Imputing Total Health Expenditure in HRS using Out of Pocket Expenditure in MEPS

Mohammad Amin Shirazian

Abstract

In this paper we impute total medical spending for the Health and Retirement Survey (HRS), Using Medical expenditure survey (MEPS). We outline the imputation procedure, which establishes a relationship between out-of-pocket and total expenditure on healthcare. We then imputed health expenditure based on this

1 Introduction

To understand the saving pattern among the elderly, we need to understand the role of medical spending in saving behavior. However, the recent literature uses out of pocket medical spending to measure this relationship. De Nardi and Fella (2017) does a complete review where they discuss the role of medical spending in explaining wealth distribution. De Nardi et al., 2025 find that medical spending is the main motive for most households but not the rich households so its explanatory power is less than bequest motive using out of pocket spending.

To understand the role of medical spending in saving behavior, we need information on total medical spending. In this paper we impute total medical spending for the Health and Retirement Survey (HRS), Using Medical expenditure survey (MEPS). We outline the imputation procedure, which establishes a relationship between out-of-pocket and total expenditure on healthcare. We then impute health expenditure based on this relationship.

Skinner (1987) uses a linear fit to impute consumption expenditure from different consumption component. Although simple, this method helps researcher to use imputed total consumption instead of food spending in their models. This approach adopted frequently (Palumbo (1999), Ziliak (1998), Browning and Leth-Petersen (2003)) since variation in food consumption explains 80 percent of total consumption. Blundell et al. (2004) use the inverse relationship introduced by Skinner (1987) and introduces adjustments that satisfies moment consistency conditions.

A common issue in these imputation techniques is measurement error Campos and Reggio (2014). In practice, imputing total non-durable expenditure from another survey causes measurement bias. The way to address this could be using instrumental variable(Blundell et al. (2008)), multiple measures (Browning and Crossley (2009)), and multiple proxies(Bollinger and Minier (2015)). Lucky for us, we do not use survey data in our process which helps us avoid the measurement error common in the survey data.

Using imputed variables instead of actual data has its own caveats. First, there is a problem of moment inference(Attanasio and Pistaferri (2014)). Second, using an imputed measure as a dependent((Crossley et al. (2022))) or independent variable Wooldridge (2010) in a regression introduces bias in the estimated coefficients. Overall, adopting any specific method depends on the focus of research.

We adopt the imputation procedure developed by Blundell et al. (2004) and describe the imputation procedure and examine the properties of the imputed health expenditure in both MEPS and HRS, comparing them to true health spending. After adjusting for differences in the distribution of out-of-pocket expenditures between the two datasets, we find that the average health expenditures in HRS closely resemble those in MEPS.

2 Imputation Procedure

Consider the following relationship between out-of-pocket and total expenditure on healthcare[1]:

$$p_i = D_i\beta + \gamma h_i + \epsilon_i \quad i \in \{M, H\} \tag{1}$$

Where p is out-of-pocket spending, D is demographic and other control variables, h is total health spending, and ϵ is the error term. Subscript $i \in \{M, H\}$ indexes sample. Subscripts M and H denote MEPS and HRS respondents. Since our goal is to impute the health expenditures, by rearranging equation (1), we can define imputed health expenditure as:

$$\hat{h}_i = \left(\frac{p_i - \left(D_i\hat{\beta}\right)}{\hat{\gamma}}\right) \tag{2}$$

For having a well-defined imputed health expenditure, we need to assume $\hat{\gamma} \neq 0$. Since we have both demographic variables and out-of-pocket expenditures in both data sets, we can use the estimators $\hat{\beta}$ and $\hat{\gamma}$ to define the imputed total health expenditure in HRS.

To understand the relationship between the true and imputed health expenditures, we can use (1) to get:

$$\hat{h}_i = D_i\frac{\left(\beta - \hat{\beta}\right)}{\hat{\gamma}} + \frac{\gamma}{\hat{\gamma}}h_i + \frac{\epsilon_i}{\hat{\gamma}} \tag{3}$$

Since, in most studies, the researchers are interested in the mean and variance behavior of the variables, With consistent estimates of γ and β, we focus on the consistency of these moments. Following Blundell et al. (2004), we define the sample mean, variance, and covariance as:

$$\bar{x} = \frac{\mathbf{1}_n' x}{n}$$
$$s_x^2 = \frac{(x - \bar{x})'(x - \bar{x})}{n}$$
$$s_{xy} = \frac{(x - \bar{x})\,(y - \bar{y})}{n}$$

$\mathbf{1}_n$ is $n \times 1$ vector of ones.

Assumption 2.1.

1. *The underlying population of donor(MEPS) and main(HRS) sample is the same.*
2. *$E[\epsilon] = 0$*
3. *$E[D'\epsilon] = 0$*
4. *$E[h'\epsilon] = 0$*

Note that under assumption 2.1, since the underlying population of both samples is the same:

1. $p\text{lim } \bar{\epsilon}_M = p\text{lim } \bar{\epsilon}_H = 0$
2. $p\text{lim } \frac{D'_M\epsilon_M}{n_M} = p\text{lim } \frac{D'_H\epsilon_H}{n_H} = 0$
3. $p\text{lim } \frac{y'_M\epsilon_M}{n_M} = p\text{lim } \frac{y'_H\epsilon_H}{n_H} = 0$

Under assumption 2.1, $\hat{\gamma}$ and $\hat{\beta}$ are consistent estimators[2]. Since the structure of the MEPS survey does not allow for a significant measurement error these assumptions are reasonable. Utilizing (3) and denotin population mean and variance by μ and σ^2:

[1] See the appendix for a complete discussion on other imputation methods.

