COLLEGE OF MANAGEMENT AND TECHNOLOGY

VIRTUAL RESEARCH SYMPOSIUM OCTOBER 2021

STATISTICAL TECHNIQUES TO IMPROVE SOFTWARE EFFORT ESTIMATION DATA QUALITY FOR COST ENGINEERS

PRESENTER NAME: DR. TOMEKA S. WILLIAMS

PROGRAM: PHD IN MANAGEMENT

CHAIR: DR. NIKUNJA SWAIN

MAY 2022

WALDEN UNIVERSITY

Presented at the 2022 ICEAA Professional Development & Training Workshop: www.iceaaonline.com/pit2022

Strategic Performance & Investment Economist

Dr. Tomeka S. Williams helps guide how best to procure and acquire engineering systems and solutions that are affordable, cost-effective, and risk-managed with ethics in mind. Dr. Williams recently worked at the JAIC to support as the lead AI/ML Strategic Performance and Investment Economist for the DoD AI CoE. She has over 20+ years of business, cost analysis & investment management experience and has worked with several federal agencies in these disciplines to support business process re-engineering, operational/functional requirement development, performance metrics strategies, planning, programming & budgeting to bring real value to stakeholders. Dr. Williams has a broad range of experience providing economic analysis and investment engineering support to federal agencies, specializing in Cost Transparency in the development of Life-cycle Cost Estimates (LCCE), Business Case Analysis (BCA), Analysis of Alternatives (AoA), and Strategic Portfolio & Financial Analytics for IT systems, Autonomous Systems, Augmentation Systems, and Artificial Intelligence. She is also the Integration & Operations Manager for the National Security Engineering Center Federally Funded Research & Development Center's (FFRDC's) Cost & Business Analytics Department, overseeing and managing ~80 economic and business analysts. Dr. Williams earned a BS degree in Business with an emphasis in Marketing Research from the North Carolina Agricultural & Technical State University and an MBA from Pepperdine University. She completed her latest studies at Walden University where she has earned a Masters of Philosophy & a Ph.D. specialized degree in Engineering Management & Economics.

Problem Statement

The specific management problem is that there is a lack of research into the techniques to handle the unreliable and incomplete data problem found in the U.S. cost estimation discipline.

Purpose

The purpose of this quantitative study was to:

-Test and measure the level of predictive accuracy of missing data theory techniques that were referenced as traditional approaches in the literature,

-Compare each theories' results to a complete data matrix used in support of the U.S. defense cost estimation discipline, and

-Determine which theories rendered incomplete and missing data sets in a single data matrix most reliable and complete under eight missing value percentages.

Significance

This study is important because "reliable and comprehensive cost data is essential to produce credible cost estimates as required in both (policy) statute and regulation" and not addressing how to physically handle missing values when building a cost estimate can no longer afford to be ignored in the context of the U.S. defense cost estimation discipline (Morin, 2017; 10 U.S. Code § 2334, 2017).

