Investigating Causal Effects of Software and Systems Engineering Effort

Anandi Hira, Jim Alstad; Mike Konrad
A.Hira@usc.edu, JAlstad@usc.edu; MDK@sei.cmu.edu
University of Southern California; Software Engineering Institute

Joint Software and IT Cost Forum 2020
September 22, 2020
Version 4

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM20-0709

Presentation Outline

- Motivation
- Causal Analysis in SW Cost

- Intro to Causal Inference
- Algorithms and Tool Used

- Datasets – COCOMO® II, COSYSMO 3.0

- Approaches and Results

- Conclusions and Questions

Motivation

- Managers are frequently faced with issues of controlling project costs
  - My cost estimate is too high. What project aspects can I modify that would most likely reduce the cost?
  - I have some money to improve my organization's performance. Changing which organizational aspects would be most likely to improve performance?
  - I need to add a new stakeholder and remove flexibility in modifying requirements to this acquisition. Is that likely to have a significant influence on cost?
- Causal Analysis is a modern technique that analyzes datasets to determine causal relationships among its variables
- **Our Research Goal:** Identify factors of software and systems engineering costs that are direct causes
  - To help manage real projects

History of Causal Analyses for Effort

Boehm – COCOMO® Models

- In-depth behavioral analyses for effort drivers
- Including COSYSMO models

Evidence-Based SW Engineering (Kitchenham et al)

- Suggests running Experiments to identify causal relationships:
  - Cause precedes effect
  - Cause covaries with effect
  - Alternative explanations are implausible

Hira et al – Unified Code Count maintenance (USC)

- Software maintenance and upgrade data
- Project data has limited scope
  - Similar projects, from a single environment

- Our difference: 2 calibration datasets (observational data) with varying values of cost drivers, application types, and project types

Presentation Outline

- Motivation
- Causal Analysis in SW Cost

- Intro to Causal Inference
- Algorithms and Tool Used

- Datasets – COCOMO® II, COSYSMO 3.0

- Approaches and Results

- Conclusions

We Employ Causal Inference as the Basis of Our Research

- **Causal Inference**
  - **Causal Search/Discovery**
    - Algorithms and domain knowledge on observational data
  - **Causal Estimation**
    - Algorithms to quantify causal influence; structural equation modeling (SEM)

Causal Discovery Algorithm Results

Result is a “causal graph”, with boxes representing variables and edges representing causal relationships.

Here are the different possible types of edges (parent-and-child relationships):

$\boxed{X_1} \longrightarrow \boxed{X_2}$	X_1 directly causes X_2
$\boxed{X_1} \text{———} \boxed{X_2}$	X_1 directly causes X_2 or X_2 directly causes X_1
$\boxed{X_1} \qquad \boxed{X_2}$	No directly causal relationship between X_1 and X_2

Causal Discovery Algorithms

PC Search

- Constraint-based algorithm
- Developed by Peter Spirtes and Clark Glymour
- Repeatedly tests dataset to determine a set of conditional independences (“constraints”) to determine where edges should not be placed
- The threshold used for determining independence from p-values is called Alpha and is set by the user
- PC’s strengths are: (1) independence is an intuitive concept; (2) PC is modular, allowing different tests to be employed, to match assumptions about data distributions.

FGES Search

- Score-based algorithm
- Developed by Christopher Meek, David Maxwell Chickering, and Joseph D. Ramsey
- Uses maximum likelihood to find the graph that best generates the dataset (has a superior model-fitness “score”)
- The score is the sum of a likelihood term and a penalty discount. The latter is set by the user to avoid over-fitting to the data.
- FGES’s strengths are: (1) resulting graph has almost all of its edges oriented; (2) score is similar to that used for model estimation.

