A New Method of Magnitude Estimation for PDM（Proportional Distribution Method）Using an Optimization Technique and Validation by Monte Carlo Simulation

Etsuji Okamoto*
Eiichi Hata*

We developed a stochastic model. Proportional Distribution Method (PDM) to estimate disease-specific costs in health insurance claims with multiple diagnoses in 1996. PDM assumes a common magnitude for each diagnostic category and distribute the cost of a claim in proportion to the magnitude for each category. We demonstrated previously that, by using arithmetic means of per diem costs of claims containing a certain diagnosis with proper correction as magnitudes, PDM was able to estimate disease-specific costs in computer-generated simulation data which mimics health insurance claims with multiple diagnoses. In this article we proposed a yet another method of magnitude estimation using Excel® Solver function for optimization and refined the established method of arithmetic means with correction by introducing a correction formula for automatic correction. A Monte Carlo simulation using 100 datasets each consisting of 1000 cases with 100 diagnostic categories, which bears little resemblance to actual claims, demonstrated that PDM had achieved near-perfect accuracy by the method using Excel® Solver function and less-perfect but acceptable accuracy by the method of arithmetic means with correction. PDM is also effective in estimating disease-specific days (in and out patient) in health insurance claims but only costs are dealt with in this article.

キーワード Disease classification, Econometrics, Health insurance claims, Optimization, Proportional Distribution Method

1. Introduction

In health economic analysis, it is sometimes necessary to retrospectively estimate costs allocated for each diagnosis out of the total cost incurred on the insured population:

* National Institute of Public Health

a reversal of explanatory and dependent variables. For example, health economists who are analyzing the cost-benefits of preventive medical programs for diabetes need to know how the programs affect the specific costs related to diabetes, and not the overall cost.

Such estimations have typically been conducted by classifying health insurance claims (hereafter, claims) into different categories
by primary diagnoses. However, attributing the entire cost to one diagnosis ignores the cost allocated for other diagnoses. Ideally, all diagnoses contained in claims should be taken into consideration allocating certain amount of resources to each diagnosis in a statistically sound way.

Two hindrances have stood in the way of analyzing claims: a technological hindrance that most claims are submitted in paper form and a methodological hindrance that there is no effective method to allow statistically sound analyses. The technological hindrance may be overcome by manual data input such as seen in model projects in Ehime prefecture (Okamoto and Tabara, 2003) and an epidemiological study in Natori city (Okamoto, 2003). It will eventually be solved by rapid computerization of claims in which all diagnoses as well as clinical procedures will be coded and submitted electronically (Figure 1).

To cope with the methodological hindrance, Okamoto proposed a method to distribute per claim costs in proportion to the “magnitude” assigned to each diagnostic category, which has initially been called “Proportional Disease Magnitude (PDM) method” (Okamoto, 1996). The basic principle behind this method is to hold constant the relative relationship of magnitudes of diagnostic categories to each other. For example, if the magnitude given to stomach cancer is twice that of diabetes, then a claim containing both stomach cancer and diabetes would allocate 2/3 of the cost on stomach cancer and 1/3 on diabetes.

Any numerical value can be used as magnitude depending on the purpose of disease-specific analysis. When Okamoto first proposed the method, he used the likelihood to be a primary diagnosis of each diagnostic category.
obtained from the Patient Survey as magnitudes. Okamoto then used median length of stay as magnitude to evaluate case-mix adjusted length of stay of hospitals (Okamoto, 2001).

Regression coefficients obtained by multivariate analysis (MVA) using the number of diagnoses in each diagnostic category in a claim as independent variables and costs and days as independent variables would appear to provide more mathematically sound estimates. However, MVA, when conducted with numerous independent variables, will almost always produce negative beta coefficients as an artifact of using a linear regression model (Wagner, Chen and Barnett, 2003) and hence not suitable as magnitudes.

