Why Does Software Cost so Much? Towards a Causal Model

Robert Stoddard
Principal Researcher and PI

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213
Agenda

- Problem within Cost Estimation
- Why Causation instead of Correlation
- Why You Must Begin with a DAG!
- How Education Misled Us
- Landscape of Causal Learning
- General Ideas for Research
Problem Statement

Today’s cost estimating relationships (CERs) predict software and program costs

However, the DoD needs to move beyond prediction and achieve models that enable cost control, if not cost reduction

This need first begins during early lifecycle cost estimation as part of the bidding, pricing and full acquisition process (should-cost, affordability, etc…), and later during management of active programs making mid-course corrections

To achieve actionable models that facilitate control over costs, one must use causal models rather than traditional prediction models
Why Traditional Cost Estimating Relationship (CERs) Prediction Models Fall Short

Most CERs are built using traditional correlation and statistical regression modeling.

However, serious concerns exist in using these methods for the development of CERs, namely:

- What if other factors not represented in the model are responsible for the cost effects (e.g. including common causes)?
- What if there are convoluted factors impacting cost (e.g. mix of causal and simple correlated factors)?
- What if cost analysts desire to interpret the regression coefficients as the degree of influence on cost?
Agenda

Problem within Cost Estimation

✅ Why Causation instead of Correlation

Why You Must Begin with a DAG!

How Education Misled Us

Landscape of Causal Learning

General Ideas for Research
Distinguishing Causation from Correlation is Important!

Just as correlation may be nonsensical...

...so may regression, built on correlation!!!

Los Angeles Times
May 12, 2014
Agenda

- Problem within Cost Estimation
- Why Causation instead of Correlation
- Why You Must Begin with a DAG!
- How Education Misled Us
- Landscape of Causal Learning
- General Ideas for Research
Why Causal Models using Regression Must Begin with a DAG

Assume the DAG below represents our theory of how the independent factors relate to the dependent factor (Y)

![DAG diagram]

- Causal effect of factor T on Y is identifiable by “adjustment”, e.g.
  we can estimate total effect of T on Y by adjustment and then regression as follows:

Two causal paths: T -> B -> Y and T -> U₂ -> Y

But must adjust on A and E before regression to block three “noncausal paths”

Blocking or Adjusting Non-Causal Paths

1. Controlling a variable

2. Stratifying a variable

3. Setting evidence on a variable

4. Observing a variable

5. Matching a variable (eg making distributions of sub-populations as similar as possible for comparison)

Quotes by Judea Pearl

“… I see **no greater impediment to scientific progress** than the prevailing practice of focusing all of our mathematical resources on probabilistic and statistical inferences while **leaving causal considerations to the mercy of intuition and good judgment.**”


“The development of **Bayesian Networks**, so people tell me, marked a turning point in the way uncertainty is handled in computer systems. For me, this development was a stepping stone towards a more profound transition, from reasoning about beliefs to reasoning about causal and counterfactual relationships.”

Judea Pearl: From Bayesian Networks to Causal and Counterfactual Reasoning
Keynote Lecture at the 2014 BayesiaLab User Conference
Recorded on September 24, 2014, in Los Angeles.
Agenda

Problem within Cost Estimation
Why Causation instead of Correlation
Why You Must Begin with a DAG!
How Education Misled Us
Landscape of Causal Learning
General Ideas for Research
How Our Education May Have Mislead Us!

1) We were taught that controlled experiments must be conducted to conclude cause and effect.

2) However, before Sir Ronald Fisher invented “designed experiments” in 1935, cause & effect was determined through basic matching methods in data.

3) Matching methods have matured in the past 30 years, and are quietly working well in social science and medical research!!

4) Today, we have a complete causal learning framework centered on DAGs and methods such as Instrumental Variables and Propensity Scoring.
Agenda

Problem within Cost Estimation

Why Causation instead of Correlation

Why You Must Begin with a DAG!

