# Evidence on economies of scale in software development

**Rajiv D Banker and Hsihui Chang**
*University of Minnesota*

**Chris F Kemerer**
*Massachusetts Institute of Technology*

**Researchers and practitioners have found it useful for cost estimation and productivity evaluation purposes to think of software development as an economic production process, whereby inputs, most notably the effort of systems development professionals, are converted into outputs (systems deliverables), often measured as the size of the delivered system. One central issue in developing such models is how to describe the production relationship between the inputs and outputs. In particular, there has been much discussion about the existence of either increasing or decreasing returns to scale. The presence or absence of scale economies at a given size are important to commercial practice in that they influence productivity. A project manager can use this knowledge to scale future projects so as to maximize the productivity of software development effort. The question of whether the software development production process should be modelled with a non-linear model is the subject of some recent controversy. This paper examines the issue of non-linearities through the analysis of 11 datasets using, in addition to standard parametric tests, new statistical tests with the non-parametric Data Envelopment Analysis (DEA) methodology. Results of this analysis support the hypothesis of significant non-linearities, and the existence of both economies and diseconomies of scale in software development.**

Keywords: scale economies, returns to scale, productivity measurement, software development, software management, data envelopment analysis, software metrics, source lines of code, function points

Researchers and practitioners have found it useful to think of software development as an economic production process, whereby inputs, most notably the effort of systems development professionals, are converted into outputs (systems deliverables), often measured as the size of the delivered system with a metric such as Source Lines of Code (SLOC)[1–4] or Function Points (FPs)[5–9]. Most commonly, such models are used either as aids in software cost estimation or in project productivity evaluations.

One central issue in developing such models is how to describe the production relationship between the inputs and outputs. In particular, there has been much discussion about the existence of either increasing or decreasing returns to scale. This discussion was most recently summarized in Banker and Kemerer, a portion of which is provided here[10].

A production process exhibits local increasing returns to scale if, at a given volume level, the marginal returns of an additional unit of input exceed the average returns. Local economies of scale are thus present when average productivity is increasing, and scale diseconomies prevail when average productivity is decreasing*. The presence or absence of scale economies at a given size are important to commercial practice in that they influence productivity, and therefore a project manager can use this knowledge to scale future projects accordingly so as to maximize the productivity of software development effort. For example, if a software project's size was in the region of decreasing returns to scale, a manager could choose to divide the project into several smaller projects in order to increase the productivity. Conversely, if a software project size was in the region of increasing returns to scale, a manager could choose to combine several similar projects into one. A

* In production economics, economies of scale are defined at specific volume levels in a production process, and are thus best described as *local*. It is therefore inappropriate to limit the characterization of a production process to only *global* economies (or diseconomies) of scale. In dealing with single input-single output production correspondences, the terms 'increasing returns to scale' and 'scale economies' may be used interchangeably.

**Table 1. Summary of sources of economies/diseconomies of scale**

| Sources of economies/diseconomies of scale[10] |
|---|
| *Economies of scale* |
| Specialization of labour |
| Learning curves |
| Software engineering tools |
| Fixed project overhead |
| *Diseconomies of scale* |
| Communication path increases |
| Complex interface requirements |
| Non-linear documentation requirements |
| Project slack |

summary of the arguments for either increasing or decreasing returns to scale in software development is listed in Table 1.

Prior empirical work had indicated increasing returns for some data sets and decreasing returns for others. In the Banker and Kemerer[10] paper it was shown that these two seemingly contradictory hypotheses could be accommodated by viewing the software production process with models that accommodate both increasing and decreasing returns to scale. This paper presented both a parametric and a non-parametric model, and used these models to analyse eight publicly available datasets, where it was shown that both increasing and decreasing returns to scale were present in five of the eight datasets.

Since that paper, researchers and practitioners have continued to show interest in this topic[11–17]. For researchers, the presence of both economies and diseconomies of scale requires the estimation of a more general model than the simple model employed in prior research. For practitioners, this approach highlights the need to consider project scale as a productivity factor in project planning.

Most recently, Kitchenham has revisited the question of scale economies[18]. In analysing six of the original eight datasets from Banker and Kemerer[10], plus four additional datasets, she concludes that '... if models are based on data from a single environment, a linear model ... is likely to be sufficient' (p 212). This is a departure, not only from the conclusions reached in Banker and Kemerer, but also from previous work that estimated either economies or diseconomies of scale, such as the COCOMO model of Boehm[1].

The current paper examines the question of non-linearities in modelling software development as an economic production process. All of the eight original datasets are analysed, along with three of the four additional datasets for a total of eleven*. In addition to the parametric regression-based tests, the data are analysed using new semi-parametric statistical tests with the non-parametric data envelopment analysis methodology. Results of this most recent analysis that are reported here continue to support the hypothesis of significant non-linearities, and hence economies and diseconomies of scale in software development.