[2] See the appendix for the proof and conditions of consistency.

$$\begin{aligned} &\operatorname{plim} \bar{\hat{h}}_M = \operatorname{plim} \bar{h}_M = \mu_h \\ &\operatorname{plim} s^2_{\hat{h}_M} = \operatorname{plim} s^2_{h_M} + \frac{1}{\gamma^2} \operatorname{plim} s^2_{\epsilon_M} = \sigma^2_h + \frac{1}{\gamma^2}\sigma^2_\epsilon \end{aligned} \tag{4}$$

The consistency of $\hat{\beta}$ implies that the terms containing $\hat{\beta}-\beta$ would be omitted in the limit. Therefore, the difference in the demographics of the two data sets would not play a significant role in imputed mean and variance consistency. The imputed variance would not converge to the true variance even if there is no endogeneity bias ($E[h'\epsilon]) = 0$). Also note that (4), implies mean consistency of $\hat{h}_h$, under assumption 2.1.

To find the mean and variance terms for HRS, Notice that equations (2) and (**??**) implies:

$$\hat{h}_H = \hat{h}_M + \frac{1}{\hat{\gamma}}\left[(p_H - p_M) - (D_H - D_M)\,\hat{\beta}\right]$$

Therefore using the equation (4), we can show that:

$$\begin{aligned} &\operatorname{plim} \bar{\hat{h}}_H = \operatorname{plim} \bar{\hat{h}}_M + \tfrac{1}{\gamma} \operatorname{plim} (\bar{p}_H - \bar{p}_M) - \tfrac{1}{\gamma} \operatorname{plim} \left(\bar{D}_H - \bar{D}_M\right) \\ &\operatorname{plim} s^2_{\hat{h}_H} = \operatorname{plim} s^2_{\hat{h}_M} + \tfrac{1}{\gamma^2} \operatorname{plim} s^2_{(p_H - p_M)} - \tfrac{1}{\gamma^2} \operatorname{plim} s^2_{(D_H\beta - D_M\beta)} \end{aligned} \tag{5}$$

Where $\bar{D}_f = \frac{\mathbf{1}'_n \times D_f \beta}{n}$ for $f \in \{H, M\}$. Consequently, under assumption 2.1:

$$\begin{aligned} &\operatorname{plim} \bar{\hat{h}}_H = \operatorname{plim} \bar{\hat{h}}_M = \mu_h \\ &\operatorname{plim} s^2_{\hat{h}_H} = \operatorname{plim} s^2_{\hat{h}_M} = \sigma^2_h + \frac{1}{\gamma^2}\sigma^2_\epsilon \end{aligned}$$

Similar underlying population assumption play a crucial role in (5). First, mean consistency depends on it. If underlying populations are different, then average imputed total spending in HRS is inconsistent, depending on the difference in out of pocket spending and demographic means of populations. Second, the bias in imputed variance of total health spending in HRS would differ from MEPS, depending on the difference in variance of control variables in underlying populations.

In practice, the assumption of similar out of pocket spending fails. Nevertheless, assuming similar demographics in underlying populations, we can make the following adjustments for difference in out of pocket spending:

$$\begin{aligned} &\hat{h}_a = \hat{h}_H - \tfrac{1}{\gamma} \operatorname{plim} (\bar{p}_H - \bar{p}_M) \\ &\operatorname{plim} \bar{\hat{h}}_a = \mu_h \\ &\operatorname{plim} s^2_{\hat{h}_a} = \sigma^2_h + \tfrac{1}{\gamma^2}\sigma^2_\epsilon \end{aligned} \tag{6}$$

3 Utilizing imputed values

Assume we want to use the imputed total health spending as a dependent variable in another regression in the main sample (HRS):

$$h = w\alpha + D\beta + u$$

where w is $n \times 1$ vector of wealth or potentially any other variable of interest. We are interested in α, yet total spending is missing in our main sample, therefor we impute it from a donor sample.

This is our main theorem. Note that as long as our assumptions hold, we can use Blundell et al. (2008) imputation method and preserve consistency.

Assumption 3.1.

1. *Assumption 2.1 holds.*

2. *The following holds in the underlying population:*

$$p = \gamma h + D\beta + \epsilon \tag{7}$$
$$h = \alpha w + D\beta + u$$

Where p, h, u and ϵ are $n \times 1$ vectors, D is $n \times k$, and β is $k \times 1$.

3. $E[h'\epsilon] = E[w'u] = 0$

4. $E[D'\epsilon] = E[D'u] = 0$

The following theorem states OLS method preserves its asymptotic properties while utilizing imputed values $\hat{h}$.

Theorem 3.1. *Consider applying OLS method to estimate α with imputed values $\hat{h}$:*

$$\hat{h}_H = w_H \alpha + D_H \beta + u_H$$

Under assumption 3.1, $\hat{\alpha}$ is consistent.

