Handling Categorical Variables in Effort Estimation

Masateru Tsunoda
Toyo University
2100 Kujirai, Kawagoe, Saitama
350-8585 Japan
tsunoda@toyo.jp

Sousuke Amasaki
Okayama Prefectural University
111 Kuboki, Soja, Okayama
719-1197 Japan
amasaki@cse.oka-pu.ac.jp

Akito Monden
Nara Institute of Science and Technology, Kansai Science City
630-0192 Japan
akito-m@is.naist.jp

ABSTRACT

Background: Accurate effort estimation is the basis of the software development project management. The linear regression model is one of the widely-used methods for the purpose. A dataset used to build a model often includes categorical variables denoting such as programming languages. Categorical variables are usually handled with two methods: the stratification and dummy variables. Those methods have a positive effect on accuracy but have shortcomings. The other handing method, the interaction and the hierarchical linear model (HLM), might be able to compensate for them. However, the two methods have not been examined in the research area. **Aim:** giving useful suggestions for handling categorical variables with the stratification, transforming dummy variables, the interaction, or HLM, when building an estimation model. **Method:** We built estimation models with the four handling methods on ISBSG, NASA, and Desharnais datasets, and compared accuracy of the methods with each other. **Results:** The most effective method was different for datasets, and the difference was statistically significant on both mean balanced relative error (*MBRE*) and mean magnitude of relative error (*MMRE*). The interaction and HLM were effective in a certain case. **Conclusions:** The stratification and transforming dummy variables should be tried at least, for obtaining an accurate model. In addition, we suggest that the application of the interaction and HLM should be considered when building the estimation model.

Categories and Subject Descriptors

D.2.9 [**Software Engineering**]: Management – *Cost estimation*, K.6.1 [**Computing Milieux**]: Project and People Management – *Staffing*

Keywords

Model-based effort estimation, dummy variable, stratification, interaction, hierarchical linear model, mixed effects.

1. INTRODUCTION

Software development effort estimation is the basis of the project management. Model-based effort estimation methods have been studied well for its quantitative nature. The linear regression is one of the widely-used methods. In model building, a past project dataset is used for parameter inference. A typical regression-based model is as follows:

$$\text{Effort} = \text{Size}^{\beta_s} x_1^{\beta_1} e^{\beta_2 x_2 + \beta_0 + \varepsilon}. \quad (1)$$

Size represents a functional size and other terms are attributes of a target project. β_k are parameters to be inferred, and ε is an error term. When β_s is greater than 1, the model signifies diseconomies of scale. When β_s is smaller than 1, it signifies economies of scale.

Log-transformed Eq. (1) is often used for parameter inference:

$$\log(\text{Effort}) = \beta_s \log(\text{Size}) + \beta_1 \log(x_1) + \beta_2 x_2 + \beta_0 + \varepsilon. \quad (2)$$

Attributes in a dataset can be classified into ratio (or ordinal) scale and nominal scale. For example, the functional size and effort are ratio scale attributes, and the programming language is the nominal scale attribute (categorical variables). In many studies on effort estimation models, categorical variables were handled with two ways: stratification [10] and dummy variables.

The stratification divides a dataset into subsets according to levels of categorical variables, and multiple models are built based on the subsets. Each model has parameters specific to a subset j:

$$\log(\text{Effort}) = \beta_{sj} \log(\text{Size}) + \beta_{1j} \log(x_1) + \beta_{2j} x_2 + \beta_{0j} + \varepsilon. \quad (3)$$

Dummy variables are used when a categorical variable is considered as predictor. A categorical variable is transformed into multiple dummy binary variables. Those are treated as ordinal scale. If a categorical variable is transformed into a dummy variable y, an equation takes the following form:

$$\log(\text{Effort}) = \beta_s \log(\text{Size}) + \beta_1 \log(x_1) + \beta_2 x_2 + \beta_0 + \beta_4 y + \varepsilon. \quad (4)$$

Dummy variables make difference only on the intercept (The model expresses either diseconomies of scale or economies of scale for all categories). Stratification is more flexible in this sense. However, stratification is disadvantageous in that its estimation accuracy may be low when an estimation model is built with small subset. In addition, these handling methods do not cover the model that a categorical variable has an effect only on economies/diseconomies of size.

