An Approximation of Skewed Healthcare Expenditure Distribution Using a Mixture Model

William W. Yu, Agency for Healthcare Research and Quality
540 Gaither Road, Rockville, MD 20850-6649

Key Words: Healthcare expenditures, skewness, mixture models, MEPS.

1. Introduction

The Medical Expenditure Panel Survey (MEPS) is designed to provide nationally representative annual estimates of health care use, expenditures, sources of payment, and insurance coverage for the U.S. civilian noninstitutionalized population. It is co-sponsored by the Agency for Healthcare Research and Quality (AHRQ) and the National Center for Health Statistics (NCHS).

The expenditure data from the MEPS have been shown to exhibit a marked positive skewness, characterized with a few high expenditure respondents and many zero expenditure respondents. Any approximation of the distribution of MEPS expenditure will need to capture the bi-modality feature of the MEPS expenditure data (i.e., the positive expenditures and the “zeros.”) A mixture model with a point mass at zero and the positive half of the real line was used to approximate the mobile communications expenditures distribution function (Yoo S, 2004). It will be applied, in this study, to approximate the distribution function for MEPS healthcare expenditures.

This mixture model captures the bi-modality feature of the MEPS expenditure distribution. When covariates were added to the model, it was found that the probability that a person has zero-expenditure significantly varies with some variables. The positive values of MEPS expenditures were assumed to follow one of the Weibull, Gamma, or Log-normal distributions. The mixture model was estimated with the maximum likelihood estimation method and evaluated with Kolmogorov-Smirnov test for goodness-of-fit.

2. MEPS Household Component

The core component for MEPS is the Household Component (HC). The MEPS-HC collects data through an overlapping panel design and data are collected through a series of five rounds of interviews over a period of two and a half years. Interviews are conducted with one member of each family who reports on the healthcare experiences of the entire family. Two calendar years of medical expenditure and utilization data are collected for each household and captured using computer-assisted personal interviews. This series of data collection rounds is launched again each subsequent year on a new sample of households to provide overlapping samples of survey data that provide continuous and current estimates of health care expenditures (Cohen JW, 1997).

The sampling frame for the MEPS-HC is drawn from respondents to the previous year’s National Health Interview Survey (NHIS), conducted by NCHS. NHIS provides a nationally representative sample of the U.S. civilian noninstitutionalized population, with over-sampling of Hispanics and blacks.

3. Source of Data

This study is based on healthcare expenditure data from the 2003 MEPS. A total of 32,681 respondents with a positive person weight were included in this analysis. Expenditures in MEPS are defined as the sum of direct payments for healthcare provided during the year, including out-of-pocket payments and payments by private insurance, Medicare, Medicaid, and other sources. Payments for over the counter drugs, alternative care services, and phone contacts with medical providers are not included in MEPS total expenditure estimates. Indirect payments unrelated to specific medical events such as Medicaid Disproportionate Share and Medicare Direct Medical Education subsidies also are not included (Cohen JW, Machlin SR, Zuvekas SH, et al., 2000).

The expenditure data included in this paper were derived from the MEPS-HC and Medical Provider Components (MPC). HC data only on expenditures were collected for non-physician office visits, dental and vision services, other medical equipment and services, and home health care not provided by an agency. MPC expenditure data were collected for office-based visits to physicians (or medical providers supervised by physicians), hospital-based events (e.g.,

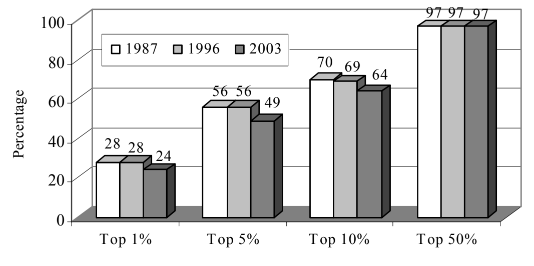
The views expressed in this paper are those of the author and no official endorsement by the Department of Health and Human Services or the Agency for Healthcare Research and Quality is intended or should be inferred. The author wishes to thank Trena Ezzati-Rice and Joel Cohen for their helpful reviews of the paper.

inpatient stays, emergency room visits and outpatient department visits), and prescribed medicines. Data on expenditures for care provided by home health agencies were collected only in the MPC. MPC data were used if complete; otherwise HC data were used if complete. Missing data for events where HC data were not complete and MPC data were not collected or not complete were derived through an imputation process (Machlin S and Dougherty D, 2004).

4. Distribution of MEPS Expenditure Data

MEPS expenditure data exhibit a marked positive skewness, with a few high expenditure cases and many low or zero expenditure cases (Yu WW, 2004, 2005). Furthermore, this skewness or concentration of medical expenditures has also been shown to be consistent over time. Figure 1 (Berk ML and Monheit AC, 2001), updated with 2003 MEPS data, shows that the concentration of healthcare expenditures among the U.S. population has remained stable: the top 1% of the population accounts for 24-28% of total expenditures, the bottom 50% of the population accounts for only 3% of total expenditures, and this degree of concentration has been consistent over time except for a slight drop of concentration for the tail of the distribution in 2003.

