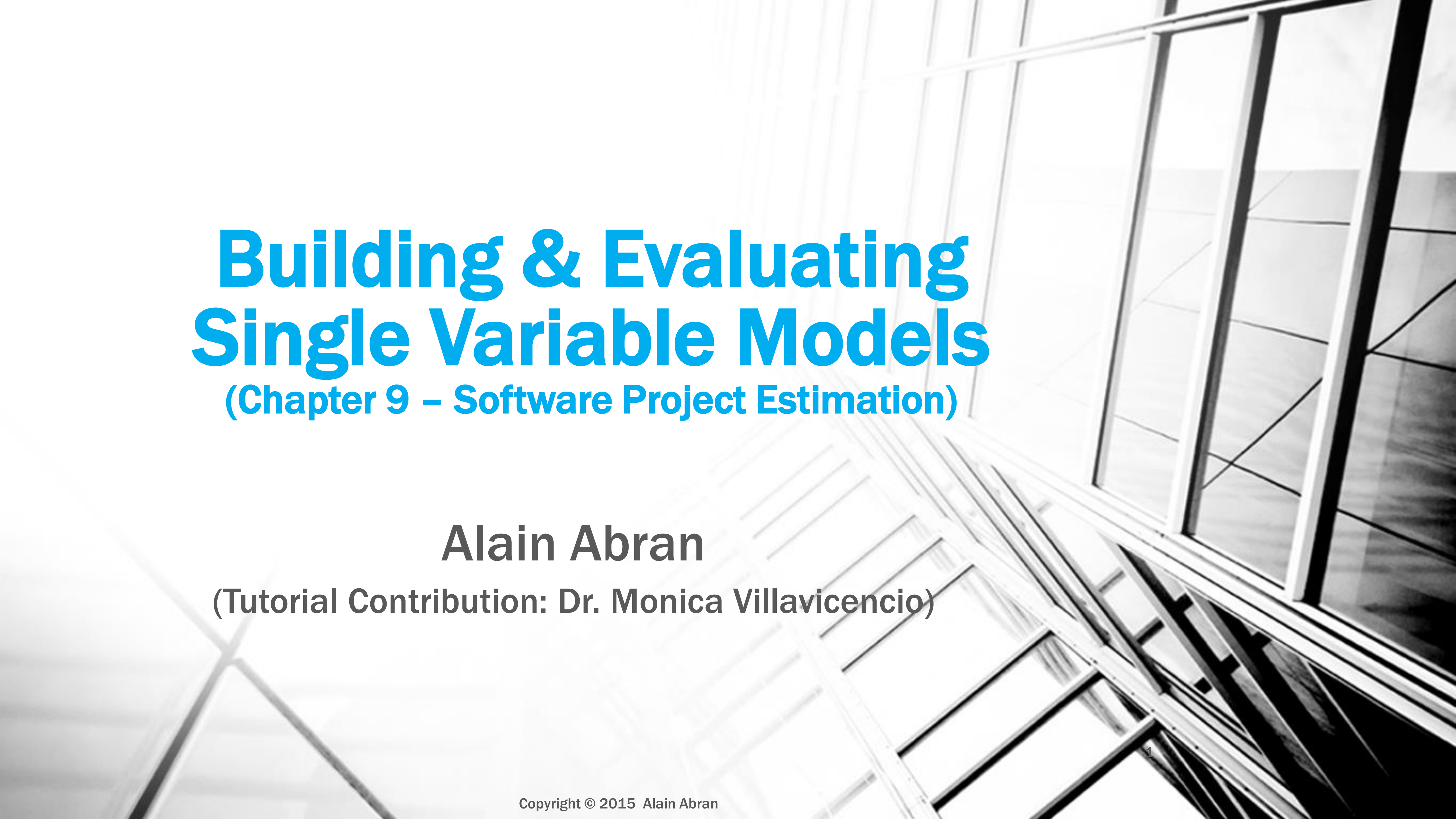


Building & Evaluating Single Variable Models

(Chapter 9 – Software Project Estimation)

Alain Abran

(Tutorial Contribution: Dr. Monica Villavicencio)

Topics covered

1. Introduction
2. One variable at a time
3. Data Preparation
4. Analysis of the Quality & Constraints of Models
5. Other Models by Programming Language

9.1 Introduction

Engineering approach

- Work-effort relationship
  - The factors are investigated and studied one at a time to avoid potentially confusing their individual contributions to the relationship.
- The models are built on the basis of:
  - Observation of past projects.
  - identification of the scale types of the variables
  - Analysis of the impact of each individual variable, one at a time.
  - Selection of relevant samples, and of samples of sufficient size.
  - Documentation and analysis of the demographics of the dataset used.
  - No extrapolation to contexts other than those from which the data were collected.
- It searches instead for models that are reasonably good within a well identified and understood set of constraints.

