Effort Prediction Models for Interval Estimation

Lefteris Angelis & Ioannis Stamelos

Programming Languages and Software Engineering Laboratory, (PLaSE Lab)

Department of Informatics
Aristotle University of Thessaloniki
Interval Estimation

- Confidence intervals for Point Estimations
  - Regression
  - Estimation by Analogy
  - Other statistical models
- Estimation of Probabilities of Predefined intervals
  - Ordinal Regression
  - BBNs
  - Machine learning methods
Advantages of interval estimation

- Safer to produce interval estimates, along with a probability distribution over the estimated intervals.
- Relying blindly on point estimates may lead easily in wrong decisions.
- An interval estimate can provide a point estimate for practical purposes.
- Intervals give information about the reliability of the estimation process.
- Intervals provide the basis for risk and what-if project analysis.
Relevant research at PLaSE Lab

● Since 1999…
  ● Statistical models for interval estimation
  ● Comparisons of various models
  ● Phd students
  ● Master theses
  ● Involved in two relevant funded research projects
  ● Most recent project: Optimization of Telecommunication Software process development (DIERGASIA)
Confidence Intervals for Estimation by Analogy (EbA)

- Bootstrap resampling methods
  - Non-parametric bootstrap (draw samples with replacement from the original sample)
  - Parametric bootstrap (draw samples from a theoretical distribution fitting well to the sample)

- The same methods were used for calibration of EbA (number of analogies, distance metric, standardization, etc)
Example (Albrecht data set)

- Non-parametric bootstrap confidence interval
- Estimate the effort of a new hypothetical project:
  - Predictors: IN = 27, OUT = 36, FILE = 20, INQ = 10.
  - Point estimation based on 2 analogies = 5.55 man months
  - \( B = 1000 \) bootstrap samples, estimation each time of the effort using 2 analogies exactly as for the point estimation
- Confidence intervals (using bootstrap distribution and kernel density smoother):
  \[ 95\% \; CI_{boot} = [3.6, 11.45] \quad 50\% \; CI_{boot} = [3.6, 7.5] \]
Confidence zones for EbA using bootstrap and jackknife (Albrecht data set)
Confidence zones for EbA using bootstrap and jackknife (Abran-Robillard data set)
BRACE
(BootstRap Based Analogy Cost Estimation)

- Typical input/output functions and file management facilities
- Definition of attributes and project characterisation
- Project/attribute management (e.g. exclusion of projects/attributes from calculations)
- Choice of options to be considered for method calibration (with and without bootstrap)
- Determination of the best attribute set (the one providing the better accuracy results according to some criterion)
- Generation of estimations for a single project (with and without bootstrap)
A Case Study in Industrial Context

- Controlling the Cost of Software Development for Supply Chain Information Systems

- **Supply Chain ISBSG Project Subset**
  - 59 projects implementing information systems for manufacturing, logistics, warehouse management, ...
  - characterised through effort, size, elapsed time, team size, project nature attributes
  - accurate project attribute measurement
  - average productivity ~ 190 FP/1000 mh

- **BRACE Application**
  - Various strategies were tried because of missing values
  - Best strategy pursued a trade-off between number of projects and attributes
  - Precision was measured through jackknife
  - Different treatment for elapsed time and max team size
Interval Estimation of the cost of a project portfolio

- Combination of:
  - EbA
  - non-parametric bootstrap
  - Stochastic Budget Simulation

- The method allows risk analysis
Example

- Cost data set: Abran – Robillard
- 21 projects - 10 variables
- 16 projects considered completed
- 5 project considered new (portfolio)
- Estimation by analogy and 1000 bootstrap samples: empirical distributions of individual projects
Histogram of the 1000 bootstrap estimates for each one of the individual efforts

Estimations for the individual project efforts based on 1000 bootstrap samples
Using the empirical distribution for simulation

- Fitting of a known theoretical distribution or a smoothing procedure like *kernel density estimation*
- Stochastic simulation sampling from fitted distributions a large number of times, adding each time the individual effort values to get the overall effort
- The entire set of the overall effort values from simulation produces the cumulative density function (useful for computing probabilities of various effort intervals)
Cumulative distribution function of the overall effort obtained by Stochastic Budget Simulation
Confidence intervals for project portfolios costs

