Imputation of Medical Out of Pocket (MOOP) Spending to CPS Records

David M. Betson
University of Notre Dame

February 2001

I. Introduction

In their final report, the NRC Panel on Poverty Measurement and Family Assistance recommended numerous changes to the method by which the US Census Bureau measures poverty.[1] The Panel sought to make recommendations that could implemented. One of the Panel's proposals was to subtract from the family's resources the amount of medical out of pocket (MOOP) spending. Given that neither the Current Population Survey (CPS) nor the preferred data set, the Survey of Income and Program Participation (SIPP), collect information on the family's medical spending, a natural question is how well can one impute this needed data from other sources to either the CPS or SIPP?[2]

The purpose of this paper is to examine the current imputation strategy, discuss its potential shortcomings, and report upon efforts to re-estimate the MOOP model on data from the Consumer Expenditure Survey (CEX).

II. Current Imputation Strategy

The current strategy to impute MOOP spending is a two step procedure. First, national control totals for MOOP spending in families headed by an individual under 65 years old (Non

[1] See Citro and Michael (1995).

Elderly) and those families headed by an individual at least 65 years old (Elderly) are determined. Second, these two aggregate amounts are then allocated to individual CPS families in a manner that reflects the distribution of MOOP spending reported in the National Medical Expenditure Survey (NMES) conducted in 1987 and subsequently 'aged' to reflect spending patterns in 1992.[3]

To formalize this procedure, let A denote the two age groups where NE signifies the non elderly families and E denotes the elderly families. In the first step, estimates of the total spending in the two age groups are determined. Let C_A denote the estimate of total national MOOP spending for the A^{th} age group.

The next step is to allocate C_A to the individual family records on the CPS file. This allocation is based upon the distribution of MOOP spending in a secondary data source such as the NMES or CEX. Using, this secondary data source, one can estimate a regression model that describes the distribution of MOOP spending. This regression model can be used to predict a level of MOOP each CPS record. The predicted values for MOOP in the CPS are not the expected value of MOOP spending for the CPS family based upon the estimated regression model. If the expected value of MOOP was used then the variation in MOOP in the CPS files would be smaller than the variation found in the secondary data source because of ignoring the unexplained errors in the imputation. To replicate the entire distribution of MOOP spending in the CPS, this unexplained variation needs to be included in the imputation procedure. This is accomplished by using the regression model to compute the expected value of the family's MOOP spending and then adding the 'unexplained error variance' through the use of a random number generator.

Let m_{fA} denoted the f^{th} family's predicted MOOP spending. The allocation of the national control totals to the individual family records is accomplished by using a proportional raking technique. In other words, the imputed MOOP value for the f^{th} family record would be equal to

---

[2] Data on an individual family's out of pocket medical spending has been collected only in three nationally representative surveys: the National Medical Expenditure Survey (NMES), the Consumer Expenditure Survey (CEX) and currently, the Medical Expenditure Panel Survey (MEPS).

$$m^{*}_{fA} = m_{fA} \times \sum_{i \in A} \frac{C_A}{m_{iA}} = m_{fA} \times S_A \ .$$

The two scaling factors (S_A), one for each age group, are computed by predicting MOOP for each record in the file and taking their sum for each age group. Then the scaling factor is expressed as the ratio of the age's group control to the sum of predicted MOOP values.

A few remarks on the current procedure are in order at this time. The importance of trying to replicate the entire distribution of MOOP spending must be stressed. Too often, the expected value of the variable is utilized for imputation. Given that MOOP spending is to be subtracted from the family's resources, the use of the expected value of MOOP will likely overstate the true proportion of families whose actual MOOP spending would place them in poverty.[4] To avoid this systematic bias in measurement, the imputation strategy must try to faithfully replicate what is known about the entire distribution of MOOP spending of similar families not just the expected amount.

Maintaining the appropriate correlation with other characteristics is equally important. Estimating the total number of individuals and families that are poor when accounting for MOOP spending, capturing the appropriate covariance between income and MOOP spending will be crucial. Accurately estimating the composition of the poverty population will depend upon how well we can reflect the covariance of demographic characteristics such as age and education with MOOP spending.

Finally, the regression approach taken in the current imputation strategy is not the only method to impute MOOP to the CPS or SIPP. Pat Doyle is investigating an alternative strategy utilizing a statistical matching technique known as 'hot decking'. Instead of predicting MOOP spending via a regression model, actual records from the secondary data source are merged onto the CPS or SIPP files. In theory, the imputation of a single variable to the primary data set (CPS or SIPP) via either method should yield approximately the same results. Any differences that

[3] A fuller description can be found in Betson (1998) which is attached to the paper.

occur will be the result of differences in the common variables taken into account via the matching process and the variables used in the regression models. This paper will not attempt to compare the relative merits of these two strategies but will focus upon the regression model approach.

III. A Critical Examination of the Current Imputation Strategy

A. Control Totals

The MOOP control totals were developed to reflect the actual amount of MOOP spending of families of a given age group in a given year. For the moment, let us assume that the primary data set to which we wish to impute MOOP is for the same year as the secondary data survey. Further, let us assume that aggregate amount of predicted MOOP is less than the control total for each age group

$$\sum_{i \in A} m_{iA} < C_A \ .$$

This is not an unanticipated result given that the predicted MOOP values should reflect not only individual MOOP spending but how the families report their actual spending to the surveys. Proportionally raking the predicted MOOP values to the controls implies that the imputation leads to estimates of actual MOOP spending.

The use of estimates of actual MOOP spending in poverty measurement is a mistake. Currently, all other sources of family resources reflect what is reported to the survey. We know that many sources of income are greatly under reported in many surveys. Even the SIPP that has better reporting of income than the CPS, has significant under reporting. If other sources of resources were similarly adjusted for under reporting then the current method would be appropriate. But since they are not adjusted, the use of estimates of actual MOOP in poverty

[4] See Betson (2000)

measurement with reported amounts of other family resources will overstate the impact of the subtraction of MOOP spending on poverty counts.

The proportional raking adjustment, S_A, assumes that under reporting of MOOP spending is a constant proportion for all family units. Assuming a constant rate of under reporting for all levels of spending is dubious and most likely leads to overstating actual MOOP spending at the higher levels of spending and understating actual MOOP spending at lower levels.

The discussion to this point has assumed that we are imputing data from one survey to another survey where both surveys are for the same time period. Unfortunately, surveys that target medical expenditures are fielded very infrequently. The National Medical Expenditure Survey (NMES) was conducted only once every ten years with the most recent being in 1987. The Medical Expenditure Panel Survey (MPES) seeks to provide more frequent and hence current estimates of what individuals and families spend on health care. While MPES collects data similar to the NMES, a decision has been made that out of pocket expenditures for medical services, supplies and prescription drugs will not provided to the public with the family's cost of health care premiums. While the files will be made public separately, no identification number will be provided to match families across the two files. The aged 1992 NMES file represents the only specially targeted survey on health care that provides both out of pocket expenditures for medical care and premium payments.

