Economics and Human Biology 19 (2015) 275–293

Contents lists available at ScienceDirect

Economics and Human Biology

journal homepage: http://www.elsevier.com/locate/ehb

Adjusting body mass for measurement error with invalid validation data☆

Charles Courtemanche [a,b], Joshua C. Pinkston [c,*], Jay Stewart [d,e]

[a] Georgia State University, United States
[b] United States
[c] University of Louisville, United States
[d] Bureau of Labor Statistics, United States
[e] IZA, Germany

ARTICLE INFO

Article history:
Received 7 October 2014
Received in revised form 24 February 2015
Accepted 30 April 2015
Available online 21 May 2015

Keywords:
Body mass index
Obesity
Measurement error
Validation data

ABSTRACT

We propose a new method for using validation data to correct self-reported weight and height in surveys that do not measure respondents. The standard correction in prior research regresses actual measures on reported values using an external validation dataset, and then uses the estimated coefficients to predict actual measures in the primary dataset. This approach requires the strong assumption that the expectations of measured weight and height conditional on the reported values are the same in both datasets. In contrast, we use percentile ranks rather than levels of reported weight and height. Our approach requires the weaker assumption that the conditional expectations of actual measures are increasing in reported values in both samples. This makes our correction more robust to differences in measurement error across surveys as long as both surveys represent the same population. We examine three nationally representative datasets and find that misreporting appears to be sensitive to differences in survey context. When we compare predicted BMI distributions using the two validation approaches, we find that the standard correction is affected by differences in misreporting while our correction is not. Finally, we present several examples that demonstrate the potential importance of our correction for future econometric analyses and estimates of obesity rates.

1. Introduction

Obesity, defined as having a body mass index (BMI) of at least 30, is associated with heart disease, diabetes, high blood pressure, stroke, and other health problems (Sturm, 2002).[1] The obesity rate among U.S. adults rose from 13% in 1960 to 34% in 2008, with obesity's annual costs reaching 112,000 lives and $190 billion in medical expenses (Flegal et al., 1998; Flegal et al., 2005; Ogden and Carroll, 2010; Cawley and Meyerhoefer, 2012). This sharp increase has prompted a large literature examining its causes and consequences.

Obtaining accurate data on weight and height has long been problematic for researchers. Medically measured weight and height are obviously ideal, but they are expensive to collect in large samples. For this reason, only one large-scale U.S. government health dataset – the National Health and Nutrition Examination Surveys (NHANES) – contains measured weights and heights. However, the NHANES has several limitations. Though large enough to produce national-level descriptive statistics, it is often too small for more sophisticated estimation. Moreover, the NHANES consists of repeated cross-sections so it does not allow for the use of panel data methods.

☆ The views expressed in this paper are the authors' and do not necessarily reflect those of the U.S. Bureau of Labor Statistics.

* Corresponding author. Tel.: +1 502 852 2342.
E-mail address: josh.pinkston@louisville.edu (J.C. Pinkston).

[1] BMI = weight in kilograms divided by height in squared meters.

http://dx.doi.org/10.1016/j.ehb.2015.04.003
1570-677X/

Finally, although the NHANES contains excellent health information, it includes a limited number of economic and demographic variables.

Other datasets include respondents' self-reported weight and height, often obtained through telephone surveys (e.g., Behavioral Risk Factor Surveillance System (BRFSS)). The use of self-reported data permits a larger sample size and broader geographic coverage, but is limited by the fact that self-reports are often subject to considerable measurement error. Some respondents may not know their current weight, while others might misreport their weight and height in an effort to adhere to social norms. Cawley (2002) finds that underweight people tend to over-report their weight while those who are heavier tend to underreport. Rowland (1990) finds similar results for weight and also documents a tendency to exaggerate height that is more pronounced among the overweight. For both sexes, these results imply that the distribution of self-reported weight is more compressed than the distribution of measured weight, and obesity rates computed from self-reported weight and height are understated. Moreover, the systematic, non-classical nature of the measurement error suggests that bias in regression estimates is possible regardless of whether BMI is an independent or dependent variable, and that the direction of the bias is unclear.

Two papers by Cawley (2002, 2004) were the first to attempt to correct for the misreporting of height and weight. Since the NHANES was not a suitable dataset for the topic of either paper, Cawley used the 1979 cohort of the National Longitudinal Survey of Youth (NLSY), which contains self-reported weight and height.[2] He attempted to correct measurement error in these variables by using the NHANES as a validation sample for the NLSY79 and applying a procedure developed by Lee and Sepanski (1995). For each race and gender group, Cawley regressed measured weight and height on the corresponding self-reports and their squares in NHANES, and then used the resulting regression estimates to predict the NLSY79 respondents' actual weights and heights.

This correction is now common in the economics-of-obesity literature. Several recent papers have used this correction when studying the impacts of obesity on labor market outcomes (e.g., Cawley and Danzinger, 2005; Gregory, 2010; Gregory and Ruhm, 2011; Majumder, 2013). It has also been used in a number of papers that examine potential economic determinants of obesity.[3]

Unfortunately, the standard validation method may not be appropriate if the amount or type of measurement error differs between the primary and validation samples. This potential problem is acknowledged in Cawley (2004) and Cawley and Burkhauser (2006). As Han et al. (2009) note, NHANES respondents should expect to be measured when they report their height and weight, while respondents in the NLSY and other commonly used datasets do not. Furthermore, Pinkston (2015) notes that interview mode (in-person vs. telephone) affects self-reported values of respondents in the NLSY – even though in-person interviewees do not expect to be measured.

Our paper develops an alternative correction for self-reported weight and height that relies on weaker assumptions about the relationship between measured and reported values in the primary and validation datasets. Instead of using the reported values, we predict actual measures using the percentile rank of reported values in their respective distributions. Our method is robust to differences across samples in the severity (or type) of measurement error as long as the *rankings* of respondents based on reported values resemble the *rankings* based on actual measures in both datasets, and both datasets represent the same population (e.g., nationally representative samples).

We illustrate our method using data from the BRFSS and the American Time Use Survey (ATUS), and show that our rank-based method produces distributions of predicted BMI that are consistent across datasets, and close to the distribution of measured BMI in the population. We also show that the standard validation approach is sensitive to differences in misreporting between the primary and validation datasets.

Finally, we illustrate how the corrections can influence regression coefficients and estimates of the prevalence of obesity. We consider basic regressions that include BMI or obesity as either a dependent or an independent variable, and compare estimates that use our correction to analogous estimates that use either no correction or the standard correction. Although our correction generally does not affect the signs of coefficient estimates or statistical significance, it can lead to important differences in the magnitudes of the estimates. We then revisit the Centers for Disease Control's (CDC's) well-known map of obesity rates by state and demonstrate that correcting the BRFSS data for measurement error dramatically increases estimated obesity rates for most states.

2. The problem of transportability and an alternative approach

Let b denote the true measures of height or weight in the population, and $\tilde{b}_j$ denote the reported value in sample j, where $j = P, V$ indicates the primary or validation dataset. The reported values, b_j, are allowed to have arbitrary (potentially non-classical) measurement error.

The standard validation approach is based on work by Lee and Sepanski (1995) (L&S in what follows) and others in the statistics literature.[4] We can distil two conditions

[2] The first paper, Cawley (2002), tests for rational addiction in caloric intake, which requires panel data. The second, Cawley (2004), examines the impact of obesity on wages, and wages are not available in the NHANES.

[3] Potential determinants of obesity studied using Cawley's correction include age (Baum and Ruhm, 2009), income (Cawley et al., 2010), unemployment rate (Ruhm, 2005), childhood socioeconomic status (Baum and Ruhm, 2009), food prices (Lakdawalla and Philipson, 2002; Chou et al., 2004; Courtemanche et al., 2015a; Goldman et al., 2011), cigarette prices (Chou et al., 2004; Baum, 2009), alcohol prices (Chou et al., 2004), food stamps (Fan, 2010; Baum, 2011), restaurant density (Chou et al., 2004), on-the-job physical activity (Lakdawalla and Philipson, 2002), smoking bans (Chou et al., 2004), urban sprawl (Plantinga and Bernell, 2007; Eid et al., 2008), and time preference (Courtemanche et al., 2015a).

[4] See Bound et al. (2001) for a brief survey of this work, and Carroll et al. (2006) for more depth.

from this literature that must be met when using validation data to correct for measurement error:

C1. There must be a *surrogate* for b. A variable, b_j^s is a surrogate for b if the distribution of y given (b, b_j^s) is the same as the distribution of y given b. In cases where b is a dependent variable that is measured with error, b_j^s is a surrogate if its distribution depends only on the true response (Carroll et al., 2006).

C2. The surrogate, b_j^s, must satisfy some form of *transportability* across datasets. Transportability is usually described as the distribution of b conditional on b_j^s being the same in both datasets; however, L&S use a weaker form of transportability, which requires $E(b|b_P^s) = E(b|b_V^s)$.[5]

The first condition simply states that a surrogate for b contains no information about the dependent variable that is not also contained in b (and possibly other observed covariates). This condition is easily satisfied by the reported values, $\tilde{b}_j$.

The second condition, transportability, is essential if the procedure used to generate predicted values in the validation dataset is to be applied to the primary dataset. C2 requires that researchers make additional assumptions about the characteristics of the primary and validation datasets. Carroll et al. (2006) note that validation data are ideally drawn from a random subsample of the primary data, and warn that transportability may not be satisfied when the validation data are drawn from external sources.

Bound et al. (2001) point out that misreporting in survey data often varies with the context of the survey. This means that transportability may not hold when the validation data are drawn from an external source. As Cawley and Burkhauser (2006) note, the NHANES collect self-reported height and weight during face-to-face interviews, while surveys used in the social sciences often collect at least some data by telephone. For example, the BRFSS, ATUS and the Panel Study of Income Dynamics (PSID) surveys are conducted by telephone. Both the 1979 and 1997 cohorts of the NLSY combine in-person and telephone interviews.[6]

Bound et al. (2001) also note that transportability requires both datasets to be representative of the same population. There is ample evidence in the literature suggesting that different populations misreport height and weight differently.[7] Therefore, we want to stress that neither the method we develop nor the standard validation approach is appropriate if the primary and validation data are not representative of the same population. As in previous work on obesity, we use datasets that are weighted to be nationally representative. Furthermore, we explicitly assume that the distribution of b, $F(b)$, does not vary across samples.

The rest of this section compares the standard validation approach and our rank-based alternative. In Section 2.1, we focus on the assumptions required for the standard approach to satisfy transportability, and discuss when those assumptions are unlikely to hold. In Section 2.2, we develop an alternative surrogate that satisfies the transportability condition under weaker assumptions.