<table>
<thead>
<tr>
<th>Portfolio number</th>
<th>Portfolio projects</th>
<th>Actual effort</th>
<th>99% CI (BS-ED method)</th>
<th>99% CI (BS-SBS method)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17, 20, 21, 5, 7</td>
<td>1147</td>
<td>533–2026</td>
<td>664–1771</td>
</tr>
<tr>
<td>2</td>
<td>6, 9, 10, 15, 17</td>
<td>1276</td>
<td>782–2013</td>
<td>890–1725</td>
</tr>
<tr>
<td>3</td>
<td>9, 12, 13, 18, 20</td>
<td>1341</td>
<td>920–2014</td>
<td>1035–1942</td>
</tr>
<tr>
<td>4</td>
<td>4, 7, 8, 13, 15</td>
<td>1414</td>
<td>1040–2027</td>
<td>1148–1791</td>
</tr>
<tr>
<td>5</td>
<td>12, 15, 16, 21, 2</td>
<td>1417</td>
<td>1089–2002</td>
<td>1136–1851</td>
</tr>
<tr>
<td>6</td>
<td>18, 21, 1, 6, 8</td>
<td>1439</td>
<td>1116–2011</td>
<td>1221–1984</td>
</tr>
<tr>
<td>7</td>
<td>15, 18, 19, 3, 5</td>
<td>1444</td>
<td>834–1945</td>
<td>899–1832</td>
</tr>
<tr>
<td>8</td>
<td>21, 3, 4, 9, 11</td>
<td>1519</td>
<td>841–2021</td>
<td>932–1786</td>
</tr>
<tr>
<td>9</td>
<td>7, 10, 11, 16, 18</td>
<td>1608</td>
<td>609–1639</td>
<td>770–1541</td>
</tr>
<tr>
<td>10</td>
<td>19, 1, 2, 7, 9</td>
<td>1617</td>
<td>712–1931</td>
<td>854–1829</td>
</tr>
<tr>
<td>11</td>
<td>20, 2, 3, 8, 10</td>
<td>1621</td>
<td>957–2147</td>
<td>1075–2119</td>
</tr>
<tr>
<td>12</td>
<td>8, 11, 12, 17, 19</td>
<td>1623</td>
<td>1135–2102</td>
<td>1313–2063</td>
</tr>
<tr>
<td>13</td>
<td>11, 14, 15, 20, 1</td>
<td>1625</td>
<td>1038–1990</td>
<td>1047–1977</td>
</tr>
<tr>
<td>14</td>
<td>10, 13, 14, 19, 21</td>
<td>1722</td>
<td>979–1865</td>
<td>999–1866</td>
</tr>
<tr>
<td>15</td>
<td>13, 16, 17, 1, 3</td>
<td>1742</td>
<td>693–1554</td>
<td>715–1550</td>
</tr>
<tr>
<td>16</td>
<td>5, 8, 9, 14, 16</td>
<td>1787</td>
<td>957–1830</td>
<td>1086–1888</td>
</tr>
<tr>
<td>17</td>
<td>14, 17, 18, 2, 4</td>
<td>1928</td>
<td>1166–2039</td>
<td>1211–2014</td>
</tr>
<tr>
<td>18</td>
<td>16, 19, 20, 4, 6</td>
<td>2014</td>
<td>1119–1804</td>
<td>1138–1799</td>
</tr>
<tr>
<td>19</td>
<td>3, 6, 7, 12, 14</td>
<td>2082</td>
<td>956–2022</td>
<td>1241–1994</td>
</tr>
<tr>
<td>20</td>
<td>1, 4, 5, 10, 12</td>
<td>2217</td>
<td>727–1987</td>
<td>1107–1976</td>
</tr>
<tr>
<td>21</td>
<td>2, 5, 6, 11, 13</td>
<td>2257</td>
<td>1008–2280</td>
<td>1200–2129</td>
</tr>
</tbody>
</table>
Estimation of predefined intervals – Ordinal Regression

- For ordinal dependent variables (the cost intervals)
- General form of the OR prediction equations:

\[ l(c_j) = \theta_j - \sum_{i=1}^{k} \beta_i x_i \]

- \( c_j \) the cumulative probability for the \( j \)-th category,

- \( l(\cdot) \) link function (usually one of the following):
  - Logit function: \( \log\left(\frac{c}{1-c}\right) \)
  - Complementary log-log function: \( \log(-\log(1-c)) \)
  - Negative log-log function: \( -\log(-\log(c)) \)
  - Probit function: \( \Phi^{-1}(c) \)
  - Cauchit function: \( \tan(\pi(c - 0.5)) \)
Application of OR to three data sets (4 categories)

