DEPARTMENT OF ECONOMICS FINANCE & ACCOUNTING

WORKING PAPERS SERIES

N263/15

Correcting for Self-Reporting Bias in BMI: A Multiple Imputation Approach.

By

Donal O'Neill

Maynooth University

NATIONAL UNIVERSITY OF IRELAND, MAYNOOTH
MAYNOOTH

CO. KILDARE

IRELAND
http://economics.nuim.ie/research/

http://repec.org/

Correcting for Self-Reporting Bias in BMI: A Multiple Imputation Approach[1].

Donal O'Neill

Maynooth University

September 2015

Abstract

Measurement error in BMI is known to be a complex process has serious consequences for traditional estimators. In this paper I examine the extent to which Stochastic Multiple Imputation approaches can successfully addressing this problem. Using both Monte Carlo simulations and real world data I show how the MI approach can provide an effective solution to measurement error in BMI in appropriate circumstances. The MI approach yields consistent estimates that efficiently use all the available data.

[1] I would like to thank participants at the Maynooth University departmental seminar series for helpful comments on an earlier draft of this paper.

1. Introduction

Obesity is an important cause of morbidity, disability and premature (WHO, 2009). Body Mass Index (BMI), defined as weight in kg/height in m^2, is the most widely used measure of obesity. However, there is a large body of evidence that shows that individuals misreport both their height and weight in surveys leading to biased estimates of BMI. Furthermore, O’Neill and Sweetman (2013) show that the popular econometric approaches that have been adopted to address problems of measurement error in BMI, such as the Regression Calibration approach and the Instrumental Variable approach, continue to exhibit significant biases. These biases reflect the non-classical nature of the measurement error in BMI. O’Neill and Sweetman (2013) propose a Multiple Imputation (MI) approach to overcome measurement error in BMI. This approach builds on work by Brownestone and Valetta (1996) who used a version of the MI approach to address concerns about measurement error in earnings, a left hand side variable in their regression analysis. In this paper I examine the extent to which the MI approach can address the measurement error in BMI, when BMI is an explanatory variable in a linear regression model. The contribution of MI to overcoming self-reported bias is illustrated using both simulations and real world data. The layout of the paper is as follows. Section 2, summarises earlier work on measurement error in BMI, with particular emphasis on the non-classical nature of the measurement error. Section 3 summarises the MI approach to missing data. Section 4 evaluates the MI approach for measurement error using Monte Carlo simulations, while section 5 applies the MI approach to a real world application examining the impact of BMI on income. Section 6 concludes.

2. Measurement Error in Self-Reported BMI

O’Neill and Sweetman (2013) provide a detailed examination of measurement error in self-reported BMI using data from both Ireland and the US. The key feature of the data sets used in their analysis is that in addition to self-reported height and weight the data also contained recorded clinical measures of height and weight. The availability of reported and recorded height and weight allowed the authors to measure the extent of measurement error in reported BMI, and also characterise the nature of this error process. O’Neill and Sweetman (2013) proceeded to examine the consequences of this error for a linear regression model relating BMI to income. Crucially they found that measurement error in self-reported BMI deviated from textbook classical error model in a two important ways. Firstly the error was correlated with the true value and secondly the error contained information about outcomes over and above that available in the true recorded measure of BMI. In the statstitics literature the second proprty is a violation of the surrogacy condition; in econometrics it is more typically referred to as differential measurement error.

The bias in the OLS estimator in these circumstances is given by

$$= \frac{\beta\{Var(X)+Cov(u,X)\}}{var(X)+var(u)+2Cov(u,X)} + \frac{Cov(u,\varepsilon)}{var(X)+var(u)+2Cov(u,X)} \qquad (1)$$

where X is the true recorded measure of BMI, u is the measurement error in reported BMI and ε is the stochastic component of the income data generating process. Differential measurement error is captured in the last term reflecting the correlation between u and ε. O’Neill and Sweetman (2013) find this correlation to be negative and substantial in bith the Irish and US data sets they examine. The regression calibration approach often used in this literature, involves using the observed X alone to impute values for the missing data. O’Neill and Sweetman (2013) show that while this approach can correct for the correlation between the measurement error in X and its true value, it fails to adequately control for the differential nature of the measurement error. In the conclusion of their paper they propose a stochastic imputation technique to adjust for differential measurement error. In the remainder of the paper we consider the effectiveness of such an approach.

