Bollinger, Christopher R.; Chandra, Amitabh

Working Paper

Iatrogenic Specification Error : A Cautionary Tale of Cleaning Data

IZA Discussion Papers, No. 1093

Provided in Cooperation with:
IZA – Institute of Labor Economics

DISCUSSION PAPER SERIES

IZA DP No. 1093

Iatrogenic Specification Error: A Cautionary Tale of Cleaning Data

Christopher R. Bollinger
Amitabh Chandra

March 2004

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Iatrogenic Specification Error: A Cautionary Tale of Cleaning Data

Christopher R. Bollinger
University of Kentucky

Amitabh Chandra
Dartmouth College,
NBER and IZA Bonn

Discussion Paper No. 1093
March 2004

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

IZA Discussion Paper No. 1093
March 2004

ABSTRACT

Iatrogenic Specification Error: A Cautionary Tale of Cleaning Data

In empirical research it is common practice to use sensible rules of thumb for cleaning data. Measurement error is often the justification for removing (trimming) or recoding (winsorizing) observations whose values lie outside a specified range. We consider a general measurement error process that nests many plausible models. Analytic results demonstrate that winsorizing and trimming are only solutions for a narrow class of measurement error processes. Indeed, for the measurement error processes found in most social-science data, such procedures can induce or exacerbate bias, and even inflate the variance estimates. We term this source of bias "Iatrogenic" (or econometrician induced) error. Monte Carlo simulations and empirical results from the Census PUMS data and 2001 CPS data demonstrate the fragility of trimming and winsorizing as solutions to measurement error in the dependent variable. Even on asymptotic variance and RMSE criteria, we are unable to find generalizable justifications for commonly used cleaning procedures.

JEL Classification: C1, J1

Keywords: measurement error models, trimming, winsorizing

Corresponding author:

Amitabh Chandra
Department of Economics
Dartmouth College
6106 Rockefeller Hall
Hanover, NH 03755
USA
Email: amitabh.chandra@dartmouth.edu

1 Introduction

Empirical researchers frequently use simple rules of thumb to clean data on the basis of the dependent variable. As an example, researchers analyzing survey reports of wages and salaries often remove observations whose value for the hourly wage is below the minimum wage or above some prespecified cutoff: sample exclusions based on wages can be found in Katz and Murphy (1992), Card and Krueger (1992), Bound and Freeman (1992), Juhn, Murphy, and Pierce (1993), and Buchinsky (1994). We cite these authors to illustrate the endorsement of this practice by leading scholars in the field. As we demonstrate in this paper, the intuitively appealing strategy of discarding certain observations is not costless and can introduce specification error in cases where no error previously existed. Given the fact that the inconsistency is exacerbated by the analyst's actions, we borrow a term from the medical literature and term this form of bias "iatrogenic" specification error. In the medical literature an iatrogenic event is an adverse reaction to a well-intentioned treatment initiated by a physician, and we believe that parameter inconsistency that is caused by the analysts well intentioned actions shares the same features of physician induced complications.

Given the widespread acceptance of this practice, the topic of "robust" estimation has received the attention of both economists and statisticians. In one the earliest formal examinations, Stigler (1977) poses an interesting question: how much have methods such as trimming, winsorizing, the Edgeworth average, or Tukey's Biweight, reduced the bias in the laboratory estimation of physical constants such as the speed of light or the density of the earth? Stigler concludes that the 10 percent trimmed mean, the smallest trimming amount considered in his study, is the most reliable estimator. In this he echoes the famous mathematician Legendre who recommended deleting those observations with errors "too large to be admissible." Stigler looks at the role of measurement error in the physical sciences; the error process may be vastly different in the social-sciences where economic agents may have strategic or cultural incentives to inflate or deflate their reports. In the econometrics literature, Angrist and Krueger (2000) apply trimming and winsorizing techniques to the matched employer-employee data from Mellow and Sider (1983). When they trim both the employer and employee wage data, they find that the correlation between the two measures improves. Interestingly, this result does not hold for reports of hours worked. On the basis of this finding they conclude that "a small amount of trimming could be beneficial." Their prescription, which summarizes the intuition and current practice of most analysts, may be summarized as:

> "Loosely speaking, winsorizing the data is desirable if the extreme values are exaggerated versions of the true values, but the true values still lie in the tails. Truncating the sample is more desirable if the extremes are mistakes that bear no resemblance to the true values. (p.1349)"

We examine this practice in detail here, using wages and earnings as a motivating example, though the results are likely to apply to other errors of measurement in survey data. We posit a general model of response error in the *dependent variable* of a linear regression model and characterize the effect of different cleaning techniques on the estimated coefficients. We demonstrate, both analytically as well as through the use of simulations, that in general there is no reason to believe that removing "obvious errors" in the dependent variable reduces bias. This is similar in spirit, to the finding of Hyslop and Imbens (2001), who examine instrumental variables approaches to solving the measurement error problem and find that they only apply to very specific measurement error processes. Our work is most closely related to that of MacDonald and Robinson (1985), who consider Bayesian estimation of an error components model in panel data when one of the error components is measurement error. They explicitly show that trimming can be thought of as an extremely dogmatic prior belief. Our paper differs in that we explicitly consider a general error process and we do not consider a panel setting. Moreover we discuss an optimal trimming approach and other classical approaches to estimation. We demonstrate that the results in Stigler (1977) do not necessarily carry over in a regression framework. Indeed, trimming or winsorizing can bias coefficient estimates by as much as 10-30 percent, and in many cases either induces bias that did not previously exist, or exacerbates the bias due to measurement error. The intuition for our result is simple: assuming that the researcher trims or winsorizes the data based on a lower bound of c or an upper bound of C, it will be shown that cleaning creates selection-bias, and this is generally worse than the effects of measurement-error in the dependent variable.

Our paper is organized as follows: Section 2 describes identification with general measurement error in the dependent variable. We use the linear projection of the mismeasured dependent variable onto the covariates to derive analytical results. In Section 3, we generalize the use of this projection to consider three specific models of measurement error (additive white noise, linear transformation and the contaminated data process) that are found in social-science data. Section 4 examines the theoretical implications of trimming the data on bias as well as the asymptotic variances of the coefficients. We prove that only in highly specialized cases, unlikely to be found in social-science data, does cleaning reduce bias. In these cases, we demonstrate that the information necessary to reduce bias leads to a simpler correction that requires fewer assumptions. This section also demonstrates that trimming will not necessarily reduce standard-errors. Section 5 presents simulation results for the cases considered analytically. We generate quasi-simulated data from the 1990 US Decennial Census to study the properties of winsorizing and trimming in a multivariate context. Finally, we present an empirical example from the March 2001 Current Population Survey (CPS). These simulations and examples support the results of the earlier two sections. Section 6 provides concluding comments. The Appendix to this paper provides detailed mathematical proofs and also considers the effects of winsorizing on bias and efficiency.

2 General Measurement Error in the Dependent Variable

To evaluate the widespread practice of "cleaning" data as described above we consider a general model for measurement error processes in the dependent variable. To keep the analysis simple, we focus on a linear regression model as the underlying structural model of interest to the researcher. Assume that the relationship between the "true" dependent variable and the covariate is described by:

$$y_i^* = x_i^{'}\beta + u_i, \tag{1}$$

Our maintained assumption is that the analyst is interested in estimates of β.[1] We assume a general process that relates the true value y_i^* to the observed value y_i :

$$y_i = h\left(y_i^*, \varepsilon_i\right), \tag{2}$$

There are six assumptions that are made for identification of the vector β and its associated covariance matrix:[2]

$$\begin{array}{lll} A1 & : & E\left[u_i|x_i\right] = 0 \\ A2 & : & x_i \text{ is a vector random variable with mean 0 and full rank second moment matrix } V_x \\ A3 & : & \text{Random Sampling} \\ A4 & : & h(.,.) \text{ has finitely many discontinuities} \\ A5 & : & \varepsilon_i \text{ is independent of } \left(y_i^*, x_i, u_i\right) \\ A6 & : & Cov\left(y_i, y_i^*\right) > 0 \end{array}$$

Regardless of the process in equation 2, one summary of the joint distribution of y_i and y_i^* is the population linear projection of y_i on y_i^* :

$$y_i = \delta + \gamma y_i^* + e_i. \tag{3}$$

Here, $\delta = E\left[y_i\right], \gamma = \frac{Cov(y_i, y_i^*)}{V\left(y_i^*\right)}$,and $E\left[e_i\right] = E\left[e_i y_i^*\right] = 0$. The linear projection is not a statement about the data generating process, but rather a summary measure of the joint distribution of $\left(y_i, y_i^*\right)$. The actual measurement process, as defined by $h\left(y_i^*, \varepsilon\right)$ may be substantially more complicated. Assumption A6 insures that $\gamma > 0$.