- Maxwell
- COCOMO81
- ISBSG 7

\[ l(c_j) = \theta_j - 2.144 \delta(\text{apl}_2) + 2.433 \delta(\text{br}_2) - d_{\text{dbm}_3} \]

\[ * d_{\text{dbm}_3} + 1.707 \delta(\text{lg}_2) + 4.260 \delta(\text{org}_2) \]

\[ - 1.045 \times \text{year} \]

for \( j = 1, 2, 3, 4, \)

where \( \delta(x) = \begin{cases} 1 & \text{if } x = 1 \\ 0 & \text{if } x = 2 \end{cases} \) and \( \theta_j = \begin{cases} 2078.154 & \text{if } j = 1 \\ 2080.423 & \text{if } j = 2 \\ 2086.031 & \text{if } j = 3 \\ 0 & \text{if } j = 4 \end{cases} \)

\[ d_{\text{dbm}_3} = \begin{cases} 3.386 & \text{if } d_{\text{dbm}_3} = 1 \\ 2.311 & \text{if } d_{\text{dbm}_3} = 2 \\ 0 & \text{if } d_{\text{dbm}_3} = 3 \end{cases} \]
Machine Learning methods for predefined intervals

Machine Learning (ML) methods
- Association Rules
- Classification and Regression Trees
- Bayesian Belief Networks

Predefined intervals in combination with ML
- Estimation of an interval
- Probabilities
- Combination of information from past historical data with expert knowledge
- Justification of the estimation
Estimation Process for ML

- Initial discretization of productivity value
- Application of the methods, estimation models
- Transformation of the interval estimate into a numeric estimate using the mean or the median point of the interval
- Calculation of MMRE, pred(25), hitrate
Estimation example

- ISBSG data set release 7
  - Data split in 3 sets according to their application type. Models predicting the productivity values of Management Information Systems (MIS), Transaction Production systems and the rest of the projects
  - Variables used:
    - Function points, time size, language type, primary programming language, organization type, database, development platform, use of methodology, business area type, implementation date.
  - Methods applied AR, CART, AR+CART (combination)
  - Example: Estimation models of MIS projects
  - Training data set: 128 projects Test data: 7 projects
Association Rules

- Probabilistic statements about the co-occurrence of certain events in a database.

\[ IF \ A1=X \ AND \ A2=Y \ THEN \ A3=Z \]

- \( A1=X \ AND \ A2=Y \) : rule body
- \( A3=Z \) : rule head
- \textbf{Confidence}: \( p(A3=Z|A1=X,A2=Y) \)
- \textbf{Support} : expresses the frequency of the rule in the whole data set.
### MIS projects

<table>
<thead>
<tr>
<th>no</th>
<th>support</th>
<th>confidence</th>
<th>rule body</th>
<th>rule head</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>3.1</td>
<td>100.0</td>
<td>BAT = OTHER and PPL in {APG, 4GL, VB, SQL, TELON, OTHER}</td>
<td>0.137 &lt; P ≤ 0.273</td>
</tr>
<tr>
<td>2.</td>
<td>9.4</td>
<td>92.3</td>
<td>PPL= ACCESS</td>
<td>0.274 &lt; P ≤ 5.353</td>
</tr>
<tr>
<td>3.</td>
<td>3.1</td>
<td>80.0</td>
<td>DP = MF and 286 &lt; FP ≤ 629 and DT in {New=development, Re- development}</td>
<td>0.066 &lt; P ≤ 0.136</td>
</tr>
<tr>
<td>4.</td>
<td>3.1</td>
<td>80.0</td>
<td>LT= 4GL and DP= PC and OT= ProfessionalServices and BAT in {Engineering, Personnel, Research&amp;Development}</td>
<td>0.274 &lt; P ≤ 0.590</td>
</tr>
<tr>
<td>5.</td>
<td>8.6</td>
<td>78.5</td>
<td>DBMS= IMS and BAT in {Banking, Accounting, Logistics, Manufacturing, Sales&amp; Marketing}</td>
<td>0.015 &lt; P ≤ 0.065</td>
</tr>
<tr>
<td>6.</td>
<td>7.8</td>
<td>76.9</td>
<td>PPL= COBOL and BAT in {Banking, Accounting, Logistics, Manufacturing, Sales&amp;Marketing}</td>
<td>0.032 &lt; P ≤ 0.065</td>
</tr>
<tr>
<td>7.</td>
<td>4.0</td>
<td>71.4</td>
<td>LT=3GL and DBMS=ORACLE</td>
<td>0.066 &lt; P ≤ 0.136</td>
</tr>
<tr>
<td>8.</td>
<td>3.1</td>
<td>66.7</td>
<td>BAT=Engineering and PPL in {ACCESS, NATURAL}</td>
<td>0.274 &lt; P ≤ 0.590</td>
</tr>
<tr>
<td>9.</td>
<td>2.4</td>
<td>58.3</td>
<td>DT=Enhancement and PPL=SQL</td>
<td>0.015 &lt; P ≤ 0.065</td>
</tr>
<tr>
<td>10.</td>
<td>3.9</td>
<td>45.5</td>
<td>LT=4GL and DBMS=ACCESS</td>
<td>0.591 &lt; P ≤ 5.353</td>
</tr>
</tbody>
</table>
AR ISBSG data set

