-10 (Winter 2004).

Cross, Steven M. *Data analysis and its impact on predicting schedule & cost risk*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2006 (ADA445179).

Department of Defense (DoD). *Cost Analysis Guidance and Procedures*. DoD Instruction 5000.4-M. Washington: GPO, 1992.

----. *Defense Acquisition Management Information Retrieval (DAMIR)*. Under Secretary of Defense, Acquisitions, Training, and Logistics USD(AT&L), online database, 2005. http://www.acq.osd.mil/damir/

----. *Program Acquisition Cost By Weapon System – Department of Defense Budget for Fiscal Years 2004/2005*. Washington: GPO, February 2003. http://www.dod.mil/comptroller/defbudget/fy2004/fy2004_weabook.pdf.

----. *Operation of the Defense Acquisition System*. DoD Instruction 5000.2. Washington: USD(AT&L), May 2003.

Drezner, Jeffrey A., Jeanne M. Jarvaise, Ron Hess, Daniel M. Norton, and Paul G. Hough. *An Analysis of Weapon System Cost Growth*. Santa Monica CA: RAND, 1993. http://www.rand.org/.

Drezner, Jeffery A. and Giles K. Smith. *An Analysis of Weapon System Acquisition Schedules*. Santa Monica CA: RAND, 1990. http://www.rand.org/.

Elkinton, Timothy L. and Todd J. Gondeck. *A Study to Determine if Newly Developed Budget at Completion (BAC) Adjustment Factors Improve the Success of Prevalent Index-Based Estimate at Completion (EAC) Techniques*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, September 1994 (ADA285004).

Genest, Daniel C. *Logistic and Multiple Regression: The Two-step Approach to Estimating Cost Growth*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2004 (ADA423097).

Gilbraeth, Robert D. *Winning at Project Management: What Works, What Fails and Why*. New York: John Wiley & Sons, 1986.

Gordon, James A. *Does a rubber baseline guarantee overruns?: A Study of Cost Performance and Contract Changes in Major Defense Acquisition Programs*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, September 1996 (ADA319331).

Government Accountability Office (GAO). *Base Operations – Challenges Confronting DoD as it Renews Emphasis on Outsourcing*. GAO/NSIAD-97-86. Washington: GPO, 1997.

----. *Major Weapon Systems Continue to Experience Cost and Schedule Problems under DOD's Revised Policy*. GAO-06-368. Washington: GPO, 2006.

----. *Assessments of Selected Major Weapon Programs*. GAO-06-391. Washington: GPO, 2006.

Hough, Paul G. *Pitfalls in Calculating Cost Growth from Selected Acquisition Reports*. Santa Monica CA: RAND, 1992. http://www.rand.org/.

JMP® Release 6.0.0. Computer software. Cary NC: SAS Institute Inc., 2005.

Jarvaise, Jeanne M., Jeffrey A. Drezner, and Daniel M. Norton. *The Defense System Cost Performance Datatabase - Cost Growth Analysis Using Selected Acquisition Reports*. Santa Monica CA: RAND, 1996.

Lucas, Brandon M. *Creating Cost Growth Models for the Engineering and Manufacturing Development Phase of Acquisition Using Logistic and Multiple Regression*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2004 (ADA422915).

McDaniel, Chris J. *Estimating Cost Growth in Engineering and Schedule Cost Categories [sic] Using a Two-pronged Regression Approach*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2004 (ADA422912).

McCrillis, John. "Derivation of Cost Variance Normalization Equation for Quantity Changes Using Learning Curve Theory." *Cost Analysis Improvement Group*. Washington: Office of the Secretary of Defense. Electronic Message. 20 Nov 2003.

Microsoft© Excel 2003. Computer software. Microsoft Corporation, 2003.

Monaco, James V. *Predicting schedule risk: A regression approach*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2005 (ADA not in Defense Technical Information Center, DTIC).

Moore, Gary W. *Estimating Procurement Cost Growth Using Logistic and Multiple Regression*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2003 (ADA413830).

Nelson, E. K. and J. E. Trageser. *Analogy Selection Methodology Study*. Fairborn OH: The Analytic Sciences Corporation, 1987.

Nystrom, Todd D. *A Comparative Study of Linear and Nonlinear Estimates at Completion Methods*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, September 1995 (ADA300643).

Phillips, Richard A. *An Analysis of Aircraft Weapon Systems Cost Growth and Implementation of Acquisition Reform Initiatives Using a hybrid Adjusted Cost Growth Model*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2004 (ADA423099).

Pletcher, John M. and Jane C. Young. *An Evaluation of Department of Defense Contractors' Cost Performance*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, September 1994 (ADA285017).

Rodrigues, Louis J. *Joint Strike Fighter Acquisition: Development Schedule Should be Changed to Reduce Risks*. Washington: General Accounting Office, March 2000. (ADA374913)

Romasz, James S. *Analysis of Cost Growth in U.S. Air Force Base Support Function Contracts*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2004 (ADA361710).