AppendixB explains the proof in detail. Nevertheless, The main idea of the proof is that as long as the underlying population is the same and assumption 3.1 holds, $p\text{lim}\, \ddot{w}'\ddot{h} = p\text{lim}\, \ddot{w}'\ddot{\hat{h}}$ where $\ddot{a} = (I - D(D'D)^{-1}D')\, a$, which implies the consistency of $\hat{\alpha}$.

4 Demographic Variables and Sample Distributions

In the Imputation Procedure section (equation (5)), it was noted that, when the demographic distributions of the two data sets are similar, a straightforward adjustment for differences in average out-of-pocket health expenditures would render imputed health spending in MEPS and HRS comparable. This section presents evidence supporting the similarity of the demographic distributions in these two data sets.

The first major distinction between the MEPS and HRS surveys is that HRS records long-term health expenditures. To mitigate this discrepancy, we restrict our analysis to individuals who have resided in nursing homes for less than one night.

Second, because out-of-pocket spending in HRS is imputed, there are notable outliers that may bias results. As recommended by the RAND HRS documentation, we exclude these outliers by retaining only observations with less than $20,000$ in out-of-pocket health spending. This adjustment removes fewer than five percent of cases. Third, since all individuals above age 85 are top-coded as 85 in MEPS, it is not possible to differentiate among those older than 85 years.

To avoid the age effect, we further limit the sample to individuals aged between 55 and 64. Lastly, due to differences in race definitions and oversampling practices aside from Black and White respondents across the two data sets, we focus exclusively on individuals identified within these two racial groups.

Table 1 provides a comparative summary of the demographics used in our estimations from both data sets. Given that HRS and the longitudinal MEPS panels each comprise two years per wave, we analyze the data in corresponding two-year periods.

The gender distribution in both data sets is highly comparable, with approximately 47 percent male and 53 percent female respondents in nearly every wave for both HRS and MEPS. The slightly higher proportion of females aligns with established patterns of lower female mortality rates.

Similarly, the racial composition remains consistent in both samples. Approximately 85 percent of participants in both surveys are White from 2006 through 2010; this percentage declines by about one point per survey year in MEPS through 2016. In HRS, the proportion of White respondents is roughly 85 percent until 2016, at which point it decreases to 80 percent.

Differences in marital status are more pronounced. In HRS, 70 percent of respondents were married prior to 2012, shifting to 69 percent thereafter. By contrast, marriage rates in MEPS decline from 66 percent in 2006 to 64 and 63 percent in subsequent waves, resulting in an overall difference of approximately five percentage points, with higher marriage prevalence in HRS.

Regarding high school education, 84 percent of HRS respondents had completed high school in 2006, with this figure gradually rising to 90 percent by 2016. MEPS respondents show similar trends: 85

Table 1: Comparing demographic means

HRS

	2006	2008	2010	2012	2014	2016
Sex	0.47	0.47	0.47	0.47	0.47	0.48
Race	0.86	0.86	0.84	0.83	0.84	0.80
Married	0.70	0.70	0.70	0.69	0.69	0.69
Finished Uni	0.27	0.27	0.30	0.31	0.32	0.32
Finished HS	0.84	0.85	0.88	0.88	0.89	0.90

MEPS

	2006	2008	2010	2012	2014	2016
Sex	0.47	0.47	0.47	0.47	0.48	0.47
Race	0.85	0.85	0.85	0.84	0.83	0.82
Married	0.66	0.64	0.64	0.63	0.62	0.64
Finished Uni	0.34	0.35	0.39	0.39	0.42	0.41
Finished HS	0.85	0.85	0.88	0.87	0.84	0.88

This table compares main regressors in the imputation process. Sex takes the value 1 if the respondent is male, and zero otherwise. Race takes the value of 1 if the respondent is white. Married represents the share of married couples and the last two variables illustrate the share of respondents with a university and high school degree.

percent had finished high school in 2006 and the next wave, increasing to 88 percent in 2010, declining to 87 and then 84 percent, before returning to 88 percent. Although the completion rates are similar, the temporal patterns differ across data sets.

University education reveals somewhat greater variation. In the HRS sample, 27 percent of respondents held a university degree in 2006 and 2010, with this share consistently increasing to 32 percent by 2016. In MEPS, the proportion is higher, beginning at 34 percent in 2006 and steadily climbing to 42 percent by 2014, then decreasing slightly to 41 percent in the subsequent wave. Overall, there is an approximate eight percentage point differential between the two samples.

Table 2 reports the results of following regression:

$$D_{it} = \alpha_{it} + \beta Meps_{it} + \epsilon_{it}$$

Where D_{it} represents demographic variables and $Meps_{it}$ indicates if the respondent belongs to MEPS dataset. As reported, the differences in sex and race between the two datasets are insignificant across all years. This finding aligns with the results presented in Table 1, where the average values of these variables remain nearly identical throughout the period.