Given that the regression model will be used to impute MOOP spending in years other than the year represented in the secondary data set (NMES), the question is how to reflect the changes in MOOP over time. The effect of changes in the number of individuals and families as well as the socio-economic composition of the population will be reflected in the out year primary data base and their inclusion in the regression model. However, differences in the cost of medical care and how individuals respond to the movement in the relative price of medical care will not be reflected in the predicted MOOP levels.

Even if families do not change their utilization of health care in response to changes in its price, multiplying the predicted MOOP values by the change in the price index for medical care will only crudely reflect how medical cost inflation affects individual families. As medical costs increase, insurance premiums will increase and employers may ask their employees to bear a larger share of their health care utilization (the actual cost or price of health care may rise faster to the family than in the economy). But as the price of utilization rises to the family, the family may choose to utilize less health care.[5] Without further research, it is not clear whether indexing predicted MOOP spending for changes in the cost of health care will over or understate MOOP spending.

Recommendation 1:

Imputation of MOOP spending to the CPS should not control the aggregate imputed amounts to an aggregate control total reflecting actual MOOP spending or administrative estimates of MOOP spending. The only scaling of imputed values from the regression model should be done to reflect differences in the costs of medical care between the time between the year of the primary data set and the secondary data set.

Recommendation 2:

After periods of health care inflation, the basic imputation should be re-estimated using secondary data from a time period closer to the year of the primary data. This is needed to capture any changes in utilization of health care and shifting of health care costs from employers to families.

[5] Estimates of the price elasticity of health care demand range from zero to –1.00. See Phelps (1997)

B. Regression Model for Allocation

The prediction of MOOP spending levels for an individual family on the CPS has been described as being the result of a regression model. To examine this characterization further, let us for the time being that all families are all similar to each other except that each family has a different level of MOOP spending. Specifically, let us assume that all the individuals have private insurance coverage, are non poor (incomes in excess of 150% of their respective poverty lines, non elderly single white individuals and all have MOOP spending. Given no differences in observed characteristics in the sample, we could assume that MOOP spending in this family group is distributed log normally, in other words,

$$ln(m_f) = \alpha + \varepsilon_f$$

where α is a constant and ε_f is a random normal variable with mean zero and standard deviation σ. Using the sample of households in the secondary data of this type, we could estimate α and σ. We will denote these estimates as *a* and *s* respectively. Next we would proceed to the primary data set and impute to each single with the same characteristics a value for MOOP spending by first drawing a random number from a standard normal random number generator, e_f, for the f^{th} family in the primary data set and imputing

$$exp\left[a + s \times e_f\right].$$

While this would have been the most straightforward way to implement a regression imputation strategy, it could not be used when the NRC Panel first received data from NMES. It was provided in tabular form (the percentage of the sample with a given set of characteristics that had values of MOOP within a given interval). Lacking data on individual families, a different estimation strategy was employed. We assumed that the underlying MOOP spending was distributed as a log-logistic random variable[6] and hence

[6] The log-logistic distribution was initially defined by Shah and Dave(1963) in a manner similar to the definition of the log normal distribution. This citation was found in Johnson and Kotz (1970).

$$Prob[m \le M] = F[M] = \frac{1}{1 + exp[-(\delta + \phi\, ln(M))]}$$

or alternatively as

$$ln\left[\frac{F[M]}{1 - F[M]}\right] = \delta + \phi\, ln(M).$$

Using the tabular information on families with the same characteristics, we had information on the cumulative probability of MOOP being less than M for various values of M. Based upon this data from the NMES, we could estimate δ and ϕ via OLS. These estimates will be denoted as *d* and *f* respectively.

To impute MOOP values, the first step would be to draw from an uniform random number generator. Let this draw be denoted as u_f for the f^{th} family. This draw represents where the f^{th} household in the MOOP distribution for families with identical characteristics. Given this 'place' in the MOOP distribution, we then compute the value for MOOP that corresponds to this percentile

$$exp\left[\frac{ln\left(\frac{u_f}{1-u_f}\right) - \delta}{\phi}\right]$$

This value is then used as the imputed MOOP value in the primary data set.

This description of the regression approach presents the closest link between this approach and statistical matching via a hot deck method. In a statistical match, one would collect all the observation in the secondary data source that 'close' to the characteristics of the family to which we wish to impute a value in the primary data set and randomly select one of these observation to append to the primary data set. While there is no need for statistical matching to do this, let us assume that the random selection is done in the following manner. First all of the similar

observations are sorted with respect to value of MOOP. Then for each observation, the percentage of similar observations with values less than that observation's MOOP is then computed for all observations that are similar to the one you want to impute a value. Then take a random number from an uniform random number generator and pick the observation whose cumulative probability is closest to the random number. This value of MOOP is used for the observation in the primary data base. This is identical to the procedure employed in the log-logistic regression approach where the only difference is the statistical description of the MOOP distribution is used instead of the actual MOOP from the secondary data source.

This discussion has assumed that all families have the same characteristics which clearly not the case. To allow for differences in the characteristics of the families to affect the imputation of MOOP spending, one could estimate separate sets of parameters (α,σ) or (δ,ϕ) for each family type.

This log-logistic regression approach was used for the preparation of the NRC Panel report. After the report was released, problems with the MOOP data were discovered. These problems were documented in Betson, Citro and Michael (2000). A new version of the MOOP data was provided that not only rectified the problems in the earlier data set but also provided the data from individuals observations that were used to compute the earlier tabular information provided to the Panel. Revisions to the log-logistic model are described in Betson (1998). However, when the new data was made available, a complete evaluation of modeling approach was not undertaken. However with the larger degrees of freedom provided by the individual data from the NMES, it is prudent to take a closer look at the regression strategy at this time.

To compare the modeling strategies, we will examine one family type: a white, non-poor, non-elderly single individual with private health care insurance and MOOP spending. In the NMES sample, there are 662 observations for this family type.[7] Examining the distribution of MOOP in this subgroup, we see that it is skewed toward zero with a long upper tail. This

observation suggests that the assumption of log normality may be a reasonable assumption.[8] The log normal approach would use the sample to estimate the mean (α) and standard deviation (σ) of the log of MOOP (lnmoop). For this subgroup, the estimates are -.784 and 1.401 respectively. Figure 1 plots the density of lnmoop implied by these estimates with a kernel estimate of the lnmoop distribution in the sample.