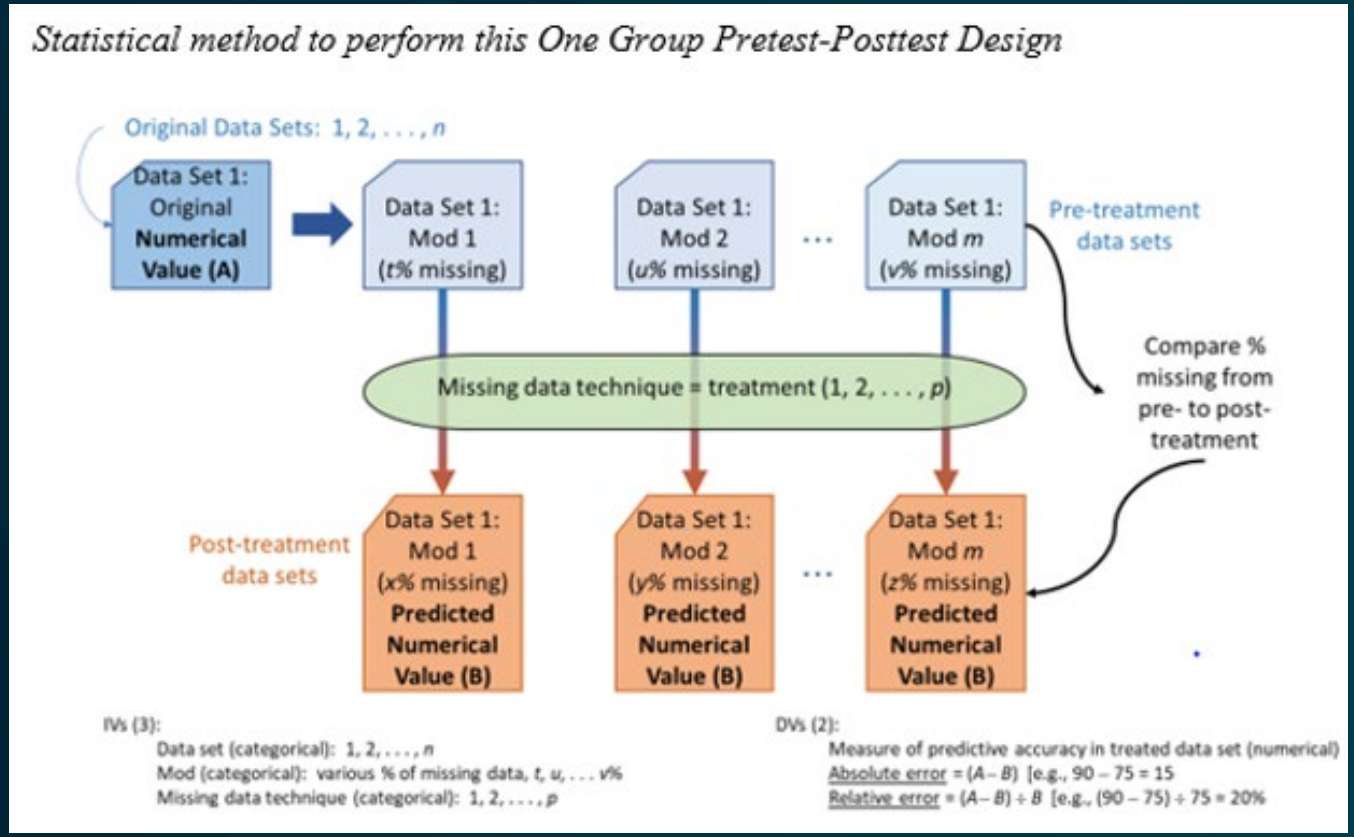
Theory or Framework

The theoretical framework for this study was grounded in missing data theory, a theory that describes the different ways in which missing data can be handled. This includes but is not limited to complete case analysis techniques such as the listwise delete and direct imputation techniques such as single imputation and multiple imputation (e.g., linear regression).

Presented at the 2022 ICEAA Professional Development & Training Workshop: www.iceaaonline.com/pit2022

RELEVANT SCHOLARSHIP

- U.S. public policy requires data reliability/completeness in cost estimation
- Effect of U.S. public policy on cost estimation data reliability/completeness
- Gap: Expanding options to handle the estimator's unreliable/incomplete data problem
- How other disciplines handle unreliable/ incomplete data problems
- Could missing data theory improve data reliability/completeness?



Research Question

To what degree can missing data theory (MDT) techniques accurately solve cost estimator's and engineering manager's unreliable and incomplete data problem when data values are missing from a representative U.S. Defense Cost Estimating data matrix?

Procedures

-Used IBM SPSS 25 Missing Value Analysis to calculate Multiple Imputation-Linear Regression (MI-LR) "Predicted Values"

-Used Microsoft Excel to calculate listwise delete (LD) and Single Imputation-Mean (SI-Mean) "Predicted Values"

-Used Microsoft Excel to record, test and measure each MDT technique

-Used IBM SPSS 25 to conduct Two Way Repeated Measures ANOVA due to LD posttest values not being a feasible technique to go through all phases of the pre-experimental research design

Participants/Secondary Data

Obtained synthetic data from a non-proprietary U.S. cost estimation data repository from a corporate university course. Used software effort data from 30 out of 55 analogous fictitious paired software resource data reports (SRDRs) programs used in an assigned case study from the corporate university's BCF 250 Software Cost Estimation Course

Data Analysis

- Data was analyzed from the results of 4,704 pre-experimental treatments in this one group pretest-posttest no control group/pre-experimental design.
- All numerical values generated from the pretest and posttest numerical values used the ratio/scale of measurement.
- I used absolute error and relative error calculations, approximation error terms, to measure the predictive accuracy of each application of traditional missing data theory techniques (Kreinovich, 2012).