The researcher is only able to observe (y_i, x_i). Substituting equation 1 into equation 3 yields:

$$y_i = \delta + x_i'\beta\gamma + \gamma u_i + e_i. \tag{4}$$

Assumption A5 insures that $Cov\left(x_i, \gamma u_i + e_i\right) = 0$ and $E\left[\gamma u_i + e_i\right] = 0$. This defines the population linear projection of y_i on x_i :

$$y_i = a + x_i'b + \eta_i \tag{5}$$

where $b = \gamma\beta$, $a = \delta$,and $\eta_i = \gamma u_i + e_i$.

[1] If the analyst is not interested in β per se, but other features of the joint distribution between y and x such as $cov(y,x)$ or $var(y|x)$, then it is possible that our results do not apply. Further analysis is required to understand the applicability of trimming, winsorizing or even rescaling for this class of problems. We thank an anonymous referee for suggesting this caveat.

[2] The mean independence assumption is stronger than necessary for identification of the vector β, but allows for a simpler analysis below. The zero mean for x_i is the usual normalization. The fourth assumption is necessary for moments to be well defined, and A6 simply ensures that the measurement error process is not so perverse that y_i is uninformative about y_i^* (covariance of zero), or that y_i and y_i^* are negatively related. Indeed, the necessary condition would simply be that the sign of the covariance were known and that the covariance is not zero. The fifth assumption is the strongest one. It implies that the measurement error process is independent of x_i and u_i except through y_i^*, and insures that $f(y_i|y_i^*) = f(y_i|y_i^*, x_i, u_i)$.

Therefore, the OLS regression of y_i on x_i yields a consistent estimate of b which is proportional to β. The parameters of interest are identified up to an unknown scaling constant. This would imply that estimates of ratios of the parameters are consistent. In some settings, identification up to scale is considered sufficient. For example, in wage regressions the coefficients on years of education and years of labor market experience can be combined to consistently identify the relative return of experience to education; a fact that might be sufficient for estimation of schooling choices. In general however, we assume the researcher is interested in recovering the parameters β. This suggests two important identification approaches: obtain information about the scaling constant γ, or obtain information about one of the elements in β. While it may be possible to obtain some consistent estimate of one element in β from auxiliary regressions or economic theory, the use of validation data may permit estimation of γ. Bound and Krueger (1992) and Bollinger (1998) have examined the structure of response error when y is the natural log of annual labor market earnings using Social Security Income data matched to the Current Population Survey. They find a point estimate for γ is 0.90. This estimate could be used in log wage models to rescale slope coefficients to account for measurement error.[3]

3 Specific Measurement Error Models

The above analysis holds for general examples of measurement error. In this section, we present three special cases of the above model. These cases are chosen because they are commonly examined or supported in the literature or lead to results in the context of this paper which are of interest. Assumptions 1-6 are maintained, additional assumptions are also imposed.

3.1 Additive White Noise:

The classical measurement error process is often assumed: $y_i = y_i^* + \varepsilon_i$, and $E[\varepsilon_i] = 0$. Indeed, the error term in regression models is often motivated as measurement error. The parameters of the linear-projection of y on y^* are $\gamma = 1$, $\delta = a = 0$, and the least squares estimates are consistent for the parameters of interest β. In this model, if y_i^* were hourly wages, it would be possible to have observations less than the minimum wage (or for that matter even negative observations) and observations above whatever threshold is deemed as a maximum. While it may be true that observations outside the acceptable region are measured with error, observations within the acceptable region are also measured with error. However, as is well known, classical measurement error does not lead to any bias– the estimated standard-errors are inflated but all statistical tests remain valid. Researchers will often point out that "standard errors are too large" because of the additional measurement error. Standard errors are meant to capture the variation in estimates due to

[3]The results in Bound and Krueger (1991) and Bollinger (1998) rely on estimates from the the 1977 and 1978 CPS-SSA matched files. It is possible that the structure of measurement-error has changed over time, thereby reducing the applicability of the rescaling option since it hinges critically on knowledge of the correct γ. Estimation of γ is further complicated by the fact that low earning repondents in the SSA data may be reporting their CPS earnings correctly. Examining these hypotheses is an important avenue for future research.

differences across samples. As long as the data generating process does not change, the sampling variation of the estimating coefficients will depend on the variation in both the structural model, as well as the variation in the error model. Hence, estimates of the standard error are not biased, but rather reflect the variation across samples for this data generating process.

3.2 Linear Measurement Error

A second case is where the data generating process is linear: $y_i = d + gy_i^* + \varepsilon_i$. Here, the parameters in the linear-projection of y on y^* are $\gamma = g$ and $\delta = d$, and the model can either have $\gamma > 1$ or $\gamma < 1$. The data generating process can lead to observations outside the "acceptable" range. Because of the values of δ and the distribution of ε_i, even if $\gamma < 1$, it is quite possible to have both observations that are "too high" and observations that are "too low". Empirical work by Bollinger (1998) and Bound and Krueger (1991) supports the possibility that $\gamma < 1$. For example, using non-parametric regression on the 1978 CPS-SSA matched data, Bollinger (1998) estimates that γ is equal to 0.91 for men and 0.97 for women. He estimates the intercepts δ to be \$1,364 and \$211 respectively. Cognitive psychologists have noted that this model, with $\gamma < 1$, will arise when respondents exhibit "regression to the mean." If survey respondents give answers that try to make them appear "average," then those below the mean report higher values, on average, while those above the mean report lower values, on average. Similarly, the hot deck procedure used by Census to impute earnings can also lead to a regression to the mean (Hirsch and Schumacher, 2001). To our knowledge, no study has found any variable with $\gamma > 1$.

3.3 Contaminated Data

A third example is a simple contaminated sample: $y_i = (y_i^*) * 1[\varepsilon_{1i} > \kappa] + (d + \varepsilon_{2i}) * 1[\varepsilon_{1i} < \kappa]$. The term $1[.]$ is the indicator function and $(\varepsilon_{1i}, \varepsilon_{2i})$ are mean zero and mutually independent. This model produces a mixture: with some probability $p = \Pr[\varepsilon_{1i} > \kappa]$, we observe the true variable y_i^*, while with probability $(1-p)$ we observe only noise: $(d + \varepsilon_{3i})$. This leads to a model where we have some correctly measured observations and some observations where the observed y has no relationship to the actual y^*. In this model $\gamma = p$ and $\delta = d(1-p)$. Again, some observations may fall outside a given range, depending on the distribution of ε_{2i} and the value of d. An important implication of this model is that estimates of the slope parameter β can be obtained if an estimate of p is available. Horowitz and Manski (1995) note that the expectation of y^* given x cannot be bounded unless information about d is available. Our analysis does not contradict this, but rather points out that in a linear model, the slopes can be identified up to the contamination rate. In many cases researchers have *a priori* bounds for the contamination rate. The bounds on the contamination rate will yield trivial bounds for the slope coefficients. If $\underline{p} < p < \overline{p}$, then the elements of β, β_j, are bounded by $\left[\frac{b_j}{\underline{p}}, \frac{b_j}{\overline{p}}\right]$.