The problems can be handled with two methods: the interaction and hierarchical linear model (HLM). The interaction [1] can handle the restriction on modeling of economies/diseconomies. The HLM [3] can mitigate the disadvantage. However, those method have rarely used in the literature. Furthermore, the effects of those four handling methods have not been evaluated comparatively.

The goal of this research is to give useful suggestions for handling categorical variable methods when building an effort estimation

ESEM'12, September 19–20, 2012, Lund, Sweden.

model based on the linear regression model. To achieve the goal, we set research questions as follows:

- RQ1. Are effort estimation accuracies of the models different from each categorical variable handling method?
- RQ2. (If the answer of RQ1 is "yes") Is there categorical variable handling method whose performance (the accuracy of the model built by the method) is the highest or the lowest in any case?
- RQ3. (If the answer of RQ2 is "no") Should we consider application of the interaction or HLM? (That is, are the accuracies of the models with the interaction or HLM higher than other methods in some cases?)

2. TREATING CATEGORICAL VARIABLE

2.1 Stratification

The stratification divides a dataset into subsets based on the values of a categorical variable. Estimation models are built with each subset. For instance, when a dataset includes a categorical variable denoting a programming language either "C" or "Java", the stratification divides the dataset into two subsets. Then, two estimation models are built with the two subsets. When a target project plans to use C, the estimation model for C is used.

The advantage of the stratification is building more flexible model than the dummy variable model. That is, with the stratification, the model of economies of scale can be built for some categories, and the model of diseconomies of scale can be built for the other categories. The disadvantage is that the estimation accuracy may be low when an estimation model is built with a small subset.

2.2 Dummy variables

Dummy variables are used to transform categorical variables into numerical variables. When a categorical variable has n categories, n - 1 dummy variables are defined. If a dummy variable corresponds with a category, its value is set to 1. If not, the value is set to 0. For example, when a categorical variable denotes a programming language either "C" or "Java", a dummy variable "C" is made. If a target project plans to use C, "C" is set to 1. If Java is used, "C" is set to 0. An estimation model using dummy variables is less flexible than a group of estimation models using stratified subsets. In the model, a categorical variable cannot have any influence on diseconomies/economies of scale.

The advantage of transforming dummy variables is that the number of required data points is smaller than the stratification, to build estimation model properly. As a rule of thumb, a linear regression model requires more than five times larger size of data points than the number of independent variables [14]. Suppose that a dataset includes b non-categorical variables and one categorical variable having a categories. The rule of thumb requires 5 ($a + b$ - 1) data points for dummy variables while 5ab data points for the stratification. The difference becomes larger as b becomes larger.

2.3 Interaction

The interaction [1] uses dummy variables so that a regression coefficient varies according to values of the dummy variables. The interaction assumes that a categorical variable itself has no effect to the dependent variable, but the combination of the categorical variable and another variable has the effect. The interaction introduces new variables made by multiplying an independent variable by dummy variables. For instance, when log(Size) denotes functional size and y denotes a dummy variable, the new variable is log(Size)y. The resultant model includes the new variable as follows:

$$\log(\text{Effort})=\beta_s\log(\text{Size})+\beta_1\log(x_1)+\beta_2x_2+\beta_0+\beta_4\log(\text{Size})y+\varepsilon. \quad (5)$$

The equation can be transformed as:

$$\log(\text{Effort})=(\beta_s+\beta_4y)\log(\text{Size})+\beta_1\log(x_1)+\beta_2x_2+\beta_0+\varepsilon. \quad (6)$$

In the equation, the relationship between log(Effort) and Size is determined by $\beta_s + \beta_4y$, and its value varies according to y. Thus, using the interaction, the model expresses diseconomies of scale and economies of scale.

To avoid multicollinearity between a main effect (log(Size)) and an interaction (log(Size)y), the average of main effect is subtracted from each value of main effect before building the model. This procedure is called as centering.