2.1. The standard validation method

Consider the regression of some dependent variable y on b and other covariates. Assume we only observe the self-reported values, $\tilde{b}_p$, in the primary dataset. In the standard approach, we would estimate

$$b = \eta(\tilde{b}_V) + \epsilon$$

using the validation dataset, and then use

$$\hat{b} = \hat{\eta}(\tilde{b}_P)$$

in place of b as an independent variable in the primary dataset.[8]

This approach assumes that $\tilde{b}_j$ is a surrogate for b, and that it is transportable. In this case, transportability is satisfied under the following assumption:

A1. The expected value of the true measure conditional on the reported value is the same in both the primary and validation datasets; i.e., $E(b|\tilde{b}_P) = E(b|\tilde{b}_V)$ if $\tilde{b}_P = \tilde{b}_V$.[9]

We would expect this assumption to hold when an internal validation sample is drawn at random from the primary sample. But when the validation data are drawn from an external dataset, transportability is satisfied only if the reporting error in $\tilde{b}_j$ is the same in the two data sources. Assumption A1 is very strong, and is less likely to hold when there are differences in interview modes (in-person vs. telephone) or other differences between the data sources. In any case, the required assumptions should be spelled out carefully and, whenever possible, evaluated empirically.

2.2. An alternative method based on percentile rank

The theory supporting the use of validation data does not require us to only use the reported measures, $\tilde{b}_j$, as surrogates for b. Any surrogate for b that satisfies transportability can be used. We exploit this fact and

[5] See Bound et al. (2001) or Carroll et al. (2006) for examples of the stronger version of transportability. Strictly speaking, L&S assume that the expectation of y conditional on b_j^s is the same in both datasets, but that reduces to $E(b|b_P^S) = E(b|b_V^S)$ in the current context.

[6] The use of telephone surveys is described under the heading "Interview Methods" in the documentation for each cohort. Pinkston (2015) notes that the reported weight of white women is especially sensitive to interview methods.

[7] This is why predictions have been made separately by race and gender group since Cawley (2002, 2004). Cawley (2004) also notes that misreporting varies by body mass. Maclean and Sikora (2014) discuss the importance of both samples having the same age distribution. Cawley and Choi (2014) find that the misreporting of height, weight and other variables differs by education group. Cawley and Burkhauser (2006) and Courtemanche et al. (2015b) note that misreporting can vary across time periods.

[8] In Cawley (2004) and other papers in the obesity literature, the dependent variable is actually regressed on a nonlinear function of the predicted values, BMI. As we discuss later, L&S argue that it would be preferable to predict the nonlinear function directly.

Additionally, L&S show that $\eta(\tilde{b}_V)$ can be a simple polynomial approximation of $E(b|\tilde{b}_V)$, making functional form assumptions (or nonparametric estimates) unnecessary.

[9] This is the weaker form of transportability used by L&S.

propose an alternative surrogate that can satisfy transportability in cases where the reported values themselves do not.

As before, we assume that the $\tilde{b}_j$ are surrogates for b. Following the previous literature, the $\tilde{b}_j$ are functions of b and a random error term that is not correlated with y. This implies that the percentile rank of $\tilde{b}_j$, given by the distribution function $G_j(\tilde{b}_j)$, is a function of b and the same random error. Therefore, $G_j(\tilde{b}_j)$ is also a surrogate for the true value, b, satisfying C1.

Our approach uses the percentile rank of the report, $\tilde{b}_j$, to generate predicted values of b. The advantage of using percentiles over levels is that it replaces A1 with a simple monotonicity assumption. Specifically, we assume:

A2. The expected value of the true measure conditional on the reported value is monotonically increasing in the reported value; i.e., $\tilde{b}'' > \tilde{b}'$, implies that $E(b|\tilde{b}'') > E(b|\tilde{b}')$.[10]

In other words, given any two people who report their weight, the person who reports the higher weight is expected to actually weigh more.[11] If A2 does not hold, then it is hard to see how respondent reports convey any useful information about actual height and weight.

To see how A2 can be satisfied in cases that violate A1, think of misreporting as arising from a mean-zero error, ε, and a bias function that represents systematic misreporting:

$$b = \tilde{b}_j + \beta_j(\tilde{b}_j) + \epsilon,$$

where the bias function, $\beta_j(\tilde{b}_j)$, describes the expected difference between a self-report and the true measure.[12] Assumption A1 requires that $\beta_V(\tilde{b}_V) = \beta_P(\tilde{b}_P)$ if $\tilde{b}_V = \tilde{b}_P$. In contrast, A2 only requires that $\tilde{b}_j + \beta_j(\tilde{b}_j)$ be monotonically increasing in both samples, which implies that[13]

$$\beta_j'(\tilde{b}_j) > -1.$$

Therefore, the bias functions can be increasing or decreasing functions of $\tilde{b}_j$, and do not even need to be monotonic, as long as they never decrease too quickly. Two simple examples of bias functions are individuals overstating their height by a fixed percent ($\beta_j(\tilde{b}_j) = -\alpha_j\tilde{b}_j$) or understating their weight by a fixed percent ($\beta_j(\tilde{b}_j) = \phi_j\tilde{b}_j$), where $0 < \alpha,\ \phi < 1$. A2 would allow α_j and ϕ_j to differ between samples, while A1 would not. A2 would even allow respondents to exaggerate their weight (or height) in one sample and understate it in the other.[14]

Now consider the unconditional distribution of true values, $F(b)$. As long as both samples are representative of the same populations, $F(b)$ does not vary between datasets. Both $F(b)$ and $G_j(\tilde{b}_j)$ are continuous, monotonically increasing functions with ranges in the interval [0,1]. This implies that for every value of$\tilde{b}_j$there is a b such that

$$F(b) = G_j(\tilde{b}_j).$$

In general, $b \neq \tilde{b}_j$; however, taking the inverse of $F(\cdot)$, we have:

$$b = F^{-1}(G_j(\tilde{b}_j)),$$

which maps reported values into the true values of b.

Note that $F^{-1}(\cdot)$ does not depend on which sample the reported values, $\tilde{b}_j$, are drawn from. It simply takes the percentile rank associated with a reported value in sample j and returns the measured value that has the same position in the distribution of measured values. As a result, $F(b|G_V(\tilde{b}_V)) = F(b|G_P(\tilde{b}_P))$. This implies that the percentile ranks, $G_j(\tilde{b}_j)$, satisfy transportability, even when the reported values do not. Therefore, the percentile rank approach satisfies both of the conditions required for the use of validation data.

3. Data and the transportability of self-reported measures

This section begins with a brief introduction to the datasets we use in our analysis. We then show that self-reported height and weight do not appear to satisfy transportability between these datasets.

3.1. Three data sets

We use two primary datasets: the BRFSS and the ATUS. The BRFSS is a telephone survey conducted by the CDC in conjunction with state health departments. It focuses on health and risky behaviors, but also contains a wide variety of demographic variables. The primary advantage of the BRFSS for obesity studies is its size. With over 300,000 respondents per year in the later waves, the BRFSS is large enough to compute reliable state-level descriptive statistics. Additionally, the large sample size makes it popular among economists seeking to identify causal effects using inherently inefficient estimation techniques such as instrumental variables.

The ATUS was designed to measure how people spend their time rather than to study health outcomes. It is a telephone survey that asks respondents to sequentially report what they did on the day prior to the interview.

[10] This is akin to the assumptions made in the principal-agent literature to allow the use of the first-order approach to solving maximization programs (see Milgrom, 1981; Rogerson, 1985). A sufficient, but not necessary, condition for this assumption is the first-order stochastic dominance of $F(b|\tilde{b}'')$ over $F(b|\tilde{b}')$.

[11] The monotonicity described by A2 is testable in the validation sample but not in the primary samples. We tested A2 for each race and gender group in NHANES using nonparametric regressions of actual height and weight on their reported values. We could not reject monotonicity in any case.

[12] Note that $b = E(b|\tilde{b}_j) + \epsilon$, so $\beta_j(\tilde{b}_j) = E(b|\tilde{b}_j) - \tilde{b}_j$. If self-reports are unbiased, $\beta_j(\tilde{b}_j) = 0$ and $b = \tilde{b}_j + \epsilon$. If respondents understate their true values, $\beta_j(\tilde{b}_j) > 0$. If they overstate their true value, $\beta_j(\tilde{b}_j) < 0$.

Alternatively, the self-report could be written as a function of the true values and a random error term, $\tilde{b}_j = \rho(b - \epsilon)$; however, our specification of the measured value simplifies the math and better describes the problem faced by the researcher.

[13] This follows from $\partial E(b|\tilde{b}_j)/\partial\tilde{b}_j = 1 + \beta_j'(\tilde{b}_j) > 0$.

[14] To give more extreme examples, A2 would hold even if respondents in the validation sample gave unbiased self-reports, $\tilde{b}_v = b - \epsilon$, while respondents in the primary sample reported values $\tilde{b}_p = ln(b - \epsilon)$. A2 would also hold if self-reports were converted to metric units in one sample, but left in standard units in the other sample.

ATUS respondents are selected from households that have completed their final month of participation in the Current Population Survey (CPS). In addition to the time diary, the ATUS includes demographic information about respondents and members of the respondent's household, and employment status information for the respondent and the respondent's spouse. In 2006, 2007, and 2008, the U.S. Department of Agriculture's Economic Research Service (ERS) sponsored the Eating and Health Module, which collects information about the respondent's health, including weight and height, and additional information on time spent eating and drinking.

Following the previous literature, our external validation data are drawn from the NHANES, which is collected by the Centers for Disease Control and Prevention to assess the health status and behaviors of children and adults in the United States. Respondents are asked their weight and height during the initial face-to-face interview administered in the respondent's home. Weight and height are then measured during a physical examination that may take place up to several weeks after the in-home interview. Respondents are told about other components of the survey before they consent to the initial interview, and they are compensated for participating.[15] We use data on race, gender, and age from the demographic background files of the 2007–2008 wave; reported values of height and weight from the Weight History questionnaire; and measures of actual height and weight from the physical examination.

For consistency, we restrict the samples from all three data sets to 2007 and 2008, and to respondents between the ages of 19 and 64. Racial categories indicate whether respondents identify as Caucasian, African-American or any other racial or ethnic group.[16] All estimation is weighted to ensure that each sample is representative of the same population.

Table 1 presents basic summary statistics for the three datasets. Even with all race and gender groups pooled together, average reported weight is higher in the NHANES (180 lbs) sample than either BRFSS (178.6) or ATUS (178). Average measured weight is roughly two pounds heavier than reported weight for the full NHANES sample.

The demographic variables in Table 1 are very similar across samples, which is consistent with the samples being representative of the same populations. The one notable exception is that black respondents appear to be underrepresented in the BRFSS. Such differences may point to shortcomings in the sample weights used; however, any shortcomings in BRFSS sample weights would affect both of the correction methods we compare. Therefore, we ignore this difference in the work that follows.