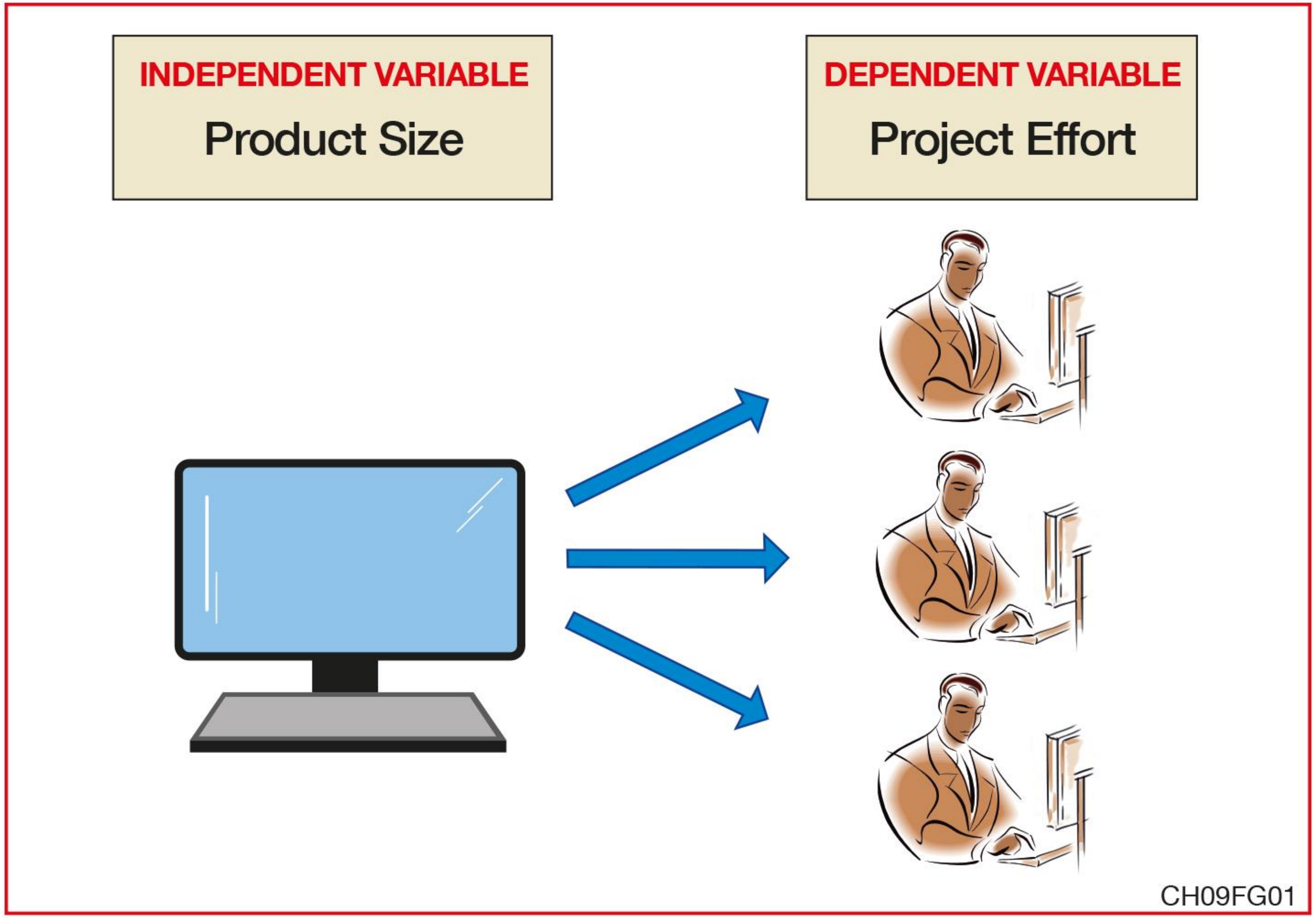
9.2 One Variable at a Time

Size: A key independent variable

- Software size is a significant driver of project effort: it has often been observed to be the key independent variable.

