Issues and Measures for Predictive Modeling of Medical Cost

Naonori Yakura[1)], Kosuke Iwasaki[2)], Yasuo Kurosawa[3)]

[1)]*Sompo Japan Insurance Inc.*

[2)]*Milliman, Inc.*

[3)]*Healthcare Frontier Japan Inc.*

Abstract

This paper discusses the issues that every researcher faces when he/she tries to model the medical cost using regression analysis and some possible measures for them. The main topics are the separation of frequency distribution and amount distribution, the systematical underestimation problem of log-transformation, variable transformation of U/V shaped variables, dummy variable, and the simulation for the distribution of predicted values.

Key words: predictive modeling, multiple linear regression, logistic regression, log-transformation, simulation

❖ Introduction

We tried to develop a predictive model of medical cost using the past year's medical cost and vital data as independent variables in addition to age and gender. Based on that experience, this paper discusses the issues that every researcher faces when he/she tries to model the medical cost using regression analysis and some possible measures for them.

❖ Issues and Measures

Features of medical cost distribution

The usual features of medical cost distribution are supposed to be the existence of many zero-cost people and the long tailed shape to the right-hand. We tried to analyze the medical cost data of a sample population consisted of more than 40,000 Japanese aged from 19 to 59 years old. About 14% of the population did not use any medical cost during the year 2002. Among the people who used any of medical cost during that year, the minimum amount for one person for one year is YEN 380 and the maximum amount is YEN 16,953,300. The simple average amount per person is around YEN 114,620. Based on these facts, the normal distribution is not considered to fit the medical cost distribution.

Frequency and amount

In order to handle the zero-cost density, one of the general strategies is to separately analyze the frequency distribution and the amount distribution. It is the very usual way we adapt in the field of property and casualty / general / non-life insurance, where the expected claim cost might be calculated based on the claim frequency ratio and the average amount of a claim. When modeling the medical cost, it is supposed to be one of the possible strategies to separately analyze the probability of the medical cost for one year being non-zero by a logistic regression and the probability distribution of annual medical cost amount for the people who use any of medical cost by a multiple linear regression.

Received: February 21, 2008
Accepted: March 28, 2008

Correspondence: N. Yakura, Sompo Japan Insurance Inc., Healthcare Business Development Dept., Shinjukunomura Bldg. 10F, 26-2, Nishi-Shinjuku 1 chome, Shinjuku-ku, Tokyo 163-0510, Japan
e-mail: NYakura@sompo-japan.co.jp

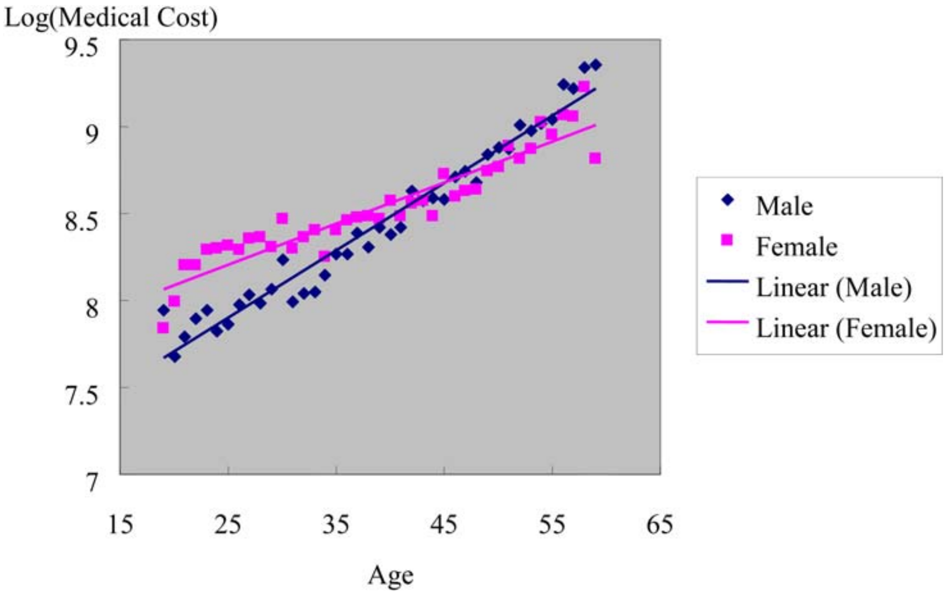

Figure 1. Sample of average Log(Medical Cost) by Age

Source: The sample population of our study (Employees of a Japanese manufacturing company, n = 43,603, medical cost = annual sum of medical cost incurred in the year of 2003).