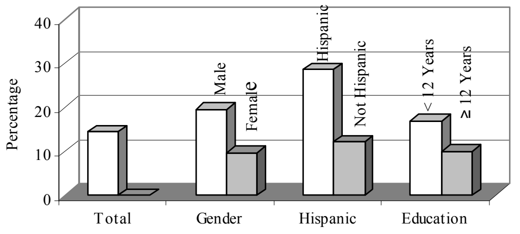
Figure 1 - Concentration of Heathcare Expenditures for the U.S. Civilian Noninstitutionalized Population



Source: 2003 Medical Expenditure Panel Survey

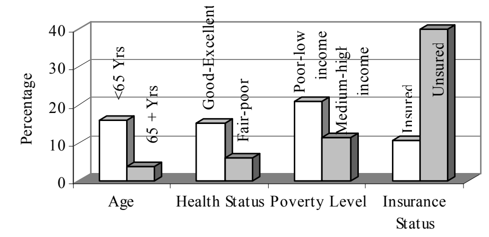
Figures 2 and 2a show the percentages of the population with zero-expenditures by gender, ethnicity, years of education, age, health status, poverty level, and insurance status. In percentage terms, 14% of the population has no healthcare expenditures. Males are more than twice as likely as females to have zero-expenditures (20% vs. 10%). Hispanics are more than twice as likely to have zero-expenditures as non-Hispanics (29% vs. 12%). Persons with less then 12 years of education are more likely to have zero-expenditures than those with 12 or more years of education (17% vs. 10%). Older (65+ yrs) persons are less likely to have zero-expenditures than younger (<65 yrs) persons. Persons with "good to excellent" health status are more than twice as likely to have zero-expenditures as persons with "fair to poor" health status (15% vs. 6%). The rate of having zero-expenditures for persons with "poor-low" income relative to the poverty level is nearly twice that of persons with "medium-high" income relative to poverty level (21% vs. 12%). Persons who are uninsured are nearly four times as likely to have zero-expenditures as persons who are insured (43% vs. 11%).

Figure 2 - Percentage of Population with Zero Heathcare Expenditures by Selected Variables



Source: 2003 Medical Expenditure Panel Survey.

Figure 2a - Percentage of Population with Zero Heathcare Expenditures by Selected Variables



Source: 2003 Medical Expenditure Panel Survey.

5. The Mixture Model

Assume the p.d.f. of MEPS total healthcare expenditures, *X,* has the following form:

$$g(X;\theta) = \begin{cases} 0 & \text{if } X < 0 \\ \delta & \text{if } X = 0 \\ (1-\delta)\, f(X;\theta) & \text{if } X > 0 \end{cases}$$

where θ is a vector of parameters and $f(x;\theta)$ is a continuous p.d.f. defined over a positive real line. It has a point mass at zero, denoted by the parameter δ. For MEPS respondents, $i = 1,2,.....,N,$ the log-likelihood of the mixture model is given by:

$$\ln L = \sum_{i=1}^{N} \{ I_i \ln[(1-\delta) f(X;\theta)] + (1-I_i) \ln \delta \}$$

where $I_i = 1$ if ith individual's expenditure is positive
$= 0$ otherwise.

For the mixture model, in order to restrict δ to lie between 0 and 1, it can be fitted as the following logistic distribution:

$$\delta = \frac{\exp(\ -\lambda\ )}{1 + \exp(\ -\lambda\ )}$$

As λ goes to -∞ and ∞, δ approaches 1 and 0, respectively. Thus, δ always ranges between 0 and 1. One can estimate the logistic component of the model with covariates by replacing λ with *y'v,* where *y* is a vector of covariates and *v* is a vector of corresponding parameters to be estimated.

The positive expenditure values can be assumed to follow one of Weibull, Gamma, or Log-normal distributions that restrict expenditure to be non-negative. If we assume that the positive expenditures follow the Weibull distribution, the p.d.f. is:

$$f(X;\theta) = f(X;\alpha,\beta) = \frac{\alpha}{\beta}(\frac{X}{\beta})^{\alpha-1}\exp(-(\frac{X}{\beta})^{\alpha}), \quad for X \geq 0$$

Covariates can also be introduced directly into the p.d.f. by replacing β with $z'\mu$, where *z* is a vector of covariates and μ is a vector of corresponding parameters to be estimated.

Table 1 below identifies the variables used in the model.