The remainder of this paper is organized as follows. The next section presents the data used in the analysis and the third section presents the models and their results. Concluding remarks are provided in the final section.

**Table 2. Dataset summary**

| Author | Date published/available | n |
|---|---|---|
| Belady-Lehman | 1979 | 33 |
| Boehm (COCOMO) | 1981 | 63 |
| Yourdon | 1981 | 17 |
| Bailey-Basili | 1981 | 19 |
| Wingfield | 1982 | 15 |
| Albrecht-Gaffney | 1983 | 24 |
| Behrens | 1983 | 22 |
| Kitchenham-Taylor | 1985 | 33 |
| Kemerer | 1987 | 17 |
| MERMAID-1 | 1992 | 81 |
| MERMAID-2 | 1992 | 30 |

## Data

The data used in this analysis come from published or otherwise publicly available sources. (Many of the datasets are available as appendices to the Conte, Dunsmore, and Shen text[19].) Table 2 summarizes the datasets used.

The above datasets include all eight datasets used originally by Banker and Kemerer, and in addition include the Kitchenham and Taylor dataset, plus two datasets from the MERMAID project, a joint collaborative project funded in part by the European Commission's ESPRIT program, whose aims included improving the process of software cost estimation[10,18,20]. It includes all of the datasets used by Kitchenham[18] except MERMAID-3, which was not made available for this research. In addition, it includes two datasets, Behrens (1983) and Albrecht and Gaffney[5], which were included in the original work, but not in Kitchenham[18].

## Models and results

### *Loglinear model*

The relationship of interest is that between the input, effort and the output, size, typically represented by the production function $y = f(x)$, where $y$ = effort and $x$ = size. A simple form of the model relating effort and project size (typically operationalized as SLOC or FPs) commonly used in the prior literature is $y = ax^b$, estimated in its logarithmic form as[1,21]:

$$\ln(y) = \beta_0 + \beta_1 \ln(x) \qquad (1)$$

where $y$ = effort, $x$ = size, $\beta_0$ corresponds to $\ln(a)$, and $\beta_1$ to $b$.

In this approach, an estimated exponent value $b$ less than 1 indicates economies of scale, while an exponent greater than 1 indicates that diseconomies of scale prevail. One problem with this simple model is that it does not allow for the possibility of increasing returns for some project and decreasing returns for others, and therefore, the general hypothesis of both increasing and decreasing returns to scale prevailing in the same dataset cannot be adequately tested with this model.

* Provision of the additional data and other continued cooperation by B. Kitchenham are gratefully acknowledged.

Kitchenham[18] estimated the loglinear model in Equation (1) above and found that the coefficient estimate of $\beta_1$ was not significantly different from one at the $\alpha = 0.05$ level for her 10 datasets. If $\beta_1 = 1$, then Equation (1) can be rewritten as $y = ax$, where $a = \exp(\beta_0)$. Based on this finding, she argued that the evidence supports a linear relationship between project size and effort. She concluded that software development production functions are adequately represented by a simple proportionate relationship $y = ax$ between size and effort.

However, there is a significant problem with this analysis. Assume for the moment that the true production function is not linear in logarithms as in Equation (1). For example, assume it is a quadratic function in logarithms. If the model is (mis)specified by omitting a relevant variable $(\ln x)^2$, the Ordinary Least Squares (OLS) estimate of the coefficient of the included variables (in this case, $\beta_1$) may be biased, and the standard error of the coefficient may be biased upward, causing inferences concerning these parameters to be inaccurate[22]. In other words, with a misspecified model that omits a correlated explanatory variable, the researcher is likely to misinterpret the evidence of non-linearity.

Therefore, $t$-tests of the $\beta_1$ coefficient of the loglinear model are not appropriate tests of the economies/diseconomies of scale hypothesis. A more suitable test would be to estimate functional forms of the production model that allow for the presence of both economies and diseconomies of scale, and perform the appropriate tests to see whether the null hypothesis can be rejected in that case.

A natural generalization of the loglinear model in Equation (1) is a logquadratic model obtained by adding another explanatory variable $(\ln x)^2$ as in the following:

$$\ln(y) = \beta_0 + \beta_1 \ln(x) + \beta_2 (\ln(x))^2 \qquad (2)$$

Conceptually, this model should allow us to test whether the dataset reflects both increasing and decreasing returns to scale for different data ranges. Empirically, the estimation of Equation (2) is precluded by severe collinearity problems[10]. The Belsley-Kuh-Welsch collinearity diagnostics are employed to check for possible collinearity problems[23]. The condition number exceeds Belsley-Kuh-Welsch's suggested limit of 20 for all 11 datasets, and ranges as high as 433, with a median of 128. The associated variance proportions also exceed 0.97 for both $\ln(x)$ and $(\ln(x))^2$ in all 11 datasets, much greater than the suggested limit of 0.4. Because of the severity of the collinearity problem, standard error estimates are biased upward and the rejection of the null hypothesis becomes less likely[22]. Therefore, the alternative of a quadratic form (rather than a logquadratic) is considered below.

