Building Models with Categorical Variables

(Chapter 10 – Software Project Estimation)

Alain Abran

(Tutorial Contribution: Dr. Monica Villavicencio)

Topics covered

1. Introduction
2. The dataset available for building the model
3. Initial Model with a Single Independent Variable
4. Regression Models with Two Independent Variables

10.1 Introduction

Models for estimating project effort

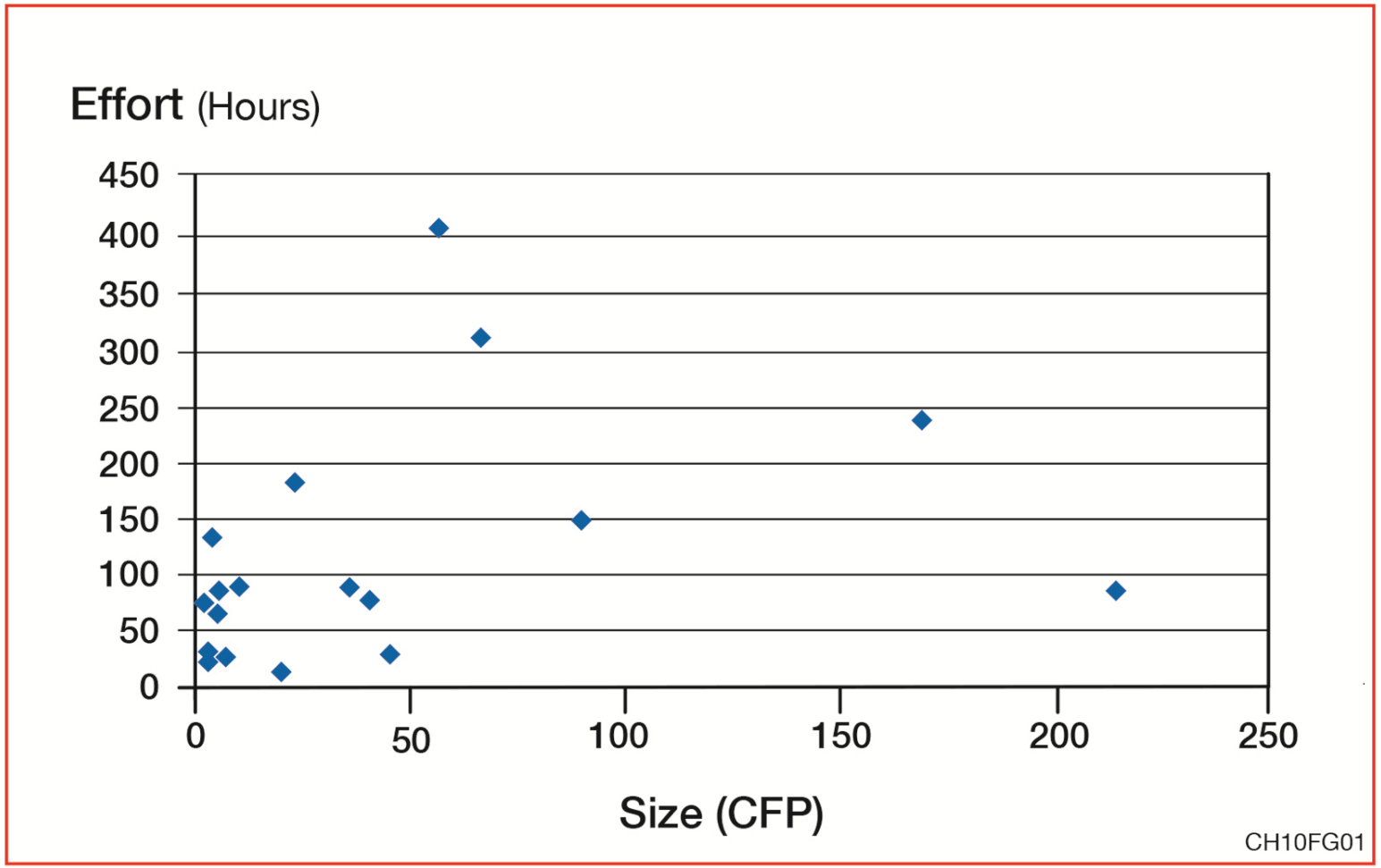
- Key factor: Software size (independent variable)
- Other factors (independent variables) not often quantifiable:
  - severe constraints on resource availability
  - functional complexity
  - technical complexity
  - low or high level of reuse,
  - etc.
- Usually, only small data sets asre available for research purposes:
  - For analyzing over 100 variables at the same time
  - Many variables are categorical

