TECHNICAL WORKING PAPER SERIES

IATROGENIC SPECIFICATION ERROR: A CAUTIONARY TALE OF CLEANING DATA

Christopher R. Bollinger
Amitabh Chandra

Technical Working Paper 289
http://www.nber.org/papers/T0289

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2003

We are grateful to Dan Black, Charlie Brown, Jon Gruber, Al Gustman, Joel Horowitz, Chuck Manski, Andrew Samwick, Douglas Staiger, Bo Honoré, participants at the Joint Statistical Meetings, seminar participants at the University of Kentucky, Princeton University, and Northwestern University for many useful comments. Chandra acknowledges generous financial support from the National Institutes of Aging, and the Rockefeller Center for Public Policy. The views expressed in this paper are those of the authors and not necessarily those of the National Bureau of Economic Research.

Iatrogenic Specification Error: A Cautionary Tale of Cleaning Data
Christopher R. Bollinger and Amitabh Chandra
NBER Technical Working Paper No. 289
February 2003
JEL No. C2, C8, J1

ABSTRACT

It is common in empirical research to use what appear to be 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. This paper considers identification in a linear model when the dependent variable is mismeasured. The results examine the common practice of trimming and winsorizing to address the identification failure. In contrast to the physical and laboratory sciences, measurement error in social science data is likely to be more complex than simply additive white noise. We consider a general measurement error process which nests many processes including the additive white noise process and a contaminated sampling process. Analytic results are only tractable under strong distributional assumptions, but demonstrate that winsorizing and trimming are only solutions for a particular class of measurement error processes. Indeed, trimming and winsorizing may induce or exacerbate bias. We term this source of bias "Iatrogenic" (or econometrician induced) error. The identification results for the general error process highlight other approaches which are more robust to distributional assumptions. Monte Carlo simulations demonstrate the fragility of trimming and winsorizing as solutions to measurement error in the dependent variable.

Christopher R. Bollinger
Department of Economics
University of Kentucky
Lexington, KY 40514
crboll@uky.edu

Amitabh Chandra
Department of Economics
Dartmouth College
Hanover, NH 03755
and NBER
amitabh.chandra@dartmouth.edu

1 Introduction

Empirical researchers frequently use apparently sensible rules of thumb to clean data. Observations where the value of the dependent variable is outside some range are typically removed. The justification for this practice is that these observations are subject to measurement error, and by removing them, the impact of measurement error is reduced. 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 only to illustrate the use, and therefore endorsement, of this practice by the leading scholars in the field. It appears to make *prima facie* sense to delete observations who values are outside of sensible bounds– for example, reported hours worked with negative values, or observations with more than 52 weeks worked in a year. 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. Moreover, rather than reduce the impact of measurement error, it can exacerbate bias caused by measurement error in the dependent variable. 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 the 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% trimmed mean, the smallest trimming amount considered in his study, is the most reliable estimator. In his conclusions he echoes a prescription made by the famous mathematician Legendre who had recommended deleting those observation whose errors were "too large to be admissible". Most recently in the econometrics literature, Angrist and Krueger (2000) apply trimming and winsorizing techniques to the matched employer-employee data that were studied by Mellow and Sider (1983). When they trim both the employer and employee wage data, they find that the correlation between the two measures improves (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, including ourselves, 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, focusing on cleaning based upon the dependent variable. We use wages and earnings as a motivating example, but clearly the results apply elsewhere to general errors of measurement in survey data (for example the type studied by Rodgers, Brown and Duncan (1993)). Similar to the work of Hyslop and Imbens (2001), we posit a general model of response error in the dependent variable of a simple linear regression model and characterize the properties of different cleaning techniques on measurement error processes. 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" 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. We demonstrate that the results in Stigler (1977) do not necessarily carry over in a regression framework. Indeed, trimming or windsorizing can bias coefficient estimates by as much as 10-30 percent. These are not second-order effects. Indeed, in many cases it either induces bias that did not previously exist, or exacerbates the bias due to measurement error. We further show that the cases where the cleaning actually results in reduced bias occur only by serendipity. If a researcher has the information to know that the data fit one of these special cases, there are other approaches which are far simpler given the availability of this information.