- Support and confidence threshold 3.125% (4 projects) and 45.5% correspondingly.
- Frequent appearing attributes are Business Area Type \((BAT)\) and Development Type \((DT)\).
- Rules for high productivity values were very few.
AR-ISBSG data set

- Estimate the project with the following values use the previous table

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>support</td>
<td>confidence</td>
<td>rule head</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>9.4</td>
<td>92.3</td>
<td>PPL_ACCESS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>3.1</td>
<td>66.7</td>
<td>BAT_Engineering+PPL in{ACCESS, NATURAL}</td>
<td>0.274&lt; P ≤5.353</td>
<td>0.274&lt; P ≤0.590</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The first two rules that provide estimation for the project are the following:
CART

- CART tree model consists of an hierarchy of:
  - Univariate binary decisions.
  - Classifies all possible cases.
  - Example: simple CART estimating productivity measured in Lines of code per hour
CART- ISBSG data set

Accounting, Banking, Sales & Marketing, Manufacturing, Logistics.

0.032 < P ≤ 0.065
C = 56.5%

Inventory, Financial, Legal, Unknown

0.274 < P ≤ 0.590
C = 45.2%

Cobol, Access, C, C++, Visual Basic, Java

0.066 < P ≤ 0.136
C = 60%

0.137 < P ≤ 0.273
C = 42.6%
The CART can be explained as follows:

- If the business area type is Accounting, Banking, Manufacturing, Sales & Marketing, Logistics then there is 56.5% probability that the productivity will be between 0.032 and 0.065 fp/hour.

- Else if the business area type is Inventory, Financial, Legal or Unknown then if:
  - the language used is C, C++, Cobol, Access, Visual Basic or Java then there is 60% probability that the productivity will be between 0.066 and 0.136 fp/hour.
  - otherwise there is 60% probability that the productivity will be between 0.066 and 0.136 fp/hour.

- If none of the previous is true then there is 45.2% probability that productivity will be between 0.274 and 0.590 fp/h.
The method exploits the advantages of AR
- pertinent relationships among the project attributes and the development
- AR is a method for descriptive modeling
- representation form of AR, is transparent

CART method on the other hand
- as a predictive modeling method
- provides a complete estimation framework
- constructs a model that classifies all projects
- CART also avoids overfitting of the model to the historical data information

The combination of the methods provides:
- Improved estimation results
- Better understanding of the problem
- Deals with the problem of AR to provide an estimation of all possible projects
- Deals with the problem of CART that are often very inaccurate
AR+CART

Estimation process

- Identify ARs describing the influence of certain project attributes on the software development productivity
- Build a CART model that will be able to classify all possible projects to a productivity
- If the new project can be estimated by the AR model with a stronger confidence value than the CART model then use that estimate
- Otherwise use the CART estimate
- AR+CART results in 10-15% accuracy improvement w.r.t. AR alone
Bayesian Belief Networks