Bootstrapping

- One bootstrap: draw a random sample (typically of size 90% or 100%) from the original dataset, with replacement
  - Run search algorithm on this sample
- Repeat this 100 or more times, and aggregate the results for all detected edges into the **edge probability table (EPT)**
  - In the EPT, each edge will show the percentage of times it was found, which reflects the fraction of data points that has this direct-causal relationship
- Stronger causal relationships appear higher in the EPT.
  - Entries further down the EPT are more likely to be due to accidental correlations
- Reduces sensitivity to small changes in dataset, improving generalizability

Graph Edges:
1. LogSize --> LogEffort [LogEffort <-- LogSize]:1.0000;
2. LSVC --> LogEffort [LSVC --> LogEffort]:0.6300;
   [no edge]:0.3700;
3. LogSize --> SCHED [LogSize <-- SCHED]:0.2300;
   [LogSize --- SCHED]:0.0800;
   [LogSize --> SCHED]:0.4900;[no edge]:0.2000;
4. DOCU --> LogEffort [DOCU --> LogEffort]:0.3800;
   [no edge]:0.6200;
5. TEAM --> ROPM [ROPM <-- TEAM]:0.3800;
   [ROPM --- TEAM]:0.1000;
   [ROPM --> TEAM]:0.2800; [no edge]:0.2400;

.

.

.

152. TRSK --> INST [INST <-- TRSK]:0.0100; [no edge]:0.9900;

Causal Estimation

- **Causal estimation** involves parameterizing the relationships appearing in the causal discovery graph and then determining what values to assign to these parameters.
  - Enables making predictions about the future values that variables will attain as a result of **hypothesized interventions** or events.
  - Causal estimation, when applied to just a single variable and its parents works like ordinary regression: **coefficients** (also called loadings) are assigned to each edge
  - A *one-unit* change in a parent, with all other variables held constant, results in a change in the child of *coefficient* units.
- The resulting model is then evaluated for **model fit**.
  - **Model fit statistics** include: Chi square (per degrees of freedom), BIC, CFI, RMSEA.
- More information can be found here:

  - [Spirtes] Spirtes, P. 2010. Introduction to causal inference. *Journal of Machine Learning Research*. 11, May (2010), 1643–1662.
  - [Kline] Kline, R. 2015. Principles and Practice of Structural Equation Modeling, Fourth Edition. Series: Methodology in the Social Sciences. The Guilford Press.

Tetrad Tool

- Implements causal algorithms
- Implemented and maintained by Center for Causal Discovery*, primarily run by Carnegie Mellon University (CMU) and University of Pittsburgh
- https://github.com/cmu-phil/tetrad
- Our results come from versions 6.5.4 (earlier results) and 6.7.0 (recent results)

*We acknowledge the Center for Causal Discovery, supported by grant U54HG008540, in maintaining the algorithms and Tetrad tool used in this research.

*https://www.ccd.pitt.edu

Presentation Outline

- Motivation
- Causal Analysis in SW Cost

- Intro to Causal Inference
- Algorithms and Tool Used

- Datasets – COCOMO® II, COSYSMO 3.0

- Approaches and Results

- Conclusions

Datasets

COCOMO® II Calibration Dataset

- 16 organizations, various application types
- Variability in all 26 variables
- 161 projects
- Boehm et al, *Software Cost Estimation with COCOMO® II* (2000)

COSYSMO 3.0 Calibration Dataset

- Covers various types of systems
  - > 2 orders of magnitude size variation
- Variability in all 18 variables
- 68 projects
- Alstad, JP, "Development of COSYSMO 3.0", *Procedia Computer Science 153* (2019)

Each dataset is reasonably representative of projects of its type

Presentation Outline

- Motivation
- Causal Analysis in SW Cost

- Intro to Causal Inference
- Algorithms and Tool Used

- Datasets – COCOMO® II, COSYSMO 3.0

- Approaches and Results

- Conclusions

Toward a Methodology for Causal Discovery on Small Samples -1

- Problem: Search did not produce structures that were very informative
  - Using strict parameter settings, few edges were found
    - Frequently only one: Size --> Effort
  - Using looser parameter settings, too many edges were found
    - Found additional plausible causes of Effort
    - But found non-plausible edges also
    - Such edges might be spurious (due to accidental correlations)—how would we know?
  - A consequence of having relatively few data points (projects)
- Systematically explored four weak-signal analysis (WSA) approaches; each uses bootstrap and sorts the EPT by **probability of no edge (PNE)**:
  - WSA 1: keep edge if PNE ≤ 40%, or < 50% and only 1 other edge orientation found
  - WSA 2: use whatever PNE nets a # of edges = # of variables. Also try two times the # of variables.
  - But is there a less arbitrary and more data-driven way to set the threshold?
  - Our final two approaches objectively set a PNE threshold. Randomized variables are added to the dataset, and edges with these variables are considered "random". The PNE near where random edges "start" is the threshold.
  - WSA 3: add a few uniformly-distributed variables between 0 and 1 ("noise variables")
  - WSA 4: add a randomly sorted copy of each variable in original dataset ("null variables")