How Education Misled Us

✔ Landscape of Causal Learning

General Ideas for Research
Why Does Software Cost So Much? Towards a Causal Model

Landscape of Causal Learning

1. Raw Observational Data
2. Statistical discovery of causal relationships to create the DAG (CMU Faculty)
3. Use theory and/or domain knowledge to hypothesize a DAG
4. Identity and magnitude of true causal parameters of cost
5. Quantifying causal relations using DAG graph surgery, Instrumental Variables and Propensity Scoring (Pearl, Elwert, Guo)
Landscape of Causal Learning

Raw Observational Data

Statistical discovery of causal relationships to create the DAG (CMU Faculty)

Use theory and/or domain knowledge to hypothesize a DAG

We use tooling from CMU called Tetrad to conduct statistical causal discovery in a data set, as depicted in the following slides.
http://www.phil.cmu.edu/tetrad/
http://www.phil.cmu.edu/tetrad/
http://www.phil.cmu.edu/tetrad/
Why Does Software Cost So Much? Towards a Causal Model
October 18, 2016
© 2016 Carnegie Mellon University

Distribution Statement A This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

CMU Causal Modeling Researchers-01

Richard Scheines (Dean, Dietrich College of Humanities & Social Sciences)
Professor of Philosophy, Machine Learning Department, and HCII
- Graphical and Statistical Causal Inference
- Philosophy of Social Science
- Foundations of Causation
- Educational Technology & Online Courses

David Danks (Department Head)
Professor of Philosophy and Psychology Department Head
- Causal learning (human and machine)
- Cognitive Science
- Philosophy of Psychology
- Philosophy of Science
- Bounded Rationality
- Decision-making

Clark Glymour
Alumni University Professor
- Philosophy of Science
- Causal Modeling
- Cognitive Science
- Machine Learning
- Automated Genomics
Landscape of Causal Learning

We use tooling from several sources to include the CMU Tetrad tool, the R programming language and a tool called Stata, to implement causal estimation, e.g. quantifying causal influence, as depicted in the next set of slides.
Causal Graph Surgery with BayesiaLab

Excerpts taken from:

Causality for Policy Assessment and Impact Analysis

Directed Acyclic Graphs and Bayesian Networks for Causal Identification and Estimation

Stefan Conrady, Managing Partner, Bayesia USA, stefan.conrady@bayesia.us
Dr. Lionel Jouffe, CEO, Bayesia
Dr. Felix Elwert, Vilas Associate Professor of Sociology, University of Wisconsin-Madison

October 27, 2014

DOI: 10.13140/2.1.2350.1763
Why Does Software Cost So Much? Towards a Causal Model

October 18, 2016

© 2016 Carnegie Mellon University

[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.
Why Does Software Cost So Much? Towards a Causal Model

October 18, 2016

© 2016 Carnegie Mellon University

Distribution Statement A: This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.
X1: Gender
X2: Treatment
X3: Outcome

Joint Probability: 50.00%
Log-Likelihood: 1
Cases: 500

X1: Gender
- 50.00% Male (1)
- 50.00% Female (0)

X2: Treatment
- 0.00% Yes (1)
- 100.00% No (0)

X3: Outcome
- 50.00% Patient Recovered (1)
- 50.00% Patient Did Not Recover (0)
### Total Effects on Target X3: Outcome

<table>
<thead>
<tr>
<th>Node</th>
<th>Value/Mean</th>
<th>Standardized Total Effects</th>
<th>Total Effects</th>
<th>G-test</th>
<th>Degrees of Freedom</th>
<th>p-value</th>
<th>G-test (Data)</th>
<th>Degrees of Freedom (Data)</th>
<th>p-value (Data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X2: Treatment</td>
<td>0.500</td>
<td>-0.101</td>
<td>-0.100</td>
<td>10.119</td>
<td>1</td>
<td>0.147%</td>
<td>23.972</td>
<td>1</td>
<td>0.000%</td>
</tr>
</tbody>
</table>

**Analysis Context**

- No Observation

**Joint Probability**: 100.00%

**Log-Likelihood**: 0

**Total Effects on Target (New graph 1b)**
Instrumental Variables

Use of an instrumental variable can assist in estimating the causal influence of a Factor A on a cost outcome Y

