National Aeronautics and Space Administration

FINANCIAL MANAGEMENT DIVISION | POLICY & GRANTS DIVISION | QUALITY ASSURANCE DIVISION | BUDGET DIVISION | STRATEGIC INVESTMENTS DIVISION | AGENCY FINANCIAL SYSTEMS OFFICE | MISSION SUPPORT OFFICE

Examining the Effects of Implementing Data-Driven Uncertainty in Cost Estimating Models

May 2023 Cost & Schedule Symposium

Victoria Nilsen

3 May 2023 | NASA Headquarters OCFO | Strategic Investments Division

Outline

- Current Application of Correlation Within Cost Estimating Models
- New Methods for Creating Data-Driven Correlation Matrices
- Previous Research: Causal Statistical Correlation
- Understanding the Correlation of Residuals Methodology
- Applying the Correlation of Residuals Methodology
- Benefits of Using PCEC
- Data-Driven Correlation Matrix Produced Using PCEC Subsystem CER Residuals
- Bias and Error of Model Using Correlation of Residuals Methodology
- Scope and Limitations
- Conclusion

Current Applications of Correlation Within Cost Estimating Models

- NASA relies upon a variety of probabilistic analysis methods to estimate the life cycle costs of various programs and projects. These are produced for a variety of reason, ranging from establishing a basis for monitoring or verifying a project's programmatic progress to informing NASA's budget requests from Congress.
- Correlation Assumptions are key assumptions within probabilistic cost analysis and often a driver for the total output or point estimate of a cost model.
- Due to the uncertain nature of correlation between random variables, NASA has had difficulty quantifying the relationships between spacecraft subsystems with specific, data-driven correlation matrices. Previously, the NASA cost analysis community has addressed this challenge by:

1. Selecting a blanket correlation matrix to address uncertainty within the model
   - The blanket value selected is usually 0.2 or 0.3
   - This is a **heuristic** that arbitrarily affects the model's probability distributions without a firm basis in statistical fact or historical project data
2. Opting out of using any correlation matrix altogether
   - This means that there is either no correlation matrix used on the data or a blanket correlation matrix of 0 values
   - In this situation, all the uncertainty within the model comes from the variability inherent to the model's underlying probability distributions

Understanding how correlation can add uncertainty into a cost model is especially important at NASA, where most missions are state-of-the-art and have limited historical data. Many statistical researchers have studied methodologies that can bridge this gap in knowledge; however, the practical application of this research is limited.

New Methods for Creating Data-Driven Correlation Matrices

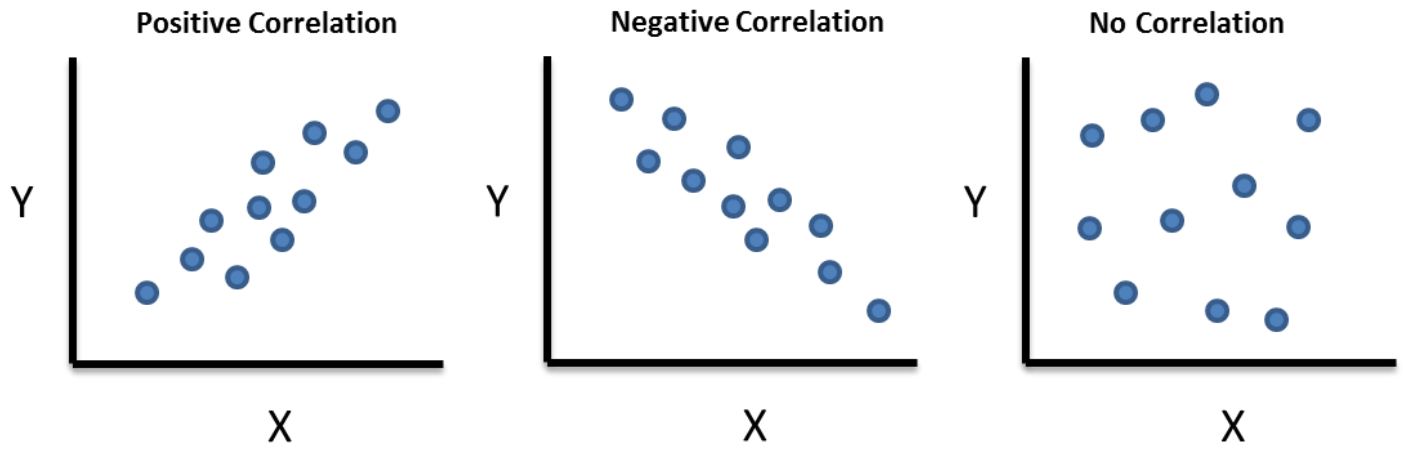
- One hypothesized method of creating data-driven correlation matrices that can be used within parametric cost analysis models to improve the accuracy of cost estimates is the "**correlation of residuals**" methodology

Correlation of Residuals Methodology

- The correlation of residuals methodology involves deriving the correlation coefficients of a model from the residuals of the regression equations for the Cost Estimating Relationships (CERs) of that model.
- This hypothesized method is based upon decades of previous research on correlation in cost models and was implemented through Latin Hypercube simulation in NASA's Project Cost Estimating Capability (PCEC) model.