### *Quadratic model*

An alternative parametric model that allows for both economies and diseconomies of scale is a quadratic model:

$$y = \gamma_0 + \gamma_1 x + \gamma_2 x^2 \qquad (3)$$

If the proportionate linear relationship $y = ax$ is an adequate representation of the production function then the result of estimating Equation (3) must be that $\gamma_0 = 0$ *and* $\gamma_2 = 0$. The results of estimating this model for the 11 datasets are shown in Table 3.

While the assumption of $\gamma_0 = 0$ cannot be rejected, as can be seen in Table 3, six of the 11 estimated coefficients $\gamma_2$ for the quadratic term are significantly different from zero. The Kitchenham assumption of linearity of the software development production function is thus rejected in six of the 11 cases with the quadratic model. Note that this is quite a different inference from that available by performing only a test of the loglinear model.

A variety of specification checks were performed for the quadratic model in Equation (3). Examination of the Belsley-Kuh-Welsch collinearity diagnostics indicated condition numbers ranging between 5.3 and 16.2 for the 11

**Table 3. OLS Estimates of quadratic model coefficients**

| Dataset | N | $\gamma_0$<br>($t$-stat) | $\gamma_1$<br>($t$-stat) | $\gamma_2$<br>($t$-stat) | $R^2$<br>(%) |
|---|---|---|---|---|---|
| Behrens | 22 | 321.667<br>(0.437) | 3.786<br>(0.385) | 0.028<br>(1.056) | 58.41 |
| Bailey-Basili | 19 | −10.802<br>(−1.359) | 3.233<br>(6.068[a]) | −0.019<br>(−3.070[a]) | 88.58 |
| Yourdon | 17 | 1.505<br>(0.072) | 1.629<br>(1.739) | −0.005<br>(−0.808) | 42.84 |
| COCOMO | 63 | −207.899<br>(−1.093) | 16.949<br>(6.279[a]) | −0.009<br>(−3.122[a]) | 57.57 |
| Albrecht-Gaffney | 24 | 8.477<br>(1.869) | −0.013<br>(−1.093) | 3.43E-05<br>(5.593[a]) | 94.95 |
| Belady-Lehman | 33 | −309.055<br>(−0.836) | 17.571<br>(3.167[a]) | −0.015<br>(−1.757) | 46.13 |
| Wingfield | 15 | 510.165<br>(0.623) | 0.082<br>(0.010) | 0.023<br>(1.307) | 71.44 |
| Kemerer | 17 | 26267.622<br>(1.945) | −58.620<br>(−2.349[a]) | 0.048<br>(4.605[a]) | 81.77 |
| MERMAID-1 | 81 | 521.646<br>(0.539) | 13.973<br>(2.717[a]) | 0.004<br>(0.745) | 54.81 |
| MERMAID-2 | 30 | −415.514<br>(−0.173) | 44.963<br>(3.764[a]) | −0.025<br>(−2.740[a]) | 39.77 |
| Kitchenham-Taylor | 33 | −33.336<br>(−1.484) | 0.012<br>(3.696[a]) | −2.71E-07<br>(−3.073[a]) | 35.10 |

$t$-statistics in parentheses: [a] indicates significant at 5%

datasets, well below the recommended cutoff of 20; thus collinearity appears not to be a problem for this model for the 11 datasets. The residuals were examined in order to determine whether they were distributed normally. The Shapiro-Wilk test for normality of residuals was rejected at the 10% level (used in the interests of being conservative) for seven of the 11 datasets[24]. The residuals were also examined for violations of the homoscedasticity assumption and other mis-specification of the estimation model, including violations of the normality assumption. White's test rejected this assumption for nine of the 11 datasets at the 10% significance level. Therefore to address this estimation problem White's heteroscedasticity-consistent estimator of the covariance matrix was also used to calculate the $t$-statistics. With the White-adjusted estimator, the null hypothesis of $\gamma_2 = 0$ is rejected at the 5% level in six of the 11 cases as before, indicating the presence of non-linearity in the data.

Given that the datasets used are secondary data from 11 different sources, the 11 tests of the null hypothesis performed above can be regarded as independent tests of a common hypothesis. Therefore, the evidence about the significance levels for the null hypothesis $\gamma_2 = 0$ can be aggregated for the 11 datasets using Fischer's exact chi-square test[25]*. With both the OLS estimates in Table 3, and the corresponding White-adjusted estimates, the cumulative evidence rejects the null hypothesis at the 0.001 significance level. It is concluded, therefore, that the data reject the use of a simple linear form to model the relation between project size and effort.