3. Multiple Imputation

Schafer (1997) provides a detailed overview of the analysis of incomplete multivariate data. The simplest way to deal with missing data is to base the final analysis only on complete case observations. However, at best this approach may be highly inefficient and result in a significant amount of valid information being excluded from the final analysis. MI is an algorithm for tackling arbitrary patterns of missing data, first proposed by Rubin (1978). It uses the fact that in any incomplete data set the observed values provide indirect evidence about the likely values of the unobserved ones. This is captured in the predictive probability distribution of the missing data given the observed data, $P\left[D_{miss} \mid D_{obs}\right]$. Rather than treating missing data as a nuisance factor to be gotten ridden of, MI views the missing data as a source of variability to be averaged over. To see this note that the observed case likelihood function, $p\left(\theta \mid D_{obs}\right)=\int p\left(\theta, D_{miss} \mid D_{obs}\right) dD_{miss}$, can be rewritten as $p\left(\theta \mid D_{obs}\right)=\int p\left(\theta \mid D_{obs}, D_{miss}\right) p\left(D_{miss} \mid D_{obs}\right) dD_{miss}$. The first term inside the integral sign is just the complete case likelihood. This is often easy to calculate, even in circumstances where the observed case likelihood may be intractable. The second term is the predictive distribution of the missing data given the observed data. This expression makes it clear that the observed likelihood is obtained by averaging the complete case likelihood over the predictive distribution of the missing data.

Evaluating the necessary integral analytically may be complicated. The MI approach solves the problem by estimating the integral using stochastic simulation techniques. In particular repeated samples are drawn from $P\left[D_{miss} \mid D_{obs}\right]$. At each draw the missing data D_{miss} is replaced (imputed) using the draw from the predictive distribution. For each of these imputed samples the parameter of interest is estimated treating the imputed data as observed. Final point estimates and standard errors are obtained by combining these estimates using rules provided by Rubin (1987). Provided the draws are from the correct predictive density (often termed proper imputations) these combined estimates are standard errors will correctly reflect the missing-data uncertainty.

The key to performing proper imputation is ensuring the imputed missing data are drawn from the correct predictive distribution. In many cases researchers rely on data augmentation approach of Tanner and Wong (1987) to achieve this. However, in the case of the normal linear regression model obtaining proper draws is more straightforward (Freedman et al. 2008). To proceed in the case of measurement error assume we have a validation data set containing true values for the outcome variable Y and the regressor X, along with mismeasured value for X which we denote X^e. Many of the proposed estimators for tackling measurement error rely on such a validation data set and validation samples often arise in medical applications where collecting true values of X for everyone would prove costly. For the remainder of the sample we only have the outcome variable Y and reported X^e. Such a data scheme is illustrated in Figure 1. The MI approach to measurement error proceeds by imputing the missing values of X using observed data Y, X^e and X (when available). Fro the normal linear regression model proper imputations may be obtained as follows.

1. For the validation sample run the regression $X = \alpha_1 + \alpha_2 X^e + \alpha_3 Y + e$.
2. From this regression obtain point estimates, the variance covariance matrix of the estimator and estimated sum of squared residuals.

$$\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3 \qquad , C = \begin{bmatrix} \hat{\text{var}}(\alpha_1) & & \\ \hat{\text{cov}}(\alpha_1,\alpha_2) & \hat{\text{var}}(\alpha_2) & \\ \hat{\text{cov}}(\alpha_1,\alpha_3) & \hat{\text{cov}}(\alpha_2,\alpha_3) & \hat{\text{var}}(\alpha_3) \end{bmatrix} \qquad \textbf{and } s^2 = \sum e_i^2$$

3. Draw parameter values from posterior parameter distributions

$$\alpha_{pos} \sim MVN\left(\hat{\alpha}, C\right), \quad \sigma^2_{pos} = \frac{\mathbf{(N}_V\mathbf{\text{-}3)}\, s^2}{\chi^2_{(\mathbf{N}_V \mathbf{\text{-}3})}}$$

4. Draw a value of the error term from $\omega_{pos} \sim N\left(0, \sigma^2_{pos}\right)$
5. Generate imputed values of X using $\hat{X}_{pos} = \alpha_{pos1} + \alpha_{pos2} X^e + \alpha_{pos3} Y + \omega_{pos}$ if X is missing and X otherwise
6. Run the final regression of interests with the observed and imputed values $Y = \beta_1 + \beta_2 \hat{X}_{pos} + \upsilon$.
7. Repeat steps 3-6 m-times and combine the m-estimates using Rubin's rules to obtain final imputation estimates. The final point estimate is simply the average of the estimates across the m imputations, while the variance of the estimator is a weighted average of the average of the m *within sample* standard errors and the variance of the estimates *across samples.*

In the following section we evaluate this approach using simulated data and compare it to other popular alternatives.