Our paper is organized as follows: Section 2 describes identification with (general) measurement error in the dependent variable. We establish that the linear projection of the mismeasured dependent variable on the covariate matrix constitutes the central equation of interest for the purpose of our analysis. In Section 3, we consider three specific models of measurement error (additive white noise, linear transformation and the contaminated data process) to understand the relationship between estimated coefficients in regression models that have these types of measurement error and the general model considered in Section 2. Section 4 examines the implications of the cleaning proposals analytically. The researcher is assumed to choose a set of values $c < E[y] < C$ and retain a data set through trimming or winsorizing such that $\{y_i, x_i | y_i \in [c, C]\}$. Under this framework we rigorously examine the bias from trimming and winsorizing, and also consider the implications of these cleaning procedures on the estimation of the asymptotic variance of the coefficients. We also discuss results for the multivariate case. We prove analytically that only in highly specialized cases does cleaning reduce bias, mostly through serendipity. In these cases, the information necessary to reduce bias leads to a simpler correction which actually requires fewer assumptions. Section 5 presents detailed Monte-Carlo simulation results for the cases considered analytically, as well as those for which no tractable

analytical results are possible. Finally, we generate quasi-simulated data from the 1990 US Decennial Census to study the properties of winsorizing and trimming on multivariate data whose covariance structure is not determined by ad hoc simulations. These simulations are found to support the results of the earlier two sections. Section 6 provides concluding comments.

2 Identification with 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. We demonstrate in the next section how this model nests common measurement error processes such as those with classical measurement error, linear transformation, and contaminated data processes of Horowitz and Manski (1995). In order to keep the analysis simple and intuitive, we focus on a linear regression model as the underlying structural model of interest to the researcher. That is, the researcher hypothesizes that the relationship between the "true" dependent variable and the covariate is described by:

$$y_i^* = \underline{x}_i^T \underline{\beta} + u_i, \tag{1}$$

combined with the usual assumptions sufficient for identification and estimation of the vector of interest $\underline{\beta}$ and its associated covariance matrix.

$$\begin{aligned} A1 &: E\left[u_i|\underline{x}_i\right] = 0 \\ A2 &: \underline{x}_i \text{ is a vector random variable with mean } \underline{0} \text{ and} \\ &\quad \text{full rank second moment matrix } V_x \\ A3 &: \text{Random Sampling} \end{aligned}$$

We note that the mean independence assumption is stronger than necessary for identification of the vector $\underline{\beta}$, but allows for a simpler analysis below. The zero mean for $\underline{x}_i$ is the usual normalization. The researcher is only able to obtain data on y_i, a mismeasured versions of the variable of interest. The focus in this paper, as motivated above, is on measurement error in the dependent variable y_i^*. At the outset, we assume a general process that relates the true value y_i^* to the observed value y_i :

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

where

$$\begin{aligned} A4 &: h(.,.) \text{ has finitely many discontinuities} \\ A5 &: \underline{\varepsilon}_i \text{ is independent of } (y_i^*, \underline{x}_i, u_i), \\ A6 &: Cov(y_i, y_i^*) > 0. \end{aligned}$$

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 $\underline{x}_i$ and u_i except through y_i^*. Formally, this insures that $f(y_i|y_i^*) = f(y_i|y_i^*, \underline{x}_i, u_i)$. 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^* :

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

The population projection defines (δ, γ) and the properties of e_i :

$$\delta = E[y_i], \tag{4}$$

$$\gamma = \frac{Cov(y_i, y_i^*)}{V(y_i^*)}, \tag{5}$$

and

$$E[e_i] = E[e_i y_i^*] = 0. \tag{6}$$

It is important to note that this is simply definitional: provided second moments exist, the linear projection of y_i on y_i^* exists as defined above. The linear projection is not a statement about the data generating process, but rather a summary measure of the joint distribution of (y_i, y_i^*). The actual measurement process, as defined by $h(y_i^*, \underline{\varepsilon})$ may be substantially more complicated. We will show that in the linear regression context the projection is the only relevant information contained in the joint distribution of (y_i, y_i^*). Assumption A6 insures that $\gamma > 0$.

The researcher is only able to observe $(y_i, \underline{x}_i)$. As is well known, OLS estimation is a consistent estimator for the population linear projection of y_i on $\underline{x}_i$. Substituting equation 1 into equation 3 yields:

$$y_i = \delta + \underline{x}_i^T \underline{\beta}\gamma + \gamma u_i + e_i. \tag{7}$$

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

$$y_i = a + \underline{x}_i^T \underline{b} + \eta_i \tag{8}$$

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

Hence, under the assumptions above, the OLS regression of y_i on $\underline{x}_i$ yields a consistent estimate of $\underline{b}$ which is proportional to $\underline{\beta}$. The parameters of interest are identified up to an unknown scaling constant. This would imply for example, that estimates of ratios of the parameters are consistent. In some situations, this may be sufficient for the conclusions of the researcher. 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. This may be sufficient to answer questions about schooling choices. Indeed in many settings, identification up to scale is considered sufficient. Estimation using prior information about the scale of parameters is largely underexplored. While it may be difficult to obtain point identification, bounding information on any one of the slope coefficients will result in bounds for all coefficients, including γ.