- Other factors can next be taken into account as additional independent variables to improve the modeling of the relationship with effort.

Product size as the key driver of project effort



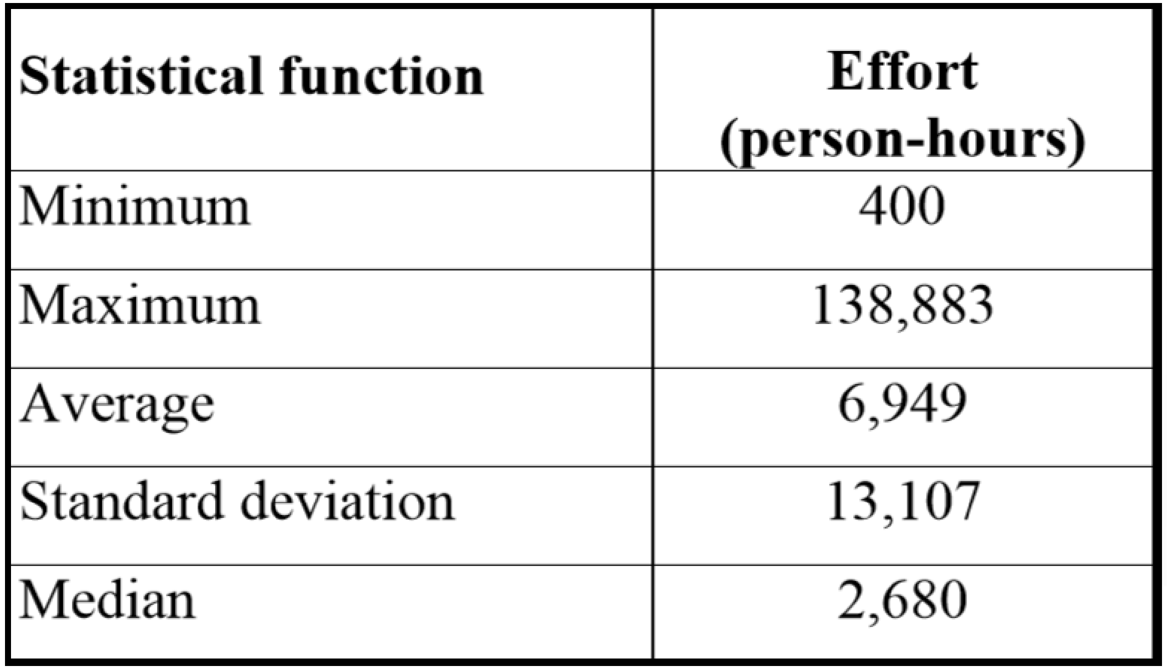
9.3 Data Preparation

Data preparation in ISBSG

An example with 789 projects from 20 countries from the Release 9 of the ISBSG repository.