2.4 Hierarchical linear model

The hierarchical linear model (HLM) [3] is used in some research areas such as social science, to analyze the dataset where data points are cohesive with some groups (e.g. countries or schools). HLM builds an estimation model using models based on subsets divided by a categorical variable and the information gained from the whole dataset. HLM makes more flexible model than the dummy variable model.

The HLM considers the errors within categories and the errors between categories and builds models whose intercept and partial regression coefficients are different from each category. The HLM presumes models as shown in Eq. (2) for each category. An intercept and partial regression coefficients on the models are expressed by the following form:

$$\beta_i = \gamma_{i0} + \mu_i. \quad (7)$$

In the equation, γ_{i0} is the average of an intercept or a partial regression coefficient of the models, and μ_i is the errors between categories. ε in Eq. (2) is the errors within categories. Eq. (7) is set to any intercept and partial regression coefficient.

HLM first builds Eq. (2) model for each category using linear regression analysis, and estimates parameters in Eq. (7). After that, based on them, HLM uses empirical Bayes to decide the intercept and partial regression coefficients in Eq. (2) for each category.

3. EXPERIMENT

3.1 Datasets

In the experiment, we build effort estimation models using categorical variable handling methods, and compare estimation accuracy among them for evaluation. We used ISBSG dataset [8], NASA dataset [2], and Desharnais dataset [7]. These datasets recorded categorical variables that have at least three categories, and have relatively many data points.

Dummy variables were made for each categorical variable. We converted a categorical variable with n levels into n-1 binary dummy variables. ISBSG dataset and NASA dataset have multiple categorical variables. We stratified the datasets according to all combinations of values of the categorical variables. For example, when variable A has m categories, and variable B bas n categories, (m - 1) (n - 1) subsets are made at most.

The ISBSG dataset (Release 9) includes 3026 data points (projects) and 99 variables. We selected projects and variables

based on the previous study [9], and excluded projects having a missing value (listwise deletion). We stratified the dataset and removed subsets which did not have at least 10 data points, because we applied 10 fold cross validation. As a result, 558 data points remained. Independent variables are: unadjusted function point and three categorical variables (development type, programming language, and development platform). Fourteen subsets and eight dummy variables were produced for the dataset.

The NASA dataset includes 93 data points. We stratified the dataset and 54 data points remained after removal of small subsets. Independent variables are: lines of code (estimated based on the function points [11]), productivity factors (six-level Likert scale), and three categorical variables (application type, system type, and development type). Three subsets and five dummy variables were produced for the dataset.

The Desharnais dataset includes 81 data points. We removed data points having a missing value, and 77 data points remained. We stratified the dataset, and all subsets remained. Independent variables are: adjusted function point, years of experience of team, years of experience of manager, and one categorical variable (programming language). Three subsets and two dummy variables were produced for the dataset.

3.2 Experimental Setting

As benchmark, we made a baseline estimation model for each dataset. The baseline models do not use any categorical variable. In the baseline estimation models, effort and size measurement were log-transformed. The handling methods were applied to the baseline models.

The interaction and HLM assumes that a relationship between effort and size measurement changes according to a categorical variable. For evaluation of the interaction, we added new variables made by multiplying functional size by dummy variables. For evaluation of HLM, we applied Eq. (7) to a partial regression coefficient of a functional size.

When a dataset had relatively many variables for sample size, we applied a variable selection based on AIC (Akaike's information criterion). The NASA dataset met the criterion in this experiment. HLM could not conduct variable selection because the HLM software we used did not support the function. Accordingly, we also performed the experiment with the NASA dataset removing productivity factors. We call it NASA FP dataset.

To evaluate the estimation accuracy, we used mean *MRE* (Magnitude of Relative Error) [6] (*MMRE*) and mean *BRE* (Balanced Relative Error) [12] (*MBRE*). Although *MRE* is widely used to evaluate effort estimation accuracy, it has biases for evaluating under estimation [4]. So we also adopted *BRE* whose evaluation is not biased, and gave weight to *BRE*. A lower value of each criterion indicates higher accuracy. The criteria were calculated for each treatment method according to 10 fold cross validation. We made training datasets and test datasets where the rate of each category is almost same as whole dataset.