In general however, we assume the researcher is interested in recovering the parameters $\underline{\beta}$. This suggests two important identification approaches: obtain information about the scaling constant γ, or obtain information about one of the elements in $\underline{\beta}$. While it may be possible to obtain some consistent estimate of one element in $\underline{\beta}$ from auxiliary regressions or economic theory, an approach with a history in the literature would be to utilize validation data to estimate γ. Bound and Kreuger (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 of γ is 0.90. This estimate could be used in log wage models to rescale slope coefficients to account for measurement error. In support of the assumptions above, with the exception of a gender variable, Bollinger (1998) finds that beyond the information contained in y_i^* the $\underline{x}_i$ variables are independent of y_i. An important point here is that *only* a consistent estimate of γ is necessary to arrive at consistent estimates of $\underline{\beta}$. As will be seen below, an optimal trimming rule requires this information as well. It further requires a complete characterization of the joint distribution of (y_i, y_i^*, x_i).

3 Specific Measurement Error Models

The above analysis holds for rather general examples of measurement error. Here we present three special cases which are interesting for at least one of three reasons: they are cases commonly examined in the

literature, they are cases empirically supported in the literature, or they have specific results in the context of this paper which are of interest.

3.1 Additive White Noise

In the first case the measurement error is additive white noise:

$$y_i = y_i^* + \varepsilon_i. \tag{9}$$

This is the traditional measurement error process often assumed. Indeed, the error (or residual) in regression models has been motivated as measurement error. It is easy to confirm that the parameters of the LP of y on y^* are $\gamma = 1$, $\delta = a = 0$. As is well known, the least squares estimates are consistent for the parameters of interest $\underline{\beta}$. Of particular interest here, is the fact that this model would imply observations that appear with error. Indeed, 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. Indeed the presence of error here does not lead to any bias. 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 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, the traditional 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:

$$y_i = d + g y_i^* + \varepsilon_i. \tag{10}$$

It is important to note that the results in the previous section do not assume this process. Here, perhaps obviously, the parameters in the LP of y on y^* are $\gamma = g$ and $\delta = d$. This model can either have $\gamma > 1$ or $\gamma < 1$. Here again, the data generating process can lead to observations outside the "acceptable" range. It is important to note that even if $\gamma < 1$, because of the values of δ and the distribution of ε_i, it is quite possible to have both observations that are "too high" and observations that are "too low". Hence observations below some minimum or above some maximum do not distinguish γ or δ from this model or any other model. 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 a $\gamma > 1$.

It is difficult to think of examples where γ should exceed one. Behaviorally, for either economic or psychological reasons, this would happen if respondents at the lower end of the distribution have an incentive to understate the true value of a variable relative to respondents at the higher end of the distribution. One example where $\gamma > 1$ may be responses to an "hours worked" question. It is possible that workers at the bottom end of the distribution work some hours at "under the table" jobs that they fail to report both to surveys and to government agencies.

3.3 Contaminated Data

A third example is a simple contaminated sample:

$$y_i = (y_i^*) * 1\left[\varepsilon_{1i} > \kappa\right] + (d + \varepsilon_{2i}) * 1\left[\varepsilon_{1i} < \kappa\right]. \tag{11}$$

The term 1[.] is the indicator function and $(\varepsilon_{1i}, \varepsilon_{2i})$ are mutually independent and independent of $(\underline{x}_i, u_i)$. This model produces a mixture: with some probability $p = \Pr\left[\varepsilon_{1i} > \kappa\right]$, 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 $\underline{\beta}$ can be obtained if an estimate of p is available. Horowitz and Manski (1995) note that the expectation of y^* given $\underline{x}$ 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 $\underline{\beta}$, β_j, are bounded by $\left[\frac{b_j}{\underline{p}}, \frac{b_j}{\overline{p}}\right]$.

One important implication for each of these processes is that they all may result in observations that fall outside of some particular range (presumably the support of y_i^*). Moreover, depending on distributions

for $\underline{\varepsilon}_i$, there is no way to distinguish between these models without further information. Corrections based upon an assumption that one of these models pervades must be supported with evidence. In general, the error process is likely to be far more complicated than any of the above. It is likely to contain some kind of mixture of these processes: some individuals have minor errors which may be simply additive white noise (or rounding noise), others may have more complicated error structure such as a linear model, while still others may give "junk" as in the contaminated sampling model. In any case, the relevant measure is the linear projection of y on y^*.