What is Correlation?

- **Correlation** - a statistical measure of association between two variables
  - Measures how strongly the variables change with each other
- There are two main statistics for measuring correlation
  1. **Pearson's Correlation** - measures linearity of a relationship between 2 random variables
  2. **Spearman's Correlation** - measures monotonicity (rank) between two random variables
- For our purposes, we will be focusing on the Pearson correlation coefficient
- Examples of positive, negative, and no correlation graphs for the Pearson Correlation coefficient can be seen below



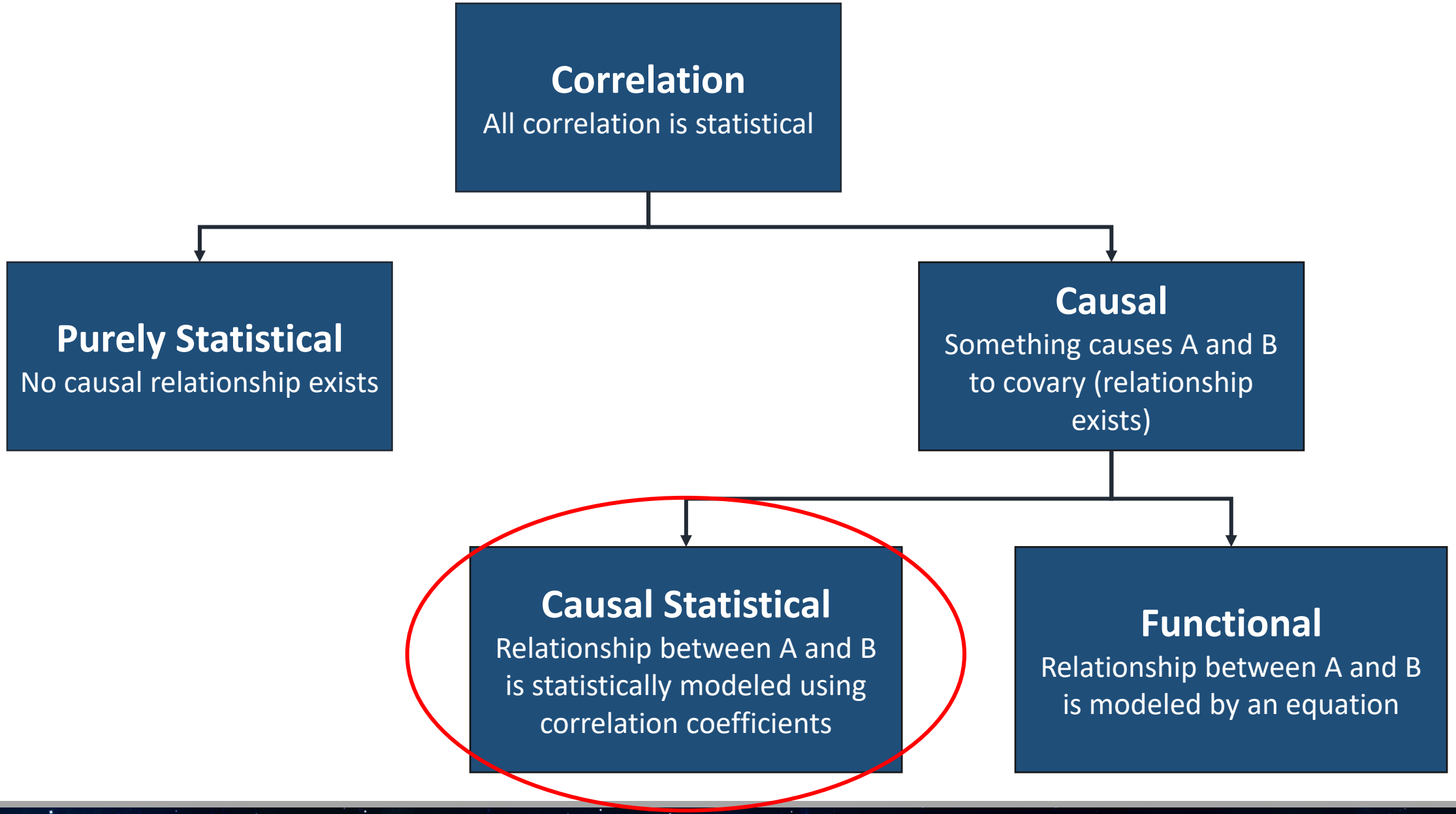
Previous Research: Causal Statistical Correlation

Correlation is inherent in the process of developing CERs

- When regression analysis is performed for parametric cost estimating models, most of the uncertainty can be explained by the functional correlation of the regression equation.
  - **Functional Correlation** results from the mathematical relationships between cost drivers and CERs and the relationship between the random variables can be modeled by an equation. In cases where functional correlation is present, correlation is usually handled without defining a correlation matrix or coefficients between CERs, since this correlation is already inherent in the model.
- Additional unexplained uncertainty remains in the model because each independent variable is correlated with cost and/or standard error to some degree. The correlation that exists between the unused independent variables is not included in the functional correlation of the regression equation
  - The remaining causal correlation is the "**causal statistical correlation**" between the unused independent variables and the standard error of the regression analysis
  - This causal statistical correlation is not inherent in the predictions of a cost model but can be implemented in the model as **a correlation matrix of all the independent variables**
    - This is what we are trying to model with the blanket 0.3 correlation. For the purposes of this research, this can be modeled as a correlation matrix of the subsystems of the typical NASA WBS
- Based on the findings of Coleman and Gupta [1994] and Raymond Covert [2006]