In 2003, two methods of magnitude estimation were proposed: Proportional Allot Estimator (PAE) by Tango (Tango, 2003) and arithmetic means with correction (AMC) by the authors. We demonstrated that with AMC, if properly corrected, PDM would yield accurate estimates with computer-generated simulation data (Okamoto and Hata, 2003) and further demonstrated a high concurrent validity between PAE and AMC with actual claims data (Okamoto and Hata, 2004). However, the AMC we proposed was proven valid only in the simulation data with average number of diagnoses in a claim was 3.87 and fell short of generalizability. To achieve generalizability, we need to develop a formula incorporating the average number of diagnoses in claims containing diagnoses of a certain diagnostic category.

In this article, we propose a formula to automate the correction process. We also propose the 3rd method of magnitude estimation: optimization using Excel® Solver and validate both methods with Monte Carlo simulation.

2. Definitions

A health insurance claim contains the following data.

- Cost (expressed in monetary value)
- The number of days (inpatient days for inpatient claims and the number of office visits for outpatient claims)
- Diagnoses (one or more. Diagnoses are classified into diagnostic categories. Note that while many claims contain no diagnoses in many diagnostic categories, some diagnostic categories may contain numerous diagnoses)

The cost and the number of days are regarded as resources allocated for the purpose of treating the diagnoses, which we want to break down into diagnostic categories.

Let individual claims be denoted by \( i \) and diagnostic categories be denoted by \( j \). Additionally total number of claims, total costs of the entire claims, total number of days of the entire claims and total number of diagnoses claimed are denoted \( R, P, D \) and \( N \) respectively.

Then \( P_i, D_i \) and \( N_i \) denote the cost, number of days and number of diagnoses in the \( i \) th claim \((1 \leq i \leq R)\), which are observable. \( N_j \) is the observable total number of diagnoses in the \( j \) th diagnostic category. \( P_j \) and \( D_j \) denote the unobservable disease-specific cost, number of days in the entire group of claims for the \( j \) th diagnosis (given 100 classifications,
I ≤ j ≤ 100), which are to be estimated by PDM.

Likewise, $P_{ij}$, $D_{ij}$ and $N_{ij}$ denote the costs, number of days and number of diagnoses of the $j$th diagnosis in the $i$th claim, respectively (note that majority of $N_{ij}$ is zero). These variables are in the following relationship.

$$P = \sum_{i=1}^{R} P_i = \frac{100}{\sum_{j=1}^{100} P_{ij}}$$
$$D = \sum_{i=1}^{R} D_i = \frac{100}{\sum_{j=1}^{100} D_{ij}}$$
$$N = \sum_{i=1}^{R} N_i = \frac{100}{\sum_{j=1}^{100} N_{ij}}$$

In actual claims data, only $P_i$, $D_i$, $N_i$, $N_{ij}$ and $N_{ij}$ are observable and all other variables such as $P_{ij}$ and $D_{ij}$ are unknown and must be estimated using only $P_i$, $D_i$ and $N_{ij}$.

3. Principles of Proportional Distribution Method (PDM)

The principles of PDM were first described by OKAMOTO (Okamoto, 1996). We describe it here at the risk of repetition.

PDM assumes a common magnitude to each diagnostic category. Let the magnitude of the $j$th diagnostic category for days and cost be $L_j$ and $M_j$ respectively. PDM will estimate $D_{ij}$ and $P_{ij}$ by distributing $D_i$ and $P_i$ in proportion to the magnitudes of diagnoses contained in the given claim:

$$D_{ij} = D_i \times \frac{N_{ij} \times L_j}{\sum_{i=1}^{100} (N_{ij} \times L_j)}$$
$$P_{ij} = P_i \times \frac{N_{ij} \times M_j}{\sum_{i=1}^{100} (N_{ij} \times M_j)}$$

The final stage of the PDM is to aggregate $D_{ij}$ and $P_{ij}$ to yield disease-specific days and costs, $D_j$ and $P_j$.

$$D_j = \sum_{i=1}^{100} (N_{ij} \times D_{ij})$$
$$P_j = \sum_{i=1}^{100} (N_{ij} \times P_{ij})$$

To avoid complexity, we will deal with only estimation of cost, $P_j$, in this article leaving estimation of disease-specific days, $D_j$, for future publication.