4 Effect of Cleaning

We assume that the researcher truncates above and below the mean. The cleaning approaches we consider are defined by

$$\{y_i, x_i | c \leq y_i \leq C\} \tag{6}$$

for known constants (c, C) such that $c < E[y] < C$. We compare the slopes obtained from a least squares projection of the cleaned y_i on x_i to those obtained from the uncleaned data regressed on the covariate (that is, relative to b the biased estimate of β from the uncensored data). Since the choice for the researcher is to "clean" or not clean, this is the relevant comparison. As noted in the section above, the slope b may be larger or smaller in magnitude than the true slope β.

4.1 Analytic Results: Trimmed Data

We first derive analytic results under the additional assumption of joint-normality.

$$A7 : (y_i, x_i) \text{ are jointly normal.}$$

As Goldberger (1981) demonstrates, the slope vector from the least squares projection of y_i on x_i in the truncated sample (where observations above C and below c are discarded) is given by:

Proposition 1 *Under assumptions 1-7, the truncated slope b^* is attenuated relative to the slope b from the least squares projection in the full sample.*

$$b^* = \left(\frac{\theta}{1 - (1-\theta)\rho^2}\right) b \tag{7}$$

with

$$\theta = \frac{V(y_i | c \leq y_i \leq C)}{V(y_i)} \tag{8}$$

and

$$\rho^2 = \frac{b^2 \sigma_x^2}{V(y_i)}. \tag{9}$$

Proof: Goldberger (1981).

As Goldberger notes, $0 \leq \frac{\theta}{1-(1-\theta)\rho^2} \leq 1$.[4] Clearly, if $\gamma \leq 1$, then the attenuation bias of the measurement error is exacerbated by the attenuation bias of the sample truncation. Hence, only if the researcher is certain that $\gamma > 1$ can truncation alleviate bias from measurement error. For this case, the optimal level is determined by finding (c, C) such that $\left(\frac{\theta}{1-(1-\theta)\lambda}\right)\gamma = 1$. With two unknown terms and only one restriction, there are many solutions. Too little truncation will fail to fully correct for the bias, while too much will overcorrect. Therefore, selecting the trimming bounds on the basis of *a priori* values for the supports of y^* (as proxied

[4] Since y_i is normally distributed, the variance of the doubly truncated distribution can be expressed (see Madalla, 1983) as:
$$V(y_i | c \leq y_i \leq C) = V(y) \left[1 + \left[\frac{\left(\frac{c-E[y_i]}{V(y_i)}\right)\phi\left(\frac{c-E[y_i]}{V(y_i)}\right) - \left(\frac{C-E[y_i]}{V(y_i)}\right)\phi\left(\frac{C-E[y_i]}{V(y_i)}\right)}{\Phi\left(\frac{C-E[y_i]}{V(y_i)}\right) - \Phi\left(\frac{c-E[y_i]}{V(y_i)}\right)}\right] - \left[\frac{\phi\left(\frac{c-E[y_i]}{V(y_i)}\right) - \phi\left(\frac{C-E[y_i]}{V(y_i)}\right)}{\Phi\left(\frac{C-E[y_i]}{V(y_i)}\right) - \Phi\left(\frac{c-E[y_i]}{V(y_i)}\right)}\right]^2\right]$$

by the 1 and 99 percentiles of the wage distribution, or trimming at the minimum wage) will not necessarily correspond to the optimal trimming rule. Other solutions exist as well. For example, if the analyst chooses $c = E[y] - c^*$ and $C = E[y] + c^*$, only the term c^* needs to be found in order to devise an optimal trimming rule. It is implicitly described in the next proposition:

Proposition 2 *Under assumptions 1-7 and $\gamma > 1$, an optimal trimming rule of the form $\{y_i, x_i | c \leq y_i \leq C\}$ with $c = E[y] - c^*$ and $C = E[y] + c^*$ may be derived implicitly as:*

$$2\left(\frac{c^*}{V(y)}\right)\left(\frac{\phi\left(\frac{c^*}{V(y)}\right)}{\Phi\left(\frac{c^*}{V(y)}\right) - \Phi\left(\frac{-c^*}{V(y)}\right)}\right) = \frac{\gamma - 1}{\gamma - \rho^2}. \tag{10}$$

Proof: see Appendix.

Because the solution involves the cdf of the standard normal distribution, there is no closed form expression. The optimal cleaning depends on the variance of the observed y, the correlation between y and x, and γ. The right hand side of (10) is increasing in γ and the left hand side of (10) is decreasing in c^*. Therefore, as γ increases, the truncation points must move closer to the mean and the data must be truncated more heavily. In order to use this approach a number of highly restrictive assumptions must be met. First, the data must be jointly normally distributed. Any discrete variables in x_i will violate this assumption. Second, the measurement error process must result in a projection equation for y_i on y_i^* where $\gamma > 1$. Finally, specific information on γ must be obtained in order to arrive at a truncation rule. Ad hoc cleaning approaches that ignore the strong nature of these assumptions may be of little value in reducing the bias in b.

The variance of the measurement error is often used as a measure of the severity of the error. Since $b^* = \left(\frac{\theta}{1-(1-\theta)\rho^2}\right) b$ under trimming, the derivative of $\frac{\theta}{1-(1-\theta)\rho^2}$ with respect to σ_ε^2 reveals how the truncation bias is effected by the measurement error. This result motivates the next proposition:

Proposition 3 *Under assumptions 1-7, the absolute value of the difference between elements of b and b^* becomes larger since, $\frac{\partial}{\partial \sigma_\varepsilon^2}\left(\frac{\theta}{1-(1-\theta)\rho^2}\right) < 0$.*

The above proposition is may appear to be *prima-facia* counterintuitive, but the intuition behind it is simple. As the measurement error becomes more severe (as measured by it variance), trimming is more likely to result in deleting observations based on the regression error instead of the measurement error, thereby causing sample-selection bias.

4.1.1 Does Trimming Reduce Standard Errors?

A second reason sometimes cited for trimming is the reduction of standard errors. To examine this procedure more rigorously we begin by noting that the truncation will introduce heteroskedasticity by reducing the variance of errors in the tails of the distribution. Therefore, the asymptotic variance of the estimated slope from the truncated data can be derived from the expression: $AV\left(\widehat{b^*}\right) = Q^{-1}E\left[\left(y_i - x_i'b^*\right)^2 x_i x_i' | c \le y_i \le C\right] Q^{-1}$, where $Q = E\left[x_i x_i^T | c \le y_i \le C\right]$.

In the appendix we demonstrate that this expression for asymptotic variance may be written as the sum of two terms:

$$\begin{aligned} AV\left(\widehat{b^*}\right) &= \theta V\left(y_i\right)\left(1 - \left(\frac{\theta}{1-(1-\theta)\rho^2}\right)\rho^2\right)Q^{-1} \\ &\quad + Q^{-1}E\left[\left(x_i'b^* - m\left(x_i\right)\right)^2 x_i x_i' | c \le y_i \le C\right] Q^{-1}. \end{aligned} \tag{11}$$

The size of the leading term is indeterminate relative to its full-sample OLS counterpart.[5] This fact is in contrast to similar comparisons for the mean (and consequently the results in Stigler (1977)). For the trimmed mean, the term $\theta V(y_i) < V(y_i)$ and therefore trimming necessarily reduces the variance in the leading term. Here, the comparison is not so straight forward. The second term is due to heteroskedasticity from the trimming and is necessarily positive definite. Hence, even if the leading term is smaller than the variance of the OLS estimate on the full sample (in a positive-definite sense), the second term may reverse, or at least mitigate that difference.

Finally, an often overlooked reason for why trimming may not reduce standard-errors is the effect of truncation on sample size. The finite sample variance of $\widehat{b^*}$ is given by

$$V\left(\widehat{b^*}\right) = \frac{AV\left(\widehat{b^*}\right)}{N * \left(\Phi\left(\frac{C-E[y_i]}{V(y_i)}\right) - \Phi\left(\frac{c-E[y_i]}{V(y_i)}\right)\right)}. \tag{12}$$

The term $\Phi\left(\frac{C-E[y_i]}{V(y_i)}\right) - \Phi\left(\frac{c-E[y_i]}{V(y_i)}\right) < 1$, measures the proportion of the sample discarded from the truncation rule, and increases in this proportion will raise estimates of the finite sample variance.