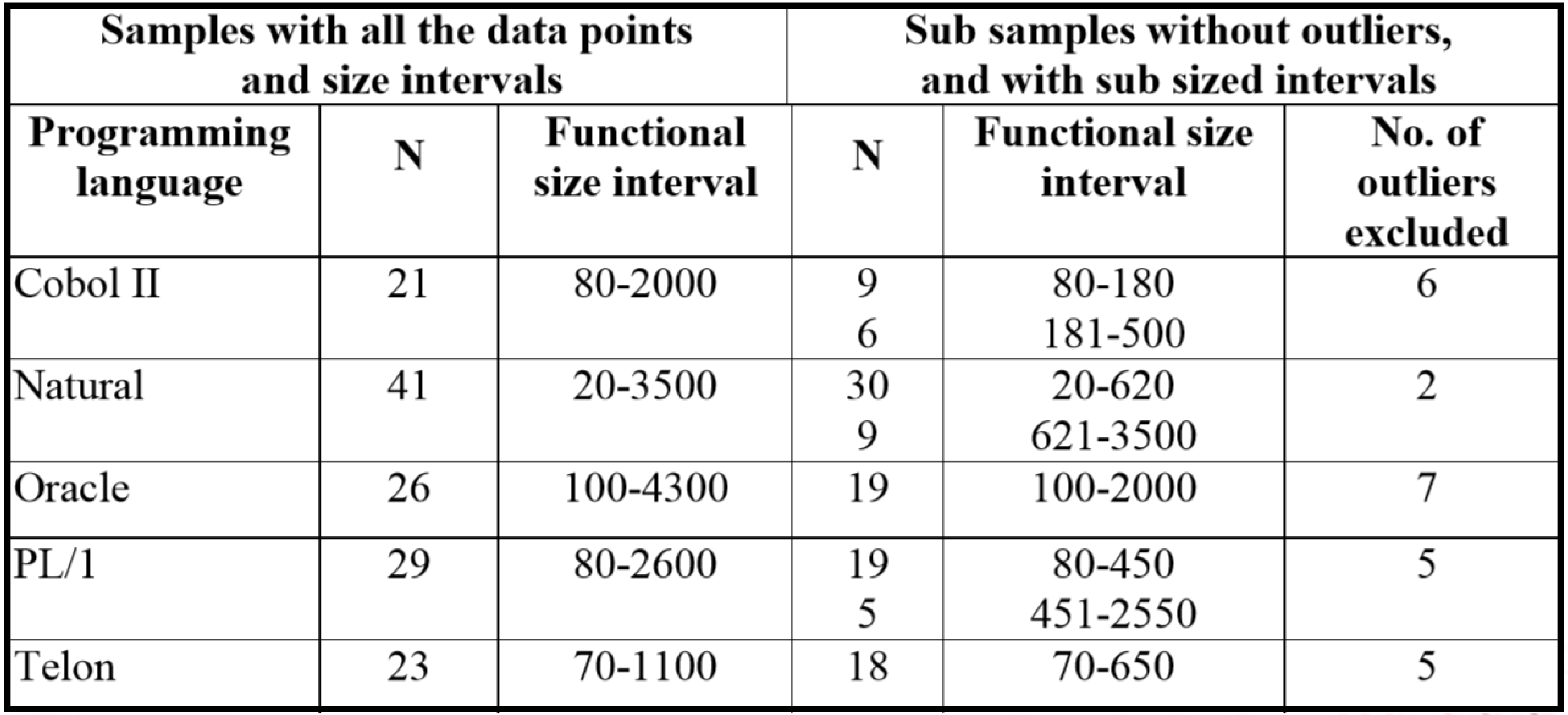
- Descriptive analysis.
- Identification of relevant samples & outliers.

Descriptive analysis of the ISBSG R9 qualified sample (N=497)



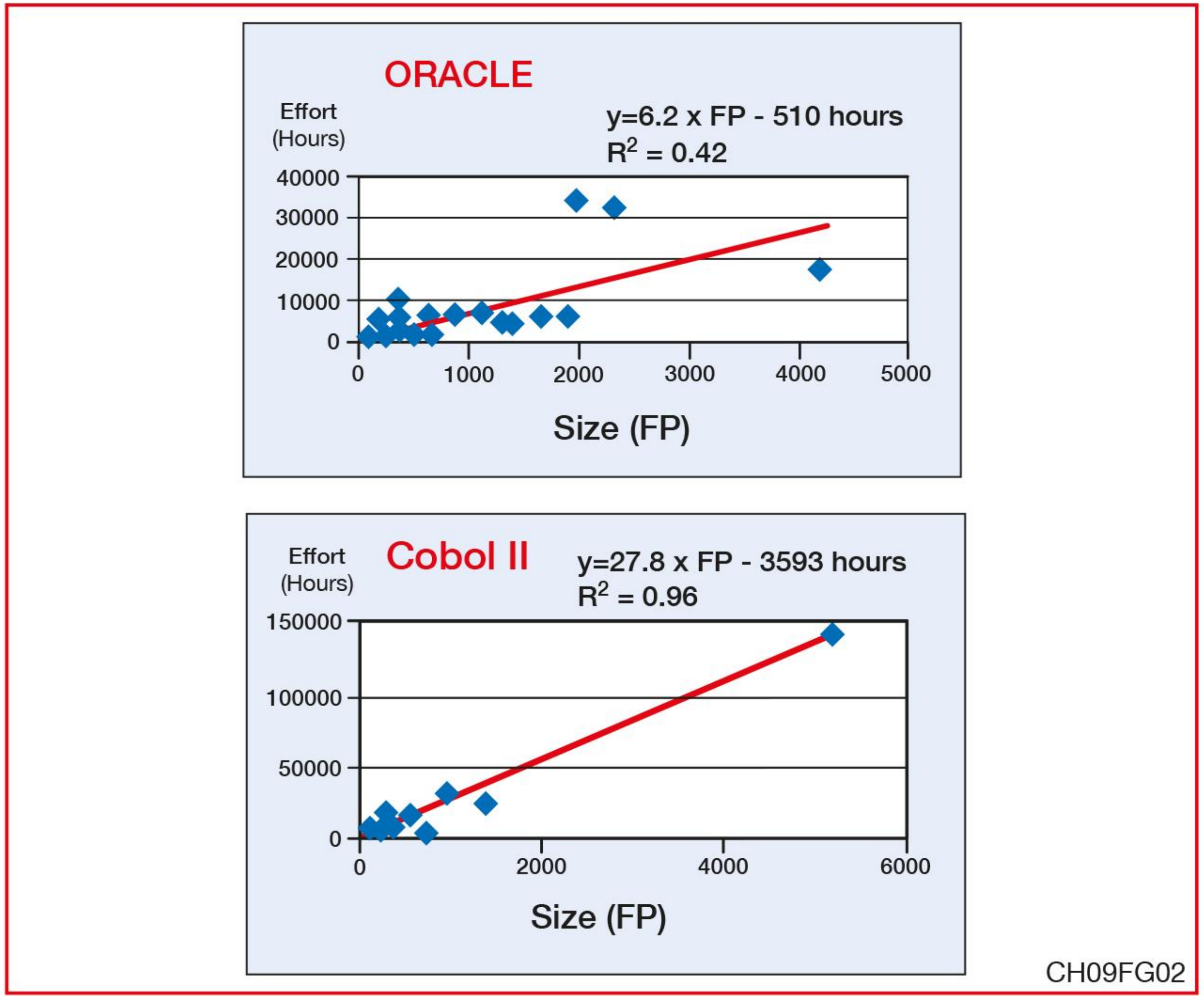
Statistical function	Effort (person-hours)
Minimum	400
Maximum	138,883
Average	6,949
Standard deviation	13,107
Median	2,680