- Directed Acyclic Graphs (DAGs)
- Express cause-effect relationships
- Nodes represent variables, arcs represent relationships
- Node Probability table (conditional dependencies)
- Bayes’ Rule:

\[
P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}
\]
a) A-node

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0.6</td>
</tr>
<tr>
<td>F</td>
<td>0.4</td>
</tr>
</tbody>
</table>

b) A-node

<table>
<thead>
<tr>
<th>B-node</th>
<th>A-node</th>
<th>T</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOW</td>
<td>0.3</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>MED</td>
<td>0.5</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>HIGH</td>
<td>0.2</td>
<td>0.15</td>
<td></td>
</tr>
</tbody>
</table>

c) A-node

<table>
<thead>
<tr>
<th>C-node</th>
<th>B-node</th>
<th>T</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LOW</td>
<td>0.7</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>MED</td>
<td>0.65</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>HIGH</td>
<td>0.4</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>LOW</td>
<td>0.3</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>MED</td>
<td>0.35</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>HIGH</td>
<td>0.6</td>
<td>0.93</td>
</tr>
</tbody>
</table>
BBN + COCOMO81 data set
BBN + COCOMO81 data set

- Variable that directly affects productivity is mode and can classify correct 34.9% of the data.
- For improved estimation results we empirically added 3 nodes as parents of productivity, PCAP, CPLX and application type.
- In the BBN we can observe the relationships among the projects attributes as well.
- The number that accompanies each arc is the estimation hitrate that each nodes classifies its child node.
- Evaluation results, Jackknife method
Interval Estimation papers

DIERGASIA

Optimization of Telecommunication Software PROCESS development

Funded by the Greek Secretariat for Research and Technology, a PAVET grant
Target of the project

- **Design, development and application of models describing software process development for telecommunication systems software.**
  - Definition of measurements
  - Definition of the appropriate observation points.
  - Selection and implementation of statistical models and software.
    - Cost estimation
    - Quality estimation
  - Specification of new improved processes.
  - Preparation for CMM-I assessment
W.P.1 – Research on software metrics and statistical methods for data processing

- **D1- Guide for measurement and statistical analysis**
  - Identification of measurements and metrics for gathering quantitative data.
  - Identification of statistical and data mining methods for qualitative analysis of data
    - Regression models, analogy based estimation, CART, Association Rules, Bayes Networks.
  - Research on requirements specification process
W.P.2 – Measurement and data collection and transformation

- **D2.1 – Company data base of measurements**
  - Placement of the appropriate audits and metric pointers at particular development points.
  - Measurement completion and result recording.
  - Data collection from measurements.
  - Data collection from SAFIRE tool.

- **D2.2 – Quality report for measurements**
  - Initial application, mining, comparison and evaluation of models – emphasis on interval models
  - Data selection
  - Pre-Processing
  - Transformation
  - Statistical analysis
W.P.3 – Modeling company's process

- **D3.1 – Mathematical models for estimation of company’s process.**
- Statistical processing of data
  - Define analysis methods
  - Apply statistical analysis
  - Interpret results
  - Discuss results
- Models development
  - Software cost estimation
  - Software quality estimation
  - Reusable code assets
  - Requirements collection
W.P.4 – A model for SDL Systems Quality Evaluation

- **D4.1 – Quality model evaluation for SDL**
  - Specification of a model for quality evaluation
  - Development of a quality rule set SDL

- **D4.2 – Tool for automated quality evaluation for SDL**
  - Development of static metrics that implement SDL rules
  - Specification of quality intervals for each metric
  - Quality evaluation based on mathematical equations
  - Project evaluation based on the quality model.
W.P.5 – Specification of the company’s processes-
evaluation of the models

- D5.1 – Company’s process quality guide
  - Models application
  - Models evaluation
  - Result comparison
  - Quality models specification
  - Models modification
W.P.6 – Software Process Improvement Certification

- **D6.1 – Software process improvement certification**
  - Software Process Improvement Certification by major clients of Teletel (Alcatel).
  - Process maturity estimation based on the standards of CMM/I.
  - Training for CMM/I.
D1.1 – Guide for measurements and statistical analysis

Part A- Software metrics.

- Introduction
- Product metrics
  - Size metrics (LOC, FP)
  - Complexity metrics (Cyclomatic complexity, nodes)
  - Halstead Metrics (vocabulary, length, volume)
  - Quality metrics (Fault metrics, reliability, maintainability)
- Process metrics
  - Software cost/ effort/ productivity estimation
- Telecommunication systems metrics
  - SAFIRE metrics
D1.1 – Guide for measurements and statistical analysis

Part B - Statistical analysis methods
- Analogy based estimation
- Regression models
- Machine learning models
  - Rules
  - CART
  - Bayes Networks

Part C - Requirements model for telecommunication systems