Causal Models for Software Cost Control (SCOPE)

Recent results from five different studies

How can we better control costs in software development and sustainment? This project is collaborating with systems and software researchers in applying causal learning to program datasets to better understand which factors can reduce costs.

DoD Problem

- DoD leadership continues to ask “Why does software cost so much?”
- DoD program offices need to know where to intervene to control software costs

Our Solution

An actionable, full causal model of software cost factors immediately useful to DoD programs and contract negotiators

Causation Vs. Correlation

To reduce costs, what causes code quality to be good or bad needs to be understood. Correlations are insufficient. For example, in the figure below, would increasing experience level improve code quality?

Results: To achieve precision, software estimation models should include both objective measures of requirements size as well as programmer-specific coding and defect factors.

Approach: Apply Causal Discovery to data from students coding to the same ten requirements specifications.

Results: For IT-type systems, only COSMIC Function Points, Programmer Capability, and Documentation-Aligns-with-Lifecycle-Needs repeatedly recur as direct causes of total effort.

Approach: Apply Causal Discovery to the results of a static code and design structure analysis to determine which type of architectural pattern violation most affects code quality.

Results: Cyclic dependency was the single architecture pattern violation affecting code quality.

Approach: Apply Causal Discovery to an existing project-survey dataset.

Results: The original analysis identified difficult requirements, stakeholder relationships, and cognitive fog; causal discovery confirmed only cognitive fog.

Approach: Apply Causal Discovery to results of 18 months of weekly surveys of software engineers from across a DoD organization to determine which factors most affect cost, schedule, and quality.

Results: Of the 20+ factors found to be highly correlated with cost, schedule, and quality, direct causal relationships were found for only two: Good Improvement Data and Stress From Overtime.

Technical Approach

Working with collaborators, we will identify and prepare datasets for causal learning to establish key cause-effect relationships among project factors and outcomes. For example, for Quality, we might have this causal graph:

The resulting causal models will then be “stitched” using CMU algorithms to create a universal causal model, but estimated and calibrated for lifecycle and super-domain. These estimated models will be the basis for improved program management.

Collaborative Approach

First, the SEI trains each collaborator to perform causal searches on their own proprietary datasets. The SEI then only needs to be provided with information about what dataset and search parameters were used as well as the resulting causal graph, which is sufficient for integrating into a universal causal model.

Summary

Causal learning has come of age from both a theoretical and tooling standpoint and provides better basis for program control than models based on correlation. Application to cost estimation requires large amounts of quality data. Now is the time to engage the larger community of systems and software researchers in deriving improved cost models that enable improved program control.