Table 2: Differences in MEPS and HRS Demographics

	2006	2008	2010	2012	2014	2016
Male	0.00	0.01	0.00	0.01	0.01	0.00
	(0.86)	(0.63)	(0.76)	(0.62)	(0.53)	(0.89)
White	-0.00	-0.01	0.00	-0.00	-0.01	-0.00
	(0.92)	(0.39)	(0.76)	(0.36)	(0.13)	(0.64)
Married	-0.05***	-0.06***	-0.06***	-0.06	-0.07***	-0.05***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
University Degree	0.07***	0.08***	0.08***	0.09***	0.09***	0.07***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Highschool Degree	-0.01	-0.02*	-0.01*	-0.02***	-0.06***	-0.03***
	(0.41)	(0.05)	(0.07)	(0.00)	(0.00)	(0.00)

This table represents the result of testing the hypothesis of difference in demographics between MEPS and HRS datasets. Numbers in parenthesis or p-values. Each number represents the value of difference between MEPS and HRS. While the number below represents p-value for statistical significance

In all years, the proportion of married individuals is higher in HRS than in MEPS, whereas a smaller share of HRS respondents appear to hold a university degree. The difference is approximately 5–7% for marital status and 7–9% for educational attainment. Although these discrepancies could theoretically pose issues, their effect appears negligible, as demonstrated in the imputation section, primarily because the corresponding regression coefficients are small.

In summary, demographics across MEPS and HRS are closely aligned. Therefore, if the assumptions articulated in previous sections are correct, any observed difference in mean values between MEPS and HRS should be attributable to disparities in out-of-pocket health expenditures.

5 Results

5.1 Imputation within MEPS

Table 3 compares the true and imputed total expenditures. According to equation (3), the difference between the true and imputed totals depends on the consistency of our estimated parameters and the exogeneity of our regressors. Our imputation method successfully reproduces the mean values, with differences of less than 0.1 in all years, and in some cases, the exact mean is predicted. This aligns with (4), where we noted that, with consistent estimators, the imputation should closely replicate the mean.

Table 3: True vs Imputed Total Expenditure - MEPS

	Data				Imputation			
	mean	sd	iqr	skewness	mean	sd	iqr	skewness
2006	8.6	1.4	1.6	-0.10	8.6	1.7	2.0	-0.19
2008	8.7	1.4	1.7	-0.13	8.8	1.8	2.1	-0.20
2010	8.7	1.5	1.9	-0.11	8.7	2.0	2.4	-0.16
2012	8.6	1.6	1.9	-0.12	8.7	2.0	2.5	-0.19
2014	8.7	1.5	1.9	-0.13	8.8	2.1	2.5	-0.22
2016	8.8	1.6	1.9	-0.11	8.9	2.3	2.8	-0.17

This table summarizes different distribution moments in actual and imputed total medical expenditure. The left box represents actual moments and the right box represents the imputed data moments. In each box, the first column represents mean, the second represents standard deviation, the third represents interquantile range and the forth represents quantile based skewness.

In terms of standard deviation, the differences are more pronounced, and our imputation does not consistently reproduce the observed standard deviation. This outcome is consistent with (4), where we noted that even when estimates are consistent, the predicted standard deviation may differ depending on the residual variance and its covariance with total spending. One potential remedy would be the adoption of an instrumental variable (IV), which lies beyond the scope of this paper.

Both interquantile ranges and standard deviations are smaller in the original data than in the imputed data, yet their ratio remains constant across all cases. This indicates that our imputation method tends to preserve the relative spread of the actual data. Heuristically, this ratio is approximately 1.2 in all years, compared to 1.3 for a normal distribution, suggesting that the dispersion is not markedly different from that of a normal distribution. Skewness levels are also comparable, although we rely on quantile-based skewness to mitigate the influence of outliers. Overall, the imputed dataset exhibits a slightly greater left skew.

5.2 Imputing total medical spending in HRS

Table 4 compares the imputed data in MEPS and HRS. Note that mean values are not exactly the same in these two data sets. In 2006 and 2008, the difference in average amount of total spending in MEPS and HRS is less than 10 percent, but then this difference increases to 30 and 40 percent in following years.

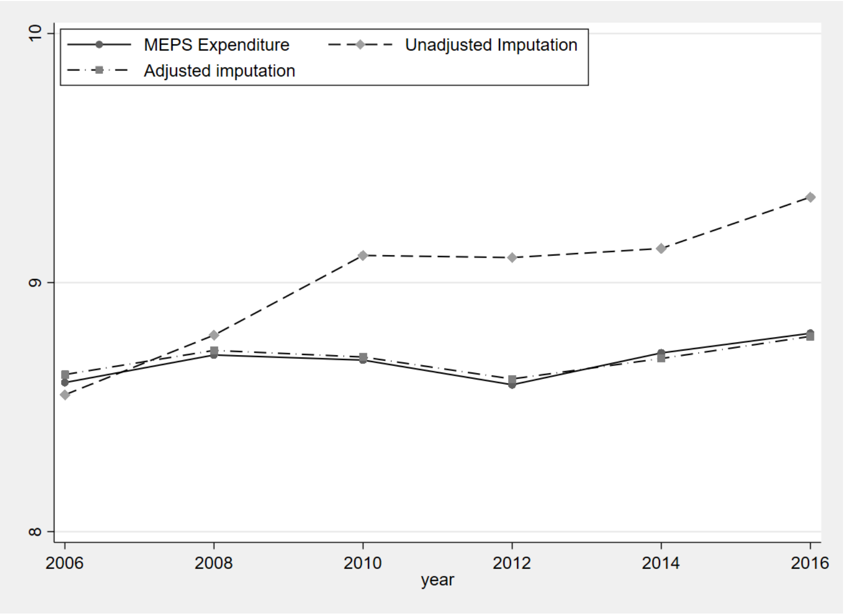
Table 4: Imputed Total Expenditure - MEPS vs. HRS