Samples by programming language (with & without outliers)



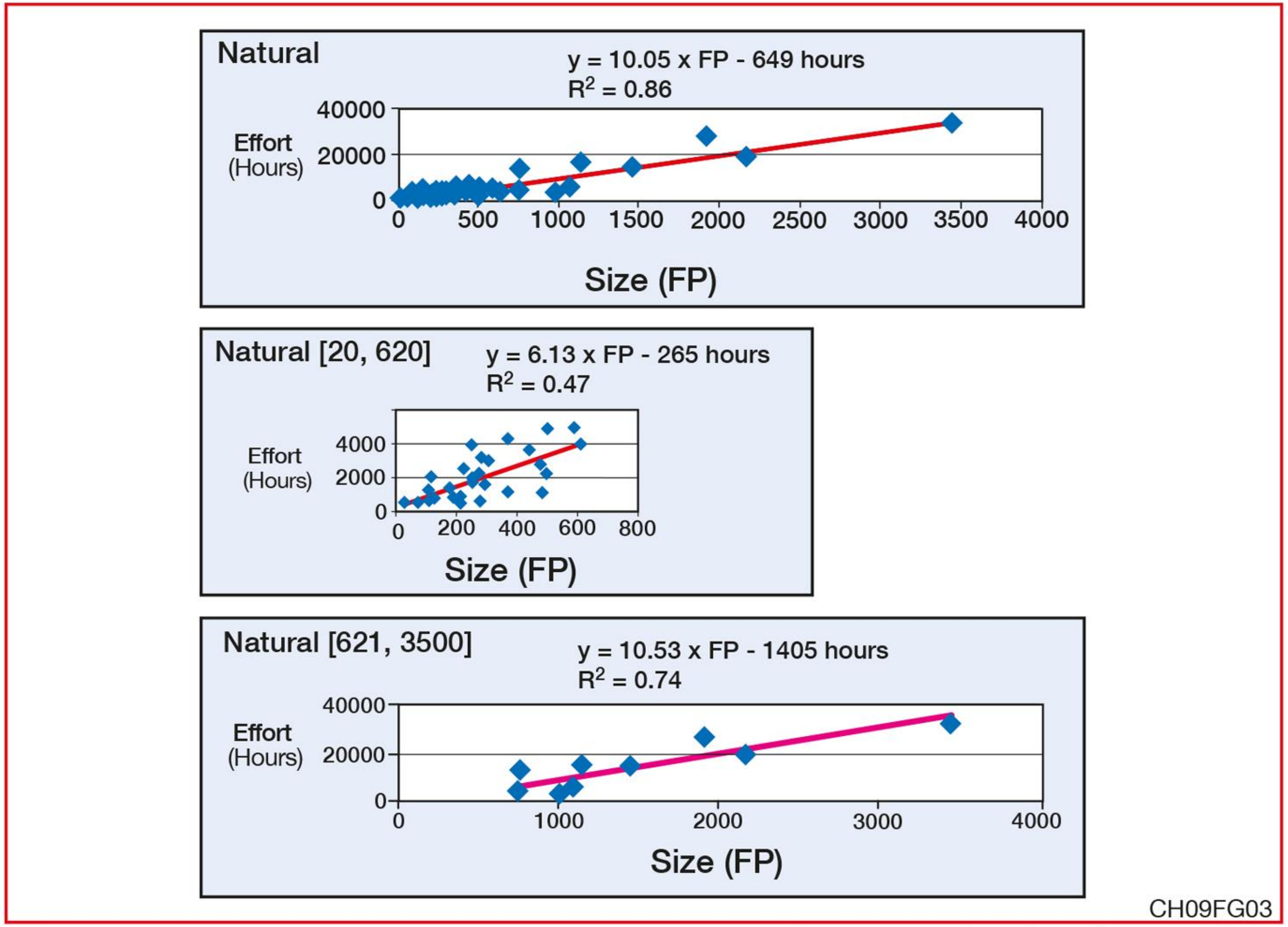
Samples with all the data points and size intervals			Sub samples without outliers, and with sub sized intervals		
Programming language	N	Functional size interval	N	Functional size interval	No. of outliers excluded
Cobol II	21	80-2000	9 6	80-180 181-500	6
Natural	41	20-3500	30 9	20-620 621-3500	2
Oracle	26	100-4300	19	100-2000	7
PL/1	29	80-2600	19 5	80-450 451-2550	5
Telon	23	70-1100	18	70-650	5