4. Estimation of Magnitudes of Diagnostic Categories

How to assign a magnitude for each diagnostic category is crucial in determining the results of PDM. In this article, we propose two methods for magnitude estimation: arithmetic means with correction (AMC) and optimization using Excel® Solver.

(1) Arithmetic Means with Correction Method

This method is similar to what we proposed and validated in our previous publication (Okamoto and Hata, 2003). The refined method we propose here is different from the previous one in two points: 1) the inflationary effects of the number of diagnoses are more properly addressed, 2) correction method was generalized leading to the full automation of correction.

a. Inflationary Effects of $D_i$ and $N_i$ on $P_i$

Cost of a claim ($P_i$) is dependent upon not only diagnoses contained but also the number of days ($D_i$) and the number of diagnoses ($N_i$). By assuming a theoretical unit cost per day per diagnosis of the $i$th claim, $U_i$, we propose the following model to illustrate their relationship.

$$P_i = D_i \times (1 + \ln(N_i)) \times U_i$$

The formula (6) implies that the per claim cost increases proportionally to the number of days and logarithmically to the number of dia-
gnoses contained.

The unit cost, \( U_i \) is expressed as follows and can be viewed as unit cost devoid of inflationary effects of \( D_i \) and \( N_i \).

\[
U_i = \frac{P_i}{D_i \times (1 + \ln(N_i))}
\]  

(7)

Arithmetic means of \( U_i \) for claims containing diagnoses in the \( j \)th diagnostic category, \( \bar{U}_j \) can be calculated using the following formula.

\[
\bar{U}_j = \frac{\sum_{i=1}^{R} (N_{ij} \times U_i)}{\sum_{i=1}^{R} N_{ij}}
\]  

(8)

b. Dilutionary Effects of the Number of Diagnoses on \( \bar{M}_j \)

Here, we have to consider the cost dilutionary effects of the average number of diagnoses in claims containing diagnoses in the \( j \)th diagnostic category. \( \bar{M}_j \) calculated in the formula (8) may be used as magnitude for a diagnostic category whose diagnoses always appear as a single diagnosis. However, for majority of diagnostic categories, \( \bar{M}_j \) calculated in the formula (8) is “dilated” by the presence of other diagnoses contained in the same claim.

Let the overall average of \( U_i \) be denoted as \( \bar{M} \), which is expressed as :

\[
\bar{M} = \frac{\sum_{i=1}^{R} (N_i \times U_i)}{\sum_{i=1}^{R} N_i}
\]  

(9)

And let the average number of diagnoses in claims with the \( j \)th diagnostic category be denoted as \( \bar{N}_j \), which is expressed as :

\[
\bar{N}_j = \frac{\sum_{i=1}^{R} (N_{ij} \times N_{ij})}{\sum_{i=1}^{R} N_{ij}}
\]  

(10)

If the magnitude of a diagnostic category is elevated by \( \Delta M \) than overall average \( \bar{M}(= \bar{M} + \Delta M) \), and the average number of diagnoses contained in the claims is \( n \), then the arithmetic mean (\( \bar{M}_j \)) will equal to \( \bar{M} + \Delta M/n \).

To calculate \( \bar{M} + \Delta M \) from \( \bar{M} + \Delta M/n \) is impossible but we propose the following correction formula to make best estimates to yield the magnitude, \( M \).

**CORRECTION FORMULA**

Correction formula for \( \bar{M}_j \) to yield the magnitude of the \( j \)th diagnostic category, \( M_j \)

\[ M_j = \bar{M}_j \times \left( \frac{\bar{M}_j}{\bar{M}} \right)^c \]  

(11)

i) if \( \bar{M}_j \geq \bar{M} \) \( C = \ln(\bar{N}_j) \)

ii) if \( \bar{M}_j < \bar{M} \) \( C = \bar{N}_j - 1 \)

(2) Optimization Method Using Excel® Solver

Excel® Solver is an add-in software included in Microsoft Excel® program. It is a program mainly to be used for linear programming. Its function is basically to change the values in the “changing cells” gradually to minimize (or maximize or converge to a certain value) the value in the “target cell”. When the value in the “target cell” is converged, i.e., does not change any more, the Solver stops its repetition. It can handle up to 200 cells as “changing cells”.