	MEPS				HRS				Mean Difference		SD Difference		Adjusted HRS
	mean	sd	iqr	skewness	mean	sd	iqr	skewness	Out of Pocket	Total	Out of Pocket	Total	mean
2006	8.64	1.72	2.00	-0.19	8.55	1.83	2.26	-0.16	-0.06	-0.09	-0.06	-0.09	8.63
2008	8.75	1.85	2.09	-0.20	8.79	1.79	2.24	-0.14	0.04	0.03	0.04	0.03	8.73
2010	8.74	2.02	2.44	-0.16	9.11	1.91	2.47	-0.10	0.27	0.36	0.27	0.36	8.70
2012	8.66	2.04	2.49	-0.19	9.10	1.93	2.47	-0.13	0.32	0.44	0.32	0.44	8.61
2014	8.78	2.11	2.55	-0.22	9.14	1.99	2.62	-0.10	0.29	0.35	0.29	0.35	8.70
2016	8.86	2.35	2.83	-0.17	9.34	2.20	2.93	-0.10	0.34	0.48	0.34	0.48	8.78

The difference in standard deviation and interquantile range is small in most years and amounts to less than 0.1 in all years. This implies that the imputed total spending in HRS mimics the spread we observe in actual MEPS data, since the imputed and actual MEPS data have similar spreads. The imputed HRS data is slightly less left skewed compared to MEPS.

Out of pocket spending strongly explains the difference in total imputed spending. Consistent with Equation (5), the last two columns of Table 4 shows that about 80 percent of the difference in mean imputed total spending is due to the difference in out of pocket spending. Similarly, the difference in standard deviation is greatly explained by the standard deviation in out of pocket spending. In the last column, we see can see the adjusted mean closely follows the imputed MEPS data as depicted in 1.

Figure 1: Total Medical Spending Mean Comparison



6 Conclusion

References

Attanasio, O. and Pistaferri, L. (2014). Consumption Inequality over the Last Half Century: Some Evidence Using the New PSID Consumption Measure. *American Economic Review*, 104(5):122–126.

Blundell, R., Pistaferri, L., and Preston, I. (2008). Consumption inequality and partial insurance. *American Economic Review*, 98(5):1887–1921.

Blundell, R. W., Pistaferri, L., and Preston, I. (2004). Imputing consumption in the psid using food demand estimates from the cex. Technical report, IFS Working Papers.

Bollinger, C. R. and Minier, J. (2015). On the Robustness of Coefficient Estimates to the Inclusion of Proxy Variables. *Journal of Econometric Methods*, 4(1):101–122.

Browning, M. and Crossley, T. (2009). Are two cheap, noisy measures better than one expensive, accurate one? *The American Economic Review*, 99(2):99–103.

Browning, M. and Leth-Petersen, S. (2003). Imputing consumption from income and wealth information. *The Economic Journal*, 113(488):F282–F301.

Campos, R. G. and Reggio, I. (2014). Measurement error in imputation procedures. *Economics Letters*, 122(2):197–202.

Crossley, T. F., Levell, P., and Poupakis, S. (2022). Regression with an imputed dependent variable. *Journal of Applied Econometrics*, 37(7):1277–1294.

De Nardi, M. and Fella, G. (2017). Saving and wealth inequality. *Review of Economic Dynamics*, 26:280–300.

De Nardi, M., French, E., Jones, J. B., and McGee, R. (2025). Why do couples and singles save during retirement? household heterogeneity and its aggregate implications. *Journal of Political Economy*, 133(3):750–792.

Palumbo, M. G. (1999). Uncertain medical expenses and precautionary saving near the end of the life cycle. *The Review of Economic Studies*, 66(2):395–421.

Skinner, J. (1987). A superior measure of consumption from the panel study of income dynamics. *Economics Letters*, 23(2):213–216.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Ziliak, J. P. (1998). Does the choice of consumption measure matter? an application to the permanent-income hypothesis. *Journal of Monetary Economics*, 41(1):201–216.

A Model Appendix

A.1 Linear Prediction

Consider two samples χ_m and χ_d with the same underlying population. The goal is to impute the variable y into our main sample χ_m from the donor sample ${\chi_d}$[3]. As mentioned in section 1, there are different methods for imputing a variable. The simplest approach is to use linear prediction. Assume the following relationship holds in the population:

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{8}$$

Where x_i largely explains y_i.