Toward a Methodology for Causal Discovery on Small Samples -2

Example Applications of WSA #2 and WSA#3

Setting of Strictness Parameter	**Minimum PNE needed to obtain desired # edges**			
	PC		FGES	
	26 Edges	52 Edges	26 Edges	52 Edges
More Strict	0.76	0.93 (58 edges)	0.49	0.79
Less Strict	0.61	0.84 (54 edges)	0.4	0.67

# Noise Variables:		1	3	10	Va-nilla
# edges in graph/ PNE	PC	15/ .86	26/ .94	26/ .94	4
	FGES	13/ .53	26/ .67	30/ .74	3
	Union	17	26	33	

COCOMO® II dataset was searched on the left and COSYSMO 3.0 dataset was searched on the right.

Toward a Methodology for Causal Discovery on Small Samples -3

Detail of current (WSA 4) approach to causal discovery on small samples:

1. Inject **null variables**: to the original dataset, append a copy of the original variables whose values have been independently randomized
2. Search with **bootstrap**: determine for each edge terminating on a null variable its PNE
3. Set a **threshold** of **10^{th} percentile** of random edges
   - i.e., 90% of edges involving a null variable will have a PNE that exceeds that threshold
4. Trim the remaining edges (on original variables) when their PNE > threshold

Factors That Are Direct Causes of Effort
Intervening on These in a Project May Improve Outcomes

COCOMO® II - Effort

- Size (SLOC)
- Team Cohesion (TEAM)
- Platform Volatility (PVOL)
- Reliability (RELY)
- Storage Constraints (STOR)
- Time Constraints (TIME)
- Product Complexity (CPLX)
- Process Maturity (PMAT)
- Risk and Architecture Resolution (RESL)

COCOMO® II - Schedule

- Size (SLOC)
- Platform Experience (PLEX)
- Schedule Constraint (SCED)
- Effort (Log_PM)

COSYSMO 3.0 - Effort

- Size
- Level of Service Requirements (LSVC)

Causal Estimation Using Tetrad

- To quantify the effects of the direct causal relationships identified in our causal graphs, we proceeded to estimation.
  - Recall that all variables in our datasets are numeric (and continuous-valued)
  - For each variable in a causal graph, Tetrad can provide estimates for the parameters governing the linear relationship with its direct causes/parents
  - For example, suppose variable C has parents A and B. Then Tetrad can provide estimated values for intercept, $coef_A$, and $coef_B$ in this equation*:

$$C = \text{intercept} + coef_A \times A + coef_B \times B$$

- For this presentation, we provide an initial estimation of our causal graphs resulting from WSA 4 applied to our two datasets

*Note: an additional parameter is estimated: the variance of the residual.

Using Tetrad to Derive Mini-Models to Produce Plausible Cost Estimates

Guided by the existing COCOMO® II and COSYSMO 3.0 estimating models' structure.

1. The structure of the estimating models does not directly conform to that needed by Tetrad. We therefore transformed the structure of each estimating equation:
   - We took the logarithm of the equation (Size -> LogSize, etc)
   - Cost drivers and scale factors are represented differently in the linear mini model.
   - Cost drivers are additive variables, which we directly included in the mini model.
   - Scale factors are multipliers of LogSize, we replaced each with the scale factor times LogSize.

   These steps allow us to use Tetrad's built-in linear regression feature.

2. We forced cost predictor independence (with the Knowledge box in Tetrad).
3. We applied WSA 4 to obtain a plausible causal graph. We discarded any variables that have no edges.
4. We used the Tetrad Estimation capability to obtain coefficients and intercepts on the graph. The mini-model was obtained by extracting the mini-estimating equation from the resulting graph.