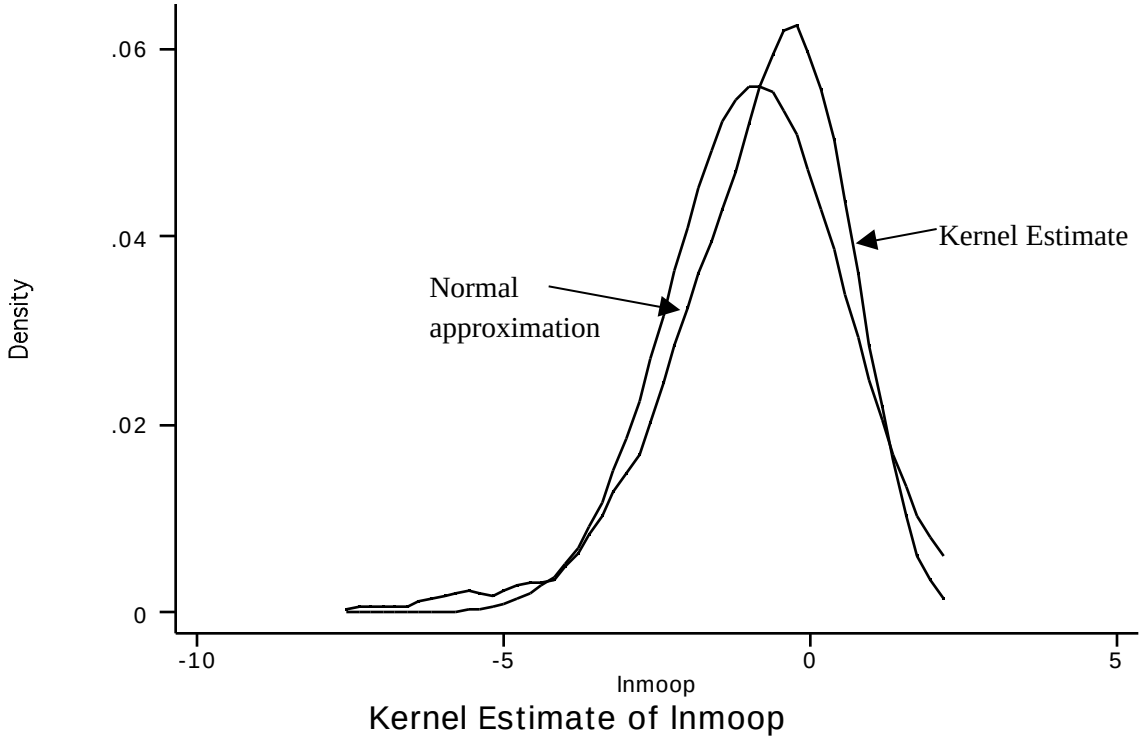


Figure 1

Figure 1 shows that the sample distribution of lnmoop is not normally distributed but is also skewed. The use of the assumption of log normality would lead to imputing too many

[7] In the sample, 60 observations of this family type do not have MOOP spending reported. In the next section, we will discuss how we plan to deal these zero observations.

[8] In the remainder of the paper, I will be analyzing the log of MOOP spending where MOOP is expressed in $1,000. Further all of the results in the paper are weighted statistics.

observations with large values of MOOP spending (note the larger or fatter upper tail of the normal approximation to lnmoop compared to the kernel estimate).

The second approach was to assume that MOOP has a log-logistic distribution. To estimate this model, the log of the ratio of the cumulative probability of MOOP for that value of MOOP over one minus the cumulative probability (lnodds) was regressed against a constant and the log of MOOP (lnmoop)[9]. The results of the regression are reported below.

```
Source |       SS       df       MS                  Number of obs =     661
---------+------------------------------               F(  1,   659) =13438.41
   Model |  2027.57039     1  2027.57039               Prob > F      =  0.0000
Residual |  99.4291262   659  .150878795               R-squared     =  0.9533
---------+------------------------------               Adj R-squared =  0.9532
   Total |  2126.99951   660  3.22272653               Root MSE      =  .38843

------------------------------------------------------------------------------
  lnodds |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  lnmoop |   1.226566   .0105808    115.924   0.000        1.20579    1.247342
   _cons |   .9599494   .0176881     54.271   0.000       .9252176    .9946811
------------------------------------------------------------------------------
```

Figure 2 plots the cumulative probability function based upon the sample observations, the log normal estimates described above and the current log-logistic estimates.

[9] The value for the cumulative probability for a given observation was computed in the following manner. For each subgroup, the observations were sorted. Then for each observation, the number of weighted observations with a value of MOOP less than or equal to the current observation's value of MOOP divided by the total number of observations was recorded as the cumulative probability.

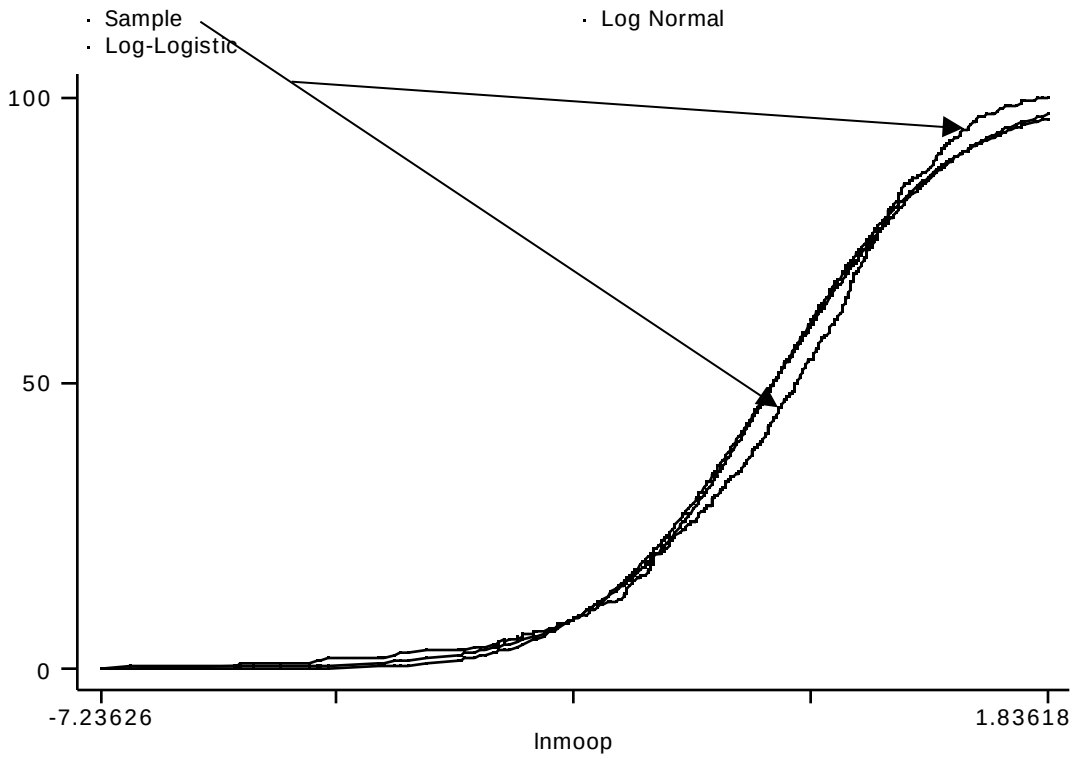


Figure 2

Figure 2 provides two important insights. First, the log normal and the log-logistic assumptions lead to almost identical cumulative probability functions. This result is not unexpected. Johnson and Kotz (1970) note that probit and logit models of discrete choice will lead to very similar results because of the similarity of cumulative probability functions of the normal and logistic distributions. Transforming the basis of the distribution to log scale should not alter this relationship. Secondly, we can conclude that our current strategy of the use of the log-logistic function will lead to too many observations with high values of MOOP spending (note that the CDFs for the log normal and log-logistic approximations lie below the sample CDF at high values of MOOP).