From the donor data set the imputed values are:

$$\hat{y}_{id} = \hat{\alpha}_d + \hat{\beta}_d X_{id}$$

[3] This is the original method proposed by Skinner (1987)

The following relationship holds:

$$y_{id} = \hat{y}_{id} + \hat{\epsilon}_{id}$$

Where $\hat{\alpha}_d$ and $\hat{\beta}_d$ are OLS estimators:

$$\hat{\alpha}_d, \hat{\beta}_d \in argmin_{\alpha,\beta} \sum_i \epsilon_{id}^2$$

Define the sample mean, variance, and covariance as:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$
$$s_x^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$$
$$s_{xy} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N}$$

Since $\bar{\hat{\epsilon}}_d = 0$, mean equivalence ($\bar{\hat{y}}_d = \bar{y}_d$) implies:

$$\hat{y}_{id} - \bar{\hat{y}}_d + \hat{\epsilon}_{id} = y_i - \bar{y}_d$$

With some manipulation:

$$\sum_i \frac{\left(\hat{y}_{id} - \bar{\hat{y}}_d\right)^2}{N_d} + 2\sum_i \frac{\left(\hat{y}_i - \bar{\hat{y}}_d\right)\hat{\epsilon}_{id}}{N_d} + \sum_i \frac{\hat{\epsilon}_{id}^2}{N_d} = \sum_i \frac{\left(y_i - \bar{y}_d\right)^2}{N_d}$$

Since $\hat{y}_i - \bar{\hat{y}}_d = \hat{\beta}_d(x_i - \bar{x})$ and $s_{x\hat{\epsilon}} = 0$, the second term on the left hand side is zero:

$$s_{\hat{y}}^2 = s_y^2 - s_{\hat{\epsilon}}^2$$

Therefore, variance is not consistent:

$$p\text{lim } s_{\hat{y}}^2 = p\text{lim } s_y^2 - p\text{lim } s_{\hat{\epsilon}}^2$$

Applying the Law of Large Numbers:

$$Var(\hat{y}) = Var(y) - Var(\epsilon)$$

Furthermore, inserting the imputation directly as a regressor could cause measurement error bias. define $\tilde{y}_i$ in the main data set:

$$\tilde{y}_i = \hat{\alpha}_d + \hat{\beta}_d X_i \tag{9}$$

Assume the following linear relationship holds in the population of study:

$$Z_i = a + b\ y_i + u_i \tag{10}$$

Where:

$$b = \frac{Cov(z,y)}{Var(y)}$$

Since y is imputed:

$$\tilde{b} = \frac{s_{\tilde{y}z}}{s_{\tilde{y}}^2}$$

Where:

$$s_{\tilde{y}z} = \hat{\beta}_d\ s_{x,z} = \frac{s_{yx}}{s_x^2} s_{xz}$$

$$s_{\hat{y}}^2 = \hat{\beta}_d^2 \; s_x^2 = \frac{s_{yx}^2}{s_x^2}$$

Therefore:

$$\tilde{b} = \frac{s_{xz}}{s_{yx}}$$

Assume OLS coefficients in the donor data set are consistent:

$$p\text{lim}\,\tilde{b} = \frac{Cov(z,x)}{Cov(y,x)}$$

from (10):

$$Cov(z,x) = b\; Cov(y,x) + Cov(u,x)$$

Therefore:

$$p\text{lim}\,\tilde{b} = b + \frac{Cov(u,x)}{Cov(y,x)}$$

The OLS estimate is consistent, if y affects z only through x (i.e. $Cov(u,x) = 0$).

A.2 Inverse linear relationship

The main advantage here is arriving at consistent estimates with imputed dependent variable.[4] Assume the following relationship holds in the population of study:

$$x_i = \alpha + \beta y_i + \epsilon_i \tag{11}$$

Using this relationship, our target variable can be defined as:

$$y_i = \frac{x_i - \alpha - \epsilon_i}{\beta}$$

with imputation values defined:

$$\hat{y}_i = \frac{x_i - \hat{\alpha}_d}{\hat{\beta}_d}$$

Replacing x_i from (11):

$$\hat{y}_i = \frac{\alpha - \hat{\alpha}_d}{\hat{\beta}_d} + \frac{\beta}{\hat{\beta}_d}\; y_i \;\; + \frac{\epsilon_i}{\hat{\beta}_d}$$

Assume consistent OLS estimates[5], mean independence[6], and mean-zero error term[7]:

$$p\text{lim}\;\bar{\hat{y}} = \mu_y$$

Where μ_y is the population mean of y. And for variance we have:

$$p\text{lim}\; s_{\hat{y}}^2 = \sigma_y^2 + \frac{\sigma_\epsilon^2}{\beta}$$

Where σ^2 denotes the population variance. Hence, imputation over estimates the variance term.

In our main data set:

[4] This method is introduced by Blundell et al. (2004).

[5] $p\text{lim}\,\hat{\beta}_d = \beta$, and $\text{plim}\,\hat{\alpha}_d = \alpha$.

[6] $E(y|\epsilon) = 0$.