Log-transformation

When the distribution is long tailed to the right-hand and not supposed to fit the normal distribution, variable transformation is to be considered. The log-transformation is one of the common ways, but we have to be careful to use it for predicting the original variable.

In the sample population of our study, the distribution of the annual medical cost amount by person has a very long right-hand tail as mentioned in Section 2.1. We introduced a log-transformed value of medical cost (i.e. log(y),hereinafter denoted as log y) as a new variable. The average of log y by age increases as the age is going up and is well fitted to the straight lines in both male and female as shown in Figure 1. Therefore, in our study, we use log y as a dependent / explained variable for the regression analysis instead of the original medical cost (y).

Though the log-transformation is used so often for the dependent variable in many areas, we have to pay a careful attention to the issue of predicting y when log y is the dependent variable. Since the exponential undoes the log, the simplest guess for predicting y is to exponentiate the fitted value of log y. But this does not work. It will systematically underestimate the expected value of y.

A general expression of a multiple linear regression model for log y is as shown in (2.1) below.

$$\log y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \qquad (2.1)$$

where x_i is an independent / explaining variable and u is a random error term.

Given the least square estimators $\hat{\beta}_i$, the fitted value of log y (herein after denoted as $\widehat{\log}\, y$) for any value of the independent variables as shown in (2.2) below.

$$\widehat{\log}\, y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k \qquad (2.2)$$

When we assume as usual that u is normally distributed with the mean = 0 and the variance = σ^2, the expected value of y for a given set of independent variables is as shown below. Please refer to the Appendix A if necessary.

$$E(y \mid x) = \exp\left(\frac{\sigma^2}{2}\right) \cdot \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$$

Therefore the expected value of y should be estimated as

$$\hat{y} = \exp\left(\frac{\hat{\sigma}^2}{2}\right) \cdot \exp\left(\widehat{\log}\, y\right) > \exp\left(\widehat{\log}\, y\right)$$

where $\hat{\sigma}^2$ is the unbiased estimator of σ^2.

Because $\hat{\sigma}^2$ is larger than 0, the exponential of $\hat{\sigma}^2$ divided by 2 is larger than 1. That is, predicting y by exponentiating $\widehat{\log}\, y$ will systematically underestimate the expected value of y and a simple adjustment is needed to be considered in predicting y.

Other variable transformation

In our study, we tried to use vital data like body mass index (BMI), blood pressure, blood sugar, etc. as independent variables. When we plotted vital data and medical cost, some of the lines became U shaped. This means that firstly medical cost decreases as BMI, for example, goes up and after that medical cost turns to increase. However the multiple linear regression model like (2.1) assumes the monotonic feature of the variables. When $\beta_1 > 0$, $\log y$ monotonically increase as x_1 goes up. When $\beta_1 < 0$, $\log y$ monotonically decreases as x_1 goes up. So we can not include the U shaped ones as independent variables into the linear model. One of the way to handle U/V shaped variables is to use the absolute value of the difference between the original value of the variables and the value of the variables which gives the lowest medical cost (the bottom of U/V shaped curve). The candidates of independent variables should be carefully reviewed and the appropriate variable transformation should be considered to satisfy the monotonicity of the independent variables assumed for the model.

Dummy variable

Another topic to be considered might be dummy variable. Gender must be a good example. It is a common way to use a dummy variable D where $D = 1$ when the person is a male and $D = 0$ when the person is a female.

When the gender is supposed to have an influence on the dependent variable, one of the ways is to add D into the model as an independent variable. Suppose that the regression coefficient for D is β_d.

$$\log y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \beta_d D + u$$

This means:

for male:

$$\log y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \beta_d + u$$

for female:

$$\log y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

that is, the difference of medical cost between male and female is always constant β_d when he and she have the same set of values of x_i whatever the values are. For example, the difference of medical cost between male and female at age 30 is to be the same as the difference between male and female at other age. When the data do not have this kind of parallelism feature, adding D into the model as an independent variable is not considered to be appropriate.