Hypothesis: CER residuals can be used to determine the subsystem correlation matrix that should be implemented into the model to mimic the behavior of causal statistical correlation and improve the accuracy of cost estimates

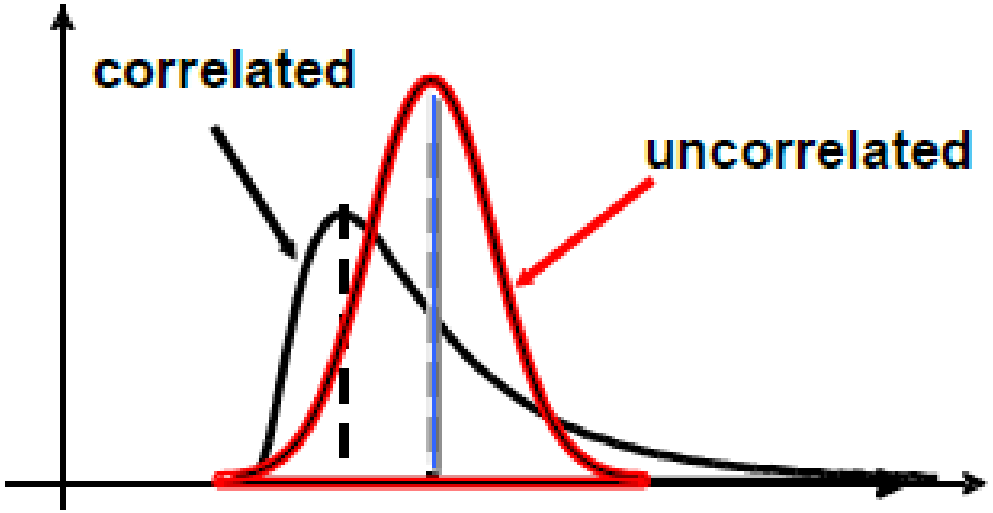
Understanding the Correlation of Residuals Methodology



Why Do We Care?

- **Statistical Answer**: Adding correlation coefficients to a model has the potential to drastically influence a continuous distribution's measures of variability (standard deviation, variance, skewness, maximum, and minimum) without having a drastic effect on the distribution's measures of central tendency (mean, median, mode).The effects on the distribution's measures of variability can increase the range of the probability density function. This means that correlation can be a huge contributor to the amount of risk in probabilistic cost estimates.

- **Simple Answer:** Adding correlation can drastically influence the range of possible cost estimates. This can decrease the probability of estimating an "average" cost and increase the probability or "risk" of estimating a "high" cost. This may not be good news for project managers trying to remain under budget, but it does provide a more realistic understanding of project cost and may be necessary information to plan for the successful completion of project activities.



8

Applying the Correlation of Residuals Methodology

For the purposes of this research, the correlation of residual methodology was applied within NASA's Project Cost Estimating Capability (PCEC). This methodology was performed within PCEC according to the following steps:

1. The correlation coefficients for each spacecraft subsystem and support function were determined by correlating the residuals of PCEC's subsystem CERs.
   - PCEC contains 20 unique spacecraft subsystems and support function CERs. The correlation coefficients between each pair of subsystem CERs were compiled into a 20x20 correlation matrix
2. The resulting correlation matrix was implemented into PCEC as an uncertainty factor influencing the model's pre-existing cost distributions using the Excel add-in Argo
3. The Latin Hypercube Sampling function of Argo was used to simulate PCEC results for 40 missions within the PCEC database.

A sensitivity analysis was also performed in order to understand how the correlation of residuals matrix compares to current standards. The steps above were repeated three additional times using the following correlation matrices:

1. A correlation matrix with a blanket value of 0
2. A correlation matrix with a blanket value of 0.3
3. A correlation matrix with a blanket value of 1