[7] $E(\epsilon) = 0$.

$$\tilde{y}_i = \frac{x_i - \hat{\alpha}_d}{\hat{\beta}_d}$$

Therefore:

$$p\text{lim } \bar{\tilde{y}} = \mu_y$$

$$p\text{lim } s^2_{\tilde{y}} = \sigma^2_y + \frac{\sigma^2_\epsilon}{\beta}$$

Imputation provides mean consistency, not variance consistency.
Finally, Assume the following relationship holds:

$$y_i = \theta_0 + \theta_1 z_i + e_i$$

Therefore, the true slope coefficient is:

$$\theta_1 = \frac{Cov(y,z)}{Var(z)}$$

Since imputed values are utilized:

$$\hat{\theta}_1 = \frac{s_{\hat{y}z}}{s^2_z}$$

The following holds by the definition of $\tilde{y}$:

$$\tilde{y}_i = \frac{s^2_y}{s_{xy}} x_i - \hat{\alpha}_d$$

Therefore:

$$s_{\tilde{y}z} = \frac{s^2_y}{s_{xy}} s_{xz}$$

Therefore if we have a large enough sample:

$$\tilde{\theta} = \frac{s^2_y}{s_{xy}} \frac{s_{xz}}{s^2_z}$$

Implying:

$$p\text{lim } \tilde{\theta} = \frac{Var(y)}{Cov(x,y)} \frac{Cov(x,z)}{Var(z)}$$

From (11):

$$\begin{aligned} Cov(x,z) &= \beta Cov(y,z) + Cov(\epsilon, z) \\ &= \tfrac{Cov(x,y)}{Var(y)} Cov(y,z) + Cov(\epsilon, z) \end{aligned}$$

Replacing in the right hand side of the $\tilde{\theta}_1$:

$$\begin{aligned} p\text{lim } \tilde{\theta}_1 &= \tfrac{Var(y)}{Cov(x,y)} \tfrac{Cov(x,y)}{Var(y)} \tfrac{Cov(y,z)}{Var(z)} + \tfrac{Var(y)}{Var(z)Cov(x,y)} \, Cov(\epsilon, z) \\ &= \theta_1 + \tfrac{Var(y)}{Var(z)Cov(x,y)} \, Cov(\epsilon, z) \end{aligned}$$

Hence, $\tilde{\theta}_1$ is consistent if z alters x entirely through y and not any other channel (i.e $Cov(z, \epsilon) = 0$).

B Proof of The Main Theorem

Assumption B.1. *6.1 The following holds.*

1. *Assumption 2.1 holds.*
2. *The following relationship holds in the population:*

$$\begin{aligned} p &= \gamma h + D\beta + \epsilon \\ h &= \alpha w + D\beta + u \end{aligned} \tag{12}$$

   Where p, h, u and ϵ are $n \times 1$ vectors, D is $n \times k$, and β is $k \times 1$.

3. $E[h'\epsilon] = E[w'u] = 0$
4. $E[D'\epsilon] = E[D'u] = 0$
5. $P_A := A(A'A)^{-1}A'$
   Where A is any $n \times k$ matrix.
6. $M_A := I - P_A$
7. $\ddot{b} := M_D\, b$

Theorem B.1. *Under assumption B.1, $\hat{\alpha}$ is consistent.*

We can rewrite (12) as:

$$\ddot{p} = \gamma\ddot{h} + M_D\epsilon \tag{13}$$

With OLS coefficient $\hat{\gamma}_M$:

$$\hat{\gamma} = (\ddot{h}'_M\ddot{h}_M)^{-1}\ddot{h}'_M\ddot{p}_M$$

Next, to find $\hat{\beta}$ from (12):

$$D_M\hat{\beta} = P_{D_M}\,(p_M - h_M\hat{\gamma} - \hat{\epsilon}_M)$$

Multiply $(D'_M D_M)^{-1}D'_M$:

$$\hat{\beta} = (D'_M D_M)^{-1}D'_M(p_M - h_M\hat{\gamma})$$

Finally to achieve imputed values we start from (12), and solve for y:

$$h = \frac{1}{\gamma}(p - D\beta - \epsilon)$$

Then imputed h is defined as:

$$\hat{h}_H = \frac{1}{\hat{\gamma}}(p_H - D_H\hat{\beta})$$

To show $\hat{\gamma}$ is consistent rewrite $\hat{\gamma}$ as:

$$\begin{aligned} \hat{\gamma} &= (\ddot{h}'_M\ddot{h}_M)^{-1}\ddot{h}'_M\ddot{h}_M\gamma + (\ddot{h}'_M\ddot{h}_M)^{-1}\ddot{h}'_M\,\epsilon_M \\ &= \gamma + (\tfrac{\ddot{h}'_M\ddot{h}_M}{n_M})^{-1}\tfrac{\ddot{h}'_M\epsilon_M}{n_M} \\ &= \gamma + (\tfrac{\ddot{h}'_M\ddot{h}_M}{n_M})^{-1}\tfrac{h'\left(I - D_M(D'_M D_M)^{-1}D'_M\right)\epsilon_M}{n_M} \end{aligned}$$

Since $E[D'\epsilon] = 0$ Therefore:

$$p\text{lim}\ \hat{\gamma} = \gamma$$

Finally, consistency of $\hat{\gamma}$, implies consistency of $\hat{\beta}$:

$$\begin{aligned} p\text{lim } \hat{\beta} &= p\text{lim}\,(D_M' D_M)^{-1} D_M' (\gamma h_M + D_M \beta + \epsilon_M - h_M \hat{\gamma}_M) \\ &= \beta + (D_M' D_M)^{-1} D_M' h_M (\gamma - p\text{lim } \hat{\gamma}) + p\text{lim}\,(\tfrac{D_M' D_M}{n})^{-1}\, p\text{lim } \tfrac{D_M' \epsilon_M}{n} \\ &= \beta \end{aligned}$$