What can be done to address this problem? The solution will require a better approximation. While this approach is ad hoc, I am suggesting that higher powers of the log of MOOP be

included in the regression model. After some experimentation, I am proposing that a cubic approximation be employed. Specifically, the regression model will now be

$$ln\left[\frac{F[M]}{1-F[M]}\right] = \delta + \sum_{n=1}^{3} \phi_n \left(ln(M)\right)^n$$

The regression results for this model are presented below

```
Source |       SS       df       MS                  Number of obs =     661
---------+------------------------------               F(  3,   657) =41417.33
   Model |  2115.81186     3  705.270619               Prob > F      =  0.0000
Residual |  11.1876545   657  .017028393               R-squared     =  0.9947
---------+------------------------------               Adj R-squared =  0.9947
   Total |  2126.99951   660  3.22272653               Root MSE      =  .13049

------------------------------------------------------------------------------
  lnodds |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  lnmoop |   1.640223   .0067654    242.442   0.000       1.626939    1.653508
   lnmp2 |   .2358787   .0043474     54.257   0.000       .2273422    .2444152
   lnmp3 |   .0217826   .0006417     33.947   0.000       .0205227    .0230426
   _cons |   .8545995   .0066664    128.195   0.000       .8415096    .8676895
------------------------------------------------------------------------------
```

In general, this approximation to the sample distribution will denoted as a ‘n order log-logistic’ distribution. Figure 3 plots the sample cumulative probability function with the log-logistic (1st order), the 2nd order and the 3rd order log-logistic approximation. This figure focuses upon MOOP spending exceeding$1,000 the top one third of the MOOP distribution.

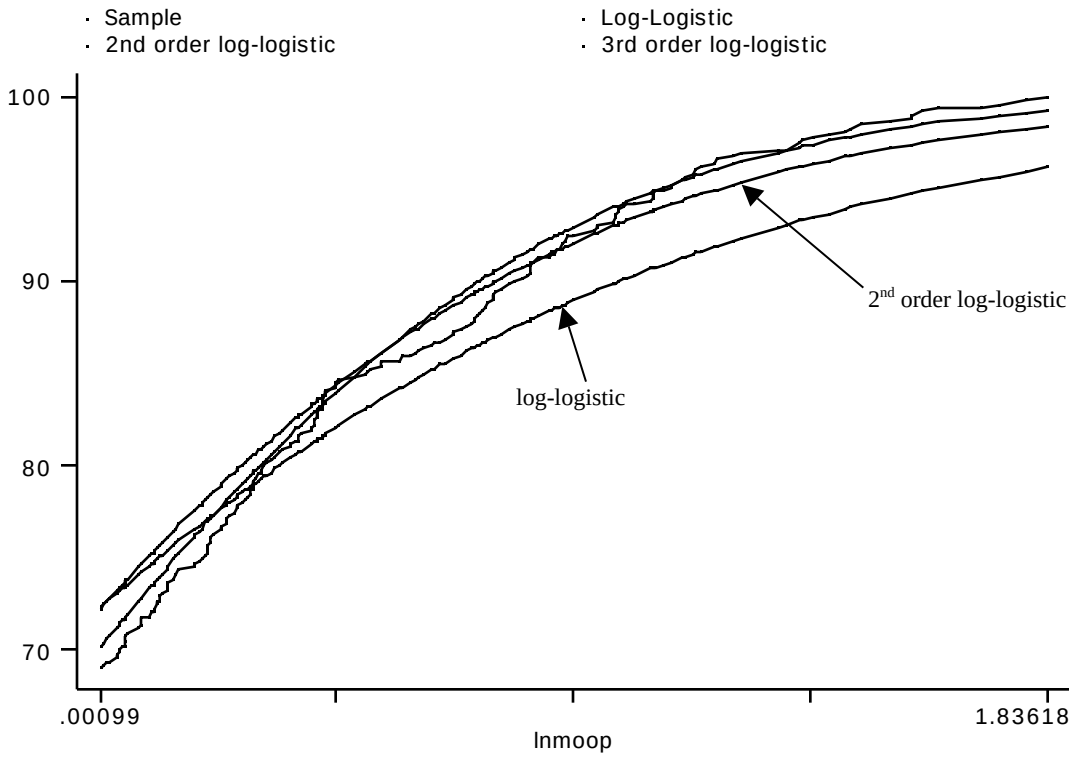


Figure 3

While employing a quadratic term improves the fit of the cumulative probability function, adding a cubic term continues to improve the fit.[10]

Even with this improvement to the regression strategy, the probability of being in the upper tail of the MOOP distribution is still overstated by the higher order smoothing strategies. My ad hoc recommendation is to limit imputation to be less than the estimated 99th percentile of the estimated MOOP distribution. This can be easily accomplished by limiting the value of the

[10] A 4th order approximation continues to improve the fit but increase in goodness of fit was judged to marginal. I should note that this was observation was subjective and not based upon any statistical test. The other consideration favoring the cubic approximation is that explicit solutions exist for cubic equations while they do not for 4th order equations. This will simplify the imputation procedure by not requiring numerical techniques for solving for M given a value of u_i.

uniform random, u_f, to a maximum value of .99. Hence once the parameters, δ, ϕ_1, ϕ_2 *and* ϕ_3 have been determined for a family type, we would impute to the f^{th} observation of the same type in the primary data set a value of M that solves the following equation

$$ln\left(\frac{min(.99,u_f)}{1-min(.99,u_f)}\right) = \delta + \sum_{n=1}^{3}\phi_n\left(ln(M_f)\right)^n$$

In summary, I would make the following recommendations.

Recommendation 3:

A 3rd order log-logistic approximation to the cumulative probability used to describe the distribution of MOOP for subgroups of the population.

Recommendation 4:

When imputing values to the primary data set, MOOP values be limited to the lower 99% of the estimated MOOP distribution.

C. Predicting Zero MOOP

We have focused upon imputing MOOP to those observations with MOOP spending. However, not all observations in the NMES sample have MOOP spending. To impute MOOP to all of the observations in the primary data set would be wrong. While estimation problems akin to sample selection bias issues are most likely present, these issues are going to be ignored. Assignment of a non zero MOOP amount to observations will be based upon the proportion of a family type that have reported MOOP in the secondary data base. Random assignment will utilize this estimated proportion in conjunction with a draw from an uniform random number generator. If *P* is the proportion of the secondary data base of a given family type then a non zero MOOP level will be assigned to the f^{th} observation in the primary data base if

$$v_f < P$$

where v_f is a draw from uniform random number generator. Otherwise, a zero value for MOOP will be assigned.