10.2 The Available Dataset

Characteristics of the data set available

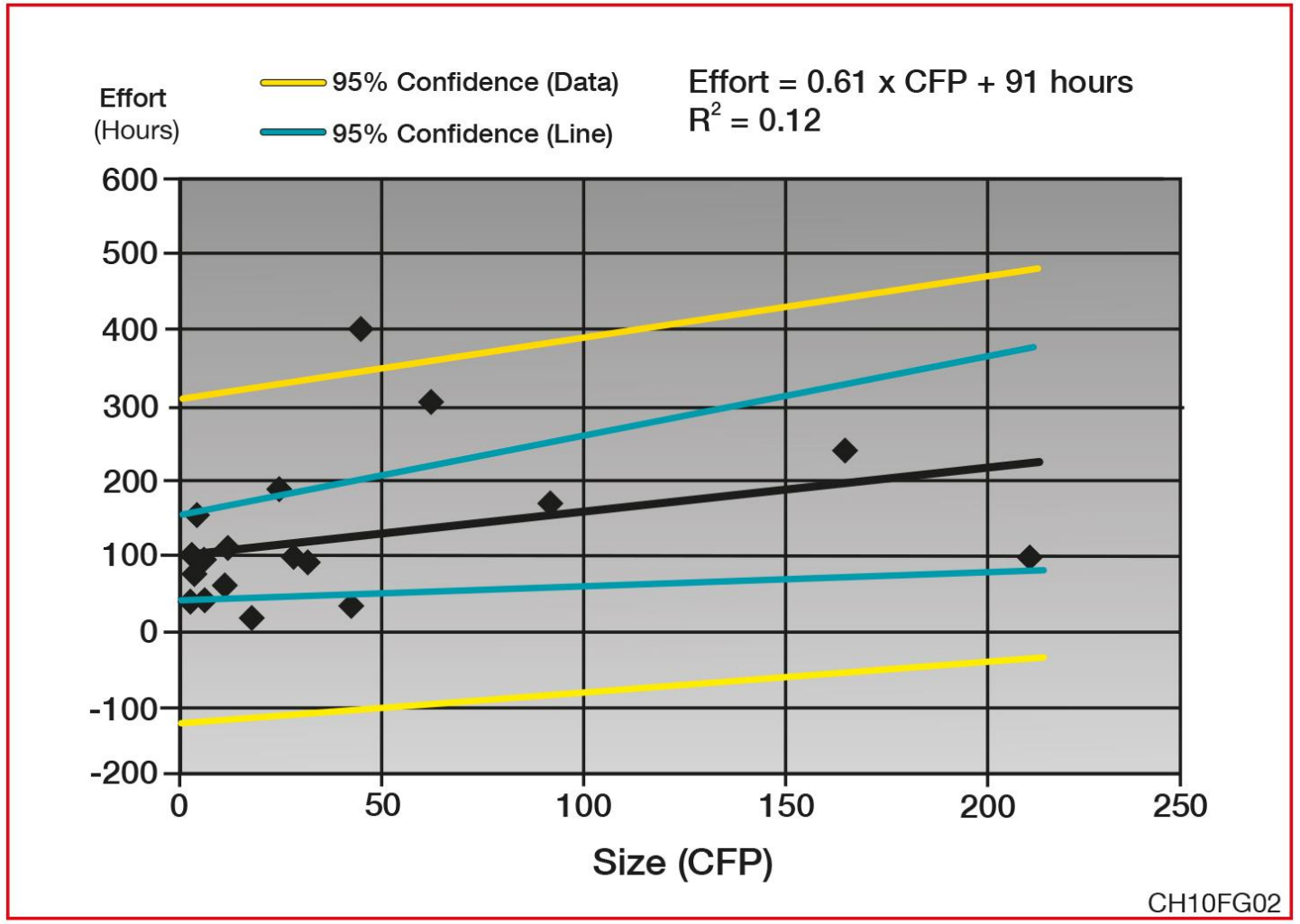
- From a single organization which designs, develops, and implements systems for the defense industry.
- The projects measured and analyzed were carried out on the same software application
- Factors that were held constant in each project in this data set:
    - software application
    - software domain
    - development and exploitation environment (platform, DBMS, testing tools, etc.)
    - programming language
    - enhancement methodology