COSYSMO 3.0 Estimation

<u>Overall Model Fit Statistics</u>

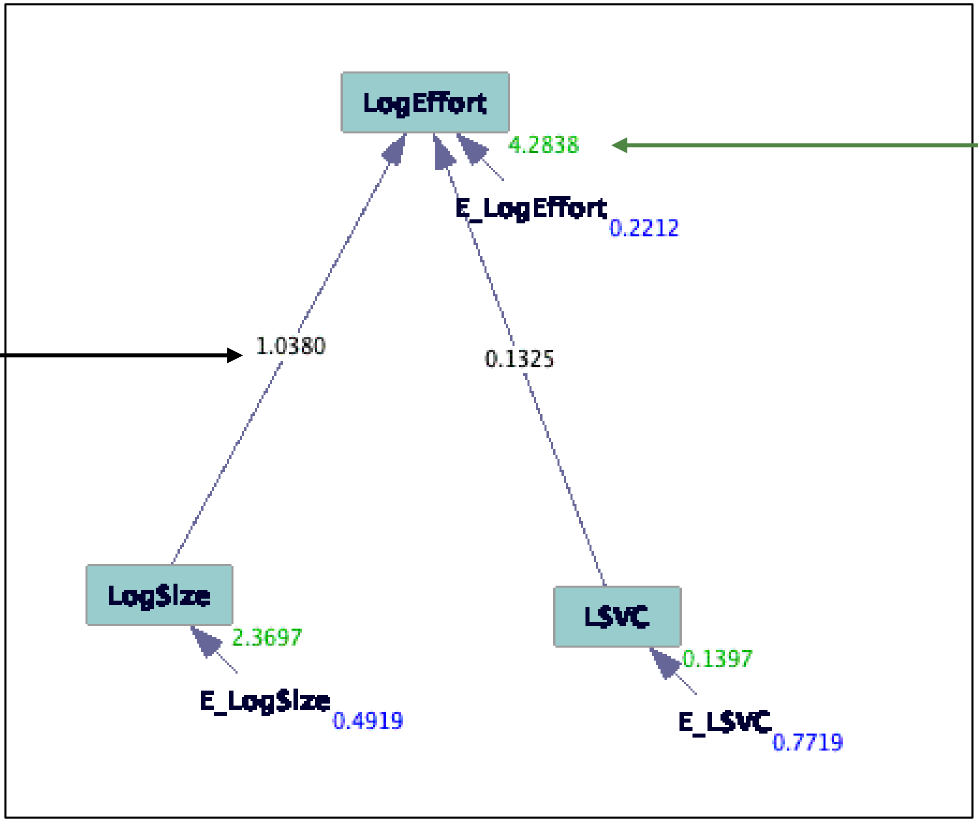
Chi-Square Test(s) of Model Fit:
P-Value = 2.3667E-5
(want > 0.01, or at the very least > 0);
Chi-Square/DF = 17.8688 (want < 5)

RMSEA (Root Mean Square Error Of Approximation):
RMSEA = 0.5018 (want < 0.08)

CFI (Comparative Fit Index):
0.9967 (want > 0.95)

Conclusion: Model fit is Poor-to-Fair.

Coefficients



Mean

How to Get a Mini-Model from that Tetrad Estimation Model

- Reading off from the Tetrad estimation model, a mini-model would be:

  $$\text{LogEffort} = 1.0380 * \text{LogSize} + 0.1325 * \text{LSVC} + 4.2838$$

- First attempt at Effort estimation:

  $$\text{Effort} = \text{Size}^{1.0380} * 1.357^{\text{LSVC}} * 10^{4.2838}$$

- That, however, doesn't work
  - The problem is that 4.2838 is the mean of the LogEffort values; however, raising 10 to that power does not yield the mean of the Effort values.
- One has to do a separate linear regression of LogEffort against LogSize and LSVC
  - That yields an exponent for 10 of 1.805, which gives this estimating equation:

    $$\text{Effort} = 63.834 * \text{Size}^{1.0380} * 1.357^{\text{LSVC}}$$

COCOMO® II Estimation

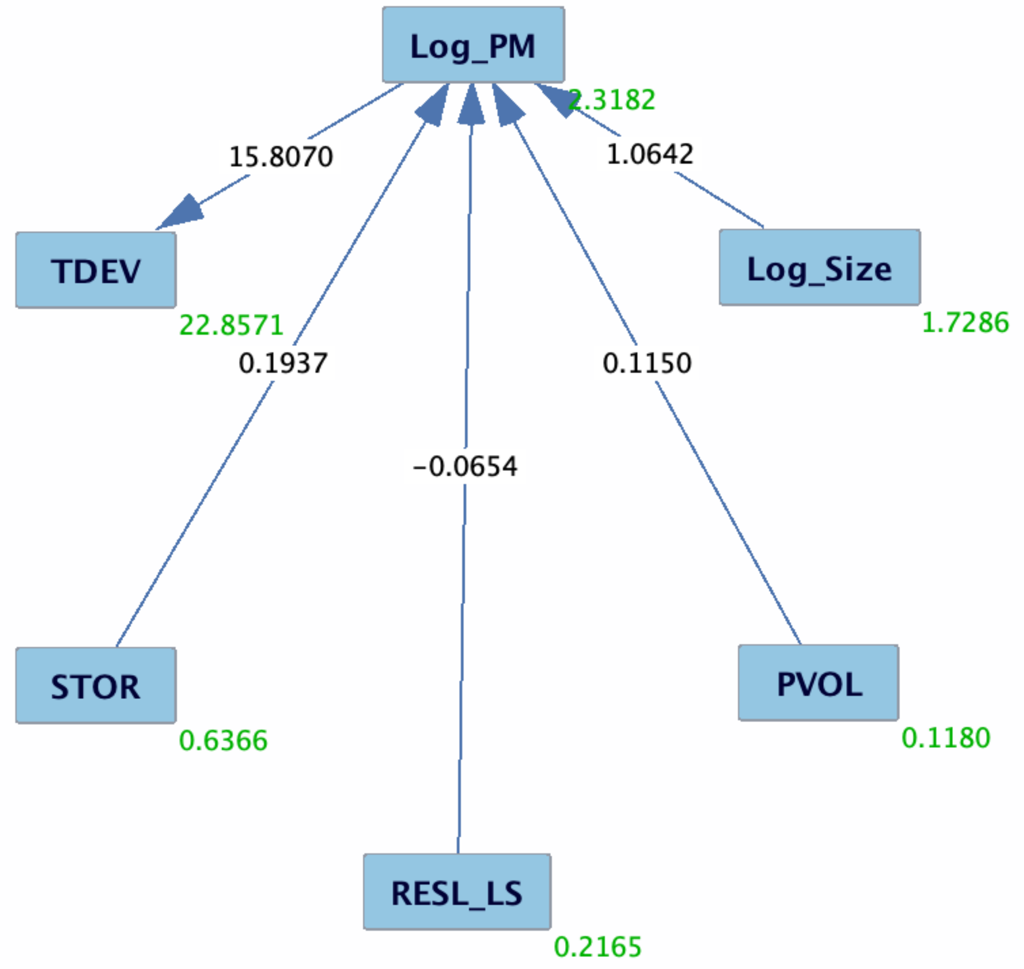
<u>Overall Model Fit Statistics</u>

Chi-Square Test(s) of Model Fit:
P-Value = 8.9571E-5
(want > 0.01, or at the very least > 0);
Chi-Square/DF = 3.58 (want < 5)

RMSEA (Root Mean Square Error Of Approximation):
RMSEA = 0.1271 (want < 0.08)

CFI (Comparative Fit Index):
0.9993 (want > 0.95)

Conclusion: Model fit is Fair-to-Good.



Prediction Accuracy: Mini-Models vs Estimating Models

COSYSMO 3.0 - Effort

	Mini-Model	Original
Max MRE	285.4%	234.8%
MMRE	45.9%	57.3%
PRED(25)	41.2%	23.5%
PRED(30)	48.5%	23.5%