D. Qualified Medicare Benefit (QMB) and MOOP

Individuals who qualify for Medicare and have incomes less than 100 percent of poverty, the Medicare program waives all cost sharing provisions and Part B premiums. For Medicare eligible individuals between 100% and 120% of poverty, Part B premiums are waved. These benefits are referred to as the Qualified Medicare Benefit (QMB). This benefit was implemented in 1990. In the current imputation procedure that uses the NMES, Part B Medicare premiums were not included in the definition of MOOP. Hence all elderly individuals are accessed a Part B premium unless they report receiving Medicaid. This procedure does not take into account the QMB portion of Medicare and leads to an overstatement of MOOP spending for this portion of the elderly population. A simple solution will be to add a Part B premium only for those elderly individual's income exceeds 120% of poverty.

QMB also waives the cost sharing provisions of Medicare eligible medical services and supplies. Given that the NMES is based upon 1987 data aged to 1992, it is doubtful that aging procedure took this provision into account. While the current imputation imputes no MOOP for elderly individuals reporting the receipt of Medicaid, not all poor elderly receive Medicaid. Hence for these individuals, the current procedure overstates their MOOP spending due to the QMB. However, to include zero MOOP for all poor elderly would also be wrong since Medicare accepts not all medical expenses. The largest single exception is prescription drugs. Since the current NMES data does not separate MOOP spending on drugs, I have chosen to continue the current practice of imputing MOOP spending to all poor elderly who do not report Medicaid.

Recommendation 5:

For those individuals over 65 years old living in a family whose income is less than 120% of poverty, no Medicare Part B premiums will be assigned.

E. Concluding Remarks

One conclusion that could be drawn is that the estimation and imputation strategy currently employed by myself and the Census Bureau produces too many observations with relatively large values for MOOP. In this section, I have proposed five recommendations aimed at improving the imputation of MOOP throughout the entire distribution. In this next section, I will discuss my re-estimation of the model on the NMES. The following section reports upon a comparison of various imputation approaches using the March 1993 CPS.

IV. Re-fitting the Model on NMES data

Before proceeding to estimate the imputation model on more recent data, I thought it would be instructive to re-fit the modified model on the economic and demographic aged NMES data. In the previous section, the case was made for the inclusion of squared and cubed terms of the log of MOOP in the model. That is the approach that will be taken in this re-estimation.

In the former version of the model, 36 separate family types were constructed for the non elderly population. These groups were based upon the insurance coverage, the family size, poverty status, and race of the family. The elderly population was subdivided into 8 groups based upon age, family size and poverty status. For each of these 44 groups, the cumulative probability was constructed by sorting the observations and computing the percentage of the group that had MOOP spending less than the observation. The cumulative probability was then transformed into the log 'odds' that is the dependent variable of the regression analysis. The previous analysis of the data was performed separately on the non elderly and elderly samples. This analysis allowed for only the main effects of the group's other characteristics to affect the estimation of the intercept *(*δ*)* and slope coefficients *(*ϕ*)*. All interaction effects between characteristics were assumed to be zero. In retrospect, this was an unfortunate assumption. Significant interaction effects where found when the 1st order log-logistic model was recently re-estimated. This lead to separate estimates of the model for each of the 44 groups. The regression estimates for the 3rd order log-logistic model are reported in Appendixes A and B.

Since the imputation of zero MOOP values has not changed, the previous estimates of the probability of having MOOP spending will be used.

V. Comparison of MOOP Imputations on the 1993 CPS

In this section, I will report upon a Monte Carlo experiment I conducted to empirically examine the consequences of the various recommendations that I have proposed. Since the NMES data represents 1992, the choice of the March 1993 CPS was ideal since the imputation would not require any out year projections. I chose three alternative imputation implementations that were the following:

Original Imputation: This is strategy that I have employed and forms the basis of the Census Bureau's imputations. This strategy uses a proportional rake to established national totals. For 1992, the control totals were $153 billion for the non elderly population and $55.5 billion for the elderly non Part B premium MOOP. The regression model was estimated for the non elderly and elderly populations separately as described in the previous section.[11] Finally, limitation were made on MOOP imputations. The maximum MOOP for a non elderly family was $8,200 while $18,000 for an elderly family. These limits represent the 99th percentile of the two populations and were provided by Pat Doyle.

No Control Totals: This implementation was identical to the previous one except that no raking was performed to 'hit' the control totals.

New Implementation: This implementation reflects recommendations 1,3, 4, and 5 made earlier. The 3rd order log-logistic model was estimated for each of the 42 different family types. Limits were placed on the maximum MOOP that was assigned. No family was assigned a MOOP that exceed the 99th percentile of the MOOP distribution for their respective family type. Elderly adults living in families whose income is less than 120% of poverty were not assigned Medicare Part B premium. And no raking was performed to achieve a control total.

For each of the three implementations, I performed 100 MOOP imputations to the entire March 1993 CPS.[12] The first variable that I examined was the mean MOOP (includes both zero and positive values) in each of the two age groups. The following table presents the Monte Carlo results for the simulations as well as the averages from the NMES (secondary file).

[11] See Betson (1998) for more a detailed description of the regression model and estimates. This paper is attached.

	Average MOOP in:	
	Non Elderly	Elderly
NMES	$1,432	$2,304
Original Imputation	$1,815	$2,600
No Control Totals	$1,735	$1,771
New Imputation	$1,398	$2,238

The use of the control totals significantly raises the average imputed MOOP from their respective averages in the original NMES file. While this difference could represent the difference between actual and reported MOOP, the differences are striking. But what is also shown is how the raking dramatically hides what a rather poor job the original regression model does in replicating the mean MOOP. Average non elderly spending is overstated while elderly spending is understated. While the previous discussion made us question the appropriateness of the model representing the upper tail of the MOOP distribution, these figures suggests it does a poor job replicating means. Given the similarity between the log-logistic and log normal models, moving toward a log normal model would not be a desirable path to follow.

The similarity of the average MOOP imputed with the New Implementation and the averages found in the NMES file are extremely comforting. They provide evidence of the gain in imputation accuracy provided by the new regression model and other recommendations.

I computed the average poverty rates for children, the elderly and for the total population for each of three implementation. For purposes of comparison, I have provided the official poverty rates for 1992. One might be concerned that the random noise in the imputation may lead to large

[12] Appendix C contains the FORTRAN source code for the new imputation routines.

variation in the poverty rates based upon the imputations. In the following table, I have also included the standard deviation of the estimated poverty rates.