In conclusion, we find that comparisons between the variance of the truncated estimates and the variance of the full sample estimates is complicated and depends on the underlying parameters of the joint distribution. It is not possible to sign this difference, even under normality. In fact, the simulations below show little or no effect of trimming on standard-errors.

4.2 Analytic Results: Winsorized Data

Rather than truncation, winsorized data are censored at the points c and C. Here, no observations are removed, but values of y_i outside of the region (c, C) are transformed as:

$$y_i^w = \begin{array}{cc} C & \text{if } y_i \ge C \\ y_i & \text{if } c < y_i < C \\ c & y_i \le c. \end{array} \tag{13}$$

[5] The leading term is comparable to the asymptotic variance expression for the OLS estimate in the full sample: $V(y_i)\left(1-\rho^2\right)E\left[\underline{x}_i\underline{x}_i^T\right]^{-1}$. Furthermore, $\theta V(y_i)\left(1 - \left(\frac{\theta}{1-(1-\theta)\rho^2}\right)\rho^2\right) \le V(y_i)\left(1-\rho^2\right)$. However, the difference $E\left[\underline{x}_i\underline{x}_i^T\right]^{-1} - E\left[\underline{x}_i\underline{x}_i^T | c \le y_i \le C\right]^{-1}$ is necessarily positive semi-definite since the variance of x_i in the truncated sample cannot be larger than the variance of the full sample (a sufficient condition here is joint normality (see Goldberger(1981)). Hence, a comparison of the leading terms is indeterminate.

In the appendix we show that the analytical results from winsorizing are similar to those obtained for trimming. The empirical results suggest winsorizing has less of an impact on the slope coefficients (relative to OLS) than truncation. Again, if $\gamma > 1$,an optimal choice of winsorizing points is available, but of course γ is unknown. Ascertaining the effect of winsorizing on the size of the standard errors is conceptually similar to the effect of trimming on standard-errors, but algebraically more difficult. We provide a discussion of this point in the appendix. One advantage winsorizing has over trimming is that the penalty of lost data does not effect the expression for the finite sample variance. But overall, winsorizing and trimming have similar effects on both slope estimates and asymptotic variance.

4.3 Cleaning Data in the General Case

The results derived in the previous section rely heavily upon normality. However, as Goldberger (1981) demonstrates, without normality some coefficients may be attenuated by truncation, while others may be inflated. Clearly, as a theoretical matter, truncation or winsorizing cannot be relied upon to adjust slope coefficients for the bias in general. In contrast, the results from Section 2 were derived under much weaker conditions.

5 Results

To suggest more general results, we present a set of Monte Carlo simulations. We draw data from the 1990 PUMS and estimate the returns to schooling, treating estimates from the full PUMS file as population parameters. The results of different cleaning procedures are gauged against these parameters. This allows for a complicated measurement model, with relationships similar to those found in typical economic data.[6]

5.1 Evidence from U.S. Decennial Census Data

We begin with evidence from quasi-simulated data drawn from the PUMS samples of the 1990 US Decennial Census. The advantage of these data is that they provide a complex multivariate distribution for analysis. We study the problem of estimating the returns to schooling, using a standard "Mincerian specification" (that is, $lnwage = \beta_0 + \beta_1 Schooling + \beta_2 Experience + \beta_3 Experience^2 + \beta_4 Black + u$) to describe the relationship between hourly wages, years of schooling, race and potential experience. We first select prime-aged men who

[6] Our working paper [Bollinger and Chandra (2003)] provides more detailed Monte Carlo simulations for the univariate and multivariate case, and data from the normal, uniform, and log-normal distributions. Of note are the results for the realistic case of $\gamma = 0.9$. We note that cleaning procedures are always dominated by not cleaning the data. Even though the bias from not cleaning the data is a little over 10 percent, the bias from trimming is uniformly greater. 1% winsorizing or the use of median regression are neutral rules with respect to point-estimates, but both procedures are dominated by not cleaning the data on the basis of a RMSE criteria. When $\gamma > 1$, a 1 percent trimming rule clearly dominates not cleaning the data. Before this conclusion is embraced too quickly by practitioners, we raise two important caveats: First, even though 1% trimming works, 5% trimming is much worse that not cleaning the data; the optimal trimming rule is therefore not a known constant and small perturbations from the optimal truncation will generate large biases relative to not cleaning the data. In fact, the "best rule" for the case of $\gamma > 1$ would be to use 5% winsorizing. Second, we reiterate the difficulty in being able to justify a behavioral model for why γ would exceed one.

are working full time in a non-agricultural industry. We remove individuals who earn less than the minimum wage. The resulting 346,900 observations consitute a sample that closely simulates the ideal population distribution assumed by many researchers. Column 1 of Table 1 presents mean and standard deviation parameters for this pseudo-population. Black men comprise 8.3 percent of the population, the mean years of potential experience is 17.67, and the mean years of schooling is 13.37. The regression parameters generated by an OLS regression on all 346,900 observations are reported in Column 2 of Table 1. For our simulations, we randomly draw samples of size 1000 from the psuedo-population of 346,900 observations. Ordinary least squares performed on these data (without any cleaning) are reported in the third column of the table. In Columns 4-8, we report the effect of alternative cleaning procedures when no measurement error has been added to the PUMS data.[7] The idea of cleaning data with no error might strike the reader as a peculiar exercise. Our motivation for doing so is to demonstrate that cleaning procedures are not benign and can introduce significant bias when they are not required; alternatively, if the degree of contamination is low, the iatrogenic error from cleaning data may be substantial. In this situation, the cleaning procedures do not generally perform better than not cleaning the data. In general, the RMSE from the cleaning procedures (including median regression) is greater than that from doing nothing. Whereas a 1% trimming rule improves the estimation of the coefficients on experience and experience squared, it is inferior to not cleaning the data as regards the estimation of the coefficients on schooling and race. Together these results confirm those from the univariate case: No cleaning procedure is neutral when applied to already clean data.

Measurement error is added to the data in Table 2. We select two values for the variance of this error using the results of Bound and Krueger (1991), who note that $Var(lnY) = 0.458$ and 0.529 with corresponding error variances are 0.083 and 0.116. This implies that the variance of the error is 18% and 22% of the total variation in lnY. Rogers, Brown and Duncan (1993) find even higher implied estimates of the variance of the measurement error. Therefore, to study empirically relevant cases we simulate measurement error whose variance is $0.1Var(wage)$ and $0.3Var(wage)$. In order to keep the reported number of results manageable we only report results from the latter simulation, but note that the results from the former are quantitatively similar [see Bollinger and Chandra (2003) for details]. In the case of additive white noise we find that trimming is once again dominated by not cleaning the data. A case can be made for a 1% winsorizing rule over not cleaning the data, but it is important to note that significant bias is introduced when the censoring rate is increased to 5%. For this case, least-squares is found to be superior to median regression. When $\gamma = 0.9$ there is no cleaning procedure that strictly dominates OLS. A 1% winsorizing rule provides superior estimates on a RMSE criteria for many coefficients but simultaneously raises the bias on others. For example, the coefficients on Exp, Exp-Sq and Black all have lower RMSE when a 1% winsorizing rule is applied, but the coefficient on schooling has a larger RMSE at the same time. When $\gamma = 1.1$ winsorizing at 1% and 5% are preferred to doing nothing. In this case, trimming procedures dominate not cleaning the data on a

[7] In the last column of the table, we report results from performing Median Regression. While not explicitly studied in our paper, we include these estimates because several readers of our paper argued that it may be viewed as an alternative to trimming and winsorizing.

RMSE criteria, but can be worse in terms of the bias component.

5.1.1 Comparing rescaling approaches to trimming approaches.

As noted in previous sections, another identification approach is to rescale the estimates. The optimal trimming rule derived in Section 4 requires both information about γ and normality. Clearly, if $\gamma < 1$, trimming or winsorizing will be dominated by the rescaling approach. Even when $\gamma > 1$, the amount of trimming or winsorizing necessary depends upon the variance of the errors (see Bollinger and Chandra (2003)). Examining the first Column of Table 2, shows that knowledge of γ alone will be sufficient to arrive at consistent estimates.