Rossetti, Matthew B. *Logistic and Multiple Regression: A Two-pronged Approach to Accurately Estimate Cost Growth in Major DoD Weapon Systems*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2004 (ADA422947).

Singleton, Pamela J. *Estimating Potential Cost Growth of the Most Probable Cost Estimate*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, September 1991 (ADA243919).

Sipple, Vincent P. *Estimating Engineering Cost Risk Using Logistic and Multiple Regression*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2002 (ADA400576).

Terry, Mark F. and Mary M. Vanderburgh. *An Analysis of Estimate at Completion Models Utilizing the Defense Acquisition Executive Summary Database*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, September 1993 (ADA275914).

Tukey, John W. *Kinds of Bootstraps and Kinds of Jackknives, Discussed in Terms of a Year of Weather-Related Data*. New Jersey: Princeton University Department of Statistics, Contract Number DAAL03-86-K-0073, April 1987 (ADA184495)

Tyson, Karen W., Bruce R. Harmon, and Daniel M. Utech. *Understanding Cost and Schedule Growth in Acquisition Programs*. Ft Belvoir VA: Institute for Defense Analyses, Paper P-2967, July 1994 (ADA284321).

Wandland, Klaes W. and Gary P. Wickman. *Analysis of Cost and Schedule Growth onSole Source and Competitive Air Force Contracts*. AirForce Institute of Technology (AU), Wright-Patterson AFB OH, September 1993(ADA273920).

Westgate, Barbara and Blaise Durante. “Cost Growth Study.” Briefing to the Chief of Staff of the Air Force. Washington: Headquarters U. S. Air Force, 21 December 2000.

Vita

Major Derek Foreman attended Texas A&M University, graduating in 1990 with a BS in Industrial Engineering and a Reserve Officer Training Corps commission. After pilot training at Laughlin AFB, Texas, he moved to Travis AFB, CA to fly the C-141. In 1996, he transitioned to the C-21 at Langley AFB, VA, followed in 1998 by a move to Dover AFB, DE and the C-5. After four more years, including stints as flight commander, assistant chief of standardization and evaluation, and the wing's chief pilot tactician, he moved to Altus AFB, OK, and took a position as the Air Mobility Command's C-5 subject matter expert and chief simulator certification pilot. At Altus, he earned an MA in Program and Acquisition management from the American Military University and continued to maintain flight qualification. From there, he attended the Air Force Institute of Technology as an Inter-developmental Education, Information Resource Management student. Major Foreman has spent over 15 years flying the Air Force mission and serving his fellow Airmen in theaters around the world.

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 074-0188*

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of the collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to an penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
23-01-2007	Master's Thesis	Jun 2006 – Mar 2007

4. TITLE AND SUBTITLE PREDICTING THE EFFECT OF LONGITUDINAL VARIABLES ON COST AND SCHEDULE PERFORMANCE	**5a. CONTRACT NUMBER** **5b. GRANT NUMBER** **5c. PROGRAM ELEMENT NUMBER**
6. AUTHOR(S) Foreman, James D., Major, USAF	**5d. PROJECT NUMBER** **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S) Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 Hobson Way, Building 640 WPAFB OH 45433-7765	**8. PERFORMING ORGANIZATION REPORT NUMBER** AFIT/GIR/ENC/07M-01
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)** **11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

12. DISTRIBUTION/AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Determining accurate cost and schedule is a crucial step to planning acquisition expenditures but history has shown that estimates are routinely low. Several researchers have attempted to forecast cost and schedule growth; we pick up this stream of research with a new approach. Our data collection and analysis focused on bringing in new data sources and added longitudinal variables to account for changes that took place over time. We assessed cost and schedule parameters for 37 major acquisition programs between Milestones II and III, resulting in 172 input variables and 5 regression models, 2 for schedule slippage and 3 for cost growth.

All five models passed statistical scrutiny and exhibited an Adjusted r^2 in excess of 0.80. The primary discriminator was the inclusion of strictly qualitative variables, taken from Selected Acquisition Report narratives and change justifications. We called these "soft" variables and coded them on a scale of 1 to 5 in the categories of funding problems, political problems, technical challenges, and contractor cost growth. Models with and without soft variables are presented to demonstrate their relative benefit. Finally, implications and implementation examples provide users a path to what-if analysis and decision-making.

15. SUBJECT TERMS
Multiple Regression, Schedule Variance, Schedule Growth, Cost Growth, Selected Acquisition Report, SAR, DoD Schedule Growth, Data Analysis, Cost Growth, Inferential Statistics, Schedule Growth in DoD Acquisition Programs, Cost Growth in DoD Acquisition Programs, Predicting Schedule Growth

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Dr. Edward D. White III
REPORT	**ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(Include area code)*
U	U	U	UU	133	(937) 255-3636, ext 4540; e-mail: Edward.White@afit.edu

Standard Form 298 (Rev: 8:98)
Prescribed by ANSI Std. Z39-18