	Poverty Rate of:		
	Children	Elderly	All Persons
Official	21.87	12.90	14.52
Old Imputation	24.90	20.68	17.72
	(.11)	(.18)	(.06)
No Control Totals	24.76	18.60	17.32
	(.11)	(.18)	(.06)
New Imputation	23.86	19.87	16.86
	(.08)	(.20)	(.05)

The use of the control totals did lead to higher poverty rates. For children, the effect of not raking the data was minor compared to the elderly. However, the raking masked the rather poor imputation of the underlying model. When the improved model is employed, less MOOP is assigned to the non elderly and more is attributed to the elderly. This shift in the distribution between the two age groups has the expected impact on poverty rates. Children's rates fall and elderly rates rise when compared to the rates produced by the previous regression model without control totals.

The standard deviations (in parenthesis) of the poverty rates show how little possible variation in the rates can be caused by imputation procedure. In my opinion, they are quite small.

VI. Imputation Model based upon CEX data

Comparison between the CEX and NMES MOOP Data

Before proceeding to the estimation of a new imputation model on CEX data, potential differences between the CEX and NMES data will be first examined. The 1987 NMES data that was used in previous analysis had been aged and weighted to reflect MOOP expenditures in 1992. To compare data from the CEX, I extracted MOOP data from the 1992 and 1993 CEX Interview Survey files.[13] The BLS collects data on a quarterly basis but not all units can be interviewed for four quarters. Instead of using each quarterly interview as independent observations, I chose to employ observations on units who had at least three completed interviews and examine only estimates of annual net medical expenditures. For those units with four interviews, annual MOOP was computed as the sum of reported MOOP from each of the quarterly interviews. For those units with only three interviews, the sum of their reported MOOP spending was multiplied by 4/3.

The CEX data collects the unit's spending on medical care services, supplies and equipment net of the amount that reimbursed by insurance or any government program as well as the cost to the unit of any health care insurance including Medicare Part B premiums. As has been already noted, the NMES data received from AHCPR did not reflect Part B premium in their definition of MOOP. The two data sets were made comparable by imputing Medicare Part B premiums to the NMES for the elderly population.

The following table compares the distribution of MOOP from both the NMES and CEX data. The entries in the table report the dollar amount of MOOP at various centiles of the distribution of the non-elderly and elderly populations separately. Performing a Kolmogrov–Smirnov test of equality of the distributions, we can reject the null hypothesis of equality for both the non-elderly

[13] I had only the interview data from the 1992 and 1993 CEX and not the files contain information on health care insurance coverage (IHCA and IHCB files) so estimating the previous model on the CEX data for these years was not feasible.

and elderly populations. Yet, the tables demonstrate that the most significant differences between the MOOP data from the CEX and NMES occur in the upper tail. In both the non-elderly population but even more clearly in the elderly population, the 99th percentile of MOOP distribution in the NMES data is significantly larger.

Distribution of MOOP from Alternative Data Sources

	10%	25%	50%	75%	90%	99%	Mean
NonElderly							
NMES	$103	$337	$955	$2,105	$3,655	$9,740	$1,590
CEX(92/3)	146	429	1,068	2,214	3,722	8,649	1,630
Difference	$43	$92	$113	$109	$67	- $1,091	$40
Elderly							
NMES	$498	$1,047	$1,779	$3,071	$5,154	$18,595	$2,812
CEX(92/3)	478	1,038	1,869	3,238	5,037	10,664	2,503
Difference	- $20	- $9	$90	$167	- $117	-$7,931	- $309

The small absolute differences in MOOP values at various points in the distribution, suggest that the CEX can indeed provide a reasonable data base for the imputation of MOOP even though it is not a survey designed specifically to collect data medical expenditures but all spending within the household.

A 'New' Imputation Model for MOOP

The primary purpose of this project was to update the MOOP imputation model by replicating the model on more recent data, namely the 1996-7 CEX. During the process of replicating the model on the CEX data, additional changes or 'improvements' were made. Some

were necessitated by data limitations, others were based upon further refinements to the specification of the model.

The analysis file was constructed using the sample of those families interviewed by the BLS from January 1996 until March 1998. Only families who had at least three completed quarterly interviews and who provided complete income responses were included in the final sample. This yielded 6,300 non-elderly observations and 1,943 observations on elderly units. The value of MOOP reflects the net medical expenditures for medical services and supplies as well as health care insurance premiums paid by the household. For those participating in the Medicare Part B, the premiums are included. Hence the revised imputation model will not need to add an amount for Part B premiums as a final step.

Once the analysis file from the 1996-7 CEX Interview Survey was constructed, a problem was encountered. While the overall sample size from the CEX was roughly similar to the aged NMES sample, the non elderly population yielded too few observations for black families to support the entire division of the population into groups based upon insurance status, income, family size and race. For some group cells, there were fewer than 4 observations with MOOP and hence the imputation model could not be estimated for those groups. Further for cells with more than 4 observations, I performed a Kolmogrov–Smirnov test of equality of the MOOP distribution between black and non-black families. Except for families who had private insurance and whose family size was four or more, I did not find any significance difference between black and non-black families. Hence except for this group, I did not make any distinction in the model for race. While the NMES sample permitted separate models to be estimated for those families with public insurance only and whose income was in excess of 150% of poverty, the CEX sample did not contain sufficient numbers of these types of families. Hence the re-estimated model did not allow for this distinction between families. In total, these restrictions implied that the model was estimated for 18 distinct groups instead of the 36 groups estimated on the NMES data.

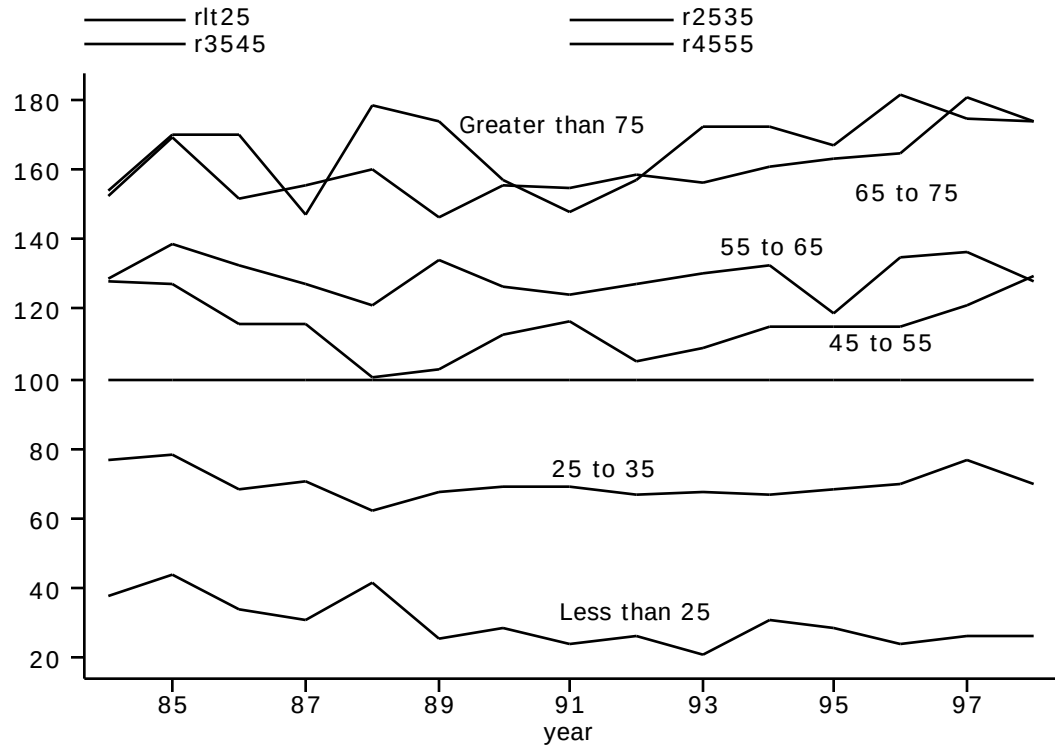
No data limitations were encountered for the sample of elderly and the model could be replicated using the more recent CEX data.