Suppose that the difference between male and female is not constant, but there is no difference at age 19 and the increase for male from age 19 is steeper than that for female. In the case like this, $(age - 19) \cdot D$ might be added to the model. When it is very difficult to formulate, we can split the data by gender and examine regression models separately. We should carefully select the way to handle the factors.

Distribution of predicted values

As mentioned in section 2.3, given the least square estimators $\hat{\beta}_i$, the fitted value of $\log y$ ($\hat{\log} y$) for any value of the independent variables as shown in (2.2). Based on the assumptions of the regression model, $\log y$ is a random/stochastic variable as shown in (2.1), and $\hat{\log} y$, the fitted result of (2.2) for a set of values of the independent variables, is the expected value of random variable $\log y$ under the set of values of independent variables and we can also estimate the variance of $\log y$. For the purpose of predicting future medical cost, the expected value is, needless to say, very important and we believe that the variance is also very important and should be taken into consideration for the probability distribution of predicted values.

When we assume that the random error term u has the normal distribution with mean = 0 and variance = σ^2, we can estimate the variance of $\hat{\log} y$ for a given set of x_i values which is called the standard error of the mean predicted value, and also the variance of $\log y$ for a given set of x_i values (hereinafter denoted as $Var[\log y]$) which is called the standard error of the individual predicted value (Please refer to the Appendix B if necessary.). These estimates are easily calculated by any modern statistical software package like SAS.

One of the possible strategies to estimate the distribution of the predicted value is a simulation. Using the fitted value $\hat{\log} y$, the standard error of the individual predicted value $Var[\log y]$ and the result of a normal random number (hereinafter denoted as RN), we can get a result of the first trial of the prediction for one person as $\hat{\log} y + \mathrm{RN} * Var[\log y]$. When we repeat the trial, for example, 10,000 times, we can get the empirical distribution of the predicted value for that person. Then we can consider not only the average of the predicted value for that person but also the upper 5% value and/or the lower 5% value.

As mentioned in section 2.2, we separately examined the probability of the medical cost for one year being non-zero by logistic regression and the probability distribution of annual medical cost amount for the people who use any of medical cost by multiple linear regression. Therefore for each trial of the prediction of future medical cost, we first determine whether the medical cost is zero or not for that specific trial, and in the case that the medical cost for that specific trial is determined not to be zero as a result of the first step, then we determine how much of the amount based on the methodology mentioned above.

The step of determining whether the medical cost is zero or not for that specific trial can be performed as follows. The typical logistic regression model is as shown in (2.3).

$$\log\frac{p}{1-p}=\gamma_0+\gamma_1x_1+\gamma_2x_2+\ldots+\gamma_kx_k+u \qquad (2.3)$$

where p is the probability of non-zero medical cost.

Given the maximum likelihood estimators $\hat{\gamma}_i$, we can get the fitted value of $\log\frac{p}{1-p}$ for any value of the independent variables and the standard error of the individual predicted value. Following the same methodology stated in the third paragraph of this section, by getting the result of a normal random number, we can get the prediction of $\log\frac{p}{1-p}$ for each trial and a simple algebra calculation gives us the predicted value of p (hereinafter denoted as $\overline{P}$) for that specific trial. One of the way to determine whether a non-zero cost case is occurred or not for that specific trial with the above predicted value of p is to introduce a uniform random number and compare the result of it (hereinafter denoted as RU) and $\overline{P}$. Suppose that $z = 1$ when RU $\leq \overline{P}$ and Z = 0 when RU $> \overline{P}$. Then z is supposed to be the result of a random number following a Bernoulli trial with the probability $\overline{P}$.

Repeating the above two-step trial, for example, 10,000 times, we can get the empirical distribution of the predicted value for that person. Also when we repeat this 10,000-time trial for each person in the population, we can get the average and the distribution of the total (or per person average) medical cost prediction for the entire population.