One concern with such results might be the effect of a small number of influential data (points) (outliers). In addition to the sensitivity analysis described earlier, data were also screened for robustness to deletion of influential observations. Outliers were deleted if they met all four of the Belsley-Kuh-Welsch[23]* criteria. Results after deleting the outliers are reported in Table 4. The linearity assumption is rejected in seven of the 11 datasets at the 5% level, eight of the 11 datasets at the 10% level, including all of the cases identified above. Therefore, the deletion of outliers improves, rather than reduces, the fit in most cases. The $R^2$ increases from 46.13% to 80.54% for the Belady-Lehman dataset when two outliers are deleted, and the absolute $t$-statistic for the $\gamma_2 = 0$ test increases from 1.757 to 5.098 (significant at the 1% level).

All the above tests were also conducted for six subsets of data obtained by partitioning the two largest datasets in this study. The COCOMO and the MERMAID-1 are the only two datasets out of the 11 that have more than 35 observations and afford statistical power for smaller subsamples. The three COCOMO modes, Organic, Semi-detached, and Embedded, proposed by Boehm to distinguish among three contexts, also serve to divide the dataset more or less on the basis of the size of the project. As a result, productivity differences between the modes are confounded with productivity differences for different project sizes due to scale economies. Only the Embedded subsample rejects the null hypothesis of $\gamma_2 = 0$, but when the outliers in the dataset are eliminated the null hypothesis is rejected for the Organic and Semi-detached subsamples. All three subsamples (E1, E2, E3) for the MERMAID-1 dataset fail to reject the null hypothesis, consistent with the results for the full MERMAID-1 sample.

Results using the White adjustment for heteroscedasticity and after the removal of outliers are similar, with the linearity assumption being rejected in seven of 11 cases at the 5%

* Only the $t$-statistics ($p$-values) are aggregated for Fischer's exact chi-square test. Each model is estimated separately by dataset (i.e., the data are not aggregated across data sites to esimate the production correspondence).

* These are implemented as the INFLUENCE option in the SAS statistical package. See SAS manual, Chapter 31, pp 676–678, or Belsey, D A, Kuh, E and Welsch, R E, *Regression diagnostics*, John Wiley (1980).

**Table 4. Summary of quadratic models with outliers removed**

| Data set | N | $\gamma_0$ | $\gamma_1$ | $\gamma_2$ | $R^2$ (%) |
|---|---|---|---|---|---|
| Behrens | 21 | 1094.218<br>(1.530) | −13.152<br>(−1.206) | 0.093<br>(2.674$^a$) | 66.83 |
| Bailey-Basili | 19 | −10.802<br>(−1.359) | 3.233<br>(6.068$^a$) | −0.019<br>(−3.070) | 88.58 |
| Yourdon | 17 | 1.505<br>(0.072) | 1.629<br>(1.739) | −0.005<br>(−0.808) | 42.84 |
| COCOMO | 58 | 29.980<br>(0.475) | 6.888<br>(2.962$^a$) | 0.0038<br>(0.384) | 62.96 |
| Albrecht-Gaffney | 23 | 7.844<br>(1.939) | −0.010<br>(−0.922) | 3.08E-05<br>(5.462$^a$) | 94.00 |
| Belady-Lehman | 31 | −39.799<br>(−0.556) | 9.755<br>(7.844$^a$) | −0.009<br>(−5.098$^a$) | 80.54 |
| Wingfield | 15 | 510.165<br>(0.623) | 0.082<br>(0.010) | 0.023<br>(1.307) | 71.44 |
| Kemerer | 17 | 26267.622<br>(1.945) | −58.620<br>(−2.349$^a$) | 0.048<br>(4.605$^a$) | 81.77 |
| MERMAID-1 | 80 | −1125.089<br>(−0.926) | 27.386<br>(3.428$^a$) | −0.016<br>(−1.501) | 44.57 |
| MERMAID-2 | 29 | 3502.969<br>(1.511) | −6.692<br>(−0.372) | 0.049<br>(2.176$^a$) | 58.73 |
| Kitchenham-Taylor | 33 | −33.336<br>(−1.484) | 0.012<br>(3.696$^a$) | −2.71E-07<br>(−3.073$^a$) | 35.10 |

$t$-statistics in parentheses: $^a$ indicates significant at 5%

level and one more at the 10% level. Therefore, the conclusion from this sensitivity analysis is that the main results are strengthened, not weakened, by the exclusion of data outliers.