The results of the re-estimation of the revised imputation model for the non-elderly population are reported in Appendix D. The results for the elderly population are presented in Appendix E.

In the present model, the age of the oldest person in the family unit is reflected in the estimation process by separating the sample into two parts: the non-elderly and the elderly populations. While this distinction is reasonable, age is not reflected in the model within the non-elderly population. Given the importance of age to the use of health care and hence MOOP, I decided to investigate whether the model should reflect the wide range of age within that subsample.

Utilizing the published CEX data from 1984 to 1998, I plotted the average MOOP for a given age interval (less than 25 years old; 25 to 35; 35 to 45; 45 to 55; 55 to 65; 65 to 75; and 75 years and older) relative to the average MOOP of those 35 to 45 years old. The time series is presented in the graph below.

Average MOOP by Age Relative to Average Spending of 35 to 45 Year Olds



While the current model allows for age differences within the elderly sample, this graph demonstrates that suppressing the impact of age on MOOP spending within the non-elderly population hides a good deal of variation that could explained by this single dimension. Given the large sample of non-elderly units that have private insurance, the age of the oldest family member was included in the model by further subdividing this group by five age categories. For the non-elderly units who had only public insurance, sample size limited our ability to employ all five age categories and groupings were based upon the results of Kolmogrov–Smirnov tests of equality of distributions. Cells that were found not to be statistically different from each other were grouped together. Finally, the non-elderly population without health insurance could not be further subdivided by age.

The results of this new imputation model for the non-elderly population are reported in Appendix F. Finally, Fortran source code for the new CEX imputation model is provided in Appendix G.

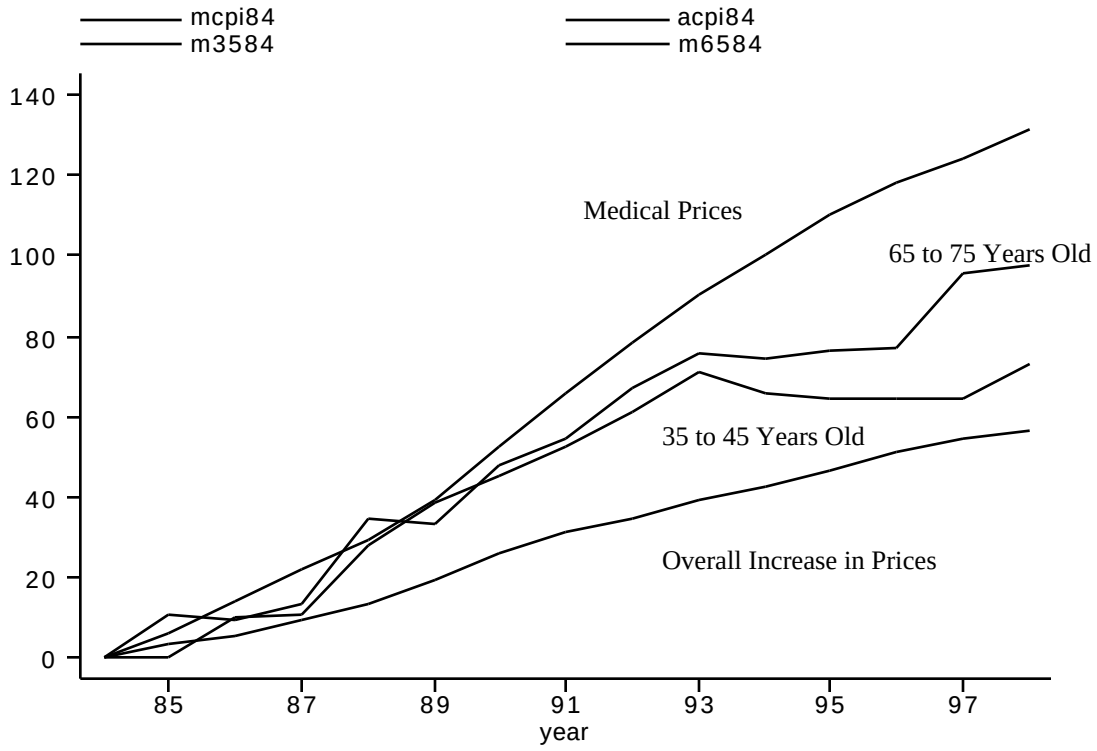
To check how well the imputation model replicates the distribution of MOOP for each of the 32 non-elderly groups and 8 elderly groups, I computed the value of MOOP implied by the separate models at the 25^{th}, 50^{th}, 75^{th}, 90^{th}, and 99^{th} percentiles to the sample data at the same percentiles. The prediction of the median (50^{th} percentile) was extremely close for all groups, however, predictions at the extreme upper tail (99^{th} percentile) tended to be 5 to 10% lower than the corresponding 99^{th} sample percentile. But in general, I found that the estimated model replicated the sample data well and recommended its use.

VII. Projecting Imputed MOOP Amounts to Other Years

The current model provides a method to impute MOOP to the CPS or any other data base. However, it should be recognized that the model produces a MOOP amount for a family of given characteristics expressed in 1997 dollars. If the year represented in the data base is different than 1997, the question of how to either inflate (later years) or deflate (earlier years than 1997) the model's MOOP amount needs to be answered?

For time being, let us consider the problem of imputing to a year later than 1997. Based upon past experience, let us assume that the price of medical care will rise faster than the overall increases in price levels. MOOP represents the net payments that family will make on health care. Some of the payments will represent the family's share of the cost of their health care utilization. Thus as the relative price of health care rises, we would expect the unit to reduce their utilization. Hence this portion of MOOP should be expected to rise but not at a rate that medical prices rise. The remainder of MOOP is primarily composed of the family's payment of their health care insurance premiums that not directly to the family's utilization of health care but overall utilization. When the medical prices rise faster than other prices, employers must decided how much of the increase in insurance costs to absorb themselves and how much to pass on their workers. Hence again, we can expect MOOP to rise but most likely more slowly than increases in the price of medical services. The point of this explanation is to show that inflating the model's MOOP amount from 1997 to some out year the CPI for medical services will most likely overstate the increase in the nominal amount of MOOP.