3.3 Results

Table 1 shows estimation accuracy of the baseline models. Figure 1 shows the difference of *MBRE* between a model with a handling method and the corresponding baseline model. Figure 2 shows the difference of *MMRE*. A model with the highest positive difference is the most accurate one. The difference of *MBRE* for the stratification on the NASA dataset was outside the figure for readability. The baseline models resulted in worse *MBRE* on the NASA FP dataset than on the NASA dataset. With the handling methods, the results had higher accuracy on the NASA FP dataset than on the NASA dataset (*MBRE* of HLM was 77.3%, and transforming dummy variables was 100.1%).

Table 1. *MBRE* and *MMRE* without categorical variables

	ISBSG	Desharnais	NASA FP	NASA
MBRE	166.1%	76.0%	150.5%	108.5%
MMRE	112.5%	61.5%	69.9%	78.4%

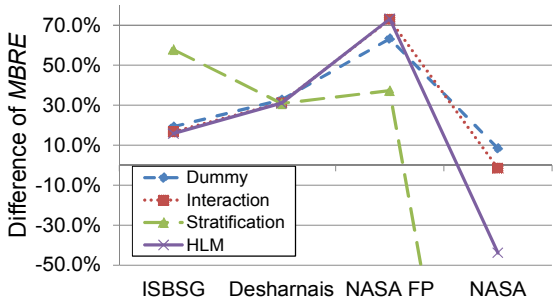

Figure 1. Difference of *MBRE* on each dataset

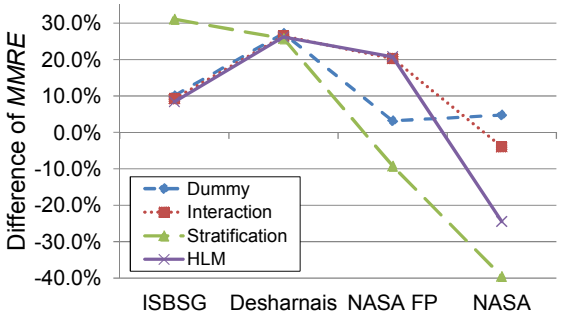

Figure 2. Difference of *MMRE* on each dataset

All but the stratification had similar tendency. The accuracy of HLM on the NASA dataset is not high. This would be because HLM did not conduct variable selection. On the contrary, the accuracy of the stratification was lower than those of the others in the NASA dataset (The difference of *MBRE* was minus 265.6%). Inclusion of many independent variables might cause the result. The stratification needs many data points when a dataset has many independent variables (see section 2.2). The result implies that choosing the stratification may result in lower performance when there are many independent variables. Note that the number of data points of each subset was enough in other datasets.

To answer RQ1 and RQ2, we confirmed the difference of estimation accuracy among the methods with the Friedman test. As a result, in the ISBSG dataset and the NASA FP dataset, the difference was statistically significant on both *MBRE* and *MMRE* (*P*-value was smaller than 0.05). We thus concluded that the answer to RQ1 is "yes" because estimation accuracy is different among the handling methods and that the answer to RQ2 is "no" because there is no handling method whose performance is always the highest or the lowest.

To answer RQ3, we confirmed the difference of estimation accuracy between alternative methods (HLM and the interaction) and common methods (dummy variables and the stratification) with the Wilcoxon signed rank test. The test did not show the statistical difference. However, on the NASA FP dataset, *MBRE* of HLM and the interaction were about 10% higher than transforming dummy variables (*MBRE* of HLM was 77.3%, the interaction was 77.5%, and transforming dummy variables was 87.2%). We think 10% difference is not ignorable. In addition, as shown in Figure 2, the difference of *MMRE* is large (The difference is about 17%). We thus concluded that the answer to RQ3 is "Yes (We should consider the interaction or HLM)."

4. RELATED WORK

There are few researches which used the interaction or HLM in the software engineering research area. However, they did not apply the methods to (ordinary) effort estimation model, and therefore their effects on effort estimation are not clear.