4 Effect of Cleaning

As noted in the introduction, the cleaning approaches we consider are defined by

$$\{y_i, x_i | c \leq y_i \leq C\} \tag{12}$$

for known constants (c, C) such that $c < E[y] < C$: the researcher truncates above and below the mean, rather than truncating so severely that all observations below (or above) the mean are removed. We compare the slopes obtained from a least squares projection of y_i on x_i using this doubly truncated data. We use the least squares projection equation 8 as a departure point. We examine a set of conditions under which an analytic comparison demonstrates that the least squares projection using the censored/truncated data will result a slope coefficient that is attenuated relative to $\underline{b}$ (see, for example, Madalla (1983)). To be clear, the result can only tell us the bias relative to $\underline{b}$ the biased estimate of $\underline{\beta}$ 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 $\underline{b}$ may be larger or smaller (in magnitude) than the true slope $\underline{\beta}$. If $\underline{b}$ is inflated relative to $\underline{\beta}$, then the cleaning may perfectly correct for the measurement error bias, but this result would simply be due to serendipity. It may reduce the bias, but not completely. Indeed, it may even go so far as to overcorrect for the inflation bias, leaving the researcher with an attenuated estimate. If $\underline{b}$ is itself already attenuated relative to $\underline{\beta}$, then the censoring will result in a deeper attenuation (although it will not reverse the sign). Indeed, without specific knowledge of the structure of the measurement error *and* distributions, it is nearly impossible to determine the net result of the cleaning so popular in empirical work. This result alone should serve as a warning: *there is no reason, a priori, to believe that this cleaning approach will reduce the bias due to measurement error.*

4.1 Analytic Results: Trimmed Data

In order to arrive at a closed form analytic result for the attenuation bias due to cleaning, additional assumptions are necessary. In the empirical section, we simulate a number of different situations where analytical results are impossible to obtain. We first invoke:

$$A7 : (y_i, \underline{x}_i) \text{ jointly Normally Distributed.}$$

This is a strong assumption, and implies that η_i is also normally distributed and homoskedastic. This significantly limits the measurement error processes. This is not a limitation of our analysis, but rather will highlight the strong assumptions necessary for cleaning to alleviate bias due to measurement error. As Goldberger (1981) demonstrates, the slope vector from the least squares projection of y_i on $\underline{x}_i$ in the truncated sample (where observations above C and below c are discarded) is given by:

Proposition 1 *Under the assumptions of normality (A7), the truncated slope $\underline{b}^*$ is attenuated relative to the slope $\underline{b}$ from the least squares projection in the full sample.*

$$\underline{b}^* = \left(\frac{\theta}{1-(1-\theta)\rho^2} \right) \underline{b} \tag{13}$$

with

$$\theta = \frac{V\left(y_i | c \leq y_i \leq C\right)}{V\left(y_i\right)} \tag{14}$$

and

$$\rho^2 = \frac{b^2 \sigma_x^2}{V\left(y_i\right)}. \tag{15}$$

Since y_i is normally distributed, the variance of the doubly truncated distribution can be expressed (see Madalla, 1983) as

$$V\left(y_i | c \leq y_i \leq C\right) = \tag{16}$$
$$V(y)\left[1 + \left[\frac{\left(\frac{c-E[y_i]}{V(y_i)}\right)\phi\left(\frac{c-E[y_i]}{V(y_i)}\right) - \left(\frac{C-E[y_i]}{V(y_i)}\right)\phi\left(\frac{C-E[y_i]}{V(y_i)}\right)}{\Phi\left(\frac{C-E[y_i]}{V(y_i)}\right) - \Phi\left(\frac{c-E[y_i]}{V(y_i)}\right)}\right] - \left[\frac{\phi\left(\frac{c-E[y_i]}{V(y_i)}\right) - \phi\left(\frac{C-E[y_i]}{V(y_i)}\right)}{\Phi\left(\frac{C-E[y_i]}{V(y_i)}\right) - \Phi\left(\frac{c-E[y_i]}{V(y_i)}\right)}\right]^2\right]$$

It is easy to confirm that $0 \leq \frac{\theta}{1-(1-\theta)\rho^2} \leq 1$ since both θ and ρ^2 are between 0 and 1. 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. When $\gamma > 1$, there exits an optimal level of truncation. 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. Ad hoc approaches do not necessarily achieve this goal. Indeed, this is an interesting point. While removing observations above and below some criterion based on an ad hoc definition of the support of y_i^* may result in the optimal trim level, other solutions exist as well. Thus, there is nothing magic about the choice of (c, C). Further, there is no reason to believe that the definition of support for y_i^* will yield the optimal trimming amount. For simplicity, only "symmetric solutions" will be examined here. We choose $c = E[y] - c^*$ and $C = E[y] + c^*$, only the term c^* needs to be found. The details of the derivation are given in the appendix, and produce the following proposition:

Proposition 2 *Because the solution involves the cdf of the standard normal distribution, there is no closed form expression. The optimal c^* is given implicitly by:*

$$2\left(\frac{c^*}{V(y)}\right)\left(\frac{\phi\left(\frac{c^*}{V(y)}\right)}{\Phi\left(\frac{c^*}{V(y)}\right) - \Phi\left(\frac{-c^*}{V(y)}\right)}\right) = \frac{\gamma - 1}{\gamma - \rho^2}. \tag{17}$$