Table 1
Summary statistics by dataset.

	BRFSS	ATUS	NHANES
Reported height	67.37	67.21	67.24
	(4.18)	(4.10)	(4.08)
Reported weight	178.6	178	180
	(44.67)	(43.60)	(45.15)
Actual height	–	–	66.9
			(3.88)
Actual weight	–	–	181.8
			(47.18)
White	0.680	0.689	0.674
	(0.466)	(0.463)	(0.469)
Black	0.103	0.114	0.119
	(0.303)	(0.318)	(0.323)
Other race/ethnicity	0.217	0.197	0.207
	(0.412)	(0.397)	(0.405)
Male	0.516	0.513	0.504
	(0.500)	(0.500)	(0.500)
Age	41.29	41.34	41.05
	(12.48)	(12.54)	(12.67)
Observations	539,072	17,721	4113

Notes: Standard deviations are in parentheses. All samples weighted to be representative of adults in the United States between the ages of 19 and 64 in the years 2007 and 2008. "Other race/ethnicity" includes all respondents who identify as Hispanic.

3.2. The transportability of self-reported height and weight

The most relevant difference in methodology between the surveys is that respondents in the BRFSS and ATUS are interviewed by telephone, while respondents in the NHANES are interviewed in person and expect to be measured in the near future.[17] It seems natural to expect misreporting to be more severe in telephone surveys than in face-to-face surveys, especially when respondents to the face-to-face survey expect to be physically examined. Pinkston (2015) points out that phone interviews in the NLSY cohorts are associated with lower reported weights for white women than in-person interviews, even though NLSY respondents have no reason to believe they will be measured in either case.

Comparing differences in misreporting between our datasets is straightforward. We expect the distribution of actual height and weight to be the same in all three datasets because they are all representative samples of the same population.[18] If the distributions of reported values are not the same in two samples that have the same distribution of actual measures, the relationships between actual and reported measures are also not the same. Therefore, it is sufficient to compare reported height and

[15] In contrast, BRFSS respondents are not compensated at all for participation. ATUS respondents are offered a small incentive only if their household did not provide a valid phone number when they participated in the CPS.

[16] All respondents who identify as Hispanic are included in the "other" category, as are all respondents who identify as anything other than African-American or Caucasian. We did not divide this category further due to the sample size in NHANES. For the sake of convenience, we refer to these groups as though they are defined by race, even though that is not strictly correct in the case of Hispanics.

[17] See the online documentation for NHANES for more detail. Although respondents are told about the physical examinations before the initial interview, any effect of that knowledge may be lessened by the amount of time between the in-home interviews and the physical examinations.

[18] The distributions of measured height or weight would only differ between samples if the surveys did not all represent the population they claim to represent, or if the measurements were somehow affected by differences in data collection. Either of these scenarios would make any use of NHANES as a validation sample suspect, and neither scenario can be ruled out because we do not observe measured values in the BRFSS or ATUS data.

weight across samples to determine whether corrections that use reported values are transportable.

Fig. 1 compares the densities of reported weight and height across samples for white, black and other women. The results cast substantial doubt on the transportability of reported weight for women in these samples. Relative to white women in NHANES, white women in both BRFSS and ATUS report weights between 120 and 150 pounds more frequently and report higher weights less frequently. The picture is similar for women of other races. We only see a difference in the reported weights of black women in the upper half of their distributions; however, self-reported height appears more sensitive to context for black women than for white or other women.

Fig. 1B makes analogous comparisons for men. Differences in reported weight are less pronounced for men than for women; but differences in height appear to be larger, especially for black and other men. In contrast to the results for women, black men and men from other races appear more likely to report higher weights when interviewed on the phone than when interviewed in person.

Tables 2A and 2B compare the self-reported measures in BRFSS and ATUS, respectively, to those from NHANES. Each table presents averages of reported height and weight, followed by the medians, 75th and 90th percentiles. The final column of each table presents non-parametric Kolmogorov–Smirnov tests for the equality of distributions between samples.

In both tables, we see that women reported lower weights in the telephone surveys than in the NHANES, with the difference being larger for black and white women than for women of other races. The differences in average reported weight are driven more by the upper tail for black and (to a lesser degree) other women than for white women.

The differences in reported distributions are again less pronounced for men. We only see differences in reported weight for black men and men of other races, and they appear to report higher weights when they cannot be observed by the interviewer. There are small differences in reported height, but they are less obvious in the summary statistics than the kernel densities.

The only cases in which Kolmogorov–Smirnov tests cannot reject the equality across samples of either measure are for white and black men in BRFSS compared to NHANES; however, we can reject the equality of reported BMI distributions (not shown) for black men. Overall, these results suggest that transportability is likely not satisfied in most cases when self-reported weight and height are used as surrogates for measured weight and height.

Finally, the minor differences in the results for BRFSS and ATUS are not surprising. Although both surveys are conducted over the telephone, they obviously differ in focus, and reporting can be influenced by something as simple as differences in the preceding questions or in the surveys' introductions. The effects of these differences could vary across demographic groups.

4. Comparing methods for predicting BMI

As noted earlier, the standard method for correcting BMI is to regress measured height and weight on reported values using NHANES data, and then use the estimated coefficients to predict the measured values in the primary dataset. Specifically, for each race and gender category, we regress measured height (or weight) on cubic polynomials in age and reported height (or weight). We then predict measured height and weight, and use those values to calculate predicted BMI.[19]

Our method is similar, but uses the percentile rank of the reports and a more flexible functional form to predict actual height and weight. Consistent with the fact that percentile ranks are (roughly) uniformly distributed between zero and one, while reported (and actual) measures are not, we found that regressing the actual measures on simple polynomials of the percentile ranks resulted in predicted values that were poor fits for the actual measures.[20] Instead, we regressed the measured values on cubic basis splines (b-splines) in percentile ranks, which resulted in a fit that was comparable to that achieved when using simple polynomials in the standard method.[21]

When comparing our percentile approach to the standard approach, we were initially concerned that the improved predictions may be due to the b-spline specification, rather than the use of percentiles. To check this, we also predicted height and weight using b-splines in the reported values. We found that adapting the standard approach to use splines resulted in no discernable change in our ability to predict actual measures from reported values, which is not surprising. All of the regressions using polynomials in reported levels had R^2 values over 0.9, leaving little room for improvement. In what follows, we compare predictions based on cubic b-splines of the percentiles to predictions based on cubic polynomials of reported levels, which are common in the literature. Our conclusions are unchanged if we also use cubic b-splines in reported levels.[22]

The fact that measured height and weight can be accurately predicted within NHANES using simple polynomial functions of the reported values illustrates an advantage of the standard validation approach. When violations of transportability are not a concern, there is nothing to be gained by transforming mismeasured variables into percentiles only to use a more complicated functional form to fit the actual measurements. For example, researchers who have access to an appropriate

[19] L&S argue that constructing a nonlinear function of mismeasured variables from predicted values of those variables may provide a useful approximation, but predicting the nonlinear function directly is preferable. In the case of BMI, this means that researchers who are interested in BMI should predict BMI directly instead of constructing it from predicted height and weight; however, we find that the mean squared error associated with the prediction of BMI is higher in NHANES when BMI is predicted directly than when it is constructed from predicted height and weight. Furthermore, predicting height and weight is more consistent with the previous obesity literature. Therefore, we do not follow the advice of L&S in this particular case.

[20] The distributions of percentile ranks differ from uniform distributions only because of clustering in the self-reports, typically around multiples of five pounds.

[21] The details of our estimation are included in Appendix.

[22] Results that use cubic b-splines in reported levels instead of polynomials are available upon request.

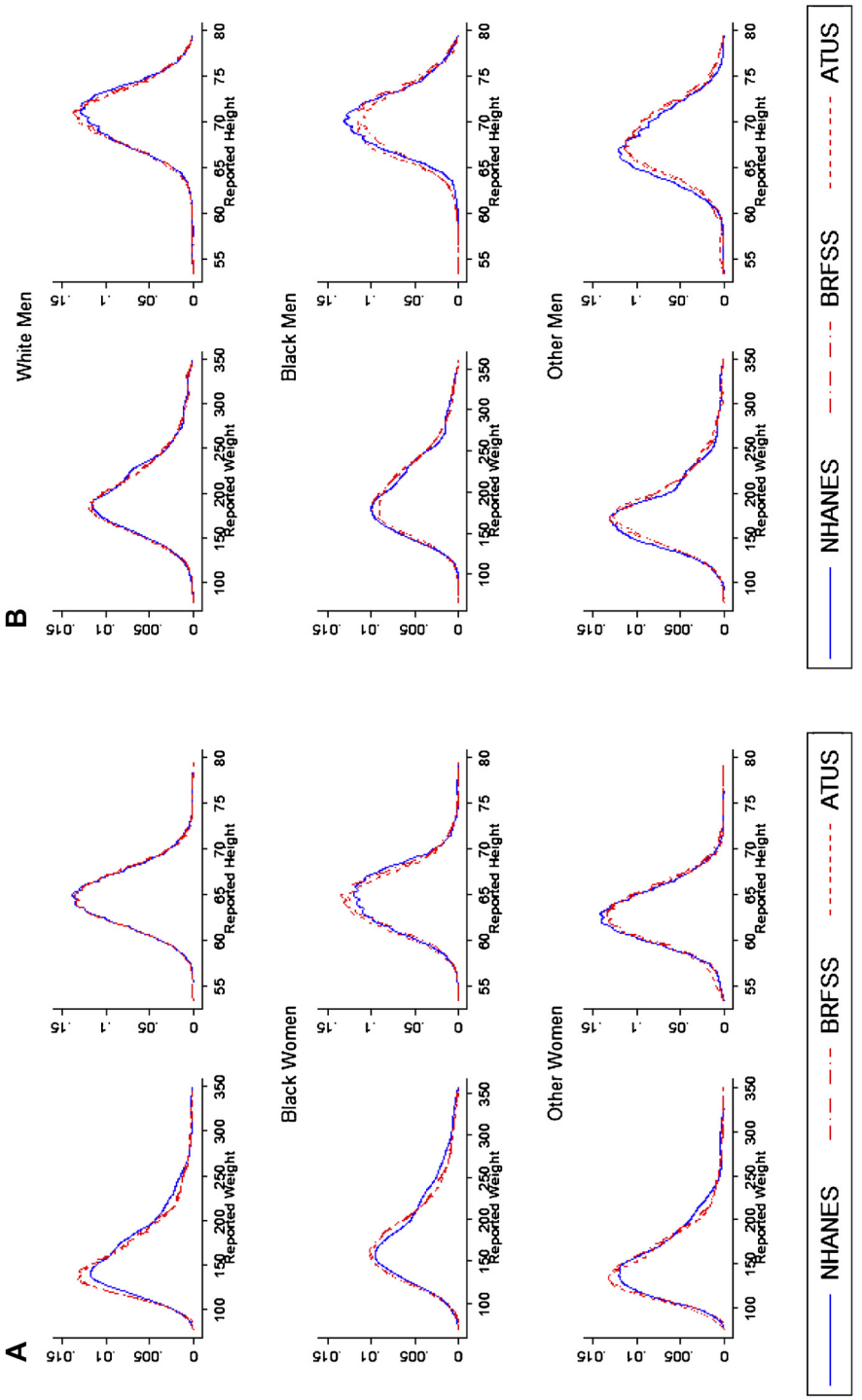


Fig. 1. Kernel densities of reported measures by race and gender.