\( M_j \) can be obtained using Excel® Solver to minimize the deviance between the observed unit cost of the \( i \)th claim (\( U_i \)) and estimated unit cost of the \( i \)th claim, \( M_i \).

\[ M_i = \frac{100 \times \sum_{i=1}^{R} (N_{ij} \times M_{ij})}{100 \times \sum_{i=1}^{R} N_{ij}} \]  

(12)

The deviance is expressed as a sum of square formula below.
\[ \sum_{i=1}^{R} (U_i - M_i)^2 \] 

The \( M_j \) estimated to minimize the above formula \( \Phi \) is used as magnitudes for the PDM. To do this, set the cell containing the formula \( \Phi \) as the target cell to be minimized and assign the cells containing \( M_j \) (100 cells) as changing cells in the Excel® Solver.

5. Validation Using Simulation Data

Accuracy of the PDM was validated using simulation data consisting of 1000 cases with 100 diagnostic categories. Estimates of disease-specific costs by the PDM were compared with “right answers” of simulation data to validate the accuracy of the PDM. The procedures were repeated 100 times (Monte Carlo simulation).

(1) Generation of Simulation Data

A set of simulation data was generated by the following procedures. All procedures were conducted using Microsoft Excel® 2003 and actual Excel® functions are shown in box. The conditions for generation were set rather arbitrarily because conditions of generating right answers are independent of the validation comparing the right answers and the estimated values.

a. Setting tentative \( \hat{M}_j \) and \( N_j \) (\( 1 \leq j \leq 100 \))

\( \hat{M}_j \) were tentatively generated to follow normal distribution with average of 100 and standard deviation of 100 using random number with non-negative constraints.

\[ \hat{M}_j = \text{NORMINV} (\text{NORMDIST} (0, 100, 100, 1) + \text{RAND} () * (1 - \text{NORMDIST} (0, 100, 100, 1)), 100, 100) \]

\( N_j \) was tentatively generated to follow uniform distribution ranging from 1 to 100. Note that these values for \( M_j \) and \( N_j \) are tentative, to be replaced by actual values generated.

\[
N_j = \text{RANDBETWEEN} (1, 100)
\]

b. Setting \( D_i \) (\( 1 \leq i \leq 100 \))

\( D_i \) was generated to follow normal distribution with average of 5 days and standard deviation of 5 days using random number with integer constraints.

\[
D_i = \text{ROUNDUP} (\text{NORMINV} (\text{NORMDIST} (0, 5, 5, 1) + \text{RAND} () * (1 - \text{NORMDIST} (0, 5, 5, 1)), 5, 5), 0)
\]

C. Generation of \( N_{ij} \) (\( 1 \leq i \leq 1000 \), \( 1 \leq j \leq 100 \))

\( N_{ij} \) was generated in a random manner so that the \( \Sigma N_{ij} \) approximates the tentative \( N_j \) (for the sake of convenience, \( N_{ij} \) was set 0 or 1).

\[
N_{ij} = \text{IF} (\text{RAND} () < N_j / 1000, 1, 0)
\]

A total of 100,000 cells were filled.

d. Generation of \( P_{ij} \) (\( 1 \leq i \leq 1000 \), \( 1 \leq j \leq 100 \))

\( P_{ij} \) was generated to follow normal distribution so that the average approximates the tentative \( \hat{M}_j \) and coefficient of variance of 0.3 (i.e. \( M_j * 0.3 \)) with non-negative constraints.

\[
P_{ij} = \text{NORMINV} (\text{NORMDIST} (0, \hat{M}_j, \hat{M}_j * 0.3, 1) + \text{RAND} () * (1 - \text{NORMDIST} (0, \hat{M}_j, \hat{M}_j * 0.3)), \hat{M}_j, \hat{M}_j * 0.3)
\]

