Online Appendix to: Linear Programming as a Baseline for Software Effort Estimation

FEDERICA SARRO, University College London
ALESSIO PETROZZIELLO, University of Portsmouth

A DESCRIPTION OF THE DATASETS USED IN THE EMPIRICAL STUDY

The Albrecht+Kemerer (AK) dataset consists of 24+15 (=39) industrial software projects coming from the Albrecht (Albrecht and Gaffney 1983) and Kemerer (Kemerer 1987) datasets, respectively. The Albrecht software projects were developed by the IBM DP service organisation and are characterised in terms of thousand Source Lines of Code (i.e., KSLOC) and Function Points count, which is essentially a weighted sum of the numbers of inputs, outputs, files, and inquiries provided to, or generated by, a software. Also, the effort required to design, develop, and test the application is provided for each of the projects in terms of number of work-hours. The Kemerer software projects are characterised by two categorical variables (Language and Hardware), two software size measures based on Source Lines of Code (i.e., KSLOC), Adjusted Function Points (i.e., AdjFP) and Raw Function Points (i.e., RAWFP), and two dependent variables, namely, the project's duration and the total effort needed to realise each of the projects and computed based on man/month. Note that the use of the KSLOC and the project's duration is usually discouraged in the construction of effort estimation models because these variables are usually correlated to the effort. Since Albrecht and Kemerer have one independent variable in common which can be used at prediction time (i.e., AdjFP), we select this to build the effort estimation model and used the variable effort as the dependent one.

The China dataset (Yun 2010) includes data of 499 projects developed by different Chinese companies. We used the basic elements used to calculate Function Points (i.e., Input, Output, Inquiry, File, Interface) as independent variables and the variable Effort as the dependent one.

The Desharnais dataset (Desharnais 1989) comprises 81 software projects derived from a Canadian software company. We considered the total effort as a dependent variable, but not the length of the code. We also excluded from our analysis the categorical variables (i.e., Language and YearEnd) and four projects that have missing values, as done in previous work (e.g., Sarro et al. (2016), Kadoda and Shepperd (2001), and Shepperd and Schofield (2000)). Therefore, we used the following independent variables: TeamExp (i.e., the team experience measured in years), ManagerExp (i.e., the manager experience measured in years), Entities (i.e., the number of the entities in the system data model), Transactions (i.e., the number of basic logical transactions in the system), and AdjustedFPs (i.e., the Adjusted Function Points).

The Finnish dataset (Shepperd et al. 1996) contains data from 38 industrial software projects developed by nine different Finnish companies. Each project is described by the dependent variable Effort, expressed in person-hours, and five other variables, among which we excluded the PROD variable since it represents the productivity expressed in terms of Effort and size, and only used

1049-331X/2018/09-ART12 $15.00
https://doi.org/10.1145/3234940

HW (i.e., the type of hardware), FP (i.e., Function Points), AR, and CO as the independent variables to build effort estimation models.

The Kitchenham dataset (Kitchenham et al. 2002) contains data from 145 maintenance and development industrial projects managed by a single outsourcing company, including effort estimates and actuals (dependent variable), and function points count (independent variable). The estimates were made as part of the company's standard project estimating process that involved producing two or more estimates for each project and selecting one estimate to be the basis of client-agreed budgets.

The Maxwell dataset (Maxwell 2002) contains 62 industrial software projects developed for one of the biggest commercial banks in Finland. We employed 17 features: Function Points (SizeFP) and 16 ordinal variables, i.e., number of different development languages used (Nlan), customer participation (T01), development environment adequacy (T02), staff availability (T03), standards used (T04), methods used (T05), tools used (T06), softwares logical complexity (T07), requirements volatility (T08), quality requirements (T09), efficiency requirements (T10), installation requirements (T11), staff analysis skills (T12), staff application knowledge (T13), staff tool skills (T14), and staff team skills (T15). As for the Desharnais dataset, we did not use categorical variables.

The Miyazaki dataset (Miyazaki et al. 1994) is composed of 48 industrial software projects developed by 20 different software companies of the Fujitsu Large Systems Users Group. For this dataset, we considered the following independent variables: SCRN (i.e., the number of different input or output screens), FORM (i.e., the number of different report forms), and FILE (i.e., the number of different record format). The dependent variable is Effort, defined as the number of person-hours needed from system design to system test, including indirect effort such as project management.

The Nasa dataset (Bailey and Basili 1981) consists of 18 software projects developed for the NASA/Goddard Space Flight Center. The projects used in our analysis are described in terms of the two independent variables Methodology (Me) and Experience (Exp), which represent, respectively, the methodologies used during design and development, and the experience of the customer and of the programmers. Effort is the dependent variable and measures the actual effort (expressed in man-months) needed to release each of the projects from the beginning of the design phase through the acceptance testing, therefore it includes the effort for programming, management, and support hours. A detailed description of these factors and the collection procedure can be found elsewhere (Bailey and Basili 1981). Note that we excluded from our analysis those factors that are unknown in the early phases of a project such as the actual number of source lines developed (Bailey and Basili 1981).

The Nasa93Coc dataset (Menzies et al. 2005) consists of 93 projects developed between 1971 and 1987 by different NASA centres. The effort (our dependent variable) was measured in calendar months of 152 hours and includes development and management hours. This dataset includes 15 COCOMO I discrete attributes (i.e., rely, data, cplx, time, stor, virt, turn, acap, aexp, pcap, vexp, lexp, modp, tool, sced), which are in the range Very Low to Extra High as defined by Boehm (1981) and software size in thousand Source Lines of Code (i.e., KSLOC), which was estimated directly or with a function point analysis.

The Telecom dataset (Shepperd and Schofield 2000) consists of 18 projects characterised by two independent variables, i.e., the number of changes made as recorded by the configuration management system (Changes) and the number of files changed by a given enhancement project (i.e., Files), and the dependent variable Effort which constitutes the actual effort. According to Shepperd and Schofield (2000), only the variable Files can be used for predictive purposes since none of the other information was available at the time the prediction was made.