Table 2A
Comparison of self-reported values: BRFSS vs. NHANES.

		Mean	Percentiles of distribution			Kolmogorov–Smirnov test (p-value)
			Median	75th	90th	
White women						
Weight	NHANES	163.907	155	185	220	0.003
	BRFSS	159.947	150	180	210	
Height	NHANES	64.828	65	67	68	0.972
	BRFSS	64.857	65	67	68	
White men						
Weight	NHANES	199.387	195	220	250	0.591
	BRFSS	200.060	195	220	250	
Height	NHANES	70.690	71	73	74	0.559
	BRFSS	70.731	71	72	74	
Black women						
Weight	NHANES	182.627	170	214	250	0.047
	BRFSS	178.571	170	200	240	
Height	NHANES	64.729	65	67	68	0.572
	BRFSS	64.793	65	67	68	
Black men						
Weight	NHANES	198.628	190	225	250	0.179
	BRFSS	201.797	195	225	260	
Height	NHANES	70.221	70	72	74	0.317
	BRFSS	70.279	70	72	74	
Other women						
Weight	NHANES	154.718	148	175	205	0.498
	BRFSS	153.868	148	172	200	
Height	NHANES	62.893	63	65	66	0.025
	BRFSS	63.171	63	65	67	
Other men						
Weight	NHANES	180.930	175	200	230	0.017
	BRFSS	183.433	179	200	230	
Height	NHANES	67.585	67	70	72	<0.001
	BRFSS	68.142	68	71	72	

Table 2B
Comparison of self-reported values: ATUS vs. NHANES.

		Mean	Percentiles of distribution			Kolmogorov–Smirnov test (p-value)
			Median	75th	90th	
White women						
Weight	NHANES	163.907	155	185	220	0.002
	ATUS	159.348	150	180	210	
Height	NHANES	64.828	65	67	68	0.993
	ATUS	64.773	65	67	68	
White men						
Weight	NHANES	199.387	195	220	250	0.270
	ATUS	197.917	190	220	250	
Height	NHANES	70.690	71	73	74	0.020
	ATUS	70.487	70	72	74	
Black women						
Weight	NHANES	182.627	170	214	250	0.017
	ATUS	176.208	170	200	237	
Height	NHANES	64.729	65	67	68	0.116
	ATUS	64.434	64	66	68	
Black men						
Weight	NHANES	198.628	190	225	250	0.331
	ATUS	198.814	195	225	257	
Height	NHANES	70.221	70	72	74	0.060
	ATUS	69.827	70	72	74	
Other women						
Weight	NHANES	154.718	148	175	205	0.001
	ATUS	150.478	140	170	197	

Table 2B (*Continued*)

		Mean	Percentiles of distribution			Kolmogorov–Smirnov test (*p*-value)
			Median	75th	90th	
Height	NHANES	62.893	63	65	66	0.764
	ATUS	62.853	63	65	66	
Other men						
Weight	NHANES	180.930	175	200	230	0.010
	ATUS	184.299	180	200	238	
Height	NHANES	67.585	67	70	72	0.036
	ATUS	67.931	68	70	72	

Notes: All samples weighted to be representative of adults in the United States between the ages of 19 and 64 in the years 2007 and 2008. "Other" includes all respondents who identify as Hispanic.

internal validation sample should use the standard method estimated with polynomials.

Fig. 2A through 2F compare kernel densities of predicted BMI using our percentile method to predictions using the standard validation approach for each race and gender group. Each figure contains four graphs, one for each prediction method and primary data set. Each graph compares predicted BMI from a primary data set to the analogous prediction from NHANES, as well as measured BMI from NHANES. At the bottom of each figure, we include results for Kolmogorov–Smirnov tests of the differences observed in each graph.

Fig. 2A shows that the standard method can produce predicted BMI values that differ significantly between samples. These differences follow the differences between samples in reported weight, and the Kolmogorov–Smirnov tests strongly reject the hypothesis that the standard method produces predicted values in ATUS or BRFSS that are equal to those in NHANES. Again, this suggests that the standard method is inappropriate in our context.[23]

In contrast, the density functions of BMI predicted using the percentile rank method are very similar across samples. The density of predicted BMI in ATUS is almost indistinguishable from the analogous density in NHANES. The Kolmogorov–Smirnov tests both have *p*-values over 0.8, providing no reason to doubt that the density functions of predicted BMI are the same across samples when our method is used.

The results for black women (Fig. 2C) and women of other races (Fig. 2E) are similar to those for white women. The densities of predicted BMI using our approach are noticeably closer to the corresponding densities in NHANES, which makes them more similar to measured BMI in NHANES (and presumably the national population). Kolmogorov–Smirnov tests again reject the equality of the standard method across data sets, but fail to reject equality when using our method.

The results for men are less striking than the results for women, but still support our approach. The density functions for white men (Fig. 2B) appear more similar when the percentile method is used than when the standard method is used, but testing the equality of distributions suggests that none of these differences are statistically significant. This is consistent with misreporting being less pronounced for white men than for other groups. Although the graphs of kernel densities for black men do not tell an obvious story, the Kolmogorov–Smirnov tests reject the standard approach and fail to reject our percentile method.[24] Finally, the standard method is rejected for men of other races when we compare the ATUS to NHANES, and the percentile method again produces BMI predictions that are more consistent across surveys.

As a robustness check, we also tested for differences in the density functions of predicted BMI between the ATUS and BRFSS (not shown). We would not expect our percentile correction to produce different distributions between the two primary samples, and we find no evidence that it does. On the other hand, we find statistically significant differences for men and women of other races in these surveys when the standard correction is used, which is consistent with the less obvious differences in context discussed above (e.g., survey content) affecting responses.[25]

5. Do these differences matter?

The results of Sections 3 and 4 suggest that using the relationship between reported and actual height and weight in the NHANES to predict BMI in data from the ATUS or BRFSS is inappropriate in most cases. Furthermore, the rank-based alternative we propose appears to work well in practice, producing predictions of BMI that are consistent across random samples of the same population. In this section we provide results that illustrate the potential importance of our adjustment to empirical work.

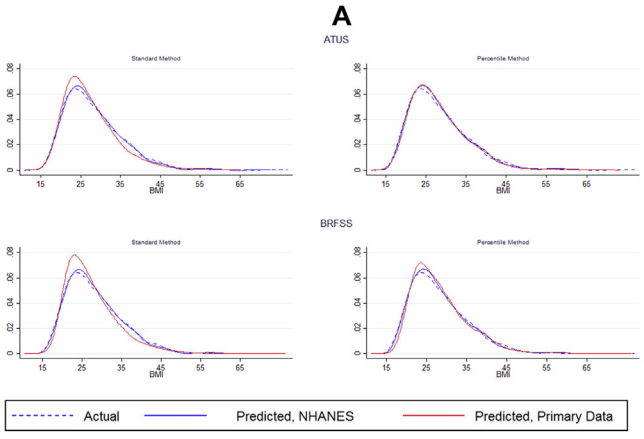
5.1. Effects of BMI correction on regression estimates

This section illustrates how our adjustment might impact empirical work that uses BMI or obesity in

[23] Recall that all three datasets should be representative of the same population. Therefore, the true BMI distributions should not vary by sample.

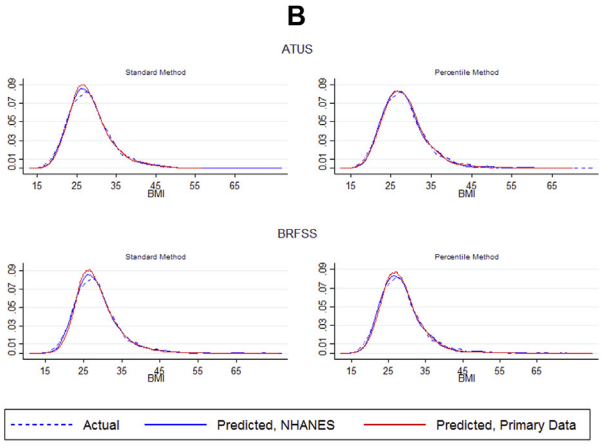
[24] Regardless of method, the BMI of black men appears to be more difficult to predict that the BMI of other groups. This is appears to be due to errors in the prediction of height for black men. We considered the possibility that this difficulty is related to the small sample size for black men; however, we found similar results when we expanded the NHANES sample to include the 2005–2006 and 2009–2010 waves.

[25] Kolmogorov–Smirnov tests reject the equality of distributions of BMI predicted using the standard correction with *p*-values less than or equal to 0.001 for other women and men. None of the other tests suggest a statistically significant difference in predicted values between any ATUS and BRFSS subsamples.