Table 1. List of Variables

Variable	Description
Totexp03	Total Healthcare Expenditures/10,000
Age	0 – 85 (Age at 12/31/2003)
Sex	1=Male, 2=Female
Race/ Ethnicity	1=Hispanic, 2=Black/not Hispanic, 3=Asian/notHispanic, 4=Other/not Hispanic
Education	0 – 17 years of education, NA/DK/Refused set to 0
Health Status	1=Excellent, 2=Very Good, 3=Good, 4=Fair, 5=Poor
Insurance Status	1=Any Private, 2=Public Only, 3=Uninsured
Poverty Level	1=Poor, 2=Near Poor, 3=Low Income, 4=Middle Income, 5=High Income
MSA	1=Non-MSA, 2=MSA
Perwt03f	Final Person Weight

6. Estimation Results

The mixture model with logistic and Weibull components is estimated by the maximum likelihood method and shown in Table 2. Estimation results for the model without covariates are presented in column 2 (Model 1). All estimates are statistically significant at P<0.01. The estimator for $\delta = e^{-\lambda}/(1 + e^{-\lambda})$ is estimated as 0.1441, the weighted proportion of zero-expenditure respondents. Given the estimates shown for the model without covariates, we can suppose that the distribution function of MEPS expenditures is given by the following:

$$G(X) = 0.1441 + 0.8559\,(1 - \exp(-(\frac{X}{0.2274})^{0.6290})),\ for\ \ X \geq 0$$

where X is MEPS total healthcare expenditures in $10,000s.

Table 2 also shows the results of estimating the mixture model with covariates. The third column (Model 2) presents the estimates in which covariates were incorporated into the parameter λ. The covariates were selected to illustrate the model's potential as an analysis tool. For example, we may conclude that a positive coefficient estimate indicates that respondents with higher values of the variable are less likely to belong to the zero-expenditure group (e.g., females are less likely to be in the zero-expenditure group than males). On the other hand, a negative coefficient estimate indicates that respondents with higher values of the variable are more likely to belong to the zero-expenditure group (e.g., respondents with more "years of education" are more likely to be in the zero-expenditure group).

In the fourth column (Model 3) the covariates are introduced directly into the p.d.f. for β. We may conclude for a given covariate that a positive coefficient estimate indicates that respondents with higher values of the variable are likely to have higher healthcare expenditures (e.g., older persons, females, persons with poorer health status). On the other hand, a negative coefficient estimate indicates that persons with higher values of the covariate are likely to have lower healthcare expenditures (e.g., persons with higher income or more years of education). The fifth column (Model 4) shows the estimation results with the set of covariates modeled for λ and β. It is noted that one may model two different sets of covariates for λ and β.

Detailed estimation results for a mixture model with logistic and lognormal components are presented in Table 3.

Table 2. Estimation Results – Mixture Model (Logistic + Weibull)

Parameters	Model 1	Model 2	Model 3	Model 4
λ	1.781354**		1.781354**	
Age		.013402**		.013402**
Sex		.859409**		.859409**
Race/Ethnicity		.320817**		.320817**
Poverty Level		.106414**		.106414**
Health Status		.283343**		.283343**
Education		-.056029**		-.056029**
Insurance Status		-.762514**		-.762514**
MSA		-.133430*		-.133430*
Intercept		.129129 (.398)		.129129 (.398)
α	.629024**	.629024**	.708810**	.708810**
β	.227366**	.227366**		
Age			.004833**	.004833**
Sex			.010658**	.010658**
Race/Ethnicity			.010020**	.010020**
Poverty Level			-.002148*	-.002148*
Health Status			.055958**	.055958**
Education			-.005302**	-.005302**
Insurance Status			-.013249**	-.013249**
MSA			.004115(.175)	.004115(.175)
Intercept			-.03808 **	-.03808 **

** indicates statistical significance at P < 0.01, * indicates statistical significance at P < 0.05, non-significant levels (≥ 0.05) are reported in the parenthesis next to the parameter estimates.

Table 3. Estimation Results – Mixture Model (Logistic + Lognormal)

Parameters	Model 1	Model 2	Model 3	Model 4
λ	1.781354**		1.781354**	
Age		.013402**		.013402**
Sex		.859409**		.859409**
Race/Ethnicity		.320817**		.320817**
Poverty Level		.106414**		.106414**
Health Status		.283343**		.283343**
Education		-.056029**		-.056029**
Insurance Status		-.762514**		-.762514**
MSA		-.133430*		-.133430*
Intercept		.129129(.398)		.129129(.398)
α	.101566**	.101566**	.094134**	.094134**
β	1.620163**	1.620163**		
Age			.004282**	.004282**
Sex			-.073287**	-.073287**
Race/Ethnicity			-.043194**	-.043194**
Poverty Level			-.042229**	-.042229**
Health Status			.101245**	.101245**
Education			-.010791**	-.010791**
Insurance Status			.093506**	.093506**
MSA			-.017370(.394)	-.017370(.394)
Intercept			1.660109**	1.660109**

** indicates statistical significance at P < 0.01, * indicates statistical significance at P < 0.05, non-significant levels (≥ 0.05) are reported in the parenthesis next to the parameter estimates.