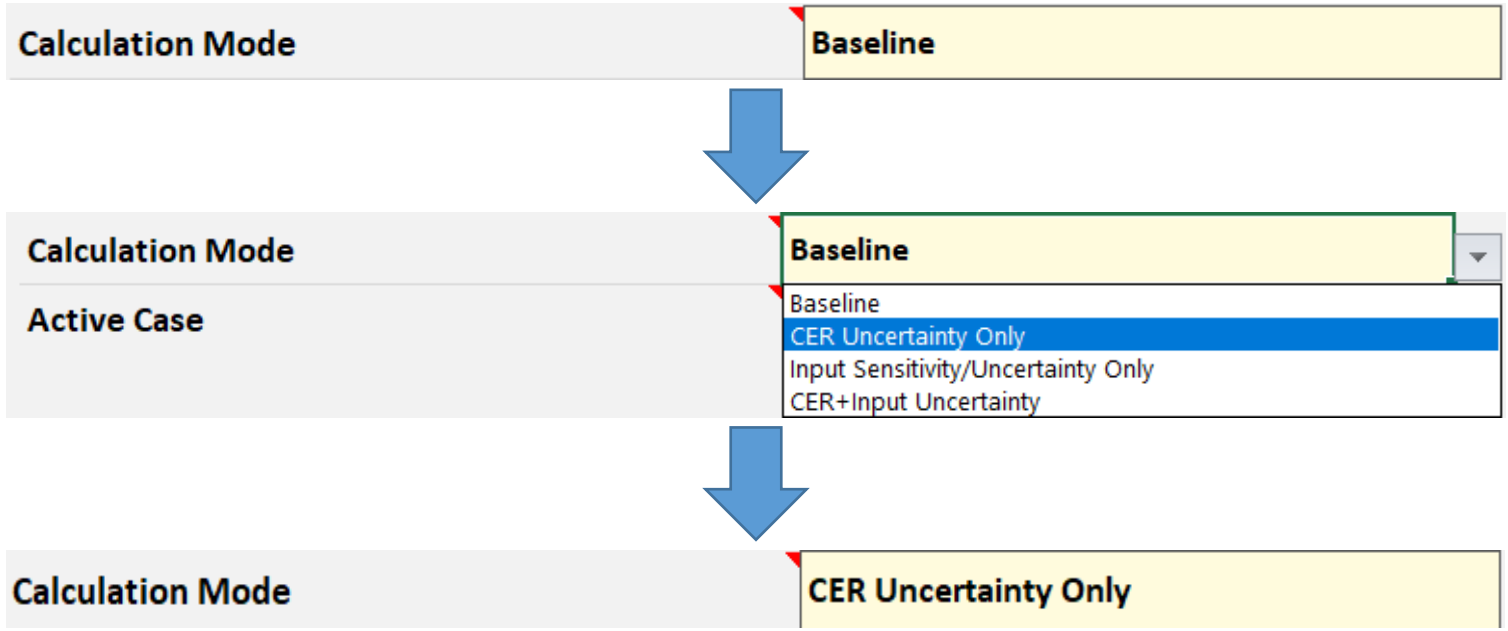
Benefits of Using PCEC

1. PCEC uses parametric estimating techniques that are very familiar to the average NASA cost estimator
   - CERs within PCEC are easily understood by PP&C community at NASA
   - Regression analysis used to create CERs is frequently tested and trustworthy
   - Follows typical NASA Work Breakdown Structure
2. CER Uncertainty calculation mode uses Argo to produce cost estimates using probability distributions
   - Student's t-distributions for each subsystem and support function
   - Easily traces to correlation matrix of subsystems and support functions
   - Compatible with Argo's Monte Carlo and Latin Hypercube simulation functions
3. PCEC Library architecture houses a database with generally complete subsystem data for 58 of NASA's near-Earth robotics spacecraft missions
   - Frequently updated
   - Standard, consistent, and current collection of spaceflight hardware data
   - Removes the need for data collection and normalization
4. Results of correlation of residuals methodology can be easily validated against actual spacecraft development data to determine if the methodology provides benefit

Subsystems & Support Functions
Attitude Control Nonrecurring Cost (NRC)
Attitude Control Recurring Cost (RC)
Command & Data Handling (CDH) NRC
CDH RC
Communications NRC
Communications RC
Electrical Power & Distribution NRC
Electrical Power & Distribution RC
Propulsion NRC
Propulsion RC
Structures & Mechanisms NRC
Structures & Mechanisms RC
Thermal Control NRC
Thermal Control RC
Project Management (PM)
Systems Engineering (SE)
Safety & Mission Assurance (SMA)
Integration & Test (I&T)
Mission Operations & Ground Data Systems Development (MOS-GDS Dev)
Phase E Mission Operations & Data Analysis (MODA PhE)

Disclaimer!

The simulation performed within PCEC uses the “CER Uncertainty Only” calculation mode instead of the “Baseline” calculation mode