Data graph - excluding 2 outliers (N=19)



10.3 Initial Model with a Single Independent Variable

Linear Regression: 1 variable = Functional Size in CFP units *(with confidence intervals - yellow & blue lines)*



Nonlinear regression models (N=19)

$Y = A \times X^B$	Power
$Y = A \times e^{(B \times X)}$	Exponential
$Y = A + B \times \ln(X)$	Logarithmic
$Y = A + B/X$	Hyperbolic 1
$= 1 / (A + (B \times X))$	Hyperbolic 2

N	A	B	R	R^2
19	43.808	0.245	0.50	0.245
19	63.067	0.006	0.39	0.15
19	44.121	29.29	0.51	0.26
19	132.463	-48.330	0.32	0.10
19	0.022	-8.8E-05	0.31	0.09

10.4 Regression Models with 2 Independent Variables

Multiple regression models with 2 independent quantitative variables

- A 2nd variable is introduced in successive linear regression models of the form y = ax + bz + c
  - Lines of code (LOC)
  - CFPs
  - The number of lines of code modified
  - The number of programs modified
- The model with 2 independent variables (functional size and number of programs modified) does not show an improvement over the simple linear regression model (R^2 value=0.12).

Multiple regression models with a categorical variable: Project difficulty (1/2)

- 4 Levels of project difficulty initially selected:
  - not difficult
  - difficult
  - very difficult
  - extremely difficult
- But with insufficient data in each of these 4 categories → a simpler classification with 2 categories only was selected:
  - low difficulty
  - high difficulty

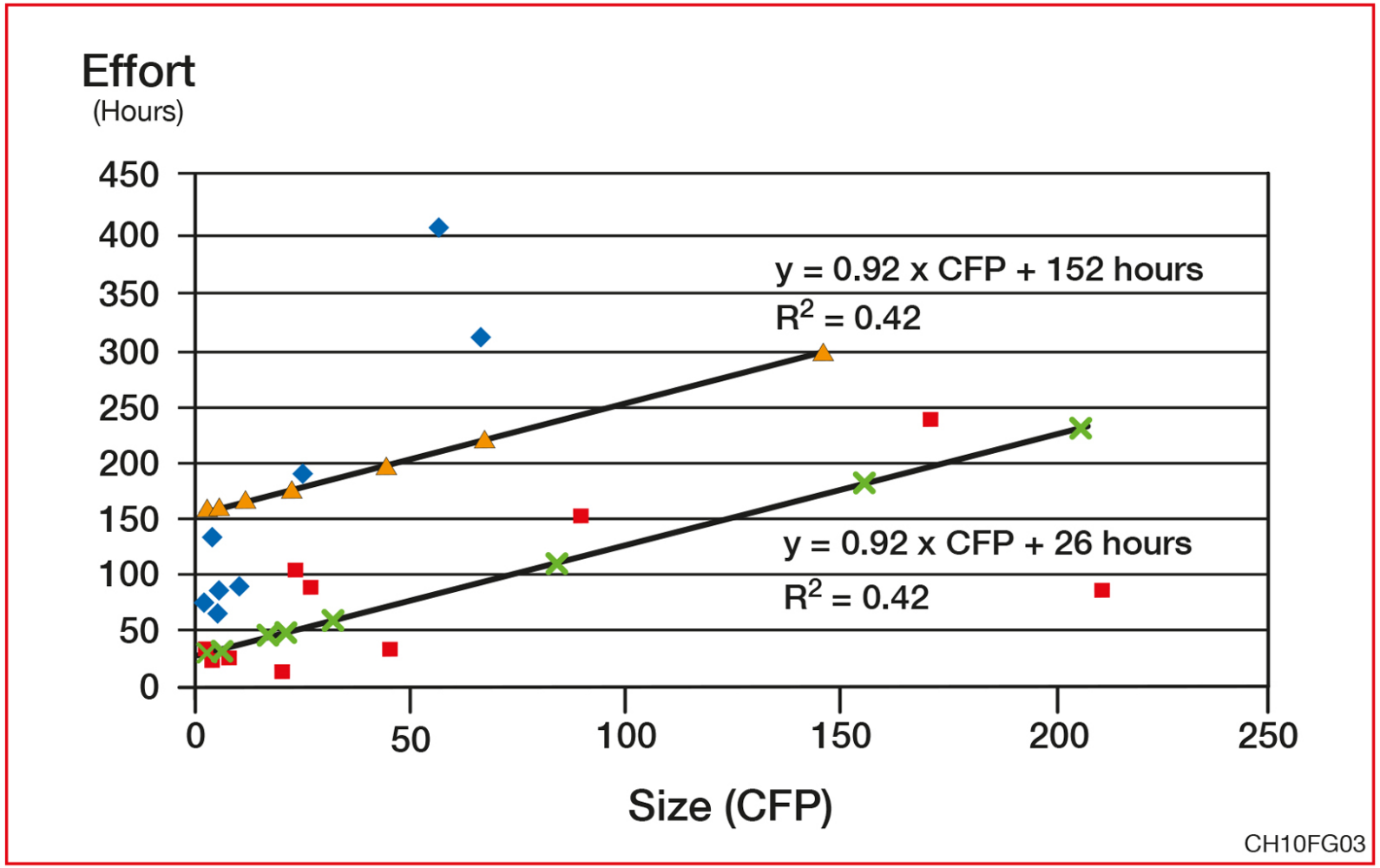
Multiple regression models with a categorical variable: Project difficulty (2/2)