To illustrate this point, I compared how since 1984, average MOOP amounts for 35 to 45 years old individuals and 65 to 75 years old individuals rose compared to the increases in medical prices and the overall CPI.

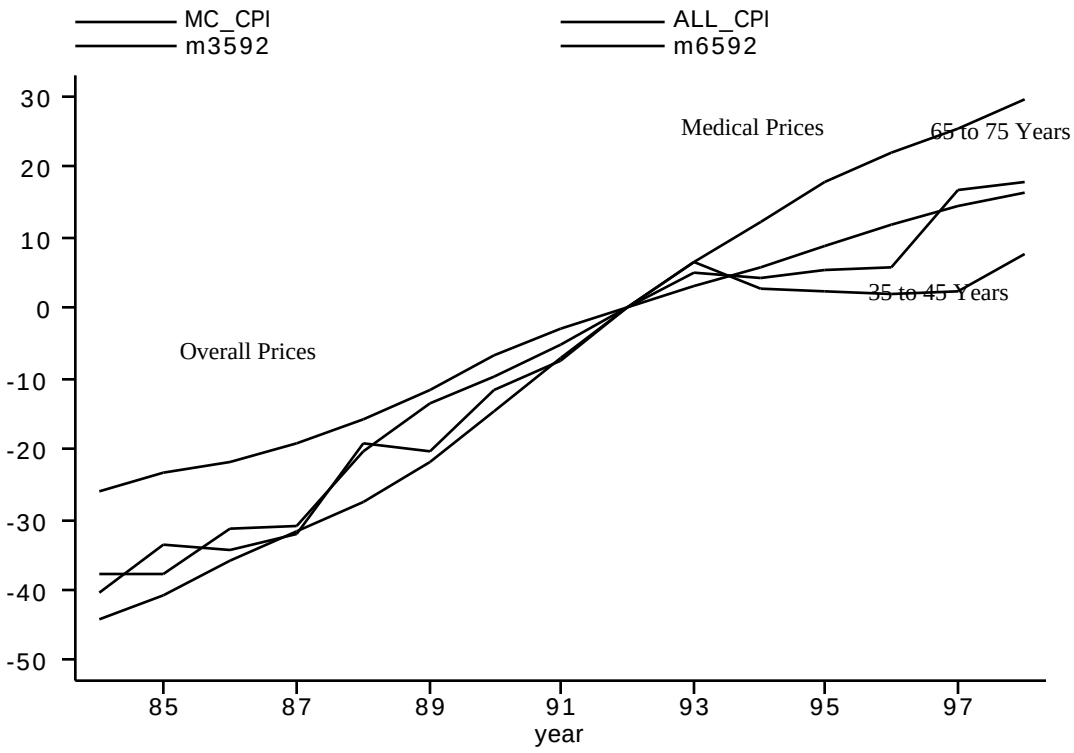


Percentage Increases of MOOP Spending and Prices Since 1984

Over the entire fourteen year period (1984 to 1998), the average MOOP of both age groups has risen slower than the increases in medical prices but faster than the increases in all prices. However since 1993, the families head by 35 to 45 year olds have been shielded from increase prices and have not seen their net medical payments rise substantially. Older families (65 to 75 year olds) have not been so fortunate. Consequently the gap between the average MOOP payments of these two age groups has grown over the time period.

This effect can also be seen in the previous graph that showed the relative MOOP payments by age over time. The relative gap of both elderly groups relative 35 to 45 year olds grow over the time period while the relative gap for the non-elderly age groups remain constant over time.[14]

What I take away from this crude analysis is that indexing the imputed MOOP amounts by medical price increases will most likely overstate MOOP in future years but understate MOOP when back casting the imputed years to earlier years. To illustrate that point directly, I have re-normalized the previous graph to reflect the base year to be 1992 – the year of our previous imputation model.



[14] The only exception is for the age group of families less than 25 years old. In this case the gap too increases but not a statistical significant rate.

As noted earlier, average MOOP for the non-elderly has grown in nominal terms since 1992 while the elderly's MOOP has grown roughly as fast as the overall rises in prices but not as fast as the increases in medical prices. Back casting the model's imputations by the deflating by the medical CPI would understate the nominal MOOP values prior to 1992.

Whether the trend in MOOP during the period from 1992 to 1997 will continue into the future is uncertain. However, one clear lesson one can draw is that inflating the 1992 NMES model values by the medical CPI clearly significantly overstates the expected nominal MOOP in later years. A more conservative approach would be to use the overall CPI to inflate imputation from the model. But in the process of back casting the model to years earlier than 1992, the use of medical CPI understates the average MOOP and hence is a more conservative modeling approach.

I experimented with a model that attempted to predict changes in nominal MOOP based upon changes in medical and overall prices. But at this time, no reliable model has been able to be estimated. In lieu of such an approach, I would suggest the following ad hoc approach. For forecasting the imputations from 1997, I would inflate the values using the overall CPI. For back casting to years prior to 1997, I would use the medical CPI. This is a very conservative approach in the sense that it will not overstate the impact of MOOP on poverty but most likely understate it.

VIII. Conclusions

In this report, I have presented an analyses of the previous modeling approach and a series of recommendations that should 'improve' the imputations of MOOP to the CPS. One concern has always been that the earlier modeling approach yielded too many observations with high or large values for MOOP. In this paper, I have provided evidence that this concern was well founded but at the same time provided suggestions that would rectify these problems. The newly estimated imputation model based upon the 1996-7 CEX data I believe represents a vast improvement over the previous modeling strategy.

References

Betson, David (1998) 'Imputation of Medical Out of Pocket (MOOP) Expenditures to CPS Analysis Files' memo, University of Notre Dame, Notre Dame, IN.

Betson, David (2000) 'Response to Bavier's Critique of the NRC Panel's Recommendations.' Appears on the Census Bureau's web site.

Betson, David, Constance Citro and Robert T. Michael (2000) 'Recent Developments for Poverty Measurement in U.S. Official Statistics' *Journal of Official Statistics.*

Citro, Constance and Robert T. Michael (1995) *Measuring Poverty : A New Approach*, National Academy Press, Washington, DC

Johnson, Norman and Samuel Kotz (1970) *Continuous Univariate Distributions – Volume 2*. John Wiley and Sons, New York, New York.

Moon, Marilyn (1996) *Medicare Now and in the Future, 2nd edition.* Urban Institute Press, Washington, DC.

Phelps, Charles, (1997) *Health Economics 2nd edition*. Addison-Wesley, Reading, MA.

Shah, B.K. and P. H. Dave (1963) "A Note on the Log-Logistic Distribution" Journal of the M.S. University of Baroda (Science Number), 12, pp. 15-20.