Even if γ is not known, the rescaling results can be used to perform sensitivity analysis. For example, researchers might ask: how sensitive to different values of γ are the conclusions we draw from our OLS estimates? The robustness of these conclusions may be examined either by placing bounds on γ, as suggested in Manski (1995), or alternatively by asking what values of γ support the conclusions typically drawn (an approach suggested in a similar context by Bollinger (2003), and Bollinger (2001)). Further, researchers may not have detailed information about γ but may have information about the likely range of γ. It is difficult to use that information for trimming and winsorizing, but it can be trivially used in a rescaling approach.

5.2 Empirical example from the CPS

We also examine data cleaning approaches using the March 2001 Current Population Survey. There are two measures of hourly wage that a researcher could exploit in these data. One is the hourly wage constructed from the reported annual earnings, weeks worked and usual hours worked. The CPS also asks the actual hourly wage for workers who are paid hourly. Most researchers do not use this variable, as the resulting sample is smaller and only represents hourly wage workers. In this context, the two measures provide an interesting comparison to examine the implications of trimming as is typically practiced. Our sample consists of males, working full time, year round in non-agricultural positions who are not self-employed. In March 2001, we find 2,626 men who are full time, year round non-agricultural hourly wage workers.

The first column of Table 3 presents the log wage regression on the reported hourly wage, while the second column uses the constructed hourly wage. We restricted the constructed hourly wage sample to contain only those workers who also reported an hourly wage, hence any difference between the two columns reflects only differences in the measurement of the hourly wage, rather than sample differences. One perspective with these results is that the first column represents the "true coefficients," while the results in column two are biased due to response error. Interestingly, most of the coefficients in column 2 are larger in magnitude than their corresponding coefficients in column 1. The exception to this is the coefficient on Bachelors degree. This is a case where trimming might be useful. However, the fact that the coefficient on Bachelors is the exception demonstrates that it is difficult to find perfect cases.

If column 1 represents the "true coefficients," then columns 3 and 4 represent attempts to correct column

2. Trimming at about 1/2 of the minimum wage is a common practice (see Angrist and Krueger, 2000). Column 3 represents this approach. Another logical correction is to trim at the minimum wage; column 4 represents this approach. Comparing column 3 with columns 1 and 2, we find that the coefficients on experience and experience squared and black are largely unaffected by the trimming, and are still "too large." The coefficient on less than high school has actually increased in magnitude, and made the bias worse. The coefficients on associates degree and graduate degree have declined in magnitude, reducing the bias, but not eliminating it. The coefficient on bachelors degree has decreased in magnitude increasing the bias in this coefficient.

Trimming at the minimum wage, represented by column 4, improves some coefficients but not others. The coefficients on experience and experience squared are now both smaller in magnitude and closer to the "ideal" column 1. The coefficient on less than high school has now declined in magnitude, but it still somewhat larger than the coefficient in column 1. The coefficient on associates degree has declined and is biased relative to the target in column 1; it now underestimates the magnitude. The coefficient on graduate degree has not changed any further and still overstates the target in column 1. The coefficient on black has declined in magnitude and is now closer to the coefficient in column 1, but is still larger in magnitude. The coefficient on bachelors degree has declined further in magnitude increasing the bias still further. The conclusion we take from this is that there is no clear advantage to trimming. While it certainly may reduce the bias for some estimates, it is simultaneously making other estimates worse. Since it is rare to have a target (as we do in this case), it is only through serendipity that one will pick the right trimming rule even if the researcher is only interested in one specific coefficient.

A second perspective on the estimates in columns 1 and 2 is that they both contain measurement error in the dependent variable. One would expect that trimmed versions of the two regressions would converge to some set of correct estimates. Columns 5 and 6 are trimmed versions of column 1. As one might expect, the reported wage has fewer observations below the trimming points than the constructed wage. In column 5 there is very little change in the coefficients. In column 6 there is little change the coefficients on experience, experience squared, or less than High School. The coefficient on Bachelors degree increases in magnitude. This is in sharp contrast to trimming the constructed wage, where the coefficient decreased in magnitude. The coefficients on graduate degree and Black both increase in magnitude slightly. We conclude that it appears only serendipitous if any coefficients converge with trimming. If one is interested in the return to a college degree, trimming is likely to be undesirable. While if one is interested in the Black-white wage gap, trimming at an even higher threshold may be desirable. The effects of trimming are unclear since we cannot even predict which direction the slope coefficients will change when we trim. Without apriori information, it is difficult or impossible to know if trimming has reduced bias, increased bias, or some of both.

6 Conclusions

The common practice of cleaning data by removing observations where the dependent variable is larger or smaller than some threshold is used in the hope of reducing the impact of measurement error. While this sounds sensible, it may make matters worse. Analytical results using normality demonstrate that cleaning strategies using truncation or winsorizing work only the case where measurement error in the dependent variable results in an upward bias on the magnitude of coefficients. This case is not empirically supported by investigations of response error in earnings data. Under certain circumstances it may be possible to achieve an optimal cleaning strategy, but if the information necessary for that result were available, a simpler approach based on rescaling the estimated coefficients would work too.

Our empirical results demonstrate that a 1 percent winsorizing rule does not alter the results in a meaningful manner, but we note that this policy is generally dominated by doing nothing to the data. Still, winsorizing is clearly better than truncation. We caution, however, that small increments to this rule (for example to 5 percent) can dramatically increase bias and render the cleaning undesirable on MSE grounds. Two important extentions we hope to address in future work is cleaning based on covariates, and cleaning based up panel data where the researcher has multiple observations on the dependent variable. In both cases, it may be possible to develop cleaning proceedures that exploit other information.

7 Appendix

Derivation of equation10

To solve for the expression in 10, begin by noting that $E[y] = \delta + +\gamma\alpha + \gamma\beta\mu_x$ and $V(y) = \gamma^2\beta^2\sigma_x^2 + \gamma^2\sigma_u^2 + \sigma_\varepsilon^2$. Additionally we simplify the analysis by only considering a symmetric truncation scheme where $c = E[y] - c^*$ and $C = E[y] + c^*$. so that only c^* need be found. Consider first the expression for θ in this case:

$$\begin{aligned}\theta &= 1 + \left[\frac{\left(\frac{c-E[y]}{V(y)}\right)\phi\left(\frac{c-E[y]}{V(y)}\right) - \left(\frac{C-E[y]}{V(y)}\right)\phi\left(\frac{C-E[y]}{V(y)}\right)}{\Phi\left(\frac{C-\delta-\gamma\alpha-\gamma\beta\mu_x}{\gamma^2\beta^2\sigma_x^2+\gamma^2\sigma_u^2+\sigma_\varepsilon^2}\right) - \Phi\left(\frac{c-E[y]}{V(y)}\right)}\right] \\ &\quad - \left[\frac{\phi\left(\frac{c-E[y]}{V(y)}\right) - \phi\left(\frac{C-E[y]}{V(y)}\right)}{\Phi\left(\frac{C-E[y]}{V(y)}\right) - \Phi\left(\frac{c-E[y]}{V(y)}\right)}\right]^2\end{aligned}$$

substituting the symmetric expressions for c, C yields

$$\begin{aligned} &= 1 + \frac{\left(\frac{-c^*}{V(y)}\right)\phi\left(\frac{-c^*}{v(y)}\right) - \left(\frac{c^*}{V(y)}\right)\phi\left(\frac{c^*}{V(y)}\right)}{\Phi\left(\frac{c^*}{V(y)}\right) - \Phi\left(\frac{-c^*}{V(y)}\right)} \\ &\quad - \left[\frac{\phi\left(\frac{-c^*}{V(y)}\right) - \phi\left(\frac{-c^*}{V(y)}\right)}{\Phi\left(\frac{c^*}{V(y)}\right) - \Phi\left(\frac{-c^*}{V(y)}\right)}\right]^2 \\ &= 1 - 2\left(\frac{c^*}{V(y)}\right)\left(\frac{\phi\left(\frac{c^*}{V(y)}\right)}{\Phi\left(\frac{c^*}{V(y)}\right) - \Phi\left(\frac{-c^*}{V(y)}\right)}\right).\end{aligned}$$