- Additive form for the 2nd variable:
  - Categorical variables can be taken into account in regression models through the addition of dummy variables. For example:
    - Difficulty = 1, for a high level
    - Difficulty = 0, for a low level
- Relationship between size and work effort (Y):

$$Y = aX + bZ + c$$

  - if $Z = 0$ $Y = aX + c$
  - if $Z = 1$ $Y = aX + (b + c)$

Additive Model



Multiplicative regression models

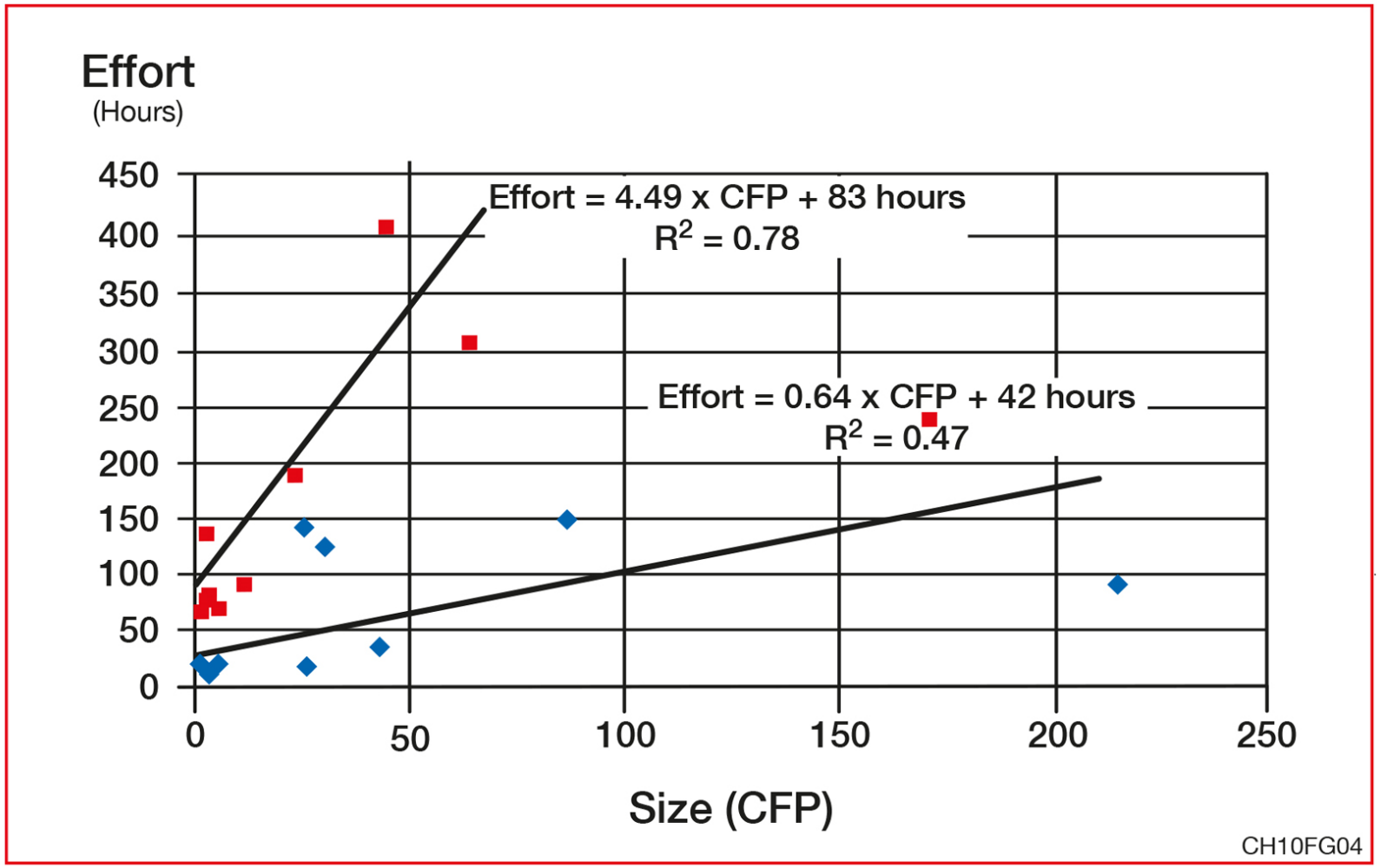
$$Y = \alpha X + \beta Z + \gamma (X \times Z) + \mu, \text{ that is,}$$

$$\text{Effort} = \alpha \text{ CFP} + \beta \text{ Difficulty} + \gamma (\text{CFP} \times \text{Difficulty}) + \mu$$