4. Monte-Carlo Simulation Results

To compare alternative approaches to adjusting for measurement error we consider the following data generating process.

$$\left.\begin{array}{l} Y = .6 + .4X + \varepsilon \\ \\ X^{e} = X + u \end{array}\right\} \quad (\varepsilon, u) \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} .84 & -.1 \\ -.1 & .5 \end{pmatrix}\right)$$

For our purposes the crucial feature of this data generating process is the differential measurement error as captured by the negative correlation between u and ε. Having generated this data for everyone I randomly replace half of the true X as missing to mimic the validation data illustrated in Figure 1. I then estimate the parameters β using a number of alternative estimators. In particular I consider the naive or denial estimator which just uses the reported X^{e}, ignoring measurement error. I also consider the standard RC approach which uses the relationship between X and X^{e} in the validation sample to impute values for X in non-validation sample. I consider an amended RC approach which adds Y as an explanatory variable to the imputation process. This is equivalent to only using the first step of the MI approach to do the imputation, treating estimated parameters as fixed truths and ignoring the stochastic component of the data generating process. I also consider the complete case analysis which in my simulations throws out half of the data. Finally I consider the stochastic MI approach outlined in Section 3. In all the simulations the sample size is 1000 and I run 1000 simulations. Within each simulation the MI approach itself is repeated m times. In these simulations I choose m=10. The results are given in Figure 2.

The dark blue curve gives the sampling distribution of the OLS estimator of the slope coefficient using true X and Y for full sample. This is included for comparison purposes and is centered at the true value of .4. To the far left of the graph the denial estimator is represented by the red distribution. As expected this estimator is badly biased with a mean value equal to .2, half of the true value. This reflects the traditional attenuation bias associated with classical measurement error plus the additional downward bias resulting from the negative differential measurement error. Moving to the right in figure 2 we have regression calibration estimator represented by the black distribution. As expected, given the results in O'Neill and Sweetman (2013), this estimator is better than the naive estimator in that it corrects for classical measurement error but fails to adjust for the downward bias due to our differential error, and so remains inconsistent.

The MI estimator is given by the green distribution. Unlike the RC estimator MI Is consistent. Furthermore it is more efficient than the complete case estimator because it uses data more effectively is more efficient than the complete case approach. Finally I compare the MI estimator to the amended RC which simply adds Y to the imputation process but fails to consider the uncertain nature of parameter estimates or the role of stochastic term in the imputation process. Although the presence of the outcome variable in the imputation model is crucial for control for differential measurement eror using this approach, these results clearly show how just adding the outcome measure to the imputation process is not sufficient. Failure to properly control for the uncertainty in the imputation process attributes too much importance to the relationship between Y and X, resulting in the amended estimator being biased in the opposite direction to the traditional RC approach.

These simulations highlight the valuable role played by MI approach when dealing with differential measurement error. It provides consistent estimates in this instance, while other popular estimators struggle. In the next section I apply these approaches to the real world example considered by O’Neill and Sweetman (2013).

5. Multiple Imputation of Self-Reported BMI.

O’Neill and Sweetman (2013) show that the popular econometric approaches that have been adopted to address problems of measurement error in BMI, such as the RC approach and the Instrumental Variable approach, continue to exhibit significant biases. This largely reflects the differential nature of the measurement error in self-reported BMI, a feature of the error process highlighted in their study using both Irish and US data. They suggested a stochastic multiple imputation approach to correct for self-reported measurement error. In this section I apply the imputation procedure discussed above to the Irish Growing Up in Ireland (GUI) survey. The GUI data tracks the development of a cohort of Irish children born between November 1997 and October 1998. The data used for this analysis are from the first wave of interviews, which were carried out between August 2007 and May 2008. In addition to self-reported measures of height and weight, the GUI also contains independent measures of the respondent’s height and weight. I refer to the latter as recorded measures and treat them as the true height and weight of the respondents. In the GUI sample the recorded measures were obtained by the interviewer in the respondent’s home at the end of the interview. The respondent was unaware that these measurements would be taken at the time they were providing their self-reported measures. Although recorded BMI is available for everyone for the purposes of exposition I only use the recorded BMI indirectly in the imputation process. This is true for both the regression calibration approach and the MI approach. I restrict the GUI sample to biological mothers of the study child who were not pregnant at the time of the study I am left with a working sample size of 6637.