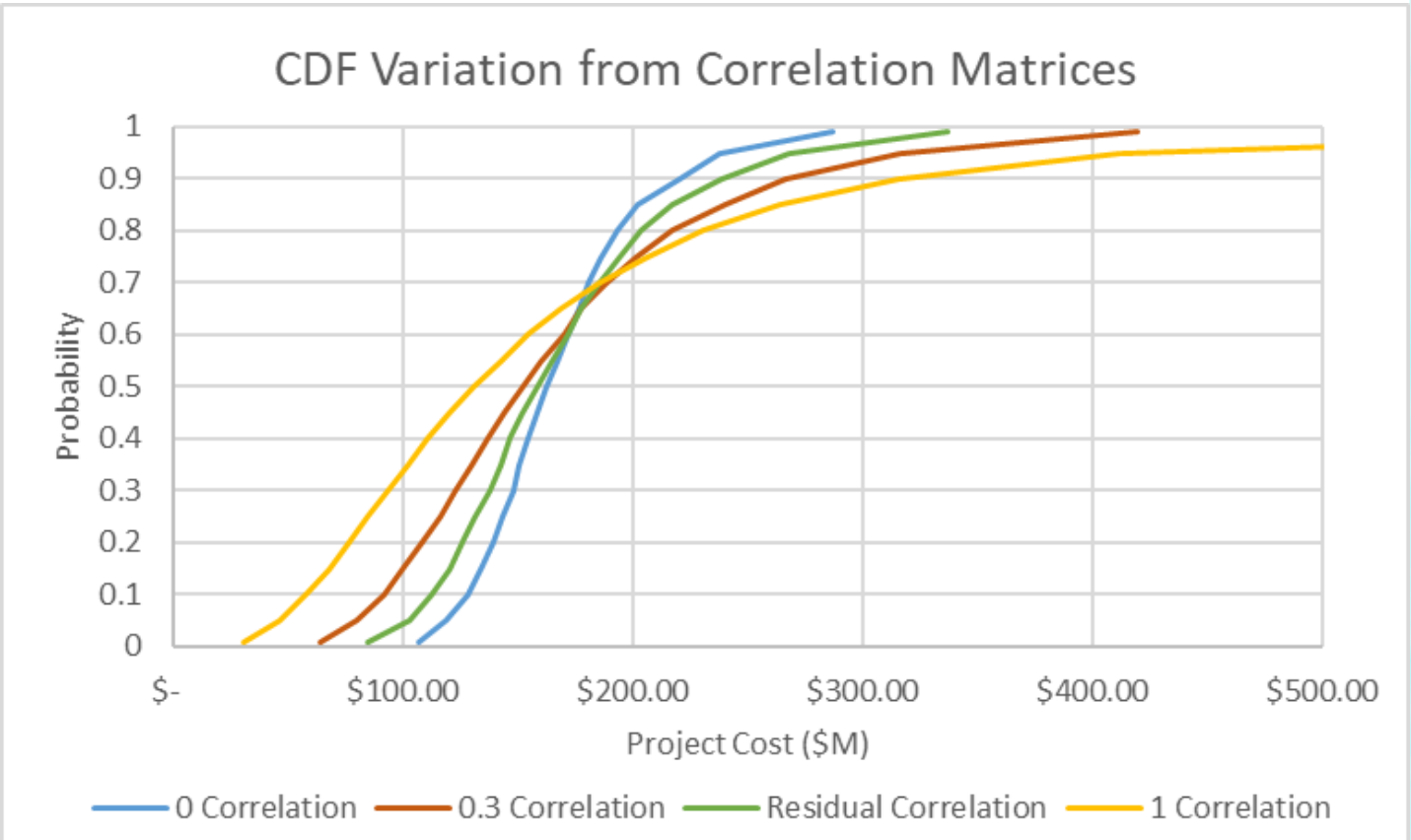
Data-Driven Correlation Matrix Produced Using PCEC Subsystem CER Residuals

- 190 unique subsystem correlation coefficients combinations
  - 120 positive, 70 negative
- Most Positive Correlation: 0.62
  - Structures & Mechanisms Recurring Cost/Electrical Power & Distribution Recurring Cost
- Most Negative Correlation: -0.47
  - Communications Recurring Cost/Mission Operations and Ground Data Systems Development
  - Command & Data Handling Nonrecurring Cost/Phase E Mission Operations & Data Analysis