Thus the optimal cleaning depends on the variance of the observed y, the correlation between y and $\underline{x}$, and γ. The right hand side of 17 is increasing in γ. The left hand side of 17 is decreasing in c^*. Thus as γ increases, the truncation points must move closer to the mean: the data must be truncated more heavily. This highlights the fact that the optimal trimming choice is not specifically a function of a "known" support for the data, since the optimal level of trimming will change with the measurement error process. In order to obtain the correct amount of truncation, the researcher must know γ, or at least have a consistent estimate. 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 $\underline{x}_i$ 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 approaches which fail to consider at least the last two requirements are highly suspect.

An interesting point here is that information about γ is necessary for obtaining an optimal trimming rule. The rule derived here relies upon the assumption of normality. Indeed, other optimal trimming rules can be derived under other distributional assumptions. However, some assumptions about distributions must be made. As will be seen below, the general impact of trimming on coefficients depends upon underlying distributions. Hence, to properly, and efficiently, trim, one must have both information about γ and information about distributions. The results in sections 2 and 3 demonstrated that with information on γ

(estimates or bounds) that consistent estimates (or bounds) for $\underline{\beta}$ could be obtained. The results in sections 2 and 3 did not require information about underlying distributions. Hence, trimming, and by extension Winsorizing (see below), require stronger information than is needed for identification. While some might argue that trimming can be done without information about γ, one can similarly argue that rescaling the slope coefficients $\underline{b}$ can be done without knowledge of γ too. In both cases, an arbitrary assumption about γ has been made: in trimming it is made implicitly through expressions like equation 17 above. When rescaling the arbitrary assumption about γ is clear to all. However, both cases are equally arbitrary if no external information about γ is available. This will be seen in the Monte Carlo results below. There is no one trimming rule that will work in all cases. There is no one rescaling rule that will work in all cases.

The variance of the measurement error is often used as a measure of the severity of the error. Hence, the derivative of $\left(\frac{\theta}{1-(1-\theta)\rho^2}\right)$ with respect to σ_ε^2 reveals how the truncation bias is effected by more or less measurement error. The results in the appendix demonstrate our next result:

Proposition 3 *Since* $\frac{\partial}{\partial\sigma_\varepsilon^2}\left(\frac{\theta}{1-(1-\theta)\rho^2}\right) < 0$ *as the measurement error becomes more severe, the bias from the truncation becomes more severe also! Thus, not only does truncation bias what would be a consistent estimate, the relationship between that bias and the severity of the measurement error is quite the opposite of what the researcher would like.*

4.1.1 Does Trimming 'Correct' Standard Errors?

A second reason often 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, hence the asymptotic variance of the estimated slope from the truncated data needs be derived from the expression

$$AV\left(\widehat{\underline{b}^*}\right) = Q^{-1}E\left[\left(y_i - \underline{x}_i^T\underline{b}^*\right)^2\underline{x}_i\underline{x}_i^T | c \leq y_i \leq C\right]Q^{-1}. \tag{18}$$

Where $Q = E\left[\underline{x}_i\underline{x}_i^T | c \leq y_i \leq C\right]$ .Under assumptions A1-A7, the conditional distribution of $y_i|\underline{x}_i$ is normal $\left(\underline{x}_i^T\underline{b}, V\left(y_i\right)\left(1-\rho^2\right)\right)$. We can use the results in Goldberger (1981) to obtain an expression for the truncated second moment matrix as:

$$Q = E\left[\underline{x}_i\underline{x}_i^T | c \leq y_i \leq C\right] = E\left[\underline{x}_i\underline{x}_i^T\right] - (1-\theta)\,E\left[\underline{x}_i\underline{x}_i^T\right]\underline{b}\underline{b}^T E\left[\underline{x}_i\underline{x}_i^T\right]. \tag{19}$$

The term $E\left[\left(y_i - \underline{x}_i^T\underline{b}^*\right)^2\underline{x}_i\underline{x}_i^T | c \leq y_i \leq C\right]$ is considered by using the law of iterated expectations. The $E\left[\left(y_i - \underline{x}_i^T\underline{b}^*\right)^2 | \underline{x}_i, c \leq y_i \leq C\right]$ 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 $\underline{x}_i^T \underline{b}^*$. Doing so yields:

$$E\left[\left(y_i - \underline{x}_i^T \underline{b}^*\right)^2 | \underline{x}_i, c \leq y_i \leq C\right] =$$
$$\theta V(y_i)\left(1 - \left(\frac{\theta}{1-(1-\theta)\rho^2}\right)\rho^2\right) + \left(\underline{x}_i^T \underline{b}^* - m(\underline{x}_i)\right)^2, \tag{20}$$

where $m(x_i)$ is the conditional mean of y_i given $\underline{x}_i$ in the truncated sample. Combining these terms yields

$$\begin{aligned} AV\left(\widehat{\underline{b}^*}\right) &= \theta V(y_i)\left(1 - \left(\frac{\theta}{1-(1-\theta)\rho^2}\right)\rho^2\right) Q^{-1} \\ &\quad + Q^{-1} E\left[\left(\underline{x}_i^T \underline{b}^* - m(\underline{x}_i)\right)^2 \underline{x}_i \underline{x}_i^T | c \leq y_i \leq C\right] Q^{-1}. \end{aligned} \tag{21}$$