42.65% (13.9%) of mothers in the GUI sample are overweight (obese) on the basis of self-reported data. However, the true numbers are 49.55% and 17.34%. This illustrates a clear tendency for individuals to underestimate their BMI in self-reported data. In addition O'Neill and Sweetman (2013) show that this error is both non-classical and differential and illustrate the consequences of such an error process by examining the relationship between income and obesity.

In this section we replicate some of the earlier result sof O'Neill and Sweetman (2013) but in addition consider the new stochastic MI approach. The results are given in Table 1. The first column reports the results based on recorded BMI and shows a significant negative correlation between BMI and income. Column 2, reports the results when self-reported BMI is used in place of recorded BMI. In contrast to what one would expect with classical measurement error use of reported BMI overstates the true effect. Column 3 shows that the traditional regression calibration approach does nothing to alleviate the bias, with the RC estimator in this case being almost identical to the simple denial estimator based only on reported BMI. The similarity of the two estimators reflects a combination of offsetting biases in both estimators and warns researchers against using the similarity of two estimators to make inferences on the presence or nature of measurement error. Adding income to RC model does not help our estimation and if anything only exaggerates the bias further. Finally, column 5 shows the results for the MI estimator. In keeping with the simulation results presented earlier the MI point estimate is very similar to the true estimate, although the standard error is somewhat larger.

These results illustrate in practice how the MI approach can be combined with internal validation data to overcome problems of self-reported BMI, despite the complicated nature of the error process.

6. Conclusion

Measurement error in BMI is known to be a complex process involving non-classical and differential measurement error. This can cause serious problems for traditional estimators, with the magnitude of the biases often uncertain and potentially opposite to what one would expect from classical textbook model. In this paper I examine the role of MI in addressing this problem. Using both Monte Carlo simulations and real world data I show how the MI approach can provide an effective solution to measurement error in BMI.

References

Angrist, J., Krueger, A., 1999. Empirical Strategies in Labor Economics. In Ashenfelter, A., Card, D., (Eds) Handbook of Labor Economics, Vol. 3A, Amsterdam and New York, Elsevier Science.

Angrist, J., Krueger, A., 2001. Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. The Journal of Economic Perspectives 15, 4, 69-86.

Bauhoff, S. 2011. Systematic self-report bias in health data: impact on estimating cross-sectional and treatment effects. Health Service Outcomes Research Methods 11, 44-53.

Biener, A., C. Meyerhoefer and J. Cawley (2014), "Estimating the Meidcal Care costs of Yotuh Obesity in the Presence of Proxy Reporting Error," mimeo Lehigh University.

Black, D., Berger, M., Scott, F., 2000. Bounding Parameter Estimates with Nonclassical Measurement Error. Journal of the American Statistical Association 95, 451, 739-748.

Bolin, K., Cawley, J., 2007. The Economics of Obesity, Advances in Health Economics and Health Services Research 17.

Bound, J., Krueger, A., 1991. The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right? Journal of Labor Economics 9, 1-24.

Bound, J., Brown, C., Duncan, G., 1994. Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data. Journal of Labor Economics 12, 345-368.

Bound, J., Jaeger, D.A., Baker, R.M., 1995. Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak. Journal of the American Statistical Association 90, 443 – 450.

Bound, J., Brown, C., Mathiowetz, N., 2001, Measurement Error in Survey Data. In Heckman, J.J., Leamer, E.E., (Eds) Handbook of Econometrics Vol. 5, North-Holland, Amsterdam, 3707-3843.

Brisbois, T., Farmer, A., McCargar, J., 2012. Early Markers of Adult Obesity: A Review. Obesity Review 13(4), 347-67.

Brownstone, D. and Valletta, 1996. Modelling Earnings Measurement Error: A Multiple Imputation Approach. The Review of Economics and Statistics, 78(4), 705-717.

Brunello, G., d'Hombres, B., 2007. Does Body Weight Affect Wages? Evidence from Europe. Economics and Human Biology 5, 1-19.

Burkhauser, R., Cawley, J., 2008. Beyond BMI: The Value of More Accurate Measures of Fatness and Obesity in Social Science Research. Journal of Health Economics 27, 519-529.