Similar estimation results for a mixture model with logistic and gamma components are not reported because they did not affect the qualitative conclusions of this study.

7. Goodness-of-Fit Test

The following Kolmogorov-Smirnov test is used to test the goodness-of-fit between the observed distribution and the distribution implied by the estimated mixture models without covariates:

$$KS = \max\left|G(x_i) - F(x_i)\right|, \quad i = 1,2,.....,n$$

where G(x) and F(x) are the respective values of the observed and implied c.d.f.s for MEPS expenditures, x. The asymptotic statistic is:

$$KS_a = KS\sqrt{n}$$

Test results for the mixture model without covariates are shown in Table 4 below:

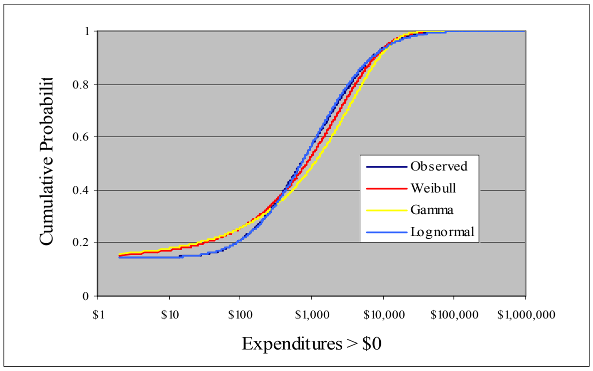
Table 4. Kolmogorov-Smirnov Test Results

Model	KS	KS√n	Ho: F(x)=G(x)*
Weibull	0.0503	4.4609	Reject at p=.05
Gamma	0.0927	8.2259	Reject at p=.05
Lognormal	0.0106	0.9413	Not reject at p=.01

* Test Statistic adapted from Table 1 (Miller LH, 1956).

Figure 3 below shows the cumulative probability plots of the observed data and that of the three (3) mixture models studied. The lognormal curve (blue) is much closer to the plot of the observed distribution (black) compared to that of the Weibull (red) and Gamma (yellow). This observation also agrees with the Kolmogorov-Smirnov goodness-of-fit test results.

Figure 3. Observed vs. Mixture Models



8. Summary

- A mixture model with a point mass at zero is proposed to capture the common bimodality feature of MEPS expenditure distribution.
- The probability of zero expenditure, represented by parameter δ, was separately identified and could be consistently estimated.
- Sets of covariates were added to the model representing the proportion of zero expenditure and the positive expenditures separately.
- The goodness-of-fit tests suggest that the MEPS expenditure data are well represented by the logistic and lognormal mixture model.
- The mixture model offers a way to approximate the distribution of MEPS expenditure data with observations of "zero" expenditure.
- Additional analysis may be needed to account for the effects of stratified multistage sampling.

9. References

Yoo, SH, "A Note on an Approximation of the Mobile Communications Expenditures Distribution Function Using a Mixture Model." Journal of Applied Statistics 2004; Vol. 31, No. 7, 747-752.

Cohen JW, "Design and methods of the Medical Expenditure Panel Survey Household Component." Rockville (MD): Agency for Health Care Policy and Research; 1997. MEPS Methodology Report No.1. AHCPR Pub. No. 97-0026.

Cohen JW, Machlin SR, Zuvekas SH, *et al.*, "Health care expenses in the United States, 1996." Rockville (MD): Agency for Healthcare Research and Quality; 2000. MEPS Research Findings 12. AHRQ Pub. No. 01-0009.

Machlin S and Dougherty D, "Overview of methodology for imputing missing expenditure data in the Medical Expenditure Panel Survey." 2004 Proceedings of the American Statistical Association, Section on Survey Research Methods, [CD-ROM], Alexandria, VA: American Statistical Association.

Yu WW, "Confidence Intervals for Skewed Healthcare Expenditure Data from the Medical Expenditure Panel Survey (MEPS)." 2005 Proceedings of the American Statistical Association, Section on Survey Research

Methods, [CD-ROM], Alexandria, VA: American Statistical Association.

Yu WW and Machlin S, "Estimation of skewed health expenditure data from the Medical Expenditure Panel Survey (MEPS)." 2004 Proceedings of the American Statistical Association, Section on Survey Research Methods, [CD-ROM], Alexandria, VA: American Statistical Association.

Berk ML and Monheit AC, "The concentration of health care expenditures, revisited." Health Affairs 2001; 20: 9-18.

Miller LH, "Table of percentage points of Kolmogorov statistics." Journal of the American Statistical Association 1956; 51, 111-121.