COCOMO II Bayesian Analysis

Sunita Chulani
(sdevnani@sunet.usc.edu)

Copyright USC-CSE 1998

Annual Research Review
March 9, 1998

Outline

- Motivation
- Research Approach
  - Modeling Methodology
  - Bayesian Analysis
- Status
  - Initial Bayesian Analysis Results
- Plans
  - Further Prediction Accuracy Improvement
- Information Sources
COnstructive COst MOdel (COCOMO)

- COCOMO published since 1981
  - Commercial Implementations of COCOMO
    - CoCoPro, CB COCOMO, COCOMOID, COSTMODL, GECOMO Plus, SECOMO, COSTAR, etc.
  - Other models based on COCOMO
    - REVIC, Gulezian
    - Intermediate model: Pred (.20) = 68%
- COCOMO II
  - Research effort started in 1994 to develop a 1990's-2000's software cost model
  - Address new processes and practices
  - COCOMO II.1999/2000
  - COCOMOII.1997 model: Pred (.20) = 46%

Calibration Challenges

- GUI builders, COTS, 4GL's, reuse, requirements breakage
  - Need to rethink size metrics
- Distributed interactive applications
  - Web-based, object-oriented, event-based
  - Middleware effects
- New process models (evolutionary, incremental, spiral)
  - Phases overlap
  - Where are cost measurement endpoints?
- Lack of good data
  - Not enough data (i.e. very little degrees of freedom)
  - Lack of dispersion
  - Multicollinearity
COCOMO II.1997 Calibration

- 83 projects
- Multiple Linear Regression
  - 10% weighted average between a-priori values and data-determined values

Develop for Reuse (RUSE)

Multiple Linear Regression Well-Suited When...

- a lot of data is available
- no data items are missing
- there are no outliers
- the predictor variables are not highly correlated
- the predictor variables have an easy interpretation when used in model

most are violated by current software engineering data
Model Plans: Affiliate Priorities
October ‘97

21 ♦ Improve accuracy of COCOMO II Model
15 ♦ Cost/schedule/quality tradeoffs
12 ♦ Sizing improvements
10 ♦ COTS integration costs
 7 ♦ Activity Distribution
 3 ♦ Life cycle tradeoff models

Solutions to Top Two Priorities

♦ Improve Accuracy of COCOMO II
  – Try Bayesian technique
    ♦ with different weighted average between data-determined and a-priori values
    ♦ Collect more data

♦ Cost/schedule/quality tradeoffs
  – Develop Quality Model and integrate with existing COCOMO II
    ♦ Use COCOMO II parameters
    ♦ Use similar Bayesian approach
Outline

- Motivation
- Research Approach
  - Modeling Methodology
  - Bayesian Analysis
- Status
  - Initial Bayesian Analysis Results
- Plans
  - Further Prediction Accuracy Improvement
- Information Sources

The Seven-Step Modeling Methodology
Literature, Behavioral Analysis (Steps 1-3)

Results of Delphi (Step 4)
Results of Sampling Data (Step 5)

Results of Bayesian Update: Using Prior and Sampling Information (Step 6)
COCOMO II Calibration Approaches

COCOMO II.1997
10% weighed-average approach

Bayesian approach - weight determined by data and prior significance

Outline

♦ Motivation
♦ Research Approach
  - Modeling Methodology
  - Bayesian Analysis
♦ Status
  - Initial Bayesian Analysis Results
♦ Plans
  - Further Prediction Accuracy Improvement
♦ Information Sources
How do we interpret g?
(from Steece’s presentation)

- When $g = 0$
  - our estimates of $\beta$ depend only on sample information
  - We can’t do this since some parameters have negative coefficients