Carroll, R.J., Ruppert, D., Stefanski, L., 1994. Measurement Error in Nonlinear Models London: Chapman and Hall.

Cawley, J., 2000. An Instrumental Variables Approach to Measuring the Effect of Body Weight on Employment Disability. Health Services Research 35, 1159–1179.

Cawley, J., 2004 The Impact of Obesity on Wages. Journal of Human Resources 39, 451-474.

Cawley, J., Meyerhoefer,C., 2012. The Medical Care Costs of Obesity: An instrumental Variables Approach. *Journal of Health Economics*, 31, 219-230.

Conor Gorber, S., M. Temblay, D. Moher and B. Gorver (2007), "A Comparison of Direct vs Self-Reported Measures for Assessing Height, Weight and Body Mass Index: A Systematic Review," Obesity Reviews, 8, pp. 307-326.

Davey Smith, G., Sterne, J., Fraser, A., Tynelius, Lawlor, D., 2009. The Association Between BMI and Mortality using Offspring BMI as an Indicator of own BMI: Large Intergenerational Mortality Study. British Medical Journal 339, b5043.

Elgar, F.J., Stewart, J.M., 2008. Validity of Self-Report Screening for Overweight and Obesity. Evidence from the Canadian Community Health Survey. Canadian Journal of Public Health 99(5), 423-7.

Elgar, F.J., Roberts, C., Tudor-Smith, C., Moore, L., 2005. Validity of Self-Reported Height and Weight and Predictors of Bias in Adolescents. Journal of Adolescent Health 37(5), 371-5.

Freedman, L., Midthune, D., Carroll, R., Kipnis, V., 2008. A comparison of Regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression, Statistics in Medicine, 27, 5195-5216.

Fuller, W., 1987. Measurement Error Models, Wiley and sons, New York.

Gottschalk, P., Huynh, M., 2010. Are Earnings Inequality and Mobility Overstated? The Impact of Nonclassical Measurement Error. The Review of Economics and Statistics 92(2), 302-315.

Guo, Y., Little, R., 2011. Regression analysis with Covariates that have Heteroscedastic Measurement Error. Statistics in Medicine 30, 2278–2294. DOI: 10.1002/sim.4261

Hyslop, D., Imbens, G., 2001. Bias from Classical and Other Forms of Measurement Error. Journal of Business And Economic Statistics 19(4), 475-481.

Johansson, E., Bockerman, P., Kiiskinen, U., Heliovaara, M., 2009. Obesity and Labour Market Success in Finland: The Difference Between Having a High BMI and Being Fat. Economics and Human Biology 7, 36-45.

Kaestner, R., Grossman, M., 2009. Effects of Weight on Children's Educational Achievement, Economics of Education Review 28, 651-661.

Kim, B., Solon, G., 2005. Implications of Mean-Reverting Measurement Error for Longitudinal Studies of Wages and Employment. Review of Economics and Statistics 87, 193-196.

Kline, B., Tobias, J., 2008, The Wages of BMI: Bayesian Analysis of a Skewed Treatment Response Model with Nonparametric Endogeneity. Journal of Applied Econometrics 23, 767-793.

Lindeboom, M., Lundborg, P., van der Klaauw, B., 2010. Assessing the Impact of Obesity on Labor Market Outcomes. Economics and Human Biology 8, 309−319.

Little, R. and D. Rubin (2002) Statistical Analysis with Missing Data (2nd Edition), Wiley and sons.

O’Neill D, Sweetman O (2013). The Consequences of Measurement Error when Estimating the Impact of Obesity on Income. *IZA Journal of Labor Economics*, 2013; **2**:3 doi:10.1186/2193-8997-2-3

Plankey, M., Stevens, J., Flegal, K., Rust, P. 1997. Prediction Equations do Not Eliminate Systematic Error in Self-Reported Body Mass Index. Obesity Research 5(4), 308-314.

Pischke, J., 1995. Measurement Error and Earnings Dynamics: Some Estimates from the PSID Validation Study. Journal of Business and Economic Statistics 13(3), 305-314.

Rigobon, R. 2003. Identification through Heteroscedasticity. The Review of Economics and Statistics. 85(4), 777-792.

Rubin (1978) “Multiple Imputation in Sample surveys,” *Proc. Survey Res. Methods Sec., American Statistical Association* 20-34.

Rubin (1987) “Multiple Imputation Non-Response in Surveys,” J.Wiley and sons New York.