### *Non-parametric DEA model*

The above parametric tests are conditioned on the maintained assumptions being true. Its estimates and statistical tests thus depend critically on the validity of specific assumptions about the structure of the model and the probability distribution function for the error term. As such, in the case of the minority of datasets where the null hypothesis of linearity cannot be rejected, it does not necessarily mean that the true model is linear. It only means that linearity does not seem to be an unreasonable approximation to the true unknown model *given* the maintained assumptions about the model specification. Similarly, rejection of the null hypothesis for the majority of the datasets means that linearity is not a reasonable approximation given the maintained assumptions about the model specification.

Non-parametric methods are employed in statistics to lessen the problem with the unknown structural assumptions maintained when using a parametric specification. Therefore, some recent developments in Data Envelopment Analysis (DEA) methodology are employed here to examine the linearity hypothesis. DEA employs a non-parametric specification to estimate the production function (the function relating inputs to outputs) from observed data. Since the DEA methodology maintains relatively few and natural assumptions for the production function, its estimates and test results are likely to be more robust than those obtained from parametric models that postulate a specific structure like a log-linear or quadratic form for the production function[26,27]. The analysis reported below employs the new DEA-based tests of Banker and Chang[28]. Two semi-parametric statistical tests are available, depending on whether the deviations of observed data from estimated production function are postulated to be distributed as exponential or half-normal. These tests are described in greater detail in the appendix. The DEA test results for returns to scale are shown in Table 5.

**Table 5. DEA Model F-tests of null hypothesis of constant returns to scale under two different inefficiency distribution specifications**

| Data set | Exponential | | Half-normal | |
|---|---|---|---|---|
| | F-stat | P > F | F-stat | P>F |
| Behrens | 1.833 | 0.024[a] | 3.234 | 0.005[a] |
| Bailey-Basili | 1.910 | 0.024[a] | 2.690 | 0.018[a] |
| Yourdon | 1.796 | 0.046[a] | 2.500 | 0.033[a] |
| COCOMO | 1.883 | 0.001[a] | 2.939 | 0.001[a] |
| Albrecht-Gaffney | 4.138 | 0.000[a] | 8.214 | 0.000[a] |
| Belady-Lehman | 1.111 | 0.334 | 1.149 | 0.346 |
| Wingfield | 1.725 | 0.070 | 2.375 | 0.052 |
| Kemerer | 1.858 | 0.037[a] | 3.053 | 0.013[a] |
| MERMAID-1 | 1.284 | 0.057 | 1.492 | 0.037[a] |
| MERMAID-2 | 1.557 | 0.044[a] | 2.154 | 0.019[a] |
| Kitchenham-Taylor | 2.006 | 0.003[a] | 2.420 | 0.007[a] |

[a] indicates significant at 5%

As shown in Table 5, with the exception for the Belady-Lehman and Wingfield datasets, the null hypothesis of constant returns to scale is rejected at the 5% significance level under one or both of two alternative statistical test procedures for all the datasets. Clearly, the results support a non-linear relationship between project size and effort.

The DEA test results reported in Table 5 indicate that non-linearity characterizes the datasets and the assumption of constant returns to scale is not sustained. New statistical tests in DEA are also employed to examine whether the data can be described well with just an increasing returns to scale model or with just a decreasing returns to scale model, rather than a model that allows for the presence of both increasing and decreasing returns to scale in different data ranges[28]. The results are reported in Table 6.

The COCOMO and Albrecht-Gaffney datasets deviate from the general model (that permits both increasing and decreasing returns to scale), because of the presence only of decreasing returns to scale. In contrast, the MERMAID-2, Kitchenham and Taylor, and Yourdon datasets deviate from

**Table 6. Summary of data envelopment analysis models**

| Null hypothesis | Non-decreasing | | Non-increasing | |
|---|---|---|---|---|
| Dataset | Exponential | Half-normal | Exponential | Half-normal |
| Behrens | 1.327 | 1.343 | 1.262 | 1.771 |
| | (0.175) | (0.247) | (0.221) | (0.093) |
| Bailey-Basili | 1.322 | 1.482 | 1.303 | 1.434 |
| | (0.196) | (0.199) | (0.208) | (0.219) |
| Yourdon | 1.070 | 1.059 | 1.606 | 2.193 |
| | (0.421) | (0.453) | (0.086) | (0.057) |
| COCOMO | 1.563 | 2.258 | 1.122 | 1.114 |
| | (0.006[a]) | (0.001[a]) | (0.259) | (0.334) |
| Albrecht-Gaffney | 4.138 | 8.298 | 1.000 | 1.000 |
| | (0.000[a]) | (0.000[a]) | (0.500) | (0.500) |
| Belady-Lehman | 1.101 | 1.131 | 1.018 | 1.014 |
| | (0.348) | (0.362) | (0.485) | (0.483) |
| Wingfield | 1.300 | 1.353 | 1.234 | 1.466 |
| | (0.237) | (0.282) | (0.284) | (0.233) |
| Kemerer | 1.294 | 1.686 | 1.306 | 1.361 |
| | (0.227) | (0.145) | (0.219) | (0.265) |
| MERMAID-1 | 1.074 | 1.133 | 1.179 | 1.269 |
| | (0.324) | (0.287) | (0.147) | (0.142) |
| MERMAID-2 | 1.000 | 1.000 | 1.557 | 2.154 |
| | (0.500) | (0.500) | (0.044[a]) | (0.019[a]) |
| Kitchenham-Taylor | 1.011 | 1.002 | 1.960 | 2.407 |
| | (0.481) | (0.497) | (0.003[a]) | (0.006[a]) |