- When $g = 1$
  - we are equally weighting prior and sample information

- When $g > 1$
  - we are giving greater weight to prior information

How do we calculate g for COCOMO II? - Step 1

Determine prior variance of all the cost drivers from the Delphi results

<table>
<thead>
<tr>
<th>Required Software Reliability (RELY)</th>
<th>Very Low</th>
<th>Low</th>
<th>Nominal</th>
<th>High</th>
<th>Very High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavioral Analysis, Relative Significance</td>
<td>0.77</td>
<td>0.88</td>
<td>1.0</td>
<td>1.15</td>
<td>1.4</td>
</tr>
</tbody>
</table>

**Initial Productivity Range = 1.40 x 1.77 = 1.87**

<table>
<thead>
<tr>
<th>Expert Judgment Delphi Round 1</th>
<th>Average</th>
<th>Std. Dev.</th>
<th>Variance</th>
<th>Max</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.06</td>
<td>0.40</td>
<td>0.16</td>
<td>3.00</td>
<td>1.75</td>
</tr>
</tbody>
</table>
How do we calculate g for COCOMO II? - Step 2

Determine sample variance of all the cost drivers from Ordinary Least Squares Regression

e.g. Variance of coefficient of RELY = 0.23
Variance of coefficient of DATA = 0.08

Initial Bayesian Analysis
Results: g = 1

<table>
<thead>
<tr>
<th>Cost Driver</th>
<th>Sample</th>
<th>Prior</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.97</td>
<td>0.92</td>
<td>0.94</td>
</tr>
<tr>
<td>B</td>
<td>0.90</td>
<td>1.01</td>
<td>0.95</td>
</tr>
<tr>
<td>RELY</td>
<td>0.39</td>
<td>1.00</td>
<td>0.70</td>
</tr>
<tr>
<td>DATA</td>
<td>1.40</td>
<td>1.00</td>
<td>1.20</td>
</tr>
<tr>
<td>DCCU</td>
<td>2.48</td>
<td>1.00</td>
<td>1.74</td>
</tr>
<tr>
<td>CPLX</td>
<td>1.26</td>
<td>1.00</td>
<td>1.14</td>
</tr>
<tr>
<td>PVOL</td>
<td>1.10</td>
<td>1.00</td>
<td>1.05</td>
</tr>
<tr>
<td>RUSE</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>TOOL</td>
<td>0.70</td>
<td>1.00</td>
<td>0.85</td>
</tr>
<tr>
<td>SITE</td>
<td>1.16</td>
<td>1.00</td>
<td>1.08</td>
</tr>
<tr>
<td>SCED</td>
<td>1.93</td>
<td>1.00</td>
<td>1.46</td>
</tr>
<tr>
<td>TIME</td>
<td>0.98</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>STOR</td>
<td>0.98</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>ACAP</td>
<td>0.94</td>
<td>1.00</td>
<td>0.97</td>
</tr>
<tr>
<td>PCAP</td>
<td>0.94</td>
<td>1.00</td>
<td>0.97</td>
</tr>
<tr>
<td>PCON</td>
<td>0.94</td>
<td>1.00</td>
<td>0.97</td>
</tr>
<tr>
<td>AEXP</td>
<td>0.71</td>
<td>1.00</td>
<td>0.85</td>
</tr>
<tr>
<td>LTEX</td>
<td>0.71</td>
<td>1.00</td>
<td>0.85</td>
</tr>
<tr>
<td>PEXP</td>
<td>0.71</td>
<td>1.00</td>
<td>0.85</td>
</tr>
<tr>
<td>PREC</td>
<td>1.39</td>
<td>1.00</td>
<td>1.20</td>
</tr>
<tr>
<td>PMAT</td>
<td>2.15</td>
<td>1.00</td>
<td>1.57</td>
</tr>
<tr>
<td>FLEX</td>
<td>0.32</td>
<td>1.00</td>
<td>0.66</td>
</tr>
<tr>
<td>RESL</td>
<td>0.32</td>
<td>1.00</td>
<td>0.66</td>
</tr>
<tr>
<td>TEAM</td>
<td>0.32</td>
<td>1.00</td>
<td>0.66</td>
</tr>
</tbody>
</table>