The instrumental variable (IV) must be closely associated with Factor A

The instrumental variable (IV) must not be associated with outcome Y except via Factor A

You can think of IV as a way to randomize the Factor A for purposes of estimating causal influence of A on Y

We use the Stata `ivregress` command to do this analysis; R also supports this analysis
### Examples of Instrumental Variable Usage

<table>
<thead>
<tr>
<th>IV Factor Example</th>
<th>Factor A Example</th>
<th>Outcome Y Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vietnam Draft Lottery</td>
<td>Military Service</td>
<td>Lifetime Earnings</td>
</tr>
<tr>
<td>Quarter of Birth</td>
<td>Years of Completed Schooling</td>
<td>Adult Earnings</td>
</tr>
<tr>
<td>Voucher distribution at graduation</td>
<td>Attending Private versus Public School</td>
<td>High School Graduation</td>
</tr>
<tr>
<td>Geographic distance to nearest hospital</td>
<td>New Emergency Treatment</td>
<td>Patient Survival</td>
</tr>
<tr>
<td>Tax on tobacco</td>
<td>Smoking</td>
<td>General Health</td>
</tr>
</tbody>
</table>

Causal Inference with Directed Graphs Training

2-Day Seminar offered by Dr. Felix Elwert, Univ of Wisconsin

Available through two channels:

Statistical Horizons
www.statisticalhorizons.com

BayesiaLab
http://www.bayesia.us/causal-inference-course-fairfax
Propensity Scoring Analysis

A modern method of “matching” data to enable causal estimation

Simple concept of propensity scoring is as follows:

1. We have observational data that is nonrandom and usually unbalanced

2. Besides the independent and dependent factor, we have additional factors that may be used to predict/explain the independent factor

3. We create a logistic regression equation predicting the independent factor; the probability prediction for each record of data is then referred to as the propensity score

4. We conduct matching of records of data based on similar propensity scores in which paired records represent quite different levels of the independent factor

5. We discard unmatched data and conduct regression on the Y outcome factor

Shenyang, Guo & Fraser, Mark; “Propensity Score Analysis”, 2nd Edition, 2015
Propensity Scoring Analysis
Agenda

Problem within Cost Estimation
Why Causation instead of Correlation
Why You Must Begin with a DAG!
How Education Misled Us
Landscape of Causal Learning
✔ General Ideas for Research
General Ideas for Research

Causal learning (discovery & estimation) is now used widely in social science, medical and economic research

Why not consider using this framework for better understanding software engineering and cost estimation?

Could we not evaluate traditional factors of software cost for true causal relevance?

Could we not move towards a unified software cost causal model?

Can we not revisit existing models using existing repositories of cost data?

Could we finally separate true causal influences of cost from spuriously-correlated factors?
Backup Reference Slides
Use of Directed, Acyclic Graphs

1. Derive testable implications of a causal model to evaluate if the model is correct

2. Understand causal identification requirements to confirm whether causality may be extracted from the data
   - Separating causal from spurious associations in the data

3. Inform use of traditional statistical techniques such as regression
   - Deciding which control variables to include versus not to include in the analysis to achieve identification of causality

Basic Concepts of DAGs

1. DAGs consist of:
   
a) **nodes** (variables),

b) **directed arrows** (possible causal relationships ordered by time), and
   
c) **missing arrows** (confident assumptions about absence of causal effects)

2. DAGs are nonparametric
   
a) No distributional assumptions

b) Linear and/or nonlinear

3. DAGs have both causal paths and non-causal (spurious) paths

Three Structures Studied in a DAG

1. Indirect Connection

2. Common Cause

3. Common Effect (Collider)

Deriving Testable Implications of a DAG

1. Uses a technique called d-Separation
   a) Algorithm to help determine which paths are causal versus non-causal
   b) Uses concept of blocking a path to stop transmission of non-causal association

2. Additional techniques employed include
   a) Graphical identification
   b) Adjustment Criterion
   c) Backdoor Criterion
   d) Frontdoor Criterion
   e) Pearl’s do-Calculus