The leading term is comparable to the asymptotic variance expression for the OLS estimate in the full sample: $V(y_i)\left(1-\rho^2\right) E\left[\underline{x}_i \underline{x}_i^T\right]^{-1}$. As can easily be confirmed

$$\theta V(y_i)\left(1 - \left(\frac{\theta}{1-(1-\theta)\rho^2}\right)\rho^2\right) \leq V(y_i)\left(1-\rho^2\right).$$

However, $\left(E\left[\underline{x}_i \underline{x}_i^T\right]^{-1} - E\left[\underline{x}_i \underline{x}_i^T | c \leq y_i \leq C\right]^{-1}\right)$ is necessarily positive semi-definite (Goldberger, 1981). Intuitively, the variance of x_i in the truncated sample cannot be larger than the variance of the full sample (a sufficient condition here is the joint normality). Hence, a comparison of the leading terms is indeterminate. This is in contrast to similar comparisons for only a mean (and consequently the results in Stigler (1977)). In that case, the term $\theta V(y_i) < V(y_i)$ and thus trimming necessarily reduces the variance in the leading term. Here, the comparison is not so straight forward.

The comparison does not end there. The second term is due to heteroskedasticity from the trimming. That term is necessarily positive definite. Hence, even if the leading term is, in a positive definite sense, smaller than the variance of the OLS estimate on the full sample, the second term may reverse, or at least mitigate that difference.

Finally, the effect of truncation on sample size must be taken into account. The finite sample variance of $\widehat{\underline{b}^*}$ is given by

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

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.

A comparison between the variance of the truncated estimates and the variance of the full sample estimates is complicated and highly dependent upon the underlying parameters of the joint distribution. It

is not possible to sign this difference, even here where the strong assumption of normality is imposed. It is not clear that trimming will provide more precise estimates under these strong assumptions. Relaxing normality will not make the result clearer.

Indeed, as the next section of the paper demonstrates, the rigorous monte-carlo simulations that we undertake below indicate little or no effect on standard errors from trimming. Hence, no large gains in precision are available. Further, precision is not the driving concern and the effects of trimming must be measured using mean squared error. Certainly, it is quite possible to arrive at examples where trimming both reduces bias (necessarily $\gamma > 1$) and reduces the variance of the estimate. We advocate trimming when these conditions are met. In order to know that these conditions are met, we must know the underlying structure of the measurement error. Again, the above results only hold for the strong assumption of joint normality. Claims when this assumption fails are highly suspect.

4.2 Analytic Results: Winsorized Data

An approach related to the truncation results above is the "winsorizing" approach (see for example Angrist and Krueger (2000)). Rather than truncation, the 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 follows

$$y_i^w = \begin{array}{cc} C & \text{if } y_i \geq C \\ y_i & \text{if } c < y_i < C \\ c & y_i \leq c. \end{array} \tag{22}$$

Here, we present the effect on the coefficient comparable to the previous subsection when the data are winsorized:

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

The results for winsorizing are, as one would expect, comparable to the results for trimming. As will be seen in the Monte-Carlo work below, Winsorizing will have less of an impact on the slope coefficients (relative to the OLS). Again, if $\gamma > 1$,an optimal choice of Winsorizing points is available. However, it requires knowledge of γ, plus the strong assumption of normality ($A7$).

The effects of Winsorizing on the variance are derived similarly to the results above:

$$AV\left(\widehat{\underline{b}^{**}}\right) = E\left[\underline{x}_i \underline{x}_i^T\right]^{-1} E\left[\left(y_i^W - \underline{x}_i \underline{b}^{**}\right)^2 \underline{x}_i \underline{x}_i^T\right] E\left[\underline{x}_i \underline{x}_i^T\right]^{-1}.$$

Here, the term $E\left[\left(y_i^W - \underline{x}_i\underline{b}^{**}\right)^2|\underline{x}_i\right]$ can be broken into three terms