♦ Equal weights to prior and sample information
Prediction Accuracies

Bayesian approach: $g = 1; 159$ observations

<table>
<thead>
<tr>
<th>Effort Prediction</th>
<th>Bayesian Approach 159 observations</th>
<th>COCOMO II.1997 83 observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRED(.20)</td>
<td>49%</td>
<td>46%</td>
</tr>
<tr>
<td>PRED(.25)</td>
<td>57%</td>
<td>49%</td>
</tr>
<tr>
<td>PRED(.30)</td>
<td>64%</td>
<td>52%</td>
</tr>
</tbody>
</table>

Outline

- **Motivation**
- **Research Approach**
  - Modeling Methodology
  - Bayesian Analysis
- **Status**
  - Initial Bayesian Analysis Results
- **Plans**
  - Further Prediction Accuracy Improvement
- **Information Sources**
Ongoing Research and Near Future Plans

- Complete Delphi Analyses to determine accurate prior distributions
- Try different g-priors (and g-prior extensions) to determine best Bayesian COCOMO II calibrated model
- Better define ambiguous parameters
  - RUSE (Required Reusability? Develop for Reuse?...)
- Collect More data (ongoing process)

COCOMO II Research Plan
Outline

♦ Motivation
♦ Research Approach
  – Modeling Methodology
  – Bayesian Analysis
♦ Status
  – Initial Bayesian Analysis Results
  – Ongoing Analysis
♦ Plans
  – Further Prediction Accuracy Improvement
♦ Information Sources

Information Sources:

• Email: cocomo-info@sunspot.usc.edu
• Web site: http://sunspot.usc.edu/COCOMOII/Cocomo.html
  – Affiliate Prospectus
  – Model Definition Manual (ver. 1.4)
  – Data Collection Form (ver. 1.6)
  – Tech Reports on COCOMO II calibration:
    http://sunspot.usc.edu/TechRpts/electronicopy.html
## Glossary

<table>
<thead>
<tr>
<th>Abrev.</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREC</td>
<td>Precendencedness</td>
</tr>
<tr>
<td>FLEX</td>
<td>Development Flexibility</td>
</tr>
<tr>
<td>RESL</td>
<td>Architecture and Risk Resolution</td>
</tr>
<tr>
<td>TEAM</td>
<td>Team cohesion</td>
</tr>
<tr>
<td>PMAT</td>
<td>Process Maturity</td>
</tr>
<tr>
<td>RELY</td>
<td>Required Software Reliability</td>
</tr>
<tr>
<td>DATA</td>
<td>Data Base Size</td>
</tr>
<tr>
<td>CPLX</td>
<td>Product Complexity</td>
</tr>
<tr>
<td>RUSE</td>
<td>Develop for Reuse</td>
</tr>
<tr>
<td>DOCU</td>
<td>Documentation Match to Life-cycle Needs</td>
</tr>
<tr>
<td>TIME</td>
<td>Time Constraint</td>
</tr>
<tr>
<td>STOR</td>
<td>Storage Constraint</td>
</tr>
<tr>
<td>PVOL</td>
<td>Platform Volatility</td>
</tr>
<tr>
<td>ACAP</td>
<td>Analyst Capability</td>
</tr>
<tr>
<td>PCAP</td>
<td>Programmer Capability</td>
</tr>
<tr>
<td>AEXP</td>
<td>Applications Experience</td>
</tr>
<tr>
<td>PEXP</td>
<td>Platform Experience</td>
</tr>
<tr>
<td>LTEX</td>
<td>Language and Tool Experience</td>
</tr>
<tr>
<td>PCON</td>
<td>Personnel Continuity</td>
</tr>
<tr>
<td>TOOL</td>
<td>Use of Software Tools</td>
</tr>
<tr>
<td>SITE</td>
<td>Multi-Site Development</td>
</tr>
<tr>
<td>SCED</td>
<td>Required Development Schedule</td>
</tr>
</tbody>
</table>

### Rating Scale
- VL = Very Low
- L = Low
- N = Nominal
- H = High
- VH = Very High
- XH = Extra High