❖ Discussion

Though the electronic data transmission of medical claims has recently been growing, most of the medical claims are still transmitted on a paper basis in Japan. So the available data items are limited, as the public health insurers usually input only the data items which are necessary for their accounting purpose and/or calculation purpose of payable benefit amount. Mainly due to this limitation of the available data, though the predictive model we developed gives a fairly good prediction on a population aggregate basis, the accuracy of the individual prediction for each person is considered to have a room to be improved for practical uses in the identification and stratification of the population for disease management programs. The Japanese government has a scheduled plan to expand the on-line transmission of medical claim data toward the year 2011. The accuracy of the model is expected to be improved as the data availability is extended.

❖ References

1) Dupont W: Statistical Modeling for Biomedical Researchers. Cambridge: Cambridge University Press, 2003: 34–107.
2) Nawata K: TSP NIYORU KEIRYOUKEIZAI-BUNSEKI NYUUMON (in Japanese): Tokyo: Asakura Publishing Co., Ltd, 2006: 30–76.
3) Wooldridge J: Introductory Econometrics: A Modern Approach: Cincinnati, Ohio: South-Western College Pub., 2002: 218–220.

❖ Appendix A

Derivation of and $E(\exp(u))$ and $E(y|x)$

Assume that $u \sim N(0, \sigma^2)$, the normal distribution with mean = 0 and variance = σ^2.

$$E(\exp(u)) = \int e^{u} \cdot f(u) du = \int e^{u} \cdot \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{u^2}{2\sigma^2}} du = \int \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{u^2}{2\sigma^2}+u} du$$

$$= \int \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{u^2-2u\sigma^2+\sigma^4}{2\sigma^2}+\frac{\sigma^2}{2}} du = e^{\frac{\sigma^2}{2}} \int \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(u-\sigma^2)^2}{2\sigma^2}} du = e^{\frac{\sigma^2}{2}} = \exp\left(\frac{\sigma^2}{2}\right)$$

From (2.1)

$\log y = \log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$

$\exp(\log(y)) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u)$

$\therefore\ y = \exp(u) \cdot \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$

$\therefore\ E(y|x) = E(\exp(u)) \cdot \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$

$= \exp\left(\frac{\sigma^2}{2}\right) \cdot \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$

❖ Appendix B

Derivation of $Var[y|x]$

Please note that the explanation below is based on a simple linear regression model for the simplicity of the notation.

The model to be considered:

$y = \alpha + \beta \cdot x + u$

where y: dependent variable
x: independent variable
α, β: unknown parameters of the population
u: random error term with mean = 0 and variance = σ^2

Suppose that a and b are the least square estimators of α and β respectively.
When we have samples (x_i, y_i) $i = 1, \ldots, n$ from the population, a and b are easily calculated as follows:

$$b = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \quad \text{(B.1)}, \qquad a = \bar{y} - b \cdot \bar{x} \quad \text{(B.2)}$$

where y_i: the value of the dependent variable for the ith person
x_i: the value of the independent variable for the ith person
n: size of the sample population

$$\bar{y} \equiv \frac{\sum y_i}{n}, \qquad \bar{x} \equiv \frac{\sum x_i}{n} \qquad \therefore \sum (x_i - \bar{x}) = 0 \qquad \text{(B.3)}$$

Step1: Transforming b expression including u_i

For each sample $i = 1, \cdots, n$,

$y_i = \alpha + \beta \cdot x_i + u_i$

$$\therefore \sum y_i = \sum(\alpha + \beta \cdot x_i + u_i) = n\alpha + \beta\sum x_i + \sum u_i$$

divide both side by n

$$\bar{y} = \alpha + \beta \cdot \bar{x} + \frac{\sum(u_i)}{n} \qquad \text{(B.4)}$$

$$\therefore \alpha = \bar{y} - \beta \cdot \bar{x} - \frac{\sum(u_i)}{n}$$

$$\therefore y_i = \bar{y} + \beta(x_i - \bar{x}) + u_i - \frac{\sum(u_i)}{n}$$

$$\therefore y_i - \bar{y} = \beta(x_i - \bar{x}) + u_i - \frac{\sum(u_i)}{n} \qquad \text{(B.5)}$$