Full dataset for Oracle and Cobol II



9.4 Analysis of the Quality & Constraints of Models

Regression analyses for projects = Natural programming language



Implications for practitioners

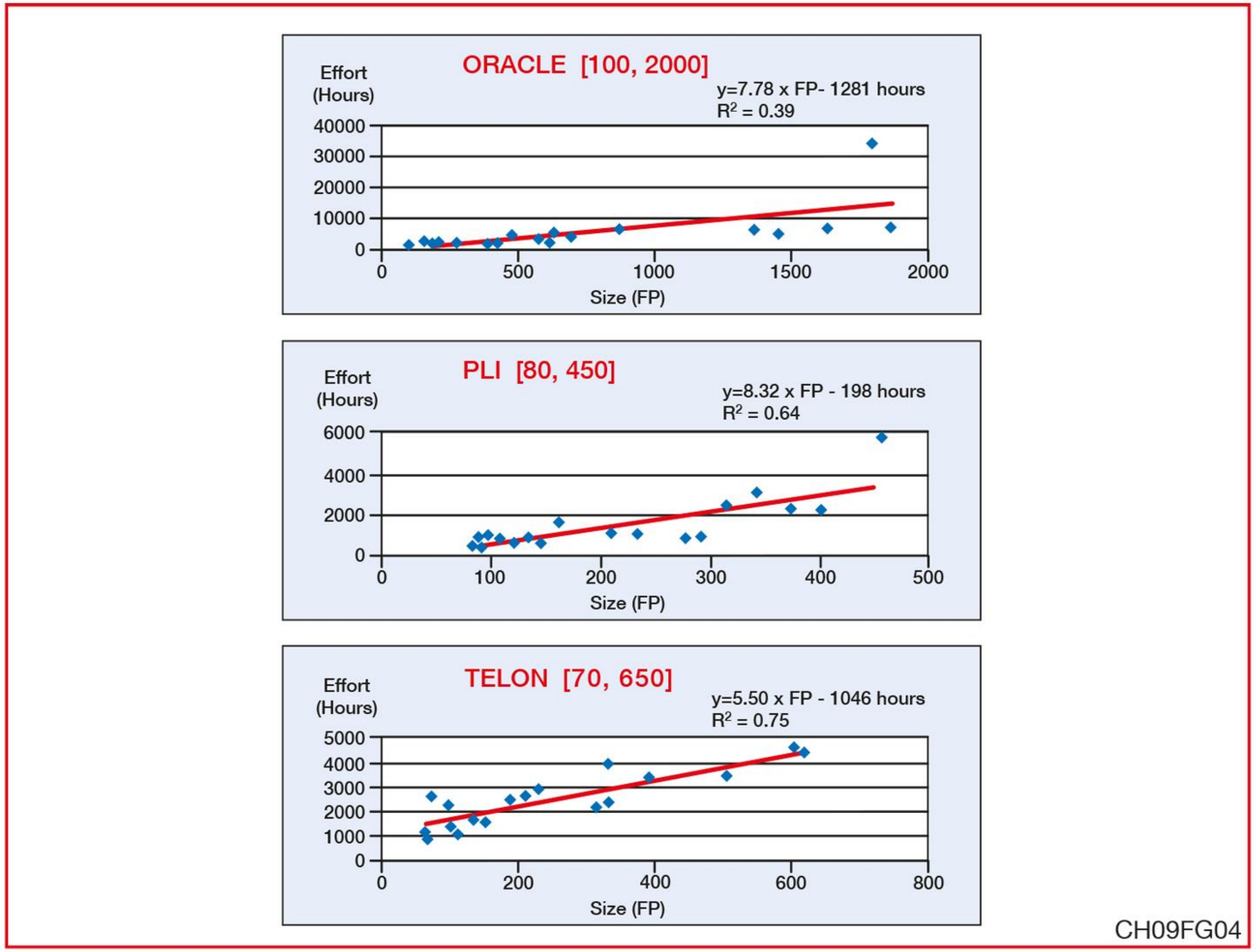
- To build models, use projects with similar characteristics (example: small projects).