*p*-values in parentheses:[a] indicates significant at 5%

the general model because of the presence of increasing returns to scale. For the remaining six datasets, the rejection of the linearity assumption is attributable to the presence of *both* increasing and decreasing returns to scale in the observed data.

As the loglinear specification in Equation (1) does not allow the possibility of increasing returns for some projects and decreasing for others in the same dataset, it is an inadequate formulation for software development production process modelling.

## Concluding remarks

The evidence from the above analysis indicates that both economies and diseconomies of scale prevail in new software development. This evidence is found to be robust, despite some recent claims to the contrary. The implications for research are that simple linear models are likely to be inadequate representations of the complexity of the software production process. The implications for practice are, as expressed in an earlier paper[10], that project managers should actively seek to use the scale size of future projects as a productivity control tool. When past project data are available, managers may choose to determine the project sizes where their organization has tended to perform most productively. Project requests that are greater than this size should be divided into several smaller projects, with the system being delivered in phases so as to maximize developer productivity. Project requests that are smaller than this size (e.g. maintenance requests), should be considered as candidates for later inclusion in a combined batch of requests which will allow project managers to take advantage of economies of scale.

It is likely that many managers already do this somewhat intuitively, and the model presented here simply provides a tool to help these managers to accomplish this with a greater degree of accuracy. For managers who are not actively using project size as a productivity lever, the models provide evidence on why they should consider this approach as a potentially useful tool.

This research was made possible through the analysis of secondary data generated, in the main, by other researchers. The analysis presented here is representative of the type of value that can be obtained through the wider dissemination of previously collected data. Given the great cost and delays in collecting such real world data, it is imperative that researchers in this area continue to make such data widely available: and given this availability it can be expected that further progress may be made in identifying factors that aid software development productivity.

## Acknowledgements

Research support from the MIT Center for Information Systems Research is gratefully acknowledged. The MERMAID-1 dataset was collected by J-M Desharnais, DMR, Montreal, Canada. The MERMAID-2 dataset was provided by B Kitchenham of the MERMAID project, City University, London, UK. Helpful comments on earlier drafts were provided by B Kitchenham, S Pfleeger and M Shepperd.

## References

1 Boehm, B W *Software engineering economics* Prentice-Hall (1981)

2 Kemerer, C F 'An empirical validation of software cost estimation models' *Comm. ACM* Vol 30 No 5 (1987) pp 416–429

3 Cusumano, M and Kemerer, C F 'A quantitative analysis of US and Japanese practice and performance in software development' *Management Science* Vol 36 No 11 (1990) pp 1384–1406

4 Jeffery, D R and Lawrence, M J 'An inter-organisational comparison of programming productivity' *(Proc. of the 4th International Conference on Software Engineering, 1979)* pp 369–377

5 Albrecht, A J and Gaffney, J 'Software function, source lines of code and development effort prediction: a software science validation' *IEEE Trans. Soft. Eng.* Vol SE-9 No 6 (1983) pp 639–648

6 Low, G C and Jeffery, D R 'Function points in the estimation and evaluation of the software process' *IEEE Trans. Soft. Eng.* Vol 16 No 1 (1990) pp 64–71

7 Symons, C R 'Function point analysis: difficulties and improvements' *IEEE Trans. Soft. Eng.* Vol 14 No 1 (1988) pp 2–11

8 Kemerer, C F 'Reliability of function points measurement: a field experiment' *Comm. ACM* Vol 36 No 2 (1993) pp 85–97

9 Kemerer, C F and Porter, B 'Improving the reliability of function point measurement: an empirical study' *IEEE Trans. Soft. Eng.* Vol 18 No 10 (1992) pp 1011–1024

10 Banker, R D and Kemerer, C F 'Scale economies in new software development' *IEEE Trans. Soft. Eng.* Vol SE-15 No 10 (1989) pp 416–429

11 Byrnes, P E, Frazier, T P and Gulledge, T R 'Returns-to-scale in software production: a comparison of approaches' in Gulledge, T R and Hutzler, W (eds) *Analytical methods in software engineering economics* Springer-Verlag (1993) pp 75–97