Schafer, J (1997) Analysis of Incomplete Multivariate Data, Chapman and Hall

Schafer, J. and M. Olsen (1998), Multiple Imputation for Multivariate missing data problems, A Data Analysts Perspective,” *Multivariate Behavioural Research,* 34(4): pp. 545-71.

Spencer, E.A., Appleby, P.N., Davey, G.K., Key, T.J., 2002. Validity of Self-Reported Height and Weight in 4808 EPIC-Oxford Participants. Public Health Nutrition 5(4), 561-565.

Tanner, M. and W. Wong (1987), The calculations of posterior distributions by Data Augmentation, *Journal of American statistical Association,* 82, 528-550.

Villanueva, E., 2001. The Validity of Self-reported Weight in US adults: a Population Based Cross-Sectional Study. BMC Public Health 1(11).

Wada, R., Tekin, E., 2010. Body Composition and Wages. Economics and Human Biology 8, 242-254.

WHO, 2009. Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks. World Health Organisation, WHO Press.

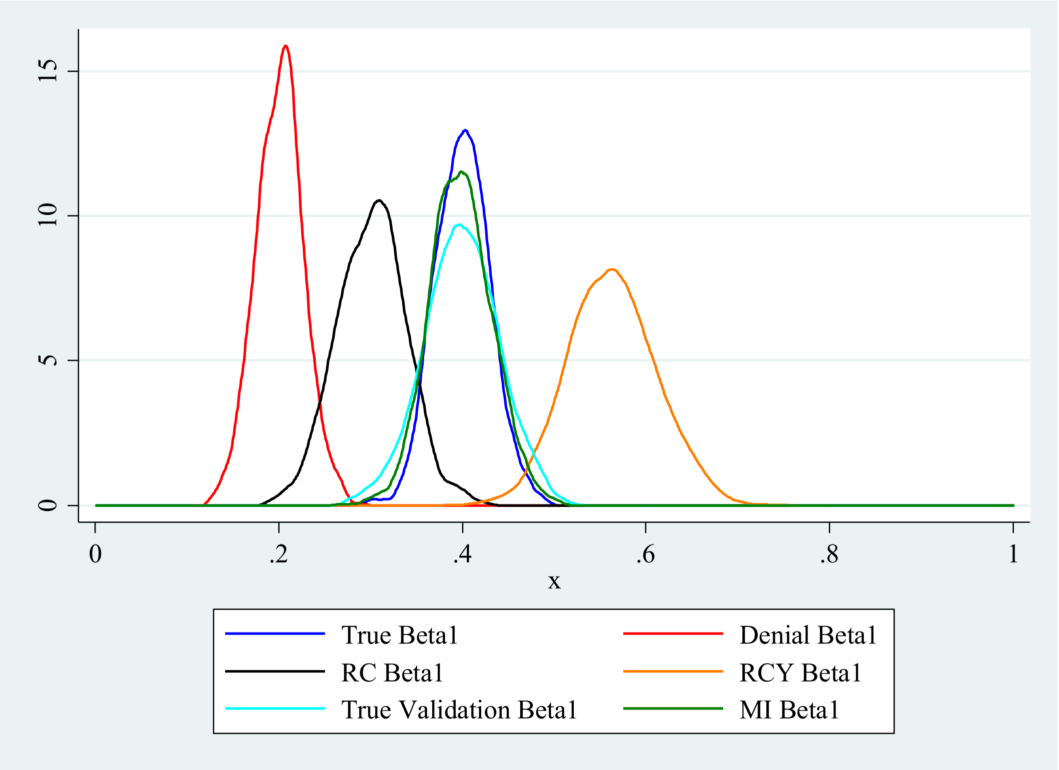
Table 1: Alternative Estimates of the relationship between BMI and Income

	OLS (True)	OLS (Self-Reported)	Regression Calibration	Regression Calibration +y	Multiple Imputation
BMI	-.82 (.09)	-.92 (.097)	-.918 (.097)	-.93 (.097)	-.83 (.103)
N	6637	6637	6637	6637	6637

Figure 1: Validation Data Set

Y	X	X^e
√	√	√
√	√	√
√	√	√
√	.	√
√	.	√
√	.	√
√	.	√

Figure 2: Monte-Carlo Simulation Results with Differential Measurement Error


x
True Beta1
Denial Beta1
RC Beta1
RCY Beta1
True Validation Beta1
MI Beta1