Subsytem Correlation Matrix (Unit)																				
	Attitude Control	Attitude Control RC	CDH NRC	CDH RC	Comm NRC	Comm RC	Power NRC	Power RC	Propulsion NRC	Populsion RC	Structures NRC	Structures RC	Thermal NRC	Thermal RC	PM	SE	MA	I&T	MOS-GDS Dev	MODA PhE
Attitude Control NRC	1.00	0.32	0.23	0.01	-0.01	-0.13	0.38	0.28	0.19	0.14	0.06	-0.15	0.11	0.04	-0.03	0.15	-0.07	0.00	0.13	-0.10
Attitude Control RC	0.32	1.00	0.09	0.06	-0.15	0.39	0.05	0.54	-0.26	0.16	-0.28	0.19	0.02	0.24	-0.09	-0.09	-0.16	-0.06	0.22	-0.37
CDH NRC	0.23	0.09	1.00	0.26	-0.09	0.04	0.14	-0.03	0.03	-0.06	0.11	0.10	0.35	0.05	0.12	0.01	-0.05	0.22	0.18	-0.47
CDH RC	0.01	0.06	0.26	1.00	0.37	0.26	-0.24	-0.12	-0.18	-0.09	0.12	0.18	0.01	0.26	0.05	0.13	0.26	0.12	-0.13	-0.14
Comm NRC	-0.01	-0.15	-0.09	0.37	1.00	0.54	0.21	-0.12	0.18	-0.27	0.39	-0.07	-0.17	0.08	0.09	0.11	-0.02	0.10	-0.02	-0.03
Comm RC	-0.13	0.39	0.04	0.26	0.54	1.00	0.19	0.45	-0.08	0.10	0.06	0.42	-0.08	0.15	0.21	0.15	0.11	0.06	-0.47	0.00
Power NRC	0.38	0.05	0.14	-0.24	0.21	0.19	1.00	0.49	0.53	0.09	0.14	0.26	-0.01	0.20	0.22	0.06	-0.02	-0.13	-0.19	-0.06
Power RC	0.28	0.54	-0.03	-0.12	-0.12	0.45	0.49	1.00	-0.02	0.52	-0.11	0.62	-0.01	0.15	0.18	0.12	0.14	-0.10	-0.09	0.10
Propulsion NRC	0.19	-0.26	0.03	-0.18	0.18	-0.08	0.53	-0.02	1.00	0.01	0.07	-0.06	-0.24	-0.03	0.13	-0.09	0.03	-0.03	0.28	0.31
Populsion RC	0.14	0.16	-0.06	-0.09	-0.27	0.10	0.09	0.52	0.01	1.00	-0.11	0.51	0.35	0.31	-0.03	0.04	0.06	-0.32	-0.20	0.34
Structures NRC	0.06	-0.28	0.11	0.12	0.39	0.06	0.14	-0.11	0.07	-0.11	1.00	0.23	0.27	-0.15	0.02	0.43	0.33	0.18	0.01	0.25
Structures RC	-0.15	0.19	0.10	0.18	-0.07	0.42	0.26	0.62	-0.06	0.51	0.23	1.00	0.15	0.26	0.20	0.28	0.24	0.04	-0.39	0.20
Thermal NRC	0.11	0.02	0.35	0.01	-0.17	-0.08	-0.01	-0.01	-0.24	0.35	0.27	0.15	1.00	0.57	0.11	0.36	-0.15	-0.29	-0.08	0.01
Thermal RC	0.04	0.24	0.05	0.26	0.08	0.15	0.20	0.15	-0.03	0.31	-0.15	0.26	0.57	1.00	0.01	0.09	-0.20	-0.43	-0.15	-0.02
PM	-0.03	-0.09	0.12	0.05	0.09	0.21	0.22	0.18	0.13	-0.03	0.02	0.20	0.11	0.01	1.00	0.52	0.31	0.25	-0.15	0.00
SE	0.15	-0.09	0.01	0.13	0.11	0.15	0.06	0.12	-0.09	0.04	0.43	0.28	0.36	0.09	0.52	1.00	0.54	0.33	-0.20	0.02
MA	-0.07	-0.16	-0.05	0.26	-0.02	0.11	-0.02	0.14	0.03	0.06	0.33	0.24	-0.15	-0.20	0.31	0.54	1.00	0.43	-0.13	0.51
I&T	0.00	-0.06	0.22	0.12	0.10	0.06	-0.13	-0.10	-0.03	-0.32	0.18	0.04	-0.29	-0.43	0.25	0.33	0.43	1.00	0.13	-0.15
MOS-GDS Dev	0.13	0.22	0.18	-0.13	-0.02	-0.47	-0.19	-0.09	0.28	-0.20	0.01	-0.39	-0.08	-0.15	-0.15	-0.20	-0.13	0.13	1.00	-0.13
MODA PhE	-0.10	-0.37	-0.47	-0.14	-0.03	0.00	-0.06	0.10	0.31	0.34	0.25	0.20	0.01	-0.02	0.00	0.02	0.51	-0.15	-0.13	1.00