A total of 100,000 cells were filled in a separate worksheet. All cells containing RAND () functions were converted to values to avoid recalculation.
e. Calculation of “Right Answers” of $\hat{M}_j$ and $P_j$

The “right answers” of $\hat{M}_j$ and $P_j$, against which the PDM estimates are to be validated, were calculated using the following formula (contrast the following formula with formula (8) used in real claims data for which $P_{ij}$ and $D_{ij}$ are unknown):

\[
\hat{M}_j = \frac{1000 \sum_{i=1}^{1000} (N_{ij} \times P_{ij})}{\sum_{i=1}^{1000} N_{ij}} \tag{46}
\]

\[
P_j = \frac{1000}{\sum_{i=1}^{1000} \left( D_i \times P_{ij} \times \frac{1 + \ln(N_{ij})}{N_i} \right)} \tag{55}
\]

\[
\hat{M}_j = \text{SUM} \left( N_{ij} \times P_{ij} \right) / \text{SUM} \left( N_{ij} \right)
\]

\[
P_j = \text{SUMPRODUCT} \left( D_i, P_{ij}, (1 + \ln(N_{ij})) / N_i \right)
\]

\[
\hat{M}_j = \text{SUMPRODUCT} \left( N_i, N_{ij} \right) / \text{SUM} \left( N_{ij} \right)
\]

\[
\hat{M}_j = \text{SUMPRODUCT} \left( P_{ij}, (1 + \ln(N_{ij})) / N_i \right) / \text{SUM} \left( N_{ij} \right)
\]

\[
\hat{M}_j = \text{SUM} \left( P_i \right) / \text{SUMPRODUCT} \left( D_i, 1 + \ln(N_{ij}) \right)
\]

\[
\hat{M}_j = \text{SUMPRODUCT} \left( M_i, 1 - P_i \right) / \text{SUM} \left( M_i \right)
\]

\[
\hat{M}_j \times (\hat{M}_j / \hat{P}) \times \ln(\hat{N}_j), \hat{M}_j \times (\hat{M}_j / \hat{P}) \times (\hat{N}_j - 1)
\]

\[
\text{target cell} = \text{SUMXMY2} \left( U_i, M_i \right)
\]

\[
P_i = \text{SUMPRODUCT} \left( M_i / \hat{P}, 1 - 1000 \times \ln(M_i / \hat{P}) \right)
\]

\[
i. Aggregation of $P_{ij}$ to Yield $P_j$

$P_j$ was calculated by aggregating $P_{ij}$ using formula (5).

(2) Monte Carlo Simulation

The procedures in (1a)–(i) were repeated to generate 100 sets of regression data. $M_j$ estimated by two different methods (AMC and the optimization method) and $P_j$ estimated by the PDM with respective magnitudes were compared with the “right answers” of $M_j$ and $P_j$ by drawing regression lines.

Four regression lines were drawn for combinations of $M_j$ and $P_j$ each estimated by the two different methods in a dataset making the total number of regression lines 400. The PDM estimates were validated by slopes of the regression line for accuracy and $R$ square values ($R^2$) for precision, both of which should be as close to one as possible.

6. Results

Averages and standard deviations of 100 sets of slopes and $R$ square values obtained by the Monte Carlo simulation are presented in Table 1 in two different methods of magnitude estimation.

PDM estimated disease-specific costs with near perfect accuracy evidenced by high slope (0.995) and $R^2$ value (0.984) of regression lines between estimated disease-specific costs and “right answers” of simulation data with the optimization method. PDM estimates with AMC were less accurate (slope was 0.865) but with good precision ($R^2 = 0.958$).

What was noteworthy was that accuracy and precision were improved from the original
magnitude estimation to the final disease-specific costs. In both methods of magnitude estimation, estimation of magnitudes was less accurate and precise, but the final estimates of disease-specific costs showed improved accuracy and precision: with AMC, the slope improved from 0.7 to 0.865 and $R^2$ from 0.799 to 0.958; with the optimization method, the slope improved from 0.974 to 0.995 and $R^2$ from 0.908 to 0.984.