- For models built with a small sample, be aware that generalizations are not possible.

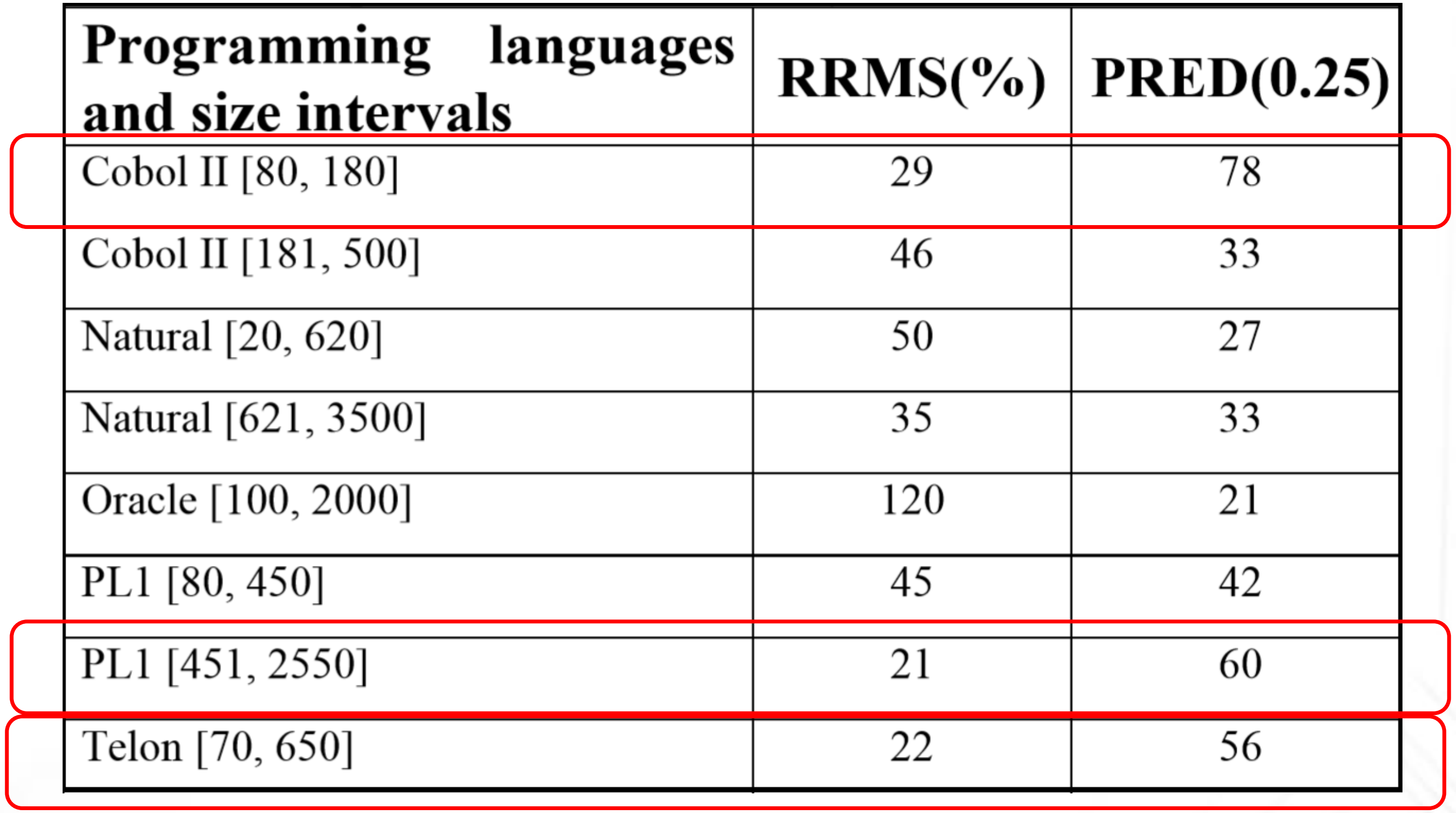
- Analyze regression coefficients (the highest coefficients do not imply that you have the best model – beware of improper dataset with outliers!).