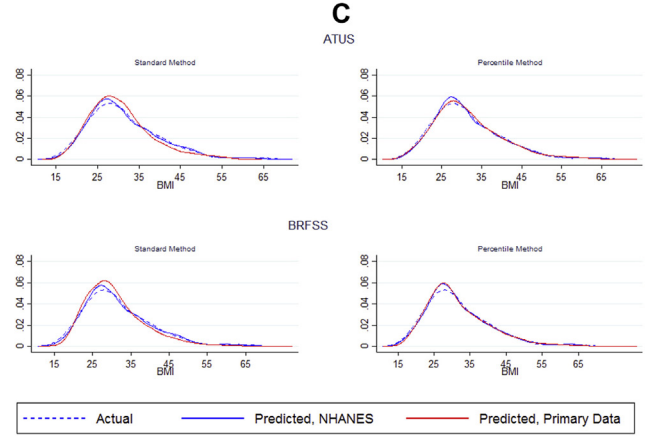
Kolmogorov-Smirnov Tests (p-values) for Equality of Density Functions

	Standard Method	Percentile Method
ATUS vs. NHANES	0.006	0.908
BRFSS vs. NHANES	0.006	0.826



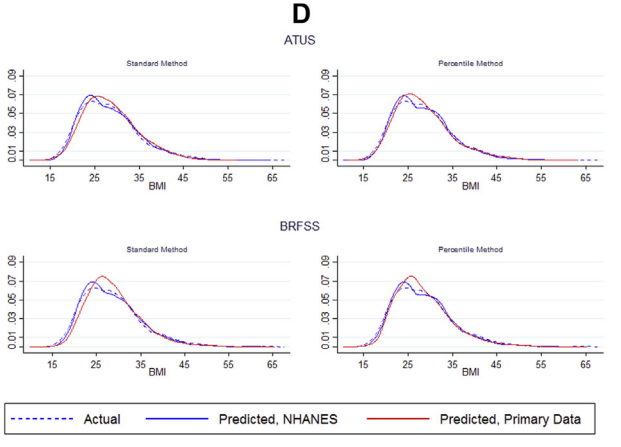
Kolmogorov-Smirnov Tests (p-values) for Equality of Density Functions

	Standard Method	Percentile Method
ATUS vs. NHANES	0.809	0.480
BRFSS vs. NHANES	0.819	0.382



Kolmogorov-Smirnov Tests (p-values) for Equality of Density Functions

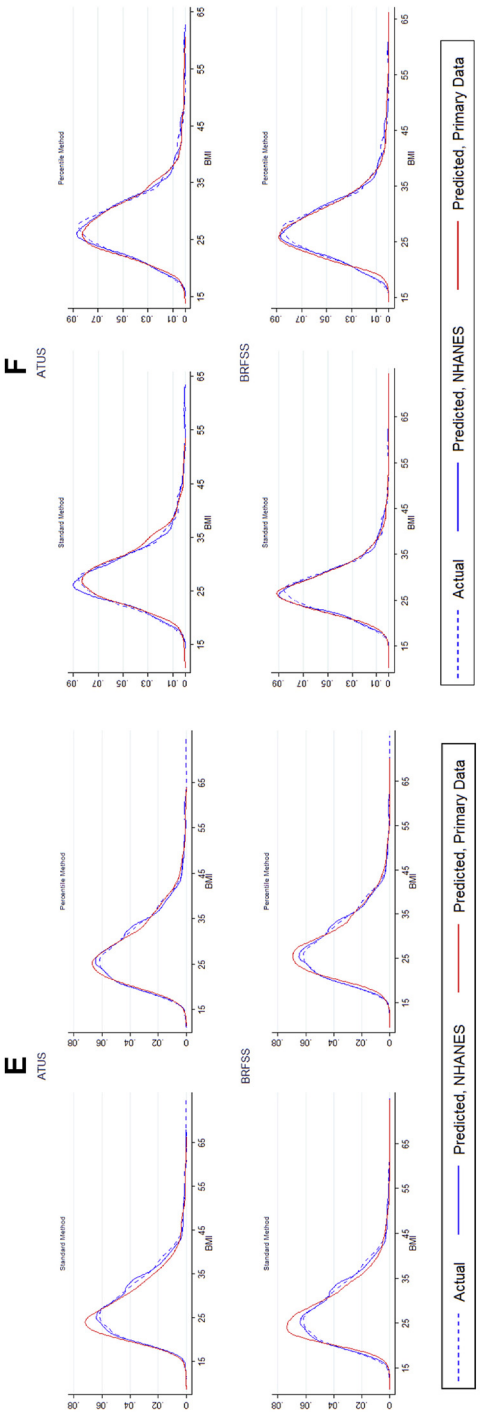
	Standard Method	Percentile Method
ATUS vs. NHANES	0.049	0.734
BRFSS vs. NHANES	0.043	0.959



Kolmogorov-Smirnov Tests (p-values) for Equality of Density Functions

	Standard Method	Percentile Method
ATUS vs. NHANES	0.075	0.312
BRFSS vs. NHANES	0.001	0.158

Fig. 2. (A) Standard vs. percentile prediction methods for white women. (B) Standard vs. percentile prediction methods for white men. (C). Standard vs. percentile prediction methods for black women. (D) Standard vs. percentile prediction methods for black men. (E) Standard vs. percentile prediction methods for other women. (F) Standard vs. percentile prediction methods for other men.



Kolmogorov-Smirnov Tests (p-values) for Equality of Density Functions

	Standard Method	Percentile Method
ATUS vs. NHANES	0.001	0.263
BRFSS vs. NHANES	0.005	0.254

Kolmogorov-Smirnov Tests (p-values) for Equality of Density Functions

	Standard Method	Percentile Method
ATUS vs. NHANES	0.018	0.461
BRFSS vs. NHANES	0.813	0.579

Fig. 2. (*Continued*).

Table 3
Average body mass by correction method & race/gender group BRFSS & ATUS compared to actual measures from NHANES.

	BMI			Overweight			Obese			Class II/III Obese		
	Self-report	Standard method	Rank-based	Self-report	Standard method	Rank-based	Self-report	Standard method	Rank-based	Self-report	Standard method	Rank-based
Full Sample		*28.51*			*67.0%*			*33.6%*			*14.4%*	
BRFSS	27.56	28.08	28.50	63.3%	65.8%	67.8%	27.2%	30.6%	33.1%	10.0%	11.9%	13.9%
ATUS	27.59	28.09	28.55	64.1%	66.2%	68.0%	27.9%	31.1%	33.1%	10.2%	11.9%	14.1%
White Women		*28.20*			*60.6%*			*33.8%*			*17.3%*	
BRFSS	26.66	27.36	28.26	51.6%	56.5%	61.2%	23.8%	28.1%	32.8%	9.9%	12.2%	15.9%
ATUS	26.70	27.39	28.24	52.2%	56.6%	61.5%	24.0%	28.4%	32.8%	10.1%	12.0%	16.2%
White Men		*28.42*			*71.2%*			*31.3%*			*10.2%*	
BRFSS	28.02	28.41	28.47	71.7%	72.1%	72.5%	28.1%	30.7%	32.0%	9.0%	10.6%	10.8%
ATUS	27.96	28.31	28.50	72.5%	72.8%	72.5%	28.4%	30.6%	31.9%	8.9%	10.5%	10.8%
Black Women		*31.50*			*76.5%*			*48.1%*			*28.7%*	
BRFSS	29.82	30.62	31.53	73.1%	76.1%	78.4%	40.8%	45.5%	48.2%	20.0%	23.5%	28.6%
ATUS	29.82	30.56	31.54	74.0%	76.6%	79.4%	41.5%	45.3%	48.5%	18.6%	20.9%	28.6%
Black Men		*28.51*			*65.8%*			*35.9%*			*14.0%*	
BRFSS	28.53	28.68	28.47	72.7%	70.5%	68.6%	33.5%	35.2%	34.8%	11.5%	14.7%	13.6%
ATUS	28.61	28.73	28.50	73.0%	71.0%	67.6%	33.9%	36.1%	33.3%	12.8%	14.7%	13.3%
Other Women		*28.39*			*64.1%*			*35.1%*			*15.7%*	
BRFSS	27.07	27.71	28.28	56.2%	61.3%	64.9%	26.2%	30.0%	33.0%	10.2%	11.9%	15.6%
ATUS	26.81	27.45	28.30	52.9%	56.9%	63.9%	25.7%	29.1%	31.5%	10.0%	11.7%	15.2%
Other Men		*28.13*			*70.8%*			*29.1%*			*9.6%*	
BRFSS	27.70	28.03	28.13	69.5%	70.5%	70.0%	26.1%	29.2%	29.7%	8.0%	8.9%	9.2%
ATUS	28.06	28.37	28.19	72.9%	74.2%	71.0%	29.4%	32.0%	30.5%	8.9%	9.6%	9.4%

Notes: Actual measures from NHANES are underlined and in italics. All samples are weighted to be representive of adults in the United States between the ages of 19 and 64.

regressions. We present one example in which BMI or obesity is an explanatory variable and one in which it is the dependent variable. An earlier version of this paper, Courtemanche et al. (2014), contains additional examples. In each case, we report the results both for the full sample and for each race × gender subgroup. None of the estimates presented in this section consider endogeneity or any other complications researchers might encounter. These results are presented for illustrative purposes only.[26]

To assist with the interpretation of results, Table 3 presents average BMI and the percent overweight, obese (BMI ≥ 30), and class II/III obese (BMI ≥ 35) by correction method and race × gender group for both the BRFSS and ATUS samples. Averages for the measured values from NHANES are included for comparison.

Both correction methods result in greater average BMI in most cases, but the differences between the two methods and reported values at the means appear to be fairly small. It is in the comparison of the upper tails of the distribution that we see important differences emerge. In both the ATUS and the BRFSS, roughly 10% of population reports a BMI of 35 or higher. Applying the standard correction raises this to 12%. Applying our rank-based correction produces estimates around 14%, which is close to the NHANES rate of 14.4%. Consistent with the hypothesis that those who are the most prone to misreporting are also the most sensitive to context, the differences between our rank-based correction and the standard correction are typically largest in cases where the standard correction appears to matter the most.

The patterns in the full sample are also observed to some degree in most of the race and gender groups, but they are most pronounced for white women. The average reported BMI for this group is 26.7. The standard correction increases average BMI to 27.4, and our correction increases it further to around 28.2, which is the same as the average measured BMI from NHANES. The differences in the correction methods are again larger at the upper tails of the distribution. For example, the overall obesity rate based on uncorrected reports is 24%. The standard correction increases the rate to 28%, and our correction increases it further to 33%. The actual rate from NHANES is 33.8%. Thus our correction still slightly underestimates the obesity rate for white women, but it does significantly better than the standard approach. Looking at class II/III obesity, the rate is around 10% when self-reported measures are used, 12% using the standard correction, and around 16% using our percentile-based correction, which again is slightly lower than the actual rate from NHANES (17.3%).

Black men are the most obvious exception to the pattern observed in the full sample. The average self-reported BMI in both the BRFSS and the ATUS samples is nearly indistinguishable from average measured BMI in

[26] We also acknowledge that measures of body composition would be preferable to BMI in many applications, as Burkhauser and Cawley (2008) and others have noted. Unfortunately, measures of body composition are rarely available in large datasets because, like measured weight and height, they are expensive to collect. Therefore, the widespread use of self-reported BMI is likely to continue, and reporting error will be a problem as long as self-reported BMI is used.

Table 4
BRFSS estimated effects of food prices on BMI, P(Obese), and P(Class II/III Obese) by race-gender group and BMI correction method.