Substitute $(y_i - \bar{y})$ in (B.1) by (B.5), and consider (B.3)

$$b = \frac{\sum\left(\beta(x_i - \bar{x}) + u_i - \frac{\sum(u_i)}{n}\right)(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$$

$$= \frac{\beta\sum(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2} + \frac{\sum(x_i - \bar{x}) \cdot u_i}{\sum(x_i - \bar{x})^2} + \frac{(\sum u_i)(\sum(x_i - \bar{x}))}{\sum(x_i - \bar{x})^2}$$

$$= \beta + \frac{\sum(x_i - \bar{x}) \cdot u_i}{\sum(x_i - \bar{x})^2} \qquad \text{(B.6)}$$

Step 2: Derivation of $E(b)$, $Var[b]$, $E(a)$, $Var[a]$

Assuming that u_i are mutually independent,

$$E(b) = E\left(\beta + \frac{\sum(x_i - \bar{x}) \cdot u_i}{\sum(x_i - \bar{x})^2}\right) = E(\beta) + \frac{\sum(x_i - \bar{x}) \cdot E(u_i)}{\sum(x_i - \bar{x})^2} = \beta$$

$$Var[b] = Var\left[\beta + \frac{\sum(x_i - \bar{x}) \cdot u_i}{\sum(x_i - \bar{x})^2}\right]$$

$$= \frac{1}{\left\{\sum(x_i - \bar{x})^2\right\}^2} \cdot \left\{(x_1 - \bar{x})^2 \cdot Var(u_1) + \cdots + (x_n - \bar{x})^2 \cdot Var(u_n)\right\}$$

$$= \frac{\sigma^2 \sum(x_i - \bar{x})^2}{\left\{\sum(x_i - \bar{x})^2\right\}^2} = \frac{\sigma^2}{\left\{\sum(x_i - \bar{x})^2\right\}}$$

When we assume that u_i are mutually independent and normally distributed, b is a sum of random variables which are mutually independent and normally distributed as shown in (B.6). Since the normal distribution has the reproductive property, b has a normal distribution.

$$\therefore b \approx N\left(\beta, \frac{\sigma^2}{\left\{\sum\left(x_i - \bar{x}\right)^2\right\}}\right)$$

Similarly, we can easily get

$E(a) = \alpha$

$$Var(a) = \frac{\sigma^2}{n} + (\bar{x})^2 \cdot Var[b]$$

Step 3: Derivation of $Var[y|x]$

The fitted value of y ($\hat{y}$) given x is $\hat{y} = a + b \cdot x$.

$\therefore E(\hat{y}|x) = E(a + b \cdot x|x) = E(a) + x \cdot E(b) = \alpha + \beta \cdot x$

In order to get a expression including u_i, substitute $\bar{y}$ in (B.2) by (B.4)

$$a = \alpha + \beta \cdot \bar{x} + \frac{\sum(u_i)}{n} - b \cdot \bar{x}$$

$$\therefore a + b \cdot x = \alpha + \beta \cdot \bar{x} + \frac{\sum(u_i)}{n} - b \cdot \bar{x} + b \cdot x$$

$$Var[\hat{y} \mid x] = Var[a + b \cdot x \mid x] = Var\left[\alpha + \beta \cdot \bar{x} + \frac{\sum(u_i)}{n} - b \cdot \bar{x} + b \cdot x\right]$$

$$= \frac{\sum Var[u_i]}{n^2} + \left(x - \bar{x}\right)^2 Var[b] = \frac{n \cdot \sigma^2}{n^2} + \left(x - \bar{x}\right)^2 Var[b] = \frac{\sigma^2}{n} + \left(x - \bar{x}\right)^2 Var[b]$$

As α, β are estimated by a, b, $y = \alpha + \beta \cdot x + u$ is estimated by $a + b \cdot x + u = \hat{y} + u$ and therefore, $Var[y|x]$ is estimated by $Var[\hat{y} + u|x]$.

$\therefore Var[y|x] = Var[\hat{y} + u|x] = Var[\hat{y}|x] + Var[u] = Var[\hat{y} \mid x] + \sigma^2$

An unbiased estimate of σ^2 is $\hat{\sigma}^2 = \dfrac{\sum\left(y_i - \left(a + b \cdot x_i\right)\right)^2}{n - 2}$,

Then $Var[y|x]$ can be estimated by $\hat{\sigma}^2 \cdot \left\{1 + \dfrac{1}{n} + \dfrac{\left(x - \bar{x}\right)^2}{\sum\left(x_i - \bar{x}\right)^2}\right\}$.