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

Combined with results from the previous section, we obtain

$$\begin{aligned}&AV\left(\widehat{\underline{b}^{**}}\right) = \\ &\left(\Phi\left(\frac{C-E[y_i]}{V(y_i)}\right)-\Phi\left(\frac{c-E[y_i]}{V(y_i)}\right)\right)\left(\theta V(y_i)\left(1-\left(\frac{\theta}{1-(1-\theta)\rho^2}\right)\right)E\left[\underline{x}_i\underline{x}_i^T\right]^{-1}QE\left[\underline{x}_i\underline{x}_i^T\right]^{-1}\right. \\ &\qquad\left.+E\left[\underline{x}_i\underline{x}_i^T\right]^{-1}E\left[\left(\underline{x}_i^T\underline{b}^{**}-m^W(\underline{x}_i)\right)^2\underline{x}_i\underline{x}_i^T|c<y_i<C\right]E\left[\underline{x}_i\underline{x}_i^T\right]^{-1}\right) \\ &\quad+\Phi\left(\frac{c-E[y_i]}{V(y_i)}\right)\left(E\left[\underline{x}_i\underline{x}_i^T\right]^{-1}E\left[\left(\left(c-\underline{x}_i\underline{b}^{**}\right)^2\underline{x}_i\underline{x}_i^T\right)|y_i\le c\right]E\left[\underline{x}_i\underline{x}_i^T\right]^{-1}\right) \\ &\quad+\left(1-\Phi\left(\frac{C-E[y_i]}{V(y_i)}\right)\right)\left(E\left[\underline{x}_i\underline{x}_i^T\right]^{-1}E\left[\left(\left(C-\underline{x}_i\underline{b}^{**}\right)^2\underline{x}_i\underline{x}_i^T\right)|y_i\ge C\right]E\left[\underline{x}_i\underline{x}_i^T\right]^{-1}\right).\end{aligned}$$

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. One advantage Winsorizing has over trimming is that the penalty of lost data does not effect the expression for the finite sample variance. Overall, Winsorizing and trimming have similar effects on both slope estimates and asymptotic variance. Again, with strong information, optimal rules can be found. However, other estimators are available with this information.

4.3 Cleaning Data in the General Case

The results derived in the previous section rely heavily upon the normality assumption (A7). However, as Goldberger (1981) shows, if the normality of x is relaxed, then the attenuation bias results for truncation do not hold (except in the simple case where x is scalar). Some coefficients may be attenuated by truncation, while others, in the same model, may be inflated by truncation. Similarly for the censoring case (winsorizing). Hence, even if the direction of the bias due to measurement error were known, without knowledge of the underlying joint distribution of (x, u, ε), the truncation or winsorizing of the data cannot be relied upon to adjust the slope coefficients for the bias. It is conventional wisdom that truncation typically results in attenuated coefficients, but this result depends heavily on the underlying distributions. The formulas given above for optimal cleaning when the parameters of the response error model are known, depend upon the joint

normality assumption. Optimal trimming (winsorizing) rules could be derived under other distributional assumptions. As a practical matter one would need to derive a rule specific to each researchers case.

In contrast, the measurement error bias results from Section 2 were derived under much weaker conditions. This implies that if information about γ is available the rescaling approach discussed previously is also available. Indeed, simple expressions for bounds and sensitivity of results are clear from the projection equation 8.

Our results demonstrate the conditions under which cleaning procedures may or may not work. However, the analytical results do not offer a sense of the degree to which cleaning affects the magnitudes of the estimated coefficients. Understanding the empirical magnitude of these effects is the subject of the next section.

5 Results

In order to examine more general results, we present a set of Monte Carlo simulations. The simulations can be divided into two groups. In the first group, we generate data with known distributions in order to examine the robustness of the results above to deviations from the distributional assumptions necessary for analytic solutions. In general, we find that the results above hold qualitatively. However, it is difficult to find optimal trimming rules, even when such rules might exists. Further, they are often not obviously related to the known distributions. In the second group of simulations, we draw data from the 1990 PUMS and estimate the returns to schooling. We treat estimates from the full PUMS file the population to be analogous to population parameters, and benchmark the results of different cleaning procedures against these parameters. This allows for a complicated model to be examined, with relationships similar to those found in typical economic data.

5.1 Monte Carlo Evidence from Simulated Data

5.1.1 Univariate Results

We discuss Monte Carlo from simulated data as evidence in support of our analytical results at two levels. First, we examine the case derived analytically above: all variables are jointly normally distributed. By parameterizing the models, the extent of the bias due to measurement error, and the specific impact of typical cleaning strategies can be seen. The analytic solutions above provide a general result, the results here provide an understanding of the relative magnitudes. Second, we study the case where the x variable is not normally distributed, and find that qualitatively the analytic results are supported. This finding further strengthens the argument against cleaning.