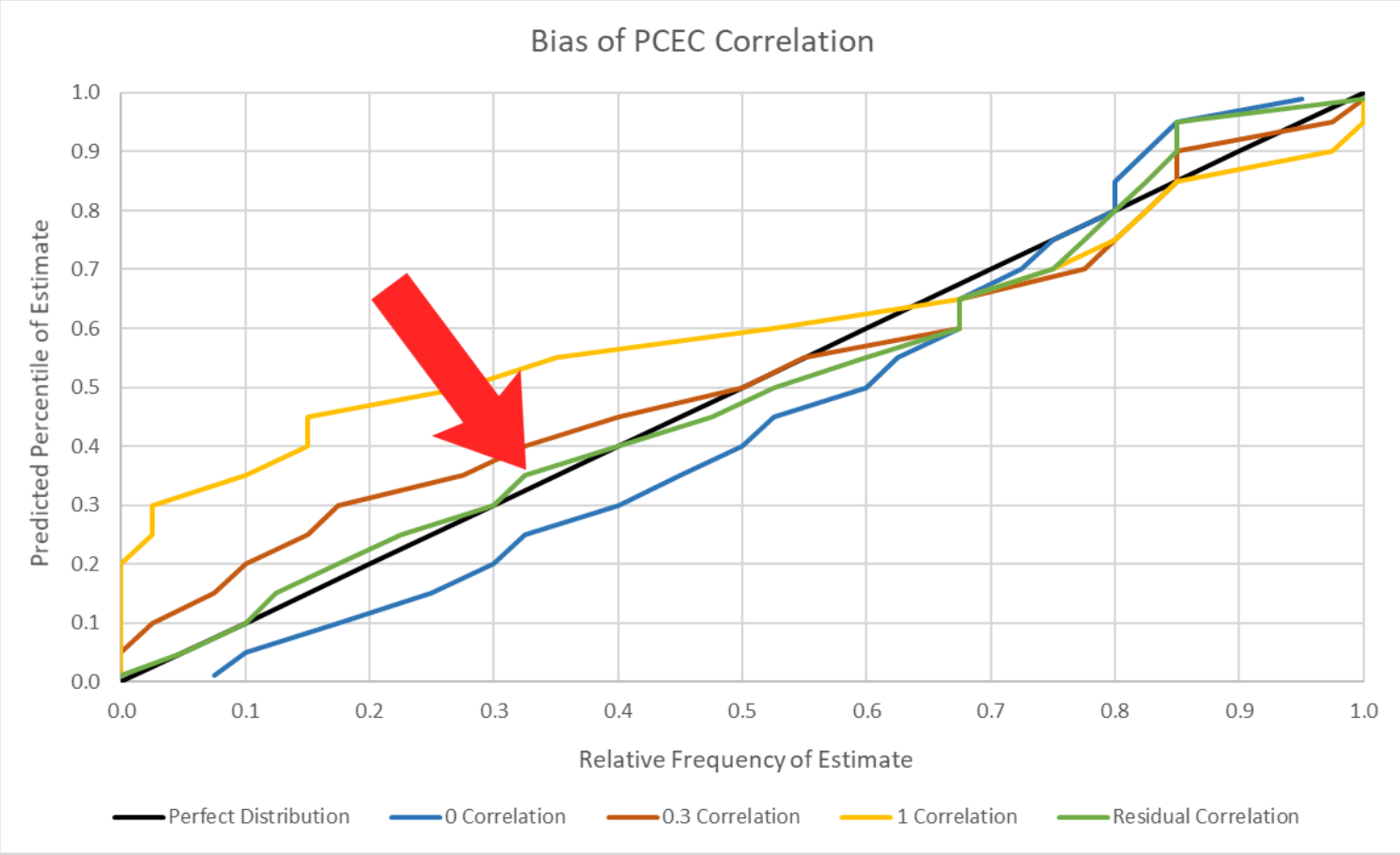
Cumulative Distribution Function

- The cumulative distribution functions (CDFs) of each simulation represent the range of possible costs that exist for each simulation
- The blanket correlation matrix of 1s produced the CDF with the widest range
- The blanket correlation matrix of 0 produced the CDF with the smallest range
- Correlation of residuals matrix CDF falls between 0 correlation simulation and 0.3 correlation
  - The **correlation of residuals** matrix has a more nuanced fit on the data than applying blanket correlation



Bias Using Correlation of Residuals Methodology

- **Bias** – tendency of a statistic to over or underestimate the population parameter that is being measured
- Calibration Charts are a technique of visually representing bias present within a model's predictive capability
  - Graphs the average observed frequency of the simulation producing an estimate result that falls below the expected percentile value against the expected frequency of each percentile
  - Compares a statistically ideal distribution of project costs against the distribution of project costs that was produced by the PCEC simulation
- The **correlation of residuals** matrix produces the line that most closely resembles the ideal distribution of project costs



Key Takeaway: The correlation of residuals matrix produced the simulation with the lowest bias

Mean Percent Error Using Correlation of Residuals Methodology

- Mean Percent Error (MPE) measures the average strength of model error
- The closer MPE is to 0%, the lower the average error in the predictive quality of the model
  - Positive MPE values indicate that on average, the observed values tend to overestimate the actual value
  - Negative MPE values indicate that on average, the observed values tend to underestimate the actual values
- The correlation of residuals matrix produces the lowest MPE

Simulation	Mean Percent Error	Missions Forecasted Within ±10% MPE	Missions Forecasted Within ±5% MPE
0 Correlation	22.08%	6	4
0.3 Correlation	87.44%	4	2
Residual Correlation	-0.13%	11	6
1 Correlation	-6.24%	11	8

Key Difference: Bias chart shows how frequently costs are being underestimated; MPE shows average strength of model error

Key Takeaway: The correlation of residuals matrix produced the simulation with the lowest error on average

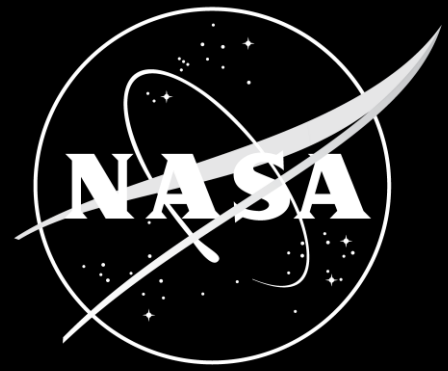
Scope and Limitations

1. This research does not attempt to determine causal factors of subsystem variation. There are many possible causal factors that could affect subsystem variation and it is outside the scope of this research to attempt to identify these factors or understand their influence
2. This research does not attempt to produce a correlation matrix that is applicable within any cost model
   - The correlation matrix in this paper was created using data from the PCEC Library and from the residuals of specific CERs within PCEC; therefore, it is not applicable in cost estimating models apart from PCEC
   - The methodology used to create this correlation matrix can be applied to other parametric models in a similar manner, however
3. This research shows how the model is correlated between subsystems according to the correlation of residuals methodology. It does not show how subsystems should be ideally correlated
4. The point estimates associated with the correlation matrix that is explored in this research are only derived from spacecraft **subsystems** and **support functions** that are referenced in the NASA Work Breakdown Structure. Therefore, the following costs are not included within the predictions that are being observed from PCEC:
   - Launch Service Vehicles
   - Instruments
   - Science