Next, noting that λ is observable and γ is assumed to be known solve $\left(\frac{\theta}{1-(1-\theta)\lambda}\right)\gamma = 1$ for θ in terms of γ and λ:

$$\theta = \frac{1-\lambda}{1-\gamma}.$$

Substituting for θ and solving, yields the implicit relationship expressed in equation (10):

$$2\left(\frac{c^*}{V(y)}\right)\left(\frac{\phi\left(\frac{c^*}{V(y)}\right)}{\Phi\left(\frac{c^*}{V(y)}\right) - \Phi\left(\frac{-c^*}{V(y)}\right)}\right) = \frac{\gamma - 1}{\gamma - \lambda.}$$

Proof of Proposition 3

To show that the bias gets worse with more variance in ε, differentiate the bias term with respect to the variance of the measurement error:

$$\begin{aligned}&\frac{\partial}{\partial\sigma_\varepsilon^2}\left(\frac{\theta}{1-(1-\theta)\rho^2}\right) \\ &= \left[\frac{\partial\theta}{\partial\sigma_\varepsilon^2}\left(1-(1-\theta)\rho^2\right) - \theta\left(\rho^2\frac{\partial\theta}{\partial\sigma_\varepsilon^2} - (1-\theta)\frac{\partial\rho^2}{\partial\sigma_\varepsilon^2}\right)\right] / \left(1-(1-\theta)\rho^2\right)^2\end{aligned}$$

This term is negative if and only if the numerator is negative. Considering only the numerator and grouping similar derivatives yields

$$\frac{\partial\theta}{\partial\sigma_\varepsilon^2}\left(1-\rho^2\right) + (1-\theta)\,\theta\frac{\partial\rho^2}{\partial\sigma_\varepsilon^2}.$$

As noted in Goldberger both ρ^2 and θ are bounded in the unit interval. Inspection of the definition of ρ^2 clearly demonstrates that $\frac{\partial \rho^2}{\partial \sigma_\varepsilon^2} < 0$. Now, consider the definition of θ: inspection reveals that this has the truncation points standardized by the mean and variance of y. Hence increasing σ_ε^2 is equivalent to increasing c and decreasing C for a truncated standard normal random variable. Since θ is also the ratio of the variance of the truncated standard normal to the variance of the untruncated standard normal (see Goldberger), increasing c and decreasing C will result in a lower variance for the truncated distribution and thus a lower θ. Hence, by inspection, $\frac{\partial \theta}{\partial \sigma_\varepsilon^2} < 0$. Combined with the bounds on ρ^2 and θ, the result is established.

Does Trimming Reduce Standard Errors

Under assumptions A1-A7, the conditional distribution of $y_i|x_i$ is $N\left(x_i'b, V\left(y_i\right)\left(1-\rho^2\right)\right)$. We can use the results in Goldberger (1981) to obtain an expression for the truncated second moment matrix as:

$$Q = E\left[x_i x_i' | c \leq y_i \leq C\right] = E\left[x_i x_i'\right] - (1-\theta) E\left[x_i x_i'\right] bb' E\left[x_i x_i'\right].$$

The term $E\left[\left(y_i - x_i'b^*\right)^2 x_i x_i' | c \leq y_i \leq C\right]$ is considered by using the law of iterated expectations. The $E\left[\left(y_i - x_i'b^*\right)^2 | x_i, c \leq y_i \leq C\right]$ can be decomposed into the variation of y_i around the conditional mean in the truncated distribution, and the squared difference between the conditional mean and the linear projection term $x_i^T b^*$. Doing so yields:

$$\begin{gathered} E\left[\left(y_i - x_i'b^*\right)^2 | x_i, c \leq y_i \leq C\right] = \\ \theta V\left(y_i\right)\left(1 - \left(\frac{\theta}{1-(1-\theta)\rho^2}\right)\rho^2\right) + \left(x_i'b^* - m\left(x_i\right)\right)^2, \end{gathered}$$

where $m\left(x_i\right)$ is the conditional mean of y_i given $\underline{x}_i$ in the truncated sample. Combining these terms produces the equation in the paper:

$$\begin{aligned} AV\left(\widehat{b^*}\right) &= \theta V\left(y_i\right)\left(1 - \left(\frac{\theta}{1-(1-\theta)\rho^2}\right)\rho^2\right) Q^{-1} \\ &\quad + Q^{-1} E\left[\left(x_i'b^* - m\left(x_i\right)\right)^2 x_i x_i' | c \leq y_i \leq C\right] Q^{-1}. \end{aligned}$$

Analytic Results: Winsorized Data

When data are winsorized, no observations are removed, but values of y_i outside of the region (c, C) are transformed as follows:

$$y_i^w = \begin{array}{cc} C & \text{if } y_i \geq C \\ y_i & \text{if } c < y_i < C \\ c & y_i \leq c. \end{array}$$

As in the section on trimming, let b represent the vector of full-sample (uncleaned) coefficients. Under bivariate normality, the winsorized coefficient vector, b^{**} is given by :

$$b^{**} = \left[\Phi\left(\frac{C - E\left[y_i\right]}{V\left(y_i\right)}\right) - \Phi\left(\frac{c - E\left[y_i\right]}{V\left(y_i\right)}\right)\right] \cdot \left[1 - \left(\frac{\phi\left(\frac{c - E[y_i]}{V(y_i)}\right) - \phi\left(\frac{C - E[y_i]}{V(y_i)}\right)}{\Phi\left(\frac{C - E[y_i]}{V(y_i)}\right) - \Phi\left(\frac{c - E[y_i]}{V(y_i)}\right)}\right)^2\right] b.$$

Does Winsorizing Reduce Standard Errors?

The effects of Winsorizing on the variance are derived similarly to the results for trimming. We start by noting that the $AV\left(\widehat{b^{**}}\right) = E\left[x_i x_i'\right]^{-1} E\left[\left(y_i^W - x_i^{'} b^{**}\right)^2 x_i x_i'\right] E\left[x_i x_i'\right]^{-1}$. Here, the term $E\left[\left(y_i^W - x_i^{'} b^{**}\right)^2 |x_i\right]$ can be broken into three terms:

$$\begin{aligned} E\left[\left(y_i^W - x_i b^{**}\right)^2 |x_i\right] = &\Phi\left(\frac{c - E\left[y_i\right]}{V\left(y_i\right)}\right)\left(c - x_i^{'} b^{**}\right)^2 \\ &+ \left(1 - \Phi\left(\frac{C - E\left[y_i\right]}{V\left(y_i\right)}\right)\right)\left(C - x_i^{'} b^{**}\right)^2 \\ &+ \left(\Phi\left(\frac{C - E\left[y_i\right]}{V\left(y_i\right)}\right) - \Phi\left(\frac{c - E\left[y_i\right]}{V\left(y_i\right)}\right)\right) E\left[\left(y_i - x_i' b^*\right)^2 |x_i, c \leq y_i \leq C\right] \end{aligned}$$

Combined with results from the previous section, we obtain

$$\begin{aligned} &AV\left(\widehat{b^{**}}\right) = \\ &\left(\Phi\left(\frac{C - E\left[y_i\right]}{V\left(y_i\right)}\right) - \Phi\left(\frac{c - E\left[y_i\right]}{V\left(y_i\right)}\right)\right)\left(\theta V\left(y_i\right)\left(1 - \left(\frac{\theta}{1 - \left(1 - \theta\right)\rho^2}\right)\right) E\left[x_i x_i'\right]^{-1} Q E\left[x_i x_i'\right]^{-1}\right. \\ &\qquad\left. + E\left[x_i x_i'\right]^{-1} E\left[\left(x_i' b^{**} - m^W\left(x_i\right)\right)^2 x_i x_i' | c < y_i < C\right] E\left[x_i x_i'\right]^{-1}\right) \\ &\quad + \Phi\left(\frac{c - E\left[y_i\right]}{V\left(y_i\right)}\right)\left(E\left[x_i x_i'\right]^{-1} E\left[\left(\left(c - x_i b^{**}\right)^2 x_i x_i'\right) |y_i \leq c\right] E\left[x_i x_i'\right]^{-1}\right) \\ &\quad + \left(1 - \Phi\left(\frac{C - E\left[y_i\right]}{V\left(y_i\right)}\right)\right)\left(E\left[x_i x_i'\right]^{-1} E\left[\left(\left(C - x_i b^{**}\right)^2 x_i x_i'\right) |y_i \geq C\right] E\left[x_i x_i'\right]^{-1}\right). \end{aligned}$$

Again, the comparison is difficult. Here, the first term will necessarily be smaller than the OLS expression. However, the second, third and fourth terms are all positive definite. As in the trimming case, the impact on standard errors depends upon the parameters of the model.