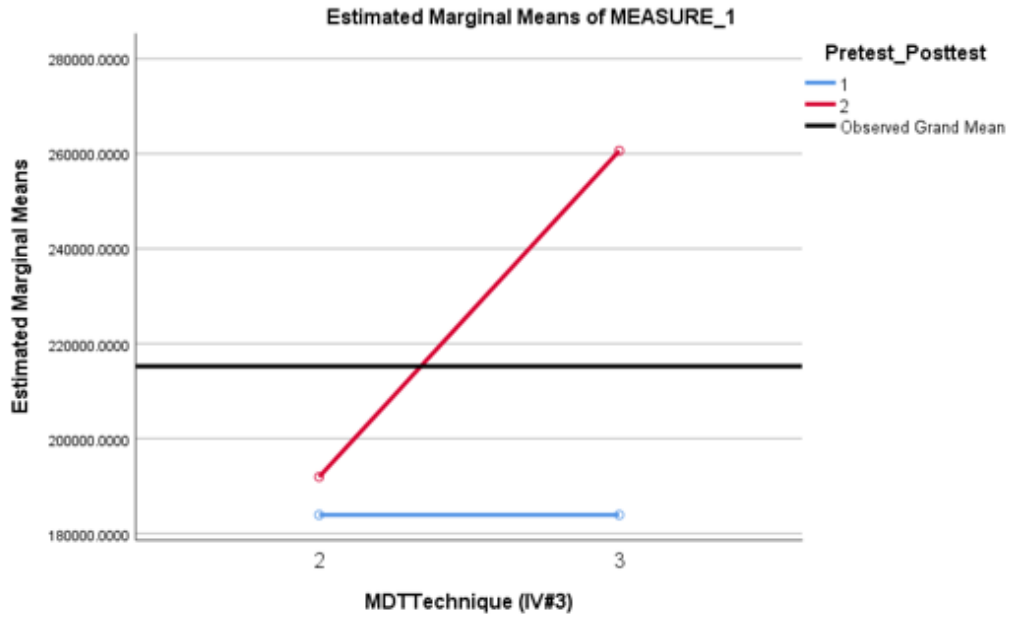
FINDINGS

- There were 28 data sets in which all 30 synthetic and analogous software programs were used as potential cases for the pre-experimental treatments, randomly resulting in 1,568 trials being conducted as an intervention per each MDT technique.
- The degree to which all three empirically tested MDT technique results came closest to the "ground truth" or "Original Numerical Value" is in the summary table on the top right.
- The estimated marginal means chart on the bottom right provides a graphical illustration that the means are not equal for the dependent variables in each MDT technique that went through all phases. If an interaction occurred between the means, the horizontal blue line and diagonal red line would have touched at least at one or two points. No interaction occurred between the two dependent variables in the SI-Mean or MI-LR techniques.

Degree to Which Missing Data Theory Techniques Can Solve the U.S. Cost Estimators' Unreliable and Incomplete Data Problem Based on Tiered Approximation Error Ranges

Tiered Predictive Accuracy Calculated Results (Approximation Error Ranges)	MDT #1 LD	MDT #2 SI-Mean	MDT#3 MI-LR
Within 80.1% or more of Original Value	N/A	47.5%	48.7%
Within 60.1% to 80.0% of Original Value	N/A	10.1%	8.9%
Within 40.1% to 60.0% of Original Value	N/A	11.8%	10.9%
Within 20.1% to 40.0% of Original Value	N/A	14.2%	12.9%
Within .001% to 20.0% of Original Value	N/A	16.4%	18.6%

Plot of the Results to Assess Interaction Between Approximation Error (Dependent Variables) of the Actual/Pretest (1) and Computed Posttest (2) Value Means



Interpretation

The key finding was that out of the three missing data theories applied, SI-Mean had the strongest level of predictive accuracy when experimental results were assessed at the individual data set level. Out of the 28 data sets results (Tables A1-A28), SI-Mean had a lower absolute and relative error in 16 data sets compared to only eight having the least amount of approximation error in MI-LR techniques. Considering many studies before me have acknowledged that multiple imputation has better prediction accuracy. Both techniques performed equally well on data sets 11, 12, and 26.

Limitations

The research design of this study was limited based on the instrumentation selected to test predictive accuracy.

Not having a control group for the one group pretest-posttest pre-experimental research design is a weakness for experimental designs.

Recommendations

This research and quantitative study provides a different perspective to address a problem, since 1972, that could potentially be mitigated by applying missing data theory as hands-on-options address the U.S. cost estimator's unreliable and incomplete data problem based on this study's pre-experimental research

Social Change Implications

This study provided an opportunity for societal change by investigating how cost estimators, engineering economists, and engineering managers could benefit from additional options that directly:

- Improve data incompleteness;
- Create better estimate predictions; and
- Ultimately reduce taxpayer funds that are spent on defense acquisition cost overruns (Schwartz & O'Connor, 2016).