12 Fenton, N E *Software metrics, a rigorous approach* Chapman & Hall (1991)

13 Gurbaxani, V and Mendelson, H 'The use of secondary analysis in MIS research' *Proc. Harvard/UCI Workshop on Survey Research in MIS*, Boston, MA, (1989)

14 Marwane, R and Mili, A 'Building tailor-made software cost model: intermediate TUCOMO' *Inf. and Soft. Technol.* Vol 33 No 3 (1991) pp 232–238

15 Richmond, W B, Seidmann, A and Whinston, A B 'Incomplete contracting issues in information systems development outsourcing' *Decision Support Systems* Vol 8 No 5 (1992) pp 459–477

16 Seiford, L M 'Models, extensions and applications of data envelopment analysis: a selected reference set' *Comput. Environ. and Urban Systems* Vol 14 (1990) pp 171–175

17 Withrow, C 'Error density and size in Ada software' *IEEE Soft.* Vol 7 No 1 (1990) pp 26–30

18 Kitchenham, B A 'Empirical studies of assumptions that underlie software cost-estimation models' *Inf. Soft. Technol.* Vol 34 No 4 (1992) pp 211–218

19 Conte, S D, Dunsmore, H E and Shen V Y *Software engineering metrics and models* Benjamin-Cummings (1986)

20 Kitchenham, B and Taylor, N R 'Software project development cost estimation' *J. Systems and Software* Vol 5 (1985) pp 267–278

21 Walston, C E and Felix, C P 'A method of programming measurement and estimation' *IBM Systems J.* Vol 16 No 1 (1977) pp 54–73

22 Judge, G G *et al. The theory and practice of econometrics* John Wiley (1985)

23 Belsley, D A, Kuh, E and Welsch, R E *Regression diagnostics* John Wiley (1980)

24 Shapiro, S S, and Wilk, M B 'An analysis of variance test for normality' *Biometrika* (1965) pp 591–611

25 Christie, A 'Aggregation of test statistics: an evaluation of the evidence on contracting size hypotheses' *J. Accounting and Economics* (1990) pp 15–44

26 Banker, R and Maindiratta, A 'Nonparametric analysis and allocative efficiencies in production' *Econometrica* (1988)

27 Mensah, Y M and Li, S 'Measuring production efficiency in a non-for-profit setting: an extension' *Accounting Review* (1993) pp 66–68

28 Banker, R D and Chang, H *Tests of returns to scale for monotone concave production functions* University of Minnesota (1993)

29 Charnes, A, Cooper, W W and Rhodes, E 'Program evaluation and managerial efficiency: an application of data envelopment analysis to program follow through' *Management Science* (1981)

30 Banker, R D, Charnes, A and Cooper, W W 'Models for the estimation of technical and scale inefficiences in data envelopment analysis' *Management Science* Vol 30 (1984) pp 1078–1092

31 Shephard, R W *The theory of cost and production functions* Princeton University Press (1970)

32 Banker, R D 'Maximum likelihood, consistency and data envelopment analysis: a statistical foundation' *Management Science* Vol 39 No 10 (1993) pp 1265–1273

# Appendix: Data Envelopment Analysis (DEA)

This appendix summarizes the tests of returns to scale described in Banker and Chang[28]. It also provides a brief overview of the Data Envelopment Analysis (DEA) methodology. (Note that a thorough introduction to DEA is beyond the scope of this appendix. Interested readers are referred to Charnes, Cooper and Rhodes[29] and Banker, Charnes and Cooper[30] for detailed explanations.) DEA is used to estimate the production function, the function relating the inputs consumed to the outputs produced. It is also used to estimate the efficiency exhibited by actual observations, such as in consuming more inputs than the minimum required to produce outputs. There are several different DEA models used in practice. The model of Charnes, Cooper and Rhodes (CCR[29]), enables the estimation of the aggregate technical and scale inefficiencies in production, and the Banker, Charnes and Cooper (BCC[30]), model allows the estimation of the technical inefficiency of an observation at the given scale of operation.

Let $Y$ and $X$ represent input and output vectors. The production possibility set can be represented as:

$$T = \{(Y,X) \mid X \geq 0 \text{ can be produced from } Y \geq 0\}. \quad (4)$$

The inefficiency is measured radially by the reciprocal of Shephard's distance function[31]. Thus, the inefficiency of an observation $(Y_0, X_0) \in T$ is given by the following function:

$$\theta\,(Y_0, X_0) = \sup\{\theta \,|\, (Y_0/\theta, X_0) \in T\} \quad (5)$$

Banker specifies the following structure for production set $T$ and the probability density function $f(\theta)$ for the inefficiency variable $\theta$[32]:

POSTULATE 1: CONVEXITY

If $(Y_1, X_1) \in T$ and $(Y_2, X_2) \in T$ then $(\lambda_1 Y_1 + \lambda_2 Y_2, \lambda_1 X_1 + \lambda_2 X_2) \in T$ for all $\lambda_1, \lambda_2 \geq 0$ such that $\lambda_1 + \lambda_2 = 1$.

