Ex-post and Ex-ante Predictive Models for Software Effort Estimation
Outline

- The effort estimation problem
- Our approach
- The organizational context
- The technical context
- The data collected
- The ex-post estimation equations
- The validation process
- The ex-ante estimation equation
- Conclusions and further research
The effort estimation problem

- Type of model
- Type of sample
- Type of scales
- Type of approach
- Estimation moment
The effort estimation problem: Type of model?

- analogy based model
- experts' opinion based model
- decomposition based model
- prediction based model
  - statistical: generally use regression technics
  - experimental: use past data to predict effort
  - theoretical: use hypothesis on the resources
  - composite: mixed approach (experts' opinion & statistical models)
The effort estimation problem: Type of sample?

- heterogeneous sample: different organizations, different technologies, etc

- homogeneous sample: same organization, same technology, etc
The effort estimation problem: Type of scales?

- mixed: from nominal to absolute scales
- homogeneous: absolute scale of variables
The effort estimation problem: Type of approach?

- **top-down**: the project as a whole
  - many cost drivers

- **bottom-up**: the project as a sum of stages
  - less cost drivers for each stage

- complementary approaches
The effort estimation problem: Estimation moment?

*ex-post estimation*

- previous software effort as a function of project observed characteristics at time $T$

  » end of project (top down: the project as a whole)
  » end of stage (bottom up: the project as a sum of stages)
The effort estimation problem: Estimation moment?

*ex-post estimation*

\[
\text{Effort (n)} = f_n(V, W) \quad \text{Effort (n+1)} = f_{n+1}(X, Y)
\]
The effort estimation problem: Estimation moment?

*ex-ante estimation*

- At the beginning of the project (top down)
- At the beginning of each stage (bottom up)
- remaining software effort estimated from project estimated characteristics at time $T$
  - expert guessing from
    - previous effort
    - specifications already known
- remaining software effort estimated from project observed characteristics at time $T$
The problem:
Estimation moment?

ex-post estimation

\[
\text{Step } n-1 \quad \overset{\text{T}}{\longleftarrow} \quad \text{Step } n \quad \overset{\text{V}}{\longleftarrow} \quad \text{Step } n+1 \quad \overset{\text{X}}{\longleftarrow} \quad \text{Step } n+2
\]

Effort (n) = \( f_n(V, W) \)
Effort (n+1) = \( f_{n+1}(X, Y) \)

ex-ante estimation

\[
\text{Step } n-1 \quad \overset{\text{T}}{\longleftarrow} \quad \text{Step } n \quad \overset{\text{V}}{\longleftarrow} \quad \text{Step } n+1 \quad \overset{\text{X}}{\longleftarrow} \quad \text{Step } n+2
\]

\[
X = g_1(T, U, V, W)
Y = g_2(T, U, V, W)
\]
Our approach:

- the model: statistical
- the sample: homogeneous
- the variables: measured on absolute scales
- the approach: bottom-up
- the estimation moment: ex-post & ex-ante
The organizational context:

- French insurance company
- 1200 employees
- Registered capital: 30 MF
- 22 regional branches
The technical context:

- client/server
- relational database
- fourth generation language and lower case
- design method: MERISE supported by PACBASE
- project management method: SDM/S methodology
The technical context: SDM/S methodology stages

- Information Systems Planning (IP)
- Definition Study (DS)
- System Design (SD)
- Detailed System Design
  - External Detailed Design (EDD)
  - Internal Detailed Design (IDD)
- Implementation
  - Programming (PRG)
  - Testing (TST)
- Installation (INS)
The data collected:

- Effort in man/days collected by project managers
- At the end of each stage of project life cycle (update of a database)
- Original data set (303 projects)
- Reduced to 110 by considering only new projects development and main evolution of existing projects
- Reduced to a dozen:
  - homogenous organizational and technical context
  - all values for independent variables
The ex-post estimation equations

- stepwise regressions

- sensitivity analysis
  - 1%
  - 5%
  - 10%

- no equation modification
The ex-post estimation equations:

<table>
<thead>
<tr>
<th>SDM/S Phases</th>
<th>Number of observations</th>
<th>Variables used in the Stepwise regressions</th>
<th>Variables retained by the stepwise regressions</th>
<th>Final equations</th>
<th>R²</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS</td>
<td>12</td>
<td>UPTOT, INT, EFF, OBJ, REL, DON, PSS, OP</td>
<td>INT</td>
<td>(1) E = 9.31 + 1.32 INT</td>
<td>0.94</td>
<td>161.73</td>
</tr>
<tr>
<td>SD</td>
<td>10</td>
<td>SCE, SORT, ENT, EFF, DON, OBJ, REL, PF, REC</td>
<td>REC, ENT</td>
<td>(2) E = - 3.19 + 0.49 REL + 1.23 REC</td>
<td>0.97</td>
<td>149.73</td>
</tr>
<tr>
<td>EDD</td>
<td>8</td>
<td>PF, REC, SORT, ENT, FIC, EFF, OBJ, REL</td>
<td>PF</td>
<td>(3) E = - 13.18 + 1.338 PF</td>
<td>0.98</td>
<td>532.27</td>
</tr>
<tr>
<td>IDD</td>
<td>13</td>
<td>TAB, FIC, SORT, ENT, MESS, MOD, EFF</td>
<td>MOD</td>
<td>(4) E = - 1.23 + 0.33 MOD</td>
<td>0.97</td>
<td>163.69</td>
</tr>
<tr>
<td>PRG</td>
<td>10</td>
<td>MOD, LIG, COMP</td>
<td>LIG, MOD</td>
<td>(5) E = + 3.03 + 0.0041 LIG - 0.94 MOD</td>
<td>0.98</td>
<td>232.90</td>
</tr>
<tr>
<td>TST</td>
<td>10</td>
<td>LIG, MOD, UTI</td>
<td>LIG</td>
<td>(6) E = 1.55 + 0.0023 LIG</td>
<td>0.88</td>
<td>63.14</td>
</tr>
<tr>
<td>INS</td>
<td>12</td>
<td>PST, UTI, SESS</td>
<td>PST</td>
<td>(7) E = 0.86 + 0.32 PST</td>
<td>0.78</td>
<td>37.01</td>
</tr>
</tbody>
</table>

© ESSEC J.Akoka, D.Briolat, I.Comyn-Wattiau
### The validation process:

<table>
<thead>
<tr>
<th>SDM/S Phases</th>
<th>Number of observations</th>
<th>Final equations</th>
<th>MRE</th>
<th>Average Error</th>
<th>Prediction (±25%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS</td>
<td>12</td>
<td>(1) $E = 9.31 + 1.32 \text{ INT}$</td>
<td>16%</td>
<td>14%</td>
<td>75%</td>
</tr>
<tr>
<td>SD</td>
<td>10</td>
<td>(2) $E = -3.19 + 0.49 \text{ REL} + 1.23 \text{ REC}$</td>
<td>23%</td>
<td>11%</td>
<td>80%</td>
</tr>
<tr>
<td>EDD</td>
<td>8</td>
<td>(3) $E = -13.18 + 1.338 \text{ PF}$</td>
<td>25%</td>
<td>9%</td>
<td>63%</td>
</tr>
<tr>
<td>IDD</td>
<td>13</td>
<td>(4) $E = -1.23 + 0.33 \text{ MOD}$</td>
<td>36%</td>
<td>15%</td>
<td>69%</td>
</tr>
<tr>
<td>PRG</td>
<td>10</td>
<td>(5) $E = +3.03 + 0.0041 \text{ LIG} - 0.94 \text{ MOD}$</td>
<td>7%</td>
<td>6%</td>
<td>100%</td>
</tr>
<tr>
<td>TST</td>
<td>10</td>
<td>(6) $E = 1.55 + 0.0023 \text{ LIG}$</td>
<td>48%</td>
<td>27%</td>
<td>50%</td>
</tr>
<tr>
<td>INS</td>
<td>12</td>
<td>(7) $E = 0.86 + 0.32 \text{ PST}$</td>
<td>24%</td>
<td>28%</td>
<td>67%</td>
</tr>
</tbody>
</table>

© ESSEC J.Akoka, D.Briolat, I.Comyn-Wattiau

9th International Forum on COCOMO and Software effort cost modeling
Definition Study Effort $f$ (Number of Interviews)

$E = 9.31 + 1.32 \text{ INT}$
External Detailed Design Effort $f$ (Number of Functional Procedures)

\[ E = -13.18 + 1.33\ PF \]

© ESSEC J.Akoka, D.Briolat, I.Comyn-Wattiau

9th International Forum on COCOMO and Software effort cost modeling
Internal Detailed Design Effort $f$ (Number of Modules)

$$E = -1.23 + 0.33 \text{ MOD}$$

© ESSEC J.Akoka, D.Briolat, I.Comyn-Wattiau

9th International Forum on COCOMO and Software effort cost modeling
Testing Effort \( f \) (Number of Lines)

\[ E = 1.55 + 0.0023 \text{ LIG} \]

© ESSEC J. Akoka, D. Briolat, I. Comyn-Wattiau

9th International Forum on COCOMO and Software effort cost modeling
Installation Effort \( f \) (Number of Workstations)

\[ E = 0.86 + 0.32 \text{ PST} \]

![Graph showing observed effort and computed effort with 25% margin]
The ex-ante estimation equation

- not all steps for the dozen remaining projects
- less successful regressions
- number of Records REC (SD Phase) function of the number of Objects OBJ (DS Phase)
  \[ REC = 4.37 \times OBJ - 36.72 \]
  \[ R^2 = 0.72 \]
  \[ F = 13.30 \]
The ex-ante estimation equation

RECord (SD) = f (OBJECT) (DS)

© ESSEC J.Akoka, D.Briolat, I.Comyn-Wattiau

9th International Forum on COCOMO and Software effort cost modeling
Further research

- Same variables for ex-post estimation equations in other firms?

- Larger sample to improve ex-ante estimation equations