9.5 Other Models by Programming Language

Directly derived ISBSG productivity models for various programming languages – within size ranges with enough data points



Performance of ISBSG regression models (on samples excluding outliers & specified size ranges)



Programming languages and size intervals	RRMS(%)	PRED(0.25)
Cobol II [80, 180]	29	78
Cobol II [181, 500]	46	33
Natural [20, 620]	50	27
Natural [621, 3500]	35	33
Oracle [100, 2000]	120	21
PL1 [80, 450]	45	42
PL1 [451, 2550]	21	60
Telon [70, 650]	22	56

Exercises

1. How would you build a model using an engineering approach?

2. Why is it important to look at the descriptive statistics of a dataset before building a model? Use Table 9.1 to explain why this is important.

3. The data points in Figure 9.2b for the programming language COBOL II leads to a model with a very high R^2 of 0.96. Explain why this R^2 for this dataset is misleading.

4. If you have a large dataset of completed projects (such as in the ISBSG repository), how would you go about determining the impact of a single cost factor? Provide an example with a specific cost factor.

5. How do you recognize a project outlier in a dataset of completed projects?

6. What is the impact of project outliers on the quality of your models? What is the impact on your next project estimate when outliers are embedded within the initial model?

Exercises

7. Look at Figure 9.3 and compare the three models. Which model is the best for estimating a project with an expected size of 400 Function Points?

8. In Table 9.3, there are a number of estimation equations with negative constants. How do you interpret this negative constant? What precaution(s) do you need to take when using models with negative constants?

9. Of the various models in Table 9.3, which have the lowest fixed costs in terms of effort, and which have the lowest variable costs in terms of effort?

10. Based on Table 4, which is better for estimation purposes, a higher RRMS or a lower PRED(0.25)?

Term Assignments

1. If your organization does not have a repository of completed projects, how can you test the relevance for your organization of using an external dataset (such as the ISBSG, or any other similar repository) in your organization? What would you recommend to an organization in such a context?

2. Think back over the past three projects you worked on. What are the 3 to 5 cost factors that had the most impact on the difference in productivity on those projects? List another 10 cost factors. What was the relative importance of the 5 most important cost factors in explaining project productivity (as compared to the other 10 you listed)?

3. Benchmark your organization's productivity model with a set of comparable projects from the ISBSG repository.

4. Select, using criteria of your choice, a subset of data from the ISBSG repository. What is the shape of the resulting graphical representation (with the Functional Size and Effort variables)? Explain.

Term Assignments

5. Three major steps are recommended for building productivity models based on historical data: data preparation, application of statistical tools, and data analysis. Document how these steps have been applied within your own organization.

6. If your organization has not built a productivity model based on past projects, select a model proposed in the literature, and carry out a similar analysis. Which step is particularly weak, and which is particularly strong?

7. Select a model documented in the literature and built from a statistical analysis. How were outliers handled in the data preparation and in the statistical analyses?