Moses et al. [13] proposed contingency (preliminary effort) estimation model, using hierarchical Bayesian model (Both the model and HLM have basically same mechanism). The model is not ordinary effort estimation model, since the independent variable is estimated effort, and the dependent variable is actual effort. Cataldo et al. [5] analyzed failures in feature-oriented software development using the logistic regression model with the interaction, and this is not effort estimation. Menzies et al. [10] showed the stratification is effective in some subsets. However, they used only one dataset and did not compare the stratification with the interaction and HLM.

5. CONCLUSIONS

In this paper, we set our research goal as giving useful suggestions to enhance the accuracy of an effort estimation model based on the linear regression analysis. We compared the estimation accuracy of linear regression models among four categorical variable handling methods. In our experiment, we applied transforming dummy variables, the stratification, the interaction, and HLM (hierarchical linear model). They were compared on three datasets in its effectiveness on estimation accuracy.

The experimental results showed that the most effective method is different for datasets. The finding suggests that the stratification and transforming dummy variables should be tried at least, for obtaining an accurate model. The experiment also suggested the interaction and HLM may have an effect in some cases. Therefore, the application of them should be considered when building the estimation model. We believe our suggestions are effective for many cases because the linear regression model is widely used for effort estimation. As future work, we will apply the handling methods to other datasets to examine how much the difference of handling methods is on the reliability of the results.

6. REFERENCES

[1] Aiken, L., West, S. 1991. *Multiple Regression: Testing and Interpreting Interactions*. SAGE Publications, Thousand Oaks, CA.

[2] Boetticher, G., Menzies, T., and Ostrand, T. 2007. *PROMISE Repository of empirical software engineering data* http:// promisedata.org/?cat=11, West Virginia University, Department of Computer Science.

[3] Bryk, A., Raudenbush, S. 1992. *Hierarchical Linear Models for Social and Behavioral Research: Applications and Data Analysis Methods*. SAGE Publications, Thousand Oaks, CA.

[4] Burgess, C., and Lefley, M. 2001. Can genetic programming improve software effort estimation? A comparative evaluation. *Journal of Information and Software Technology* 43, 14, 863-873.

[5] Cataldo, M., and Herbsleb, J. 2011. Factors leading to integration failures in global feature-oriented development: an empirical analysis. *In Proc. of the International Conference on Software Engineering (ICSE 2011)*, 161-170.

[6] Conte, S., Dunsmore, H., and Shen, V. 1986. *Software Engineering, Metrics and Models*. Benjamin/Cummings.

[7] Desharnais, J. 1989. *Analyse Statistique de la Productivitie des Projets Informatique a Partie de la Technique des Point des Function*. Master Thesis. University of Montreal.

[8] International Software Benchmarking Standards Group (ISBSG). 2004. *ISBSG Estimating: Benchmarking and research suite*. ISBSG.

[9] Lokan, C., and Mendes, E. 2006. Cross-company and single-company effort models using the ISBSG Database: a further replicated study. *In Proc. of the international symposium on Empirical software engineering (ISESE 2006)*, 75-84.

[10] Menzies, T., Chen, Z., Hihn, J., and Lum., K. 2006. Selecting Best Practices for Effort Estimation. *IEEE Trans. Softw. Eng.* 32, 11 (Nov. 2006), 883-895.

[11] Menzies, T., Port, D., Chen, Z., Hihn, J., and Stukes, S.: Validation methods for calibrating software effort models. *In Proc. of the international conference on Software engineering (ICSE 2005)*, 587-595.

[12] Miyazaki, Y., Terakado, M., Ozaki, K., and Nozaki, H. 1994. Robust Regression for Developing Software Estimation Models. *J. Syst. Softw*. 27, 1 (October 1994), 3-16.

[13] Moses, J., and Farrow, M. 2003. A Procedure for Assessing the Influence of Problem Domain on Effort Estimation Consistency. *Software Quality Control* 11, 4 (November 2003), 283-300.

[14] Tan, H. B., Zhao, Y., and Zhang, H. 2009. Conceptual data model-based software size estimation for information systems. *ACM Trans. Softw. Eng. Methodol*. 19, 2, Article 4 (October 2009), 37 pages.