	BMI			Obese			Class II/III Obese		
	Self-report	Standard method	Rank-based	Self-report	Standard method	Rank-based	Self-report	Standard method	Rank-based
Whole sample	**−0.764** **(0.170)**	**−0.858** **(0.187)**	**−0.937** **(0.189)**	**−0.050** **(0.010)**	**−0.058** **(0.011)**	**−0.060** **(0.012)**	**−0.025** **(0.007)**	**−0.032** **(0.009)**	**−0.041** **(0.010)**
White women	**−0.704** **(0.216)**	**−0.786** **(0.218)**	**−0.890** **(0.226)**	**−0.042** **(0.012)**	**−0.046** **(0.012)**	**−0.051** **(0.013)**	**−0.021** **(0.008)**	**−0.027** **(0.008)**	**−0.038** **(0.009)**
White men	*−0.415* *(0.191)*	*−0.479* *(0.193)*	*−0.510* *(0.208)*	*−0.032* *(0.015)*	*−0.037* *(0.015)*	**−0.038** **(0.015)**	−0.013 (0.011)	−0.014 (0.010)	−0.017 (0.011)
Black women	*−0.894* *(0.368)*	*−0.871* *(0.362)*	*−0.991* *(0.401)*	−0.029 (0.029)	−0.020 (0.033)	−0.026 (0.029)	**−0.050** **(0.017)**	**−0.049** **(0.018)**	**−0.067** **(0.021)**
Black men	**−1.565** **(0.229)**	**−1.785** **(0.258)**	**−1.721** **(0.276)**	**−0.130** **(0.019)**	**−0.129** **(0.022)**	**−0.105** **(0.019)**	**−0.055** **(0.011)**	**−0.077** **(0.014)**	**−0.081** **(0.011)**
Other women	**−1.478** **(0.321)**	**−1.539** **(0.301)**	**−1.640** **(0.334)**	**−0.098** **(0.022)**	**−0.113** **(0.026)**	**−0.119** **(0.025)**	**−0.037** **(0.010)**	**−0.047** **(0.013)**	**−0.061** **(0.013)**
Other men	*−0.710* *(0.286)*	*−0.855* *(0.355)*	**−0.975** **(0.364)**	*−0.042* *(0.017)*	**−0.063** **(0.021)**	**−0.061** **(0.019)**	*−0.028* *(0.011)*	*−0.036* *(0.014)*	*−0.052* *(0.020)*

Notes: All cells report estimated effects of $1 increase in the state food price basket in the corresponding regression; the average food price is $2.56. Average marginal effects are reported in the probit regressions for obesity and Class II/III obesity. Standard errors, heteroskedasticity-robust and clustered by state, are in parentheses. Bold indicates statistically significant at 1% level, while *italics* indicates 5% level. The regressions include control variables for race, education, marital status, age, inflation-adjusted household income, and a dummy for whether the year was 2008. Sampling weights are used.

NHANES for black men. Furthermore, while black men appear to under-report obesity and class II/III obesity slightly, the incidence of self-reported overweight status is higher than the actual incidence for black men.

5.1.1. BMI as a dependent variable

For our first empirical example, we use the BRFSS to evaluate the relationship between state-level food prices and BMI.[27] Table 4 reports coefficient estimates of interest from OLS regressions of BMI on food price and other basic control variables, as well as probit estimates (presented as average marginal effects) of the effects of food prices on the probabilities of being obese or class II/III obese. In each case, results are presented with no adjustment, adjustment using the standard approach, and then adjustment using our percentile method. The control variables include race (dummies for non-Hispanic black and non-Hispanic white), education (dummies for some high school, high school graduate, some college, and four-year college degree or greater), marital status (dummies for married, divorced, and widowed), age, inflation-adjusted household income, and a dummy for whether the year was 2008. The state food price measure is computed from city-level data from the Council for Community and Economic Research (formerly American Chamber of Commerce Researchers Association) Cost of Living Index.[28] The sample average of the food price measure is $2.56, so a one-dollar increase in food price represents approximately a 40% increase relative to the mean.

The results presented in Table 4 suggest that neither the standard correction nor our correction affect the conclusion that higher food prices are associated with lower body mass; however, the choice of correction method has potentially important implications for the magnitudes of those coefficients. For the full sample and most subgroups, the standard correction leads to larger magnitudes than no correction, while our correction leads to even larger magnitudes than the standard correction. This is consistent with the aforementioned result that measurement error in self-reported weight and height serves to compress the BMI distribution. The more measurement error is eliminated, the more "stretched out" the BMI distribution becomes. In other words, the smaller the measurement error, the larger the change in BMI that is associated with a given change in food prices. For the 0–1 variables, this stretching out of the distribution increases the number of individuals who reported a lower BMI but are now categorized as obese or class II/III obese. It is therefore not surprising that our correction, which purges more measurement error, would lead to larger magnitudes than the standard correction, which purges less measurement error. Of course, both corrections lead to larger magnitudes than using no correction.

More specifically, consider the whole-sample regressions. The effect of food price on BMI using our percentile correction is 23% larger (−0.937 compared to −0.764) than that using no correction, and 9% larger (−0.937 compared to −0.858) than that using the standard levels-based correction. For obesity, the estimated effect using our

[27] In an earlier version of this paper, Courtemanche et al. (2014), we replicated this analysis using the ATUS and found similar results.

[28] Following Chou et al. (2004), for each city we average over the prices of each grocery food item, weighting by the C2ER shares of each item's importance in the basket of goods. We then define state prices as the population-weighted average of the prices in the state's C2ER markets. Prices are in 2008 dollars.

approach is 20% larger than with no correction, and 3.4% larger than with the standard correction. The differences are most striking, however, for class II/III obesity. The estimates with our correction are 64% and 28% larger than those using no correction and the standard correction. The finding that mitigating measurement error matters most for class II/III obesity makes sense in light of the aforementioned results from the literature that the extent of misreporting of both weight and height increases as weight increases (Rowland, 1990; Cawley, 2002). In other words, correcting measurement error leads to the largest increases in BMI in the right tail of the distribution, where the class II/III obesity cutoff lies.

Accurately estimating effects on class II/III obesity is vital, as a recent meta-analysis found that an increased risk of mortality from high BMI does not begin until crossing the class II/III obesity threshold (Flegal et al., 2013). To illustrate, suppose we are interested in predicting lives saved from a calorie tax that raises the price of the food basket by $1. Obesity is estimated to cause 112,000 deaths per year (Flegal et al., 2005), and Flegal et al.'s (2013) results suggest it is reasonable to attribute all of the premature mortality from obesity to class II/III obesity. The class II/III obesity rate in the 2007–2008 NHANES is 14.4%. Using our percentile-rank correction, a $1 increase in food prices would reduce class II/III obesity by an estimated 28.5%, compared to 22.2% using the standard correction and 17.4% using no correction. Multiplying these numbers by the annual deaths from obesity, the estimated lives saved from the hypothetical policy are 31,889 using our correction, compared to 24,889 using the standard correction and 19,444 with no correction. Therefore, the chosen correction method can lead to important differences in policy implications, even if conclusions about sign and statistical significance are unaffected.

Turning to the subsamples, we observe the same general pattern of the magnitudes increasing as measurement error is purged for all groups except black women, black men, and men of a race other than white or black. Even for these three groups, however, our correction still leads to the largest magnitudes for class II/III obesity – substantially larger for black women and other men. For BMI, our correction increases the food-price effect most substantially relative to the standard correction for other men (14% larger estimated food price effect), black women (14% larger), and white women (13%). For class II/III obesity our correction increases the estimated food price effect by over 20% relative to the standard correction for all groups except black men. The largest changes are among other men (44%), white women (41%), and black women (37%).

5.1.2. BMI as an explanatory variable

We next turn to an examination of the implications of our rank-based correction in regressions with a weight-related independent variable. Using the BRFSS, we consider a question of broad interest to epidemiologists and health policy researchers: the impact of obesity on diabetes. We estimate probit models with a dummy for whether the individual has ever been diagnosed with diabetes as the dependent variable; either BMI, obese, or class II/III obese as the independent variable of interest; and the same set of controls as our analysis in Section 5.1.1. In the earlier working paper version (Courtemanche et al., 2014), we also present an example that considers the effects of BMI, obesity, and class II/III obesity on the probability of being disabled using the ATUS data.

Average marginal effects of BMI, obesity, and class II/III obesity on P(Diabetes) are reported in Table 5. In all regressions, the association between BMI, obesity, or class II/III obesity and diabetes is positive and statistically significant, so the correction method again does not

Table 5
BRFSS estimated effects of BMI, obesity, and Class II/III obesity on P(Diabetes) by race-gender group and BMI correction method.

	BMI			Obese			Class II/III Obese		
	Self-report	Standard method	Rank-based	Self-report	Standard method	Rank-based	Self-report	Standard method	Rank-based
Whole sample	**0.0042**	**0.0041**	**0.0041**	**0.069**	**0.065**	**0.062**	**0.107**	**0.102**	**0.096**
	(0.0001)	**(0.0001)**	**(0.0001)**	**(0.002)**	**(0.003)**	**(0.002)**	**(0.005)**	**(0.005)**	**(0.004)**
White women	**0.0036**	**0.0035**	**0.0034**	**0.073**	**0.068**	**0.063**	**0.103**	**0.098**	**0.088**
	(0.0001)	**(0.0001)**	**(0.0001)**	**(0.002)**	**(0.002)**	**(0.002)**	**(0.003)**	**(0.003)**	**(0.003)**
White men	**0.0044**	**0.0042**	**0.0044**	**0.065**	**0.062**	**0.061**	**0.114**	**0.107**	**0.105**
	(0.0002)	**(0.0002)**	**(0.0002)**	**(0.003)**	**(0.002)**	**(0.002)**	**(0.004)**	**(0.004)**	**(0.004)**
Black women	**0.0051**	**0.0050**	**0.0046**	**0.079**	**0.077**	**0.072**	**0.100**	**0.098**	**0.097**
	(0.0002)	**(0.0002)**	**(0.0002)**	**(0.004)**	**(0.004)**	**(0.004)**	**(0.007)**	**(0.007)**	**(0.006)**
Black men	**0.0064**	**0.0058**	**0.0060**	**0.081**	**0.073**	**0.078**	**0.125**	**0.116**	**0.121**
	(0.0005)	**(0.0004)**	**(0.0004)**	**(0.008)**	**(0.008)**	**(0.009)**	**(0.013)**	**(0.015)**	**(0.013)**
Other women	**0.0040**	**0.0043**	**0.0043**	**0.072**	**0.067**	**0.068**	**0.094**	**0.093**	**0.093**
	(0.0002)	**(0.0002)**	**(0.0003)**	**(0.007)**	**(0.006)**	**(0.006)**	**(0.013)**	**(0.012)**	**(0.010)**
Other men	**0.0040**	**0.0041**	**0.0041**	**0.057**	**0.052**	**0.050**	**0.111**	**0.099**	**0.083**
	(0.0004)	**(0.0004)**	**(0.0005)**	**(0.005)**	**(0.006)**	**(0.005)**	**(0.017)**	**(0.021)**	**(0.017)**

Notes: All regressions are probits; the cells report average marginal effects of BMI and average effects of a switch from 0 to 1 in obesity and Class II/III obesity status. The dependent variable is a dummy forever being diagnosed with diabetes; its sample mean is 0.065. Standard errors, heteroskedasticity-robust and clustered by state, are in parentheses. Bold indicates statistically significant at 1% level. The regressions include control variables for race (dummies for non-Hispanic black and non-Hispanic white), education (dummies for some high school but no degree, high school degree but no further, some college but no degree, and four-year college degree or greater), marital status (dummies for married, divorced, and widowed), age, inflation-adjusted household income, and a dummy for whether the year was 2008. Sampling weights are used.