POSTULATE 2: MONOTONICITY

If $(Y_1, X_1) \in T$, $Y_2 \geq Y_1$ and $X_2 \leq X_1$ then $(Y_2, X_2) \in T$.

POSTULATE 3: ENVELOPMENT

If $\theta > 1$ then $f(\theta) = 0$.

POSTULATE 4: LIKELIHOOD OF EFFICIENT PERFORMANCE

If $\delta > 0$ then $\int_1^{1+\delta} f(\theta)d\theta > 0$.

Let $(Y_j, X_j)$, $j = 1, \ldots J$, be the observed output-input vectors for $J$ observations. To estimate the pure technical efficiency $e^{BCC} = 1/\theta(Y_0, X_0)$ of an observation $(Y_0, X_0)$, the following DEA model of Banker, Charnes and Cooper[30] (BCC hereafter) is employed:

$$\theta(Y_0, X_0) = \text{Max } \theta \quad (6.0)$$

subject to

$$\sum_{j=1}^{J} \lambda_j Y_j - Y_0/\theta \leq 0 \quad (6.1)$$

$$\sum_{j=1}^{J} \lambda_j X_j \geq X_0 \quad (6.2)$$

$$\sum_{j=1}^{J} \lambda_j = 1 \quad (6.3)$$

$$\theta \text{ and } \lambda_j \geq 0 \quad (6.4)$$

This model is solved as a linear program in the variables $e = 1/\theta$ and $\lambda_j$. Banker[32] shows that the DEA estimator of $\theta$ using the BCC model described in Equation (6) is statistically consistent and the asymptotic empirical distribution of the DEA estimates retrieves the true distribution of $\theta$ under the maintained assumptions embodied in the four postulates. These postulates are logically consistent with both increasing and decreasing returns to scale; they do not impose constant returns to scale. Such an inefficiency measure estimated using the BCC model is referred to as $\theta^B$.

If an additional condition that the production set exhibits constant returns to scale is imposed, then the so-called CCR efficiency estimates $e^{CCR} = 1/\theta$ are obtained by solving the above linear program, except that the objective function in Equation (6.0) is maximized subject only to constraints in Equations (6.1) (6.2) and (6.4), the constraint in Equation (6.3) is deleted. The maintained assumptions now also include the following postulate:

POSTULATE 5: CONSTANT RETURNS TO SCALE

If $(Y, X) \in T$ then $(kY, kX) \in T$ for any $k > 0$.

The CCR estimator is also statistically consistent under the maintained assumptions reflected in postulate 1, 2, 3, 4, and 5. The CCR inefficiency estimate is referred to as $\theta^c$.

As the DEA estimator is statistically consistent, under the null hypothesis of constant returns to scale the asymptotic empirical distributions of DEA estimates of $\theta^B$ and of $\theta^C$ are identical, each recovering the true distribution of $\theta$. The asymptotic correspondence between the empirical distributions of $\theta^B$ and $\theta^C$ under the null hypothesis of constant returns to scale motivates the following two semi-parametric statistical tests[28]:

(I) If the inefficiency variable $\theta$ follows the exponential distribution over the range of values from one to infinity with a parameter $\sigma$, then asymptotically both the sums

$$\sum_{j=1}^{J} 2(\theta_j^B - 1)/\sigma \text{ and } \sum_{j=1}^{J} 2(\theta_j^C - 1)/\sigma$$

follow the chi-square distribution with 2N degrees of freedom. Therefore, the test statistic for the null hypothesis of constant returns to scale is given by

$$\sum_{j=i}^{J} (\theta_i^c - 1)/ \sum_{j=i}^{J} (\theta_j^B - 1),$$

which asymptotically obeys the F-distribution with (2N, 2N) degrees of freedom.

(II) If the inefficiency variable is distributed half-normally as $|N(1,\sigma^2)|$ then both

$$\sum_{j=1}^{J} (\theta_j^B - 1)^2/\sigma^2 \text{ and } \sum_{j=1}^{J} (\theta_j^C - 1)^2/\sigma^2$$

follow the chi-square distribution with N degrees of freedom. Therefore, the test statistic for the null hypothesis of constant returns to scale is given by

$$\sum_{j=i}^{J} (\theta_j^c - 1)^2/ \sum_{j=i}^{J} (\theta_j^B - 1)^2$$

which follows the F-distribution with (N, N) degrees of freedom.

See Banker and Chang[28] for an additional Smirnov-type non-parametric test of constant returns to scale. Semi-parametric and non-parametric tests resembling those described here are also presented by Banker and Chang to test for increasing or decreasing returns to scale.