This suggests the robust tendency of PDM: even if the original magnitudes are not so accurate and precise, PDM eventually gets to the more accurate and precise results in many times, though not always (The authors did encounter a case when $R^2$ values deteriorated from magnitude estimation to final estimation of disease-specific costs. The authors’ impression is that such cases are rare, but how often such reversals occur warrants further investigation).

7. Discussion

PDM (originally called Proportional Disease Magnitude method) was developed to a Windows-compatible computer program and is now distributed as a freeware. The magnitude estimation method of AMC validated in this article is incorporated into its version 3. We could demonstrate in this article that yet another method of magnitude estimation using Excel® Solver is likely to yield even more accurate and precise estimates.

However we emphasize that it is premature to jump onto this new method of magnitude estimation. In the first place, it is subject to limitations of Excel® and its add-in software, Solver. Excel® can handle up to 65,536 data and Solver can handle up to 200 changing cells. The capacity of Solver is enough to accommodate the typical 119 diagnostic categories commonly used for classification of health insurance claims in Japan. Still, these limiting factors make this method of magnitude estimation unsuitable when it has to analyze a large size claims data which numbers more than a billion a year particularly after full computerization demand quick and speedy analysis.

Another reason is that the optimization method will inevitably yield numerous zero magnitudes and resultant zero disease-specific costs. Given the nature of claims, it is hard to accept that many diagnoses do not cost at all.
Given these limitations and acceptability as well as the speed of data processing, the AMC will continue to be a practical and acceptable method for magnitude estimation of PDM in the foreseeable future.

Sponsors of Research

This study was supported by Grant-in-Aid for Scientific Research (C) “Development of data mining technique to detect drug adverse events from health insurance claims (theme No. 14570373, Principal investigator: Etsuji Okamoto)”

References


(2004年7月20日受付, 2004年10月12日採用)

Correspondence：岡本啓司（Etsuji Okamoto）
atoz@nih.go.jp
PDM（比例配分）法のための重み推計の新手法
とモンテカルロシミュレーションによる検証

岡本 悦司*
畑 榊一 *

複数病名の記載されることの多いレセプトの日数と医療費の合計に占める傷病別日数と傷病別医療費を推計する Proportional Distribution Method（比例配分法。岡本が1996年に発表した時は Proportional Disease Magnitude、傷病マグニチュード按分法と呼んだがこの名称の方がわかりやすいため改称した）法について妥当性と精度の改善を試み、モンテカルロ法によって検証した。

PDM法とは、傷病分類ごとに共通の重み（マグニチュード）を想定し、各レセプトの日数と点数を重みに比例して配分し、最終的に傷病ごとの日数と点数を合計し傷病別医療費を推計する。そのため重み推計法が重要となるが、今回は従来からある平均値補正法に加えて新しい Excel® ソルバーを用いた最適化法の 2 つの重み推計法を試みた。平均値補正法では、傷病数による補正法をより一般化する補正式を考案し導入した。なお日数の分析は点数とは異なった原理が必要なので別稿にゆずる。

コンピュータで生成したシミュレーションデータによる検証結果は、傷病別医療費の推計値と正解との回帰直線の傾きで妥当性を、決定係数で精度をそれぞれ評価した。その結果、最適化法による重みを用いた場合はほぼ完全な推計となり、平均値補正法もそれに近い良好な推計値が得られた。また重み推計値の妥当性と精度は低くてもその重みを PDM 法にかけると最終的な傷病別医療費の推計値の妥当性と精度は向上し、PDM 法原理の有する頑健な傾向を示唆した。

最適化法は優れているが制約も多く、レセプトが電子化され迅速かつ大量の傷病分析が要求されると対応できない。平均値補正法は最適化法よりもやや劣るとはいえ十分かっ許容できる妥当性と精度を有している。

キーワード：レセプト，最適化，PDM 法，エコノメトリクス，傷病分析

* 国立保健医療科学院