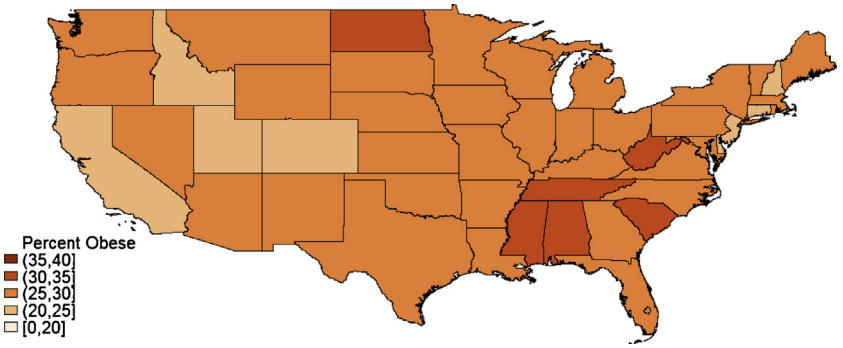


Fig. 3. Prevalence of obesity by state among adults ages 19–64 self-reported BMI.

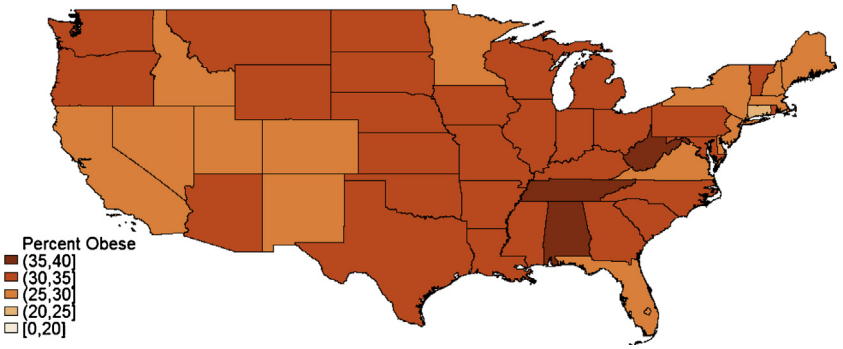


Fig. 4. Prevalence of obesity by state among adults ages 19–64 standard validation method.

influence the general conclusions. However, important differences again emerge in the magnitudes. In most regressions, the standard correction leads to smaller magnitudes than no correction, while our correction leads to even smaller magnitudes than the standard correction. This is the opposite of the pattern observed when the weight-related variable was the outcome, but is again consistent with the observation that measurement error compresses the BMI distribution. Correcting measurement error leads to a larger change in BMI being associated with a given change in diabetes, and therefore a smaller coefficient estimate when BMI is an explanatory variable.

Turning to the results for the full sample, the most interesting observation is that the corrections have only a minimal effect on the estimated relationship between BMI and diabetes, but more substantial effects on the estimates for obesity and class II/III obesity. Being obese is estimated to increase P(Diabetes) by 6.9 percentage points using no correction, 6.5 percentage points using the standard levels correction, and 6.2 percentage points using our percentile correction. Our correction therefore leads to a 10% smaller magnitude than no correction, and a 5% smaller magnitude than the standard correction.

For most subsamples, we observe the same pattern – the estimated effects of obesity and class II/III obesity on P(diabetes) decrease with the amount of measurement error purged. The most notable exceptions are for black men, but they are also the only group for which the incidence of class II/III obesity appears to be lower when our correction is used than it is when the standard correction is used. In the BMI regressions, our correction makes the biggest difference for black women (10% smaller in magnitude compared to no correction, 8% smaller compared to the standard correction). In the obesity and class II/III obesity regressions, our correction is most consequential for white women. The estimated effect of obesity on P(diabetes) using our correction is 14% smaller than using no correction, and 7% smaller than using the standard correction. For class II/III obesity, these numbers are 10% and 6%. It is not surprising that the correction is important for white women since they are the group for which underreported weight is the most common.

5.2. Obesity maps

Finally, we demonstrate the effect of measurement error corrections on the estimated prevalence of obesity in the United States. Fig. 3 is inspired by the well-known obesity maps produced by the CDC using data from BFRSS.[29] Fig. 4 is similar to Fig. 3, but uses the standard method to correct for measurement error before calculating the prevalence of

[29] There are minor differences between our maps and those produced by the CDC. While the CDC map considers all states and uses data on all adults, we focus on adults between the ages of 19 and 64 in the continental United States. We also pool data from 2007 and 2008 into one map instead of creating separate maps for each year.

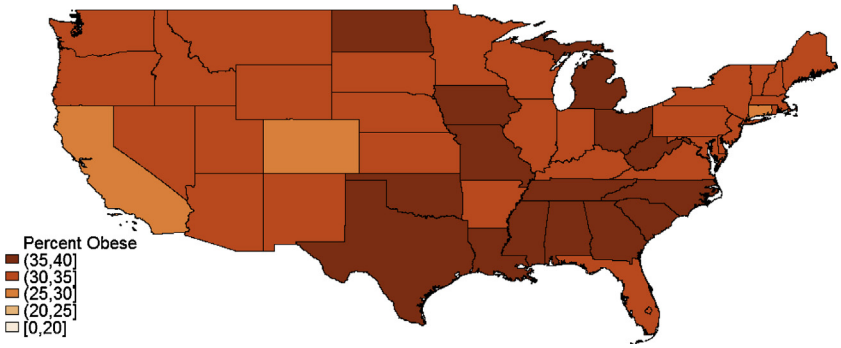


Fig. 5. Prevalence of obesity by state among adults ages 19–64 percentile validation method.

Table 6
Summary of state-level obesity prevalence results.

	Reported obesity	Standard correction	Percentile correction
Number of states in each range of obesity prevalence			
20–25%	8	1	0
25–30%	34	15	3
30–35%	6	29	30
35–40%	0	3	15
Number of states moving to a higher prevalence category, relative to reported prevalence			
No change	–	12	0
Up 1 category	–	36	34
Up 2 categories	–	0	14

Notes: This table summarizes results presented in Figs. 3–5. The data are from the 2007 and 2008 BRFSS. Observations are limited to the 48 states in the continental United States.

obesity. Fig. 5 shows the map after our rank-based method has been used.

These three figures demonstrate that correcting for measurement error has dramatic effects on the estimated prevalence of obesity. Looking at the nation as a whole, we find that 27.3% of the population reports being obese, 30.8% are found to be obese using the standard correction, and 33.3% are found to be obese using our correction.

Table 6 summarizes the pattern of obesity prevalence in Figs. 3–5. When no correction is made, the obesity rates for most states (34) are between 25% and 30%, while no state has a rate over 35%. With the standard correction, most states (29) fall into the 30–35% range and three fall into the 35–40% range. Using our correction, the modal obesity rate is still in the 30–35% range, but now 15 states have obesity rates in the 35–40% range. Furthermore, we find at least a quarter of the adult population in every state is obese when our correction is applied.

It is also interesting to note the number of states that move to a higher interval of obesity prevalence when we correct for measurement error. When using the standard correction, 36 states move to the next highest prevalence interval, while 12 states do not change ranges. Using our approach, no state remains in the same prevalence interval and 14 move up two intervals, which is consistent with a median increase in the prevalence of obesity (not shown) of just over six percentage points.

The effects of correcting for measurement error also vary by state.[30] When we apply our correction, the largest absolute increases in prevalence are in West Virginia and Iowa (7.5 and 7.1 percentage points). The smallest increases, 5.0 and 5.2 points, are in New York and Rhode Island. Relative increases in obesity prevalence, compared to no correction, were largest (over 27%) in Delaware and Utah, and smallest (around 17%) in Mississippi and South Carolina. Finally, the corrections increase the variance between states (not shown) from 5.88 to 6.21 with the standard correction and to 6.53 with our correction.

6. Conclusion

Since Cawley (2002, 2004) it has been common in the economics-of-obesity literature to correct for reporting error in height and weight by using data from the NHANES as an external validation sample for the authors' primary data. The standard approach regresses the measured values on the reported values using NHANES data, and then uses the estimated coefficients to predict height and weight in the primary dataset. This approach relies on the

[30] Differences between states in the amount of misreporting warrant further attention. If prevalence estimates based on uncorrected BRFFS data are used to determine funding for public health initiatives, differences in measurement error could result in misallocated funds.

assumption that the misreporting of height and weight is the same in both surveys, even though interview methods and context often differ between surveys.

We propose an alternative correction that does not require misreporting to be the same across samples. Instead, we assume that if person A says she weighs more than an otherwise similar person B, the conditional expectation of A's actual weight is higher than B's. This assumption implies that the relationship between the percentile rank of a respondent's reported height or weight and her measured height or weight will be the same in both samples, as long as the samples are representative of the same population. We can then use percentile ranks of reported values, in place of the reported values themselves, to predict measured values.

The implementation of our rank-based correction is similar to that of the standard approach. Other than calculating the percentiles, the only difference is that our regressions require functional forms of the percentile ranks that are more flexible than the polynomials that are commonly used with self-reported values. The result is a measurement-error correction that is more robust than the standard approach, while still being easy to implement.

To illustrate the value of our correction, we compare data from the NHANES and two nationally representative telephone surveys, the BRFSS and the ATUS. Since all three datasets are representative of the same population, we would expect the distributions of reported height and weight to be the same if there are no survey effects; however, we repeatedly reject the hypothesis that the distributions of these self-reports are the same across datasets. Furthermore, our results suggest that misreporting is more sensitive to context in cases (e.g., the weight of white women) where misreporting is more severe in the NHANES sample.

When we compare predictions of BMI using the standard approach and our rank-based approach, we find that the standard approach predicts statistically significant differences in the distributions of BMI between samples that are representative of the same populations, and those differences reflect the differences in respondent reporting. We find no evidence of such differences between samples when BMI is predicted using our percentile-rank method. In other words, the standard approach appears to be biased by differences between samples in the misreporting it aims to correct, while our approach is robust to such differences.

We also consider how corrections for misreported height and weight might affect empirical research. We find that the estimated prevalence of obesity and class II/III obesity is higher when our rank-based correction is used than when either the standard correction or no correction is used. Next, we present examples of regression estimates with body mass as either a dependent or an independent variable. In each of these examples, we find differences in coefficient estimates that are consistent with measurement error compressing the upper tail of the distribution. The differences in coefficient estimates between our correction and the standard correction are often similar in size to the differences in estimates between the standard correction and uncorrected reports. These differences are often economically significant and could affect the conclusions researchers and policymakers draw.