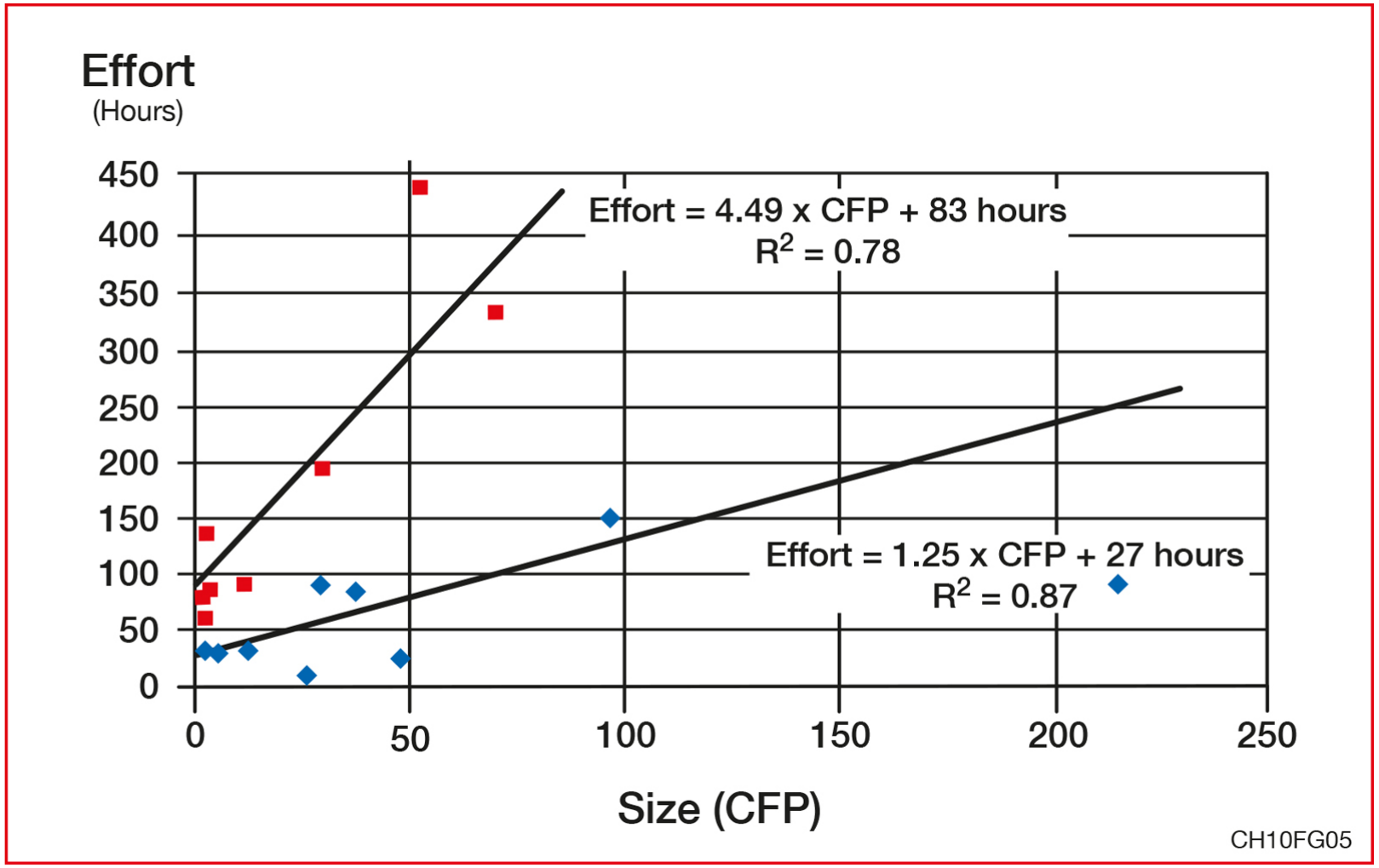
If difficulty $= 0 \rightarrow$ Effort $= \alpha$ CFP $+ \mu$.

If difficulty $= 1 \rightarrow$ Effort $= (\alpha + \gamma)$ CFP$+ (\mu + \beta)$.

Multiplicative Model (n=19)



Multiplicative Model (n=18)



Exercises

ID number	Effort in hours	Total functional size of the program (CFP)	Functional size of the modification ($CFP_{modified}$)	Difficulty rating (2-level: Low, High)
1	88	360	216	L
2	956	984	618	L
3	148	123	89	L
4	66	40	3	H
5	83	16	3	H
6	34	18	7	L
7	96	120	21	L
8	84	88	25	L
9	31	151	42	L
10	409	75	46	H
11	30	36	2	L
12	140	7	2	H
13	308	125	67	H
14	244	232	173	L
15	188	53	25	H
16	34	44	1	L
17	73	22	1	H
18	27	6	1	L
19	91	53	8	H
20	13	37	19	L
21	724	248	157	-

Excercises

1. Discuss the difference between a quantitative variable, a categorical variable, and a nominative variable. Discuss how you can correctly handle each of these types of variable in a regression model.

2. The table above represents the full dataset of the case study presented in this chapter. Identify the 2 projects considered as outliers, using both graphical analysis and statistical tests.

3. Delete the 5 largest projects. What is the impact of this on the linear regression model? Delete the 5 smallest projects. What is the impact of this on the linear regression model? Comment on the difference between these various regression models. Based on what you observed, what would you recommend to your management; that is, what productivity model should be used, and under what conditions?

Exercises

4. Repeat exercise 3, this time using an exponential regression technique.

5. Change the project difficulty category of project 10 (from high to low), and rerun the additive and multiplicative regression models.

6. Build a multiplicative regression model taking into account both the quantitative variables: the total size of programs (column 3), and the size of the modified functionality (column 4).

Term Assignments

1. Using project data collected in your organization, identify categorical variables and build both additive and multiplicative regression models.
2. Using project data from the ISBSG, identify categorical variables and build both additive and multiplicative regression models.