Next, we show that $M_D \hat{h}$ is consistent as long as our assumptions hold:

$$M_{D_H} \hat{h}_H = \frac{1}{\hat{\gamma}}\, M_{D_H} \left(p_H - D_H \hat{\beta}\right)$$

using $M_{D_H}\, P_{D_H} = 0$:

$$M_{D_H} \hat{h}_H = \frac{1}{\hat{\gamma}}\, \ddot{p}_H$$

Replacing $\ddot{p}$ from (13):

$$M_{D_H} \hat{h}_H = \frac{\gamma}{\hat{\gamma}} \ddot{h}_H + \frac{n_H}{\hat{\gamma}} \frac{M_{D_H} \epsilon_H}{n_H}$$

Therefore, by consistency of $\hat{\gamma}_M$ and since $E[D'\epsilon] = E[\epsilon] = 0$:

$$p\text{lim}\, M_{D_H} \hat{h}_H = \ddot{h}_H$$

Next, we show that using imputed values would not bias OLS estimate of α in (12), as long as OLS estimates are unbiased themselves. Rewrite (12) as:

$$\ddot{h} = \alpha \ddot{w} + \ddot{u}$$

Then OLS estimate of α is consistent:

$$\begin{aligned} p\text{lim}\, \hat{\alpha} &= p\text{lim}(\ddot{w}'\ddot{w})^{-1} \ddot{w}' \ddot{h} \\ &= \alpha + p\text{lim}(\ddot{w}'\ddot{w})^{-1} w' [u - D(D'D)^{-1} D'u] \\ &= \alpha \end{aligned}$$

Finally, Since $p\text{lim}\, M_{D_H} \hat{h}_H = \ddot{h}$, utilizing imputed values would not bias OLS estimates:

$$\begin{aligned} p\text{lim } \hat{\alpha} &= p\text{lim}(\ddot{w}'\ddot{w})^{-1} \ddot{w}' M_{D_H} \hat{h}_H \\ &= (\ddot{w}'\ddot{w})^{-1} \ddot{w}'\ p\text{lim}\, M_{D_H} \hat{h}_H \\ &= p\text{lim}(\ddot{w}'\ddot{w})^{-1} \ddot{w}'\, \ddot{h} \\ &= \alpha \end{aligned}$$

C Data Appendix

Since 1996, the Medical Expenditure Panel Survey has conducted large-scale surveys of individuals, families, and their healthcare providers (doctors, hospitals, pharmacies, etc.). MEPS collects data on the specific health services that Americans use, how frequently they use them, how much they cost, and how they are paid for. Furthermore, it collects data about the cost, scope, and breadth of U.S. workers' health insurance.

Currently, MEPS consists of two major components: the Household Component and the Insurance Component. Employer-based health insurance data can be found in the Insurance Component of the survey. Data from households and their members in the Household Component are complemented by data from their medical providers. This attractive property of the MEPS distinguishes it from representative surveys with measurement errors, which is why we believe there is no need for instruments to avoid measurement error bias.

In this study, we use the household component of MEPS. A nationally representative subsample of households that participated in the National Health Interview Survey (conducted by the National Center for Health Statistics) is used in the Household Component (HC) to collect data from families and individuals in selected communities across the United States. In the household interviews, MEPS collects detailed information about each household member, including demographic characteristics such as sex, race, marital status, and education. It also contains information on both totals and out-of-pocket health spending.

Through an overlapping panel design, the MEPS-HC collects data from a nationally representative sample of households. Data are collected for two calendar years for each panel of sample households selected each year. A total of five rounds of interviews are conducted for each panel over a two-and-a-half-year period. Health care expenditures are estimated continuously and annually at both individual and household levels. To be consistent with HRS sampling (described next), we use MEPS Longitudinal Data Files. These files contain the information on the respondents in each MEPS household panel for two years.

Health and Retirement Survey (HRS) is a longitudinal study funded by the National Institute on Aging and the Social Security Administration that surveys a representative sample of approximately 20,000 people in America. Research about the challenges and opportunities of aging can be addressed through the HRS's unique and in-depth interviews, which provide a growing library of multidisciplinary data.

With derived variables covering a wide range of topics, the RAND HRS Longitudinal File is a cleaned, easy-to-use version of the Health and Retirement Study (HRS). In addition to imputations for out-of-pocket medical expenditures developed at RAND, it includes demographics such as sex, race, marital status, and education.

The oldest cohort in HRS is the AHEAD cohort that was born before 1924, although they are not the first sampled group. This group was initially studied in The Study of Assets and Health Dynamics Among the Oldest Old program. Children of depression (CODA) are the second oldest cohort born between 1924 and 1930. Then we have the first interviewed group, HRS, with their first interview in 1992. War Babies (WB), Early Baby Boomers (EBB), Mid Baby Boomers (MBB), and Late Baby Boomers are the youngest cohorts. Each cohort is interviewed in some initial wave and subsequently every two years. To keep the relevance of the data, notice that after every three waves (six years), a new cohort is added to the data.