Conclusion

- The correlation of residuals methodology has demonstrated improvement in the predictive capability of PCEC compared to the current standards of no correlation or a blanket 0.3 correlation matrix
  - Bias chart indicates that the correlation of residuals methodology produces the line that most closely resembles the ideal statistical distribution of project costs
    - Less bias in PCEC than current correlation standards
    - Less likely to overestimate or underestimate project costs than the current correlation standards
  - MPE calculations indicate that the correlation of residuals methodology also improves the average accuracy of cost estimates
- These results indicate that the correlation of residuals methodology is likely to capture a more **realistic** distribution of project cost performance than current standards
  - This does not mean that the estimates produced are going to be lower, but they will likely be more accurate and reflective of real-world spacecraft development costs

Key Takeaway: The correlation of residuals methodology can be used on other parametric cost estimating models in the future to improve the accuracy and precision of cost estimates and paint a more realistic picture of the possible range of project costs.



www.nasa.gov

Backup

Previous Research on Creating Data-Driven Correlation Matrices

Derivative Correlation: Coleman & Gupta [1994]

Findings of this research:

1. In parametric cost estimating, there are many types of functional correlations that exist and can result from the mathematical relationships between various combinations of cost drivers and CERs
   - When functional correlation is present, correlation is usually handled without defining a correlation matrix since this correlation is already inherent in the model
2. Functional correlation that is observed in the data due to the production of functional dependencies within CERs also produces a "derivative" correlation among variables that are not jointly observed
   - Derivative correlation arises as a natural outcome and is inevitable

Causal Statistical Correlation: Covert [2006]

Findings of this research:

1. CER residuals can be used to determine the correlation values that should be implemented into the model to mimic the behavior of derivative correlation and improve the accuracy of cost estimates
   - When CERs are created, most of the uncertainty in the model can be explained through the functional correlation of the regression equation
   - Additional unexplained uncertainty remains in the model because each independent variable is correlated with cost and/or standard error to some degree
   - The correlation that exists between the unused independent variables is not included in the functional correlation of the regression equation
   - The remaining causal correlation is the "causal statistical correlation" between the unused independent variables and the standard error of the regression analysis
     - This causal statistical correlation is not inherent in the predictions of a cost model but can be implemented in the model as **a correlation matrix of all the independent variables**
- For the purposes of this research, this can be modeled as a correlation matrix of the subsystems of the typical NASA WBS

Causal Statistical Correlation

- **Causal Statistical Correlation** – correlation that is causal in nature without a functional relationship defined in the model
  - We know that two random variables covary, but we have not modeled the relationship with an equation
  - We use correlation coefficients to mimic their behavior
- Correlation starts when we develop CERs
  - When Regression Analysis is performed, most of the uncertainty in the model can be explained by the functional correlation of the regression equation.
  - Additional unexplained uncertainty remains in the model because each independent variable is correlated with cost and/or standard error to some degree. How do we find the remaining causal correlation?
    - The remaining causal correlation is the causal statistical correlation that can be modeled using the correlation coefficients for the subsystems of the WBS
      - This is what we are trying to model with the blanket 0.3 correlation
      - Based on the findings of Coleman and Gupta [1994] and Raymond Covert [2006]

Hypothesis: Correlation Coefficients for the subsystems can be derived through analysis of the residuals from the regression equation

References

[1]Cost Estimating Handbook 4th ed., vol. 1. Hampton, VA: NASA Cost Analysis Division, 2015, p. 25.

[2] NASA Project Cost Estimating Capability Help Documentation, NASA Engineering Cost Office, pp. 5–12.

[3] Scientific and Technical Information Program and S. M. Terrell, Marshall Space Flight Center, 2018.

[4] A. Prince, S. Hayes, M. Jacobs, and B. Alford, “PCEC v 2.3 Robotic Mission CER Development Process and Validation,” 2014.

[5] B. Alford and S. Hayes, “Modeling Regression Error,” International Cost Estimating and Analysis Association, Mar. 2015.

[6] A. Smith, “Examination of Functional Correlation and Its Impacts on Risk Analysis” in Society of Cost Estimating and Analysis, 2007.

[7] C. Smart, “Robust Default Correlation for Cost Risk Analysis,” p. 4.

[8] R. Coleman and S. Gupta, “On Overview of Correlation and Functional Dependencies in Cost Risk Analysis” in 28th Annual DoD Cost Analysis Symposium, 1994.

[9] S. Book, “Why Correlation Matters in Cost Estimating” in 32nd Annual DoD Cost Analysis Symposium, 1999.

[10] R. Covert, “Correlations in Cost Risk Analysis” in 2006 Annual SCEA Conference, 2006.

Acknowledgments

I would like to thank Brian Alford, Shawn Hayes, and Mark Jacobs for their help in understanding PCEC's near-Earth robotics CERs and the possible applications of Argo within these CERs. I would also like to thank Charley Hunt and James Johnson for collecting much of the historical and statistical research necessary in the development of this paper.