References

[1] Angrist, Joshua D. and Alan B. Krueger, 2000. "Empirical Strategies in Labor Economics," in Orley Ashenfelter and David Card (Eds.) *Handbook of Labor Economics*, Vol 3A (Elsevier Science).

[2] Black, Dan A., Mark C. Berger and Frank A. Scott, 2000. "Bounding Parameter Estimates with Non-classical Measurement Error," *Journal of the American Statistical Association* 95: 739-48.

[3] Bollinger, Christopher R., 1996. "Bounding Mean Regressions When A Binary Regressor is Mismeasured," *Journal of Econometrics 73: 387-399.*

[4] Bollinger, Christopher R., 2003 "Measurement Error in Human Capital and the Black -White Wage Differential," *Review of Economics and Statistics85: 578-587.*

[5] Bollinger, Christopher and Martin H. David. 1997. "Modeling Food Stamp Participation in the Presence of Reporting Errors," *Journal of the American Statistical Association 92: 827-35.*

[6] Bollinger, Christopher, 1998. "Measurement Error in the Current Population Survey: A Nonparametric Look," *Journal of Labor Economics* 16(3): *57-71.*

[7] Bollinger, Christopher and Amitabh Chandra. 2003. "Iatrogenic Specification Error" NBER Technical Working Paper 289, Cambridge, MA.

[8] Bound, John and Alan B. Krueger, 1991. "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics* 9: 1-24.

[9] Bound, John and Richard Freeman, 1992. "What Went Wrong? The Erosion of Relative Earnings and Employment Among Black Men in the 1980s," *Quarterly Journal of Economics* 107(1), February: 201-32.

[10] Bound, John, Charles Brown, Greg J. Duncan and Willard L. Rodgers, 1994. "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data," *Journal of Labor Economics* 12: 345-68.

[11] Buchinsky, Moche, 1994. "Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression," *Econometrica* 62: 405-58.

[12] Card, David and Alan B. Krueger, 1992a. "School Quality and Black-White Relative Earnings: A Direct Assessment," *Quarterly Journal of Economics* 107, February: 151-200.

[13] Fuller, Wayne A. 1987. *Measurement Error Models*. John Wiley and Sons. (New York, NY).

[14] Goldberger, Arthur S., 1981, "Linear Regression after Selection," *Journal of Econometrics* 15(3): 357-66.

[15] Hirsch, Barry T. and Edward J. Schumacher, 2001. “Match Bias in Wage Gap Estimates Due to Earnings Imputations,” unpublished manuscript.

[16] Horowitz, Joel L. and Charles F. Manski, 1995. “Identification and Robustness with Contaminated and Corrupted Data,” *Econometrica* 63(2): 281-302.

[17] Hyslop, Dean R. and Guido W. Imbens, 2001. “Bias From Classical and Other Forms of Measurement Error,” Journal of Business and Economic Statistics 19(4): 475-481.

[18] Juhn, Chinhui, Kevin M. Murphy and Brooks Pierce, 1993. “Wage Inequality and the Rise in the Returns to Skill,” *Journal of Political Economy* 101: 410-42.

[19] Katz, Lawrence and Kevin M. Murphy, 1992. “Changes in Relative Wages 1963-1987,” *Quarterly Journal of Economics* 107(1): 35-78.

[20] MacDonald, Glenn M. and Robinson Chris, 1985. “Cautionary Tales About Arbitrary Deletion of Observations; or, Throwing the Variance out with the Bathwater,” *Journal of Labor Economics* 3(2): 124-52.

[21] Maddala, G. S., 1983. *Limited Dependent and Qualitative Variables in Econometrics* (Cambridge University Press).

[22] Manski, Charles F., 1995. *Identification Problems in the Social Sciences* (Harvard University Press).

[23] Mellow, Wesley and Hal Sider, 1983. “Accuracy of Response in Labor Market Surveys: Evidence and Implications,” *Journal of Labor Economics* 1: 331-44.

[24] Rodgers, Willard L., Charles C. Brown and Greg J. Duncan, 1993. “Errors in Survey Reports of Earnings, Hours Worked and Hourly Wages,” *Journal of the American Statistical Association* 88: 1208-18.

[25] Stigler, Stephen M., 1977. “Do Robust Estimators work with Real Data?” *Annals of Statistics* 5(6): 1055-98.

Table 1: Effect of Cleaning Procedures on Uncorrupted Data, Evidence from the Returns to Schooling in 1990 PUMS Data

	Mean	Population β	No Cleaning	Trim 1%	Trim 5%	Wins 1%	Wins 5%	Median
Schooling	13.3741	**0.092**	0.0918	0.0878	0.0704	0.0910	0.0856	0.1001
SE (Yrs of Schooling)	2.2189	-	0.0078	0.0072	0.0068	0.0075	0.0070	0.0086
RMSE	-	-	0.0078	0.0083	0.0227	0.0076	0.0095	0.0118
Potential Experience	17.6700	**0.0374**	0.0375	0.0361	0.0294	0.0372	0.0353	0.0402
SE (Schooling)	8.5976	-	0.0084	0.0079	0.0069	0.0082	0.0076	0.0095
RMSE	-	-	0.0084	0.0080	0.0106	0.0082	0.0079	0.0099
Pot. Exp. Sq /100	3.8639	**-0.0535**	-0.0538	-0.0522	-0.0419	-0.0535	-0.0508	-0.0565
SE (Pot. Exp)	3.3643	-	0.0218	0.0202	0.0175	0.0211	0.0196	0.0246
RMSE	-	-	0.0218	0.0203	0.0210	0.0211	0.0198	0.0248
Black (1= yes)	0.0831	**-0.1419**	-0.1416	-0.1374	-0.1072	-0.1417	-0.1339	-0.1640
SE (Black)	0.2762	-	0.0597	0.0544	0.0492	0.0579	0.0531	0.0697
RMSE	-	-	0.0597	0.0545	0.0602	0.0579	0.0538	0.0731
Constant	-	**0.8608**	0.8639	0.9293	1.2376	0.8746	0.9649	0.7280
SE	-	-	0.1275	0.1195	0.1112	0.1229	0.1150	0.1430

Dependent variable is ln hourly wage. PUMS data are restricted to white (non-hispanic) and black men in the 1990 PUMS files of the Decennial Census who are aged 25-55 during the census reference week. Nonworkers and repondents with hourly wages less than $3.35 in 1989 (the nominal value of the minimum wage) are deleted from the analysis. Column (1) reports means and standard deviations for this sample of 346,900 individuals, and column 2 reports the parameters from estimating the model: ln wage=β_0 +β_1 Schooling +β_2 Exp+β_3 Exp^2 +β_4 Black + u on this sample. Reported estimates in other columns are empirical sample moments from 1,000 replications each with a sample size of 1,000.