There are a few caveats we wish to discuss. First of all, we must stress that neither the standard approach nor our proposed alternative should be used every time researchers encounter self-reported height and weight. We argue that our rank-based method should be used when the validation and primary samples are representative of the same population and there is concern that misreporting differs across samples. However, the standard approach is appropriate when this concern is not present, such as when the validation data are a random subsample of the primary data. One such example is the Physical Measures subsample of the Health and Retirement Study. The use of an internal validation sample should obviously be preferred to an external validation sample, even when our correction is used; however, such data are rarely available to social scientists.

There are also cases in which we believe no use of validation data is likely to reliably address the measurement error concern. In particular, both the standard validation method and our alternative method rely on the assumption that the validation sample and the primary sample are representative of the same population.[31] Researchers who doubt that this assumption holds in their data should be wary of using either validation method. Obvious examples include when the primary and validation datasets cover different time periods or geographic locations, or sample different age ranges. In such cases, it is unclear whether measurement error would be minimized by the standard correction, our correction, or no correction. One strategy would be to restrict the validation sample to match the primary sample as closely as possible, and then verify that the conclusions reached are similar using each of the three approaches.[32] Additionally, it is unclear that either correction method reliably improves on uncorrected measures when using primary data from surveys, such as the NLSY cohorts, that include non-random combinations of in-person and telephone interviews.[33] In some applications, researchers may need to adjust for differing patterns of misreporting within their sample before any correction is applied using external validation data.

More broadly, our work should be seen as a warning (or reminder) that external validation data should be used with caution. The reporting error we wish to correct can be sensitive to differences in interview context, and even our rank-based method requires stronger assumptions than the use of an internal validation sample would. Furthermore, context could vary in ways that are

[31] Differences in misreporting between populations would affect the standard correction, and differences in the distributions of actual height and weight would affect both corrections. See Section 2 for more detail.

[32] For instance, if the primary data includes only 18–39 year olds, researchers should drop those 40 and older from the NHANES when constructing the validation sample. Researchers should also match their primary data to NHANES surveys from the closest years available.

[33] We invite readers who still doubt the potential effects of context on misreporting to look at how self-reported weight varies with interview mode in the NLSY79 or NLSY97.

less obvious than differences in interview method or sample age range. For example, there is no reason to assume that misreporting in surveys that span decades, such as the BRFSS or NLSY79, has remained constant over time as waistlines expanded and social norms changed.[34]

Appendix. How to use our percentile-based correction

Although our percentile-based approach may sound more complicated to implement than the standard approach, we want to assure the reader that the added complication is trivial. Our approach adds two simple steps to the standard approach. Each of those two steps requires one line of code in Stata.

First, we find the percentile rank of the reported measure in the relevant subsample. This is easily done in Stata using the "cumul" command by dataset, race or ethnic group, and gender. Using reported weight as an example, we would have:

bysort *dataset race sex* : cumul *reported_weight* [aw = *samp_wt*], *g*(*wt_rank*) equal

where *wt_rank* is the newly created percentile rank in the distribution of *reported_weight* for the subsample determined by the indicators *dataset, race* and *sex*.

Actual weight must then be regressed on a flexible function of *wt_rank*, and a polynomial in age. We found that simple polynomials in *wt_rank* were not flexible enough to predict actual weight, and used cubic basis splines in *wt_rank* instead. The second step our method adds to the standard approach generates the splines with the user-written command "bspline".[35] For example,

bspline, xvar(*wt_rank*) p(3) gen(*wt_spline*)
knots(0, .05, .1, .25, .5, .75, .9, .95, 1)

where *wt_spline* is the prefix of the generated splines.[36]

From this point on, our approach closely resembles the standard approach. Our splines simply replace the polynomial in reported weight that previous authors have used:

reg *actual_weight* (i.*race*i.*sex*)(c.*age*c.*age*c.*age*c.*wt_spline*∗)
[aw = *samp_wt*], nocons

Weight is predicted in both the validation sample, which contains *actual_weight*, and the primary sample. Finally, the process can be repeated for height, allowing a predicted BMI measure to be constructed.[37]

[34] Courtemanche et al. (2015b) apply our correction to BRFSS year-by-year for exactly this reason. Cawley and Burkhauser (2006) also acknowledge this problem.

[35] The command "bspline" was written by Roger Newson. Documentation and code can be found here: http://econpapers.repec.org/RePEc:boc:bocode:s411701.

[36] The number and spacing of knots can be adjusted as needed to improve the fit of the predicted values.

[37] Alternatively, BMI itself could be predicted directly following the same approach.

References

Baum, C., 2009. The effects of cigarette costs on BMI and obesity. Health Econ. 18 (1), 3–19.
Baum, C., 2011. Effects of food stamps on obesity. South. Econ. J. 77 (3), 623–651.
Baum, C., Ruhm, C., 2009. Age, socioeconomic status and obesity growth. J. Health Econ. 28 (3), 635–648.
Bound, J., Brown, C., Mathiowetz, N., 2001. Measurement error in survey data. Handb. Econom. 5, 3705–3843.
Burkhauser, R., Cawley, J., 2008. Beyond BMI: the value of more accurate measures of fatness and obesity in social science research. J. Health Econ. 27 (2), 519–529.
Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M., 2006. Measurement Error in Nonlinear Models: A Modern Perspective, 2nd ed. Chapman & Hall, Boca Raton, FL.
Cawley, J., 2002. Addiction and the Consumption of Calories: Implications for Obesity. (Unpublished manuscript)Cornell University.
Cawley, J., 2004. The impact of obesity on wages. J. Hum. Resour. 39 (2), 451–474.
Cawley, J., Burkhauser, R., 2006. Beyond BMI: The Value of More Accurate Measures of Fatness and Obesity in Social Science Research. NBER Working Paper No. 12291.
Cawley, J., Choi, A., 2014. Health Disparities Across Education: The Role of Differential Reporting Error.
Cawley, J., Danzinger, S., 2005. Morbid obesity and the transition from welfare to work. J. Policy Anal. Manage. 24 (4), 727–743.
Cawley, J., Meyerhoefer, C., 2012. The medical care costs of obesity: an instrumental variables approach. J. Health Econ. 31 (1), 219–230.
Cawley, J., Moran, J., Simon, K., 2010. The impact of income on the weight of elderly Americans. Health Econ. 19 (8), 979–993.
Chou, S.-Y., Grossman, M., Saffer, H., 2004. An economic analysis of adult obesity: results from the Behavioral Risk Factor Surveillance System. J. Health Econ. 23 (3), 565–587.
Courtemanche, C., Pinkston, J.C., Stewart, J., 2014. Adjusting Body Mass for Measurement Error with Invalid Validation Data. NBER Working Paper 19928.
Courtemanche, C., Heutel, G., McAlvanah, P., 2015a. Impatience, incentives, and obesity. Econ. J. 125 (582), 1–31.
Courtemanche, C., Pinkston, J., Ruhm, C., Wehby, G., 2015b. Can Changing Economic Factors Explain the Rise in Obesity. NBER Working Paper No. 20892.
Eid, J., Overman, H., Puga, D., Turner, M., 2008. Fat city: questioning the relationship between urban sprawl and obesity. J. Urban Econ. 63 (2), 385–404.
Fan, M., 2010. Do food stamps contribute to obesity in low-income women? Evidence from the National Longitudinal Survey of Youth 1979. Am. J. Agric. Econ. 92 (4), 1165–1180.
Flegal, K., Graubard, B., Williamson, D., Gail, M., 2005. Excess deaths associated with underweight, overweight, and obesity. J. Am. Med. Assoc. 293 (15), 1861–1867.
Flegal, K., Carroll, M., Kuczmarski, R., Johnson, C., 1998. Overweight and obesity in the United States: prevalence and trends, 1960–1994. Int. J. Obes. Relat. Metab. Disord. 22 (1), 39–47.
Flegal, K., Kit, B., Orpana, H., Graubard, B., 2013. Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis. J. Am. Med. Assoc. 309 (1), 71–82.
Goldman, D., Lakdawalla, D., Zheng, Y., 2011. Food prices and the dynamics of body weight. In: Grossman, M., Mocan, N. (Eds.), Economic Aspects of Obesity. University of Chicago Press, Chicago, IL.
Gregory, C., 2010. Wages, BMI, and Age. (Unpublished manuscript)United States Department of Agriculture Economic Research Service.
Gregory, C., Ruhm, C., 2011. Where does the wage penalty bite? In: Michael, G., Mocan, N. (Eds.), Economic Aspects of Obesity. University of Chicago Press, Chicago, IL.
Han, E., Norton, E.C., Stearns, S.C., 2009. Weight and wages: fat versus lean paychecks. Health Econ. 18 (5), 535–548.
Lakdawalla, D., Philipson, T., 2002. The Growth of Obesity and Technological Change: A Theoretical and Empirical Investigation. National Bureau of Economic Research Working Paper No. 8965.
Lee, L.-F., Sepanski, J.H., 1995. Estimation of linear and nonlinear errors-in-variables models using validation data. J. Am. Stat. Assoc. 90 (429), 130–140.
Maclean, J.C., Sikora, A., 2014. Personality Traits and Body Weight: Evidence From the Health and Retirement Study.
Majumder, M.A., 2013. Does obesity matter for wages? Evidence from the United States. Econ. Pap. 32 (2), 200–217.

Milgrom, P.R., 1981. Good news and bad news: representation theorems and applications. Bell J. Econ. 12 (2), 380–391.

Ogden, C.L., Carroll, M.D., 2010. Prevalence of Overweight, Obesity, and Extreme Obesity Among Adults: United States, Trends 1960–1962 Through 2007–2008. National Center for Health Statistics Report, http://www.cdc.gov/nchs/data/hestat/obesity_adult_07_08/obesity_adult_07_08.htm.

Pinkston, J.C., 2015. The Dynamic Effects of Obesity on the Wages of Young Workers. University of Louisville, Mimeo. Available at SSRN: http://ssrn.com/abstract=2537554.

Plantinga, A., Bernell, S., 2007. The association between urban sprawl and obesity: is it a two-way street? J. Reg. Sci. 47 (5), 857–879.

Rogerson, W.P., 1985. The first-order approach to principal-agent problems. Econometrica 53 (6), 1357–1367.

Rowland, M.L., 1990. Self-reported weight and height. Am. J. Clin. Nutr. 52 (6), 1125–1133.

Ruhm, C., 2005. Healthy living in hard times. J. Health Econ. 24 (2), 341–363.

Sturm, R., 2002. The effects of obesity, smoking, and drinking on medical problems and costs. Health affair 21 (2), 245–253.