COCOMO® II - Effort

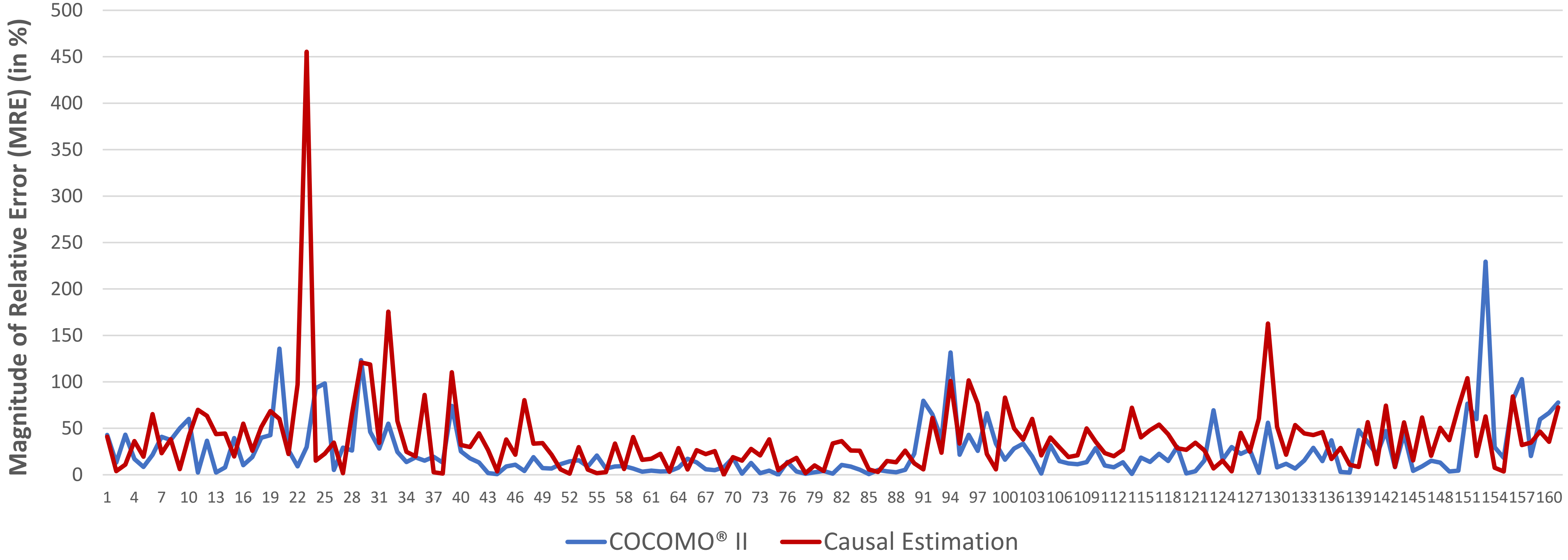
	Mini-Model	Original
Max MRE	455.4%	229.41%
MMRE	38.64%	25.67%
PRED(25)	44.72%	67.08%
PRED(30)	52.8%	74.53%

COCOMO® II - Schedule

	Mini-Model*	Original
Max MRE	628.6%	130.95%
MMRE	42.28%	50.88%
PRED(25)	45.34%	9.94%
PRED(30)	52.8%	12.42%

*Note: Analysis done with TDEV; but realized Log(TDEV) might have been better.

COCOMO® II vs Mini-Model Effort MRE Comparison



Using more factors for a cost estimate (as with the full model) tends to reduce the frequency of way-off predictions (of course, on any given project, either model might be more accurate). The advantage of the mini-model is that it uses just the factors, among many, that are more likely to drive cost and schedule.

Presentation Outline

- Motivation
- Causal Analysis in SW Cost

- Intro to Causal Inference
- Algorithms and Tool Used

- Datasets – COCOMO® II, COSYSMO 3.0

- Approaches and Results

- Conclusions

Conclusion – Causal Discovery

- Straightforward use of causal discovery algorithms may result in little information about cost-causing factors
  - Relatively small datasets (# of cases) compared to # of variables
- Weak-Signal Approaches enhanced results
  - Identified additional causes of effort and duration, while minimizing spurious correlations
  - Established a principled approach (methodology) to determining what cutoff to use for trimming results of a bootstrapped search (based on null variables and EPT)
- We identify (on slide 19) specific direct causes, where action has been shown statistically to affect the cost or schedule
  - The data we used considered multiple application types and multiple organizations
  - We also investigated choice of Tetrad search algorithm and parameter values

Conclusion – Causal Estimation

- We developed a methodology (slide 20) for generating cost estimation mini-models based on datasets that deliver plausible results
- Observation
  - Modestly fitting with inferior predictions compared to original model
- Further Research
  - More investigation in alternative estimation approaches could produce more effective Tetrad-based models for use

Considerations for Future Research

- Expensive, prohibitive experiments of acquisition factors could be obviated by use of causal methods
  - Revisit effort factors, potentially reducing number of required dataset characteristics
  - Causal discovery can also help identify additional measures to provide insight into what drives cost
- Acquisition researchers can integrate causal conclusions for holistic model
- Identify and prioritize research funding towards causal research outcomes worthy of investment in repeatability and reproducibility studies
- Causal research findings more confidently tested by acquisition program interventions with less risk of waste

Contact Information for Questions

Anandi Hira

A.Hira@usc.edu

Jim Alstad

JAlstad@usc.edu

Mike Konrad

MDK@sei.cmu.edu