REFERENCES

- Allison, P. D. (2002). Quantitative applications in the social sciences: missing data. Thousand Oaks, CA: Sage. doi:10.4135/9781412985079
- Christensen, D. S., & Gordon, J. A. (1998). Does a rubber baseline guarantee cost overruns on defense acquisition contracts? Project Management Journal, 29(3), 43-51. doi: 10.1177/875697289802900307
- Defense Acquisition University. (2018a). Business: Cost estimating courses. Retrieved from https://www.dau.mil/cop/ce/Pages/Course.aspx
- Defense Acquisition University. (2018b). Policy for business: Cost estimating. Retrieved from https://www.dau.mil/policy#Business%20Cost%20Estimating|All|All|All||recent
- Deloitte. (2016). Cost overruns persist in major defense programs. Retrieved from https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/cost-overruns-persist-in-major-defense-programs.html

REFERENCES

- Government Accountability Office. (1972). Theory and Practice of Cost Estimating for Major Acquisitions (Report No. B-163508) Washington, D.C.: U.S. Government Printing Office. Retrieved from http://www.gao.gov/assets/210/200036.pdf
- Government Accountability Office. (2009). GAO Cost Estimating and Assessment Guide: Best Practices for Developing and Managing Capital Program Costs (Report No. GAO-09-3SP). Washington, D.C.: U.S. Government Printing Office. Retrieved from https://www.gao.gov/new.items/d093sp.pdf
- Government Accountability Office. (2020). GAO Cost Estimating and Assessment Guide: Best Practices for Developing and Managing Program Costs (Report No. GAO-20-195G). Washington, D.C.: U.S. Government Printing Office. Retrieved from https://www.gao.gov/assets/710/705312.pdf
- Graham, J. W. (2012). Missing data theory. In Missing Data (pp. 3-46). Springer, New York, NY. doi: https://doi.org/10.1007/978-1-4614-4018-5_1

REFERENCES

- International Cost Estimation and Analysis Association. (2019). International Cost Estimation and Analysis Association Testable Topics List. Retrieved from http://www.iceaaonline.com/ready/wpcontent/uploads/2014/02/testableTopicsList.pdf
- Joint Agency Cost Estimating Relationship (CER) Development Handbook. (2018, February 9). Retrieved from https://www.asafm.army.mil/Portals/72/Documents/Offices/CE/CER%20Developement%20Handbook.pdf
- Kreinovich, V. (2012). How to define relative approximation error of an interval estimate: a proposal. Retrieved from http://www.cs.utep.edu/vladik/2012/tr12-37.pdf
- Little, R. J., & Rubin, D. B. (1987). Statistical analysis with missing data (Vol. 793). Hoboken, New Jersey: John Wiley & Sons.
- Little, R. J., & Rubin, D. B. (2002). Statistical analysis with missing data (Vol.2). Hoboken, New Jersey: John Wiley & Sons.

REFERENCES

- Little, R. J., & Rubin, D. B. (2020). Statistical analysis with missing data (Vol.). Hoboken, New Jersey: John Wiley & Sons.
- Office of the Under Secretary of Defense (OUSD) for Acquisition, Technology and Logistics (AT&L). (2019). Acquisition resources and analysis (ARA) directorate MDAP and MAIS list. Retrieved from https://www.acq.osd.mil/ara/documents/mdap_mais_program_list.pdf
- Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592. doi:10.1093/biomet/63.3.581
- Saeed, A., Butt, W. H., Kazmi, F., & Arif, M. (2018, February). Survey of software development effort estimation techniques. In Proceedings of the 2018 7th International Conference on Software and Computer Applications (pp. 82-86). ACM. doi:10.1145/3185089.3185140

REFERENCES

- Seo, Y. S., Yoon, K. A., & Bae, D. H. (2009, December). Improving the accuracy of software effort estimation based on multiple least square regression models by estimation error-based data partitioning. In 2009 16th Asia- Pacific Software Engineering Conference (pp. 3-10). IEEE. doi:10.1109/APSEC.2009.57
- Valerdi, R., Dabkowski, M., & Dixit, I. (2015). Reliability improvement of major defense acquisition program cost estimates—Mapping DoDAF to COSYSMO. Systems Engineering, 18(5), 530-547. doi:10.1002/sys.21327

COLLEGE OF MANAGEMENT AND TECHNOLOGY

VIRTUAL RESEARCH SYMPOSIUM OCTOBER 2021

FINAL REMARKS

QUESTIONS AND ANSWERS

OCTOBER 28-29, 2021

WALDEN UNIVERSITY