Table 2: Effect of Cleaning Procedures on Corrupted Data, Evidence from the 1990 PUMS

	No Cleaning	Trim 1%	Trim 5%	Wins 1%	Wins 5%	Median
Error Model: ln wage = ln wage* + e; var (e) = 0.3 x var (wage)						
Schooling	0.0925	0.0883	0.0704	0.0916	0.0859	0.1001
SE (Schooling)	0.0081	0.0074	0.0068	0.0078	0.0072	0.0091
RMSE	0.0082	0.0084	0.0227	0.0078	0.0095	0.0121
Pot. Exp	0.0377	0.0363	0.0294	0.0374	0.0354	0.0402
SE (Pot Exp)	0.0084	0.0078	0.0070	0.0081	0.0075	0.0096
RMSE	0.0084	0.0079	0.0107	0.0081	0.0078	0.0100
Pot. Exp. Sq /100	-0.0539	-0.0524	-0.0419	-0.0537	-0.0510	-0.0566
SE (Pot. Exp Sq)	0.0215	0.0201	0.0178	0.0209	0.0193	0.0248
RMSE	0.0215	0.0202	0.0213	0.0209	0.0195	0.0250
Black (1= yes)	-0.1426	-0.1394	-0.1097	-0.1426	-0.1346	-0.1616
SE (Black)	0.0639	0.0578	0.0513	0.0616	0.0564	0.0730
RMSE	0.0639	0.0578	0.0606	0.0616	0.0569	0.0756
Constant	0.8535	0.9222	1.2396	0.8667	0.9621	0.7281
SE	0.1345	0.1237	0.1147	0.1289	0.1203	0.1502
Error Model: ln wage = 0.9 ln wage* + e; var (e) = 0.3 x var (wage)						
Schooling	0.0833	0.0795	0.0634	0.0825	0.0773	0.0901
SE (Schooling)	0.0074	0.0067	0.0062	0.0071	0.0066	0.0082
RMSE	0.0115	0.0142	0.0294	0.0119	0.0161	0.0084
Pot. Exp	0.0339	0.0327	0.0264	0.0337	0.0318	0.0360
SE (Pot Exp)	0.0075	0.0071	0.0061	0.0073	0.0068	0.0087
RMSE	0.0083	0.0086	0.0127	0.0083	0.0088	0.0088
Pot. Exp. Sq /100	-0.0485	-0.0471	-0.0375	-0.0483	-0.0459	-0.0502
SE (Pot. Exp Sq)	0.0195	0.0182	0.0158	0.0190	0.0175	0.0225
RMSE	0.0201	0.0193	0.0225	0.0197	0.0191	0.0227
Black (1= yes)	-0.1276	-0.1250	-0.0976	-0.1276	-0.1205	-0.1439
SE (Black)	0.0576	0.0518	0.0458	0.0554	0.0507	0.0641
RMSE	0.0594	0.0545	0.0637	0.0572	0.0551	0.0641
Constant	0.7672	0.8295	1.1165	0.7789	0.8651	0.6558
SE	0.1208	0.1108	0.1020	0.1158	0.1075	0.1351
Error Model: ln wage = 1.1 ln wage* + e; var (e) = 0.3 x var (wage)						
Schooling	0.1018	0.0972	0.0775	0.1008	0.0945	0.1104
SE (Schooling)	0.0090	0.0082	0.0075	0.0087	0.0080	0.0100
RMSE	0.0132	0.0097	0.0164	0.0123	0.0084	0.0208
Pot. Exp	0.0415	0.0400	0.0325	0.0412	0.0390	0.0441
SE (Pot Exp)	0.0092	0.0087	0.0075	0.0090	0.0083	0.0106
RMSE	0.0101	0.0090	0.0089	0.0097	0.0085	0.0125
Pot. Exp. Sq /100	-0.0594	-0.0577	-0.0466	-0.0591	-0.0562	-0.0616
SE (Pot. Exp Sq)	0.0239	0.0224	0.0192	0.0232	0.0214	0.0275
RMSE	0.0246	0.0227	0.0205	0.0238	0.0216	0.0286
Black (1= yes)	-0.1560	-0.1527	-0.1204	-0.1562	-0.1478	-0.1789
SE (Black)	0.0702	0.0635	0.0554	0.0675	0.0618	0.0803
RMSE	0.0716	0.0644	0.0594	0.0690	0.0621	0.0884
Constant	0.9380	1.0126	1.3612	0.9522	1.0567	0.7982
SE	0.1480	0.1352	0.1238	0.1418	0.1321	0.1656

Dependent variable is ln hourly wage. PUMS data are restricted to white (non-hispanic) and black men in the 1990 PUMS files of the Decennial Census who are aged 25-55 during the census reference week (n=346,900). Nonworkers and repondents with hourly wages less than $3.35 in 1989 (the nominal value of the minimum wage) are deleted from the analysis, and measurement error is added to observed lnwage using the specified error models. Reported estimates are empirical sample moments from 1,000 replications each with a sample size of 1,000. The variance of ln(wage)=0.3144.

Table 3: Effect of Cleaning Procedures: Evidence from the March 2001 CPS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Baseline Regressions		**Constructed Hourly Wage**		**Reported Hourly Wage**		**Constructed Hourly Wage**		**Reported Hourly Wage**	
	Reported Hourly Wage	Constructed Hourly Wage	Trim at 1/2 min wage	Trim at min wage	Trim at 1/2 min wage	Trim at min wage	Wins at 1/2 min wage	Wins at the min wage	Wins at 1/2 min wage	Wins at min wage
Potential Exp	0.028	0.033	0.033	0.030	0.028	0.028	0.033	0.032	0.028	0.028
	(0.002)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)
Potential Exp ^{2}	-0.045	-0.051	-0.051	-0.046	-0.046	-0.045	-0.052	-0.049	-0.045	-0.045
	(0.005)	(0.006)	(0.006)	(0.005)	(0.005)	(0.004)	(0.006)	(0.005)	(0.005)	(0.005)
Less than HS (1=Yes)	-0.261	-0.287	-0.292	-0.280	-0.252	-0.253	-0.290	-0.285	-0.258	-0.256
	(0.024)	(0.031)	(0.029)	(0.028)	(0.024)	(0.023)	(0.030)	(0.028)	(0.024)	(0.023)
Associates Deg (1=Yes)	0.170	0.191	0.187	0.166	0.168	0.168	0.191	0.184	0.170	0.169
	(0.028)	(0.037)	(0.034)	(0.032)	(0.028)	(0.027)	(0.035)	(0.034)	(0.028)	(0.028)
Bachelors Deg (1=Yes)	0.254	0.242	0.235	0.211	0.253	0.270	0.238	0.232	0.254	0.258
	(0.028)	(0.037)	(0.034)	(0.032)	(0.028)	(0.027)	(0.035)	(0.034)	(0.028)	(0.028)
Graduate Deg (1=Yes)	0.377	0.590	0.575	0.575	0.375	0.450	0.584	0.577	0.377	0.398
	(0.058)	(0.076)	(0.070)	(0.066)	(0.057)	(0.056)	(0.072)	(0.069)	(0.058)	(0.056)
Black (1=Yes)	-0.087	-0.135	-0.136	-0.113	-0.090	-0.093	-0.138	-0.130	-0.088	-0.090
	(0.025)	(0.033)	(0.031)	(0.030)	(0.025)	(0.024)	(0.032)	(0.030)	(0.025)	(0.025)
Constant	2.240	2.197	2.220	2.282	2.238	2.241	2.203	2.230	2.240	2.240
	(0.025)	(0.032)	(0.030)	(0.029)	(0.024)	(0.024)	(0.031)	(0.029)	(0.025)	(0.024)
Observations	2626	2626	2612	2534	2622	2605	2626	2626	2626	2626

Sample is drawn from the March 2001 CPS. The sample consists of males who are not self-employed, working full time, year round in non-agricultural positions, who were paid hourly. Reported hourly wage refers to respondent's report of hourly wage (true wage); constructed hourly wage is constructed using annual earnings, hours and weeks worked (and therefore constitutes a noisy measure of wage). The correlation between the two wage measures is 0.39 (and 0.54 in logs). The standard-deviations for actual and constructed hourly pay are 7.4 and 12.5 respectively (0.45 and 0.58 in logs). Reported hourly pay ranged from $1 to $99, whereas constructed hourly wages ranged from $0.02-$184.13.