The first set of results is reported in Table 1. Here, we report the effect of different cleaning procedures on clean (uncorrupted) data that were generated using the model $y = 1 + x + u$ where $u \tilde{} N(0, 1)$.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 Table 1 we report Monte Carlo results from estimating the basic model using 1,000 replications each with a sample size of 1,000. For each replication we resample the covariate, the measurement error, and the regression error. The first panel indicates that with a normally distributed covariate, trimming and winsorizing both lead to bias: the bias from 1% trimming is almost 7% and that from 1% winsorizing is 2% (the true value of the coefficient on x is known to be one). The bias grows dramatically as the degree to which the data are truncated or censored increases, and approaches 25% with a 5% trimming rule. None of the cleaning procedures are neutral when the data are uncontaminated. While not explicitly discussed in the analytical section of our paper, we also report the results from performing median regression on the data as an alternative cleaning procedure. In general, median regression produces results that are very similar to standard least squares regressions (although it is computationally much more intensive and is inferior to least-squares on a RMSE criterion).

Similar results are obtained regardless of the distribution of the covariate (the distributions of the measurement error and regression error continue to be drawn from $N(0, 1)$). When the covariate is uniformly distributed, 1% trimming introduces a bias of over 10%, while the bias from 1% winsorizing stays at 2 percent. While the pattern of results demonstrating the non-neutrality of trimming is completely stable, it is not the case that winsorizing dominates trimming as a general rule– as the last panel of the table demonstrates, when the covariate is log normally distributed, 1% winsorizing introduces a bias of over 20 percent (and 5% winsorizing attenuates the coefficient by 45%). Here, trimming is certainly less harmful than winsorizing, although the strategy of not cleaning the data at all strictly dominates the other two.

In Table 2 we repeat the exercise but now introduce different forms of measurement error. Since the results of the second section proved that the linear projection of y on y^* is the central equation of interest for assessing the improvements produced by cleaning, we restrict our analysis to three values of γ that parameterize the linear projection ($\gamma = 1, \gamma = 0.9$ and $\gamma = 1.1$). Here, it is obvious that in the case of additive white noise, all cleaning procedures are inferior to doing nothing. This is true both in terms of bias, as well as in terms of variance, implying that the strategy of not cleaning the data dominates on a RMSE criteria. Furthermore, we find support for our theoretical conjecture that as the variance of the measurement error increases, the bias from trimming increases as well. Winsorizing is immune to this

problem– while dominated by the "doing nothing" strategy, the bias from 1% and 5% winsorizing is stable regardless of the variance of the underlying measurement error.

When we consider the realistic case of $\gamma = 0.9$ (as was estimated by Bollinger (1998) using the structure of measurement error in earnings) we see that once again the cleaning procedures are always dominated by the choice of 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. As such, we find support for our "iatrogenic" characterization of cleaning processes. For the most part, 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. Finally, in the last panel we note the cases where trimming works: 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, especially in the light of winsorizing robustness to the variance of the measurement error. Second, as stated in Section 2 of the paper, it is very difficult to find examples of measurement error where $\gamma > 1$.Therefore, before applying the prescription, it is key that the analyst be able to justify the $\gamma > 1$ model.

5.1.2 Multivariate Results and Departures from Normality

In the three panels of Table 3, we study the effects of cleaning procedures on multiple regression with a non-normal distribution for the regressors. In Table 3A we construct a normally distributed covariate and include its squared term (the squared term will not have a normal distribution). In Panel 3B, the covariate has an exponential distribution and in Panel 3C it is log-normally distributed. In all three tables, the population parameters on the coefficient of the covariate and its squared term are both set to one. The results from Table 3A are striking. When $\gamma = 1$ or $\gamma < 0$, there is no cleaning procedure that improves estimates over not cleaning the data. In fact, as demonstrated in the univariate case, the bias from cleaning procedures is an increasing function of the variance of the measurement error. In the case where $\gamma = 1.1$ (Table 3C), a 1% trimming rule is preferred to doing nothing, but this result is undone at the 5% level. unlike the univariate case studied in Table 2, it is no longer the case that a 5% winsorizing rule is the best cleaning procedure– for the case in which $\gamma = 1.1$,a 5% winsorizing rule introduces much more bias than doing nothing, and a 1 percent winsorizing rule does the best.

Unfortunately, it appears to be impossible to derive generalizable results for the optimal trimming rule

in the multivariate case. As can be seen in Table 3B winsorizing the data at any level wrecks havoc with the data when the X's are exponentially distributed. Analogous to the univariate case, cleaning procedures do not reduce bias (even on a RMSE criteria) when $\gamma = 0.9$. In this non-normal cases, a limited case can be made for 1% trimming when $\gamma = 1.1$ and the variance of the measurement error is small. However, even in this case, as the variance of the measurement error increases, all the gains from trimming disappear implying that this is not a general result. In the case of Table 3C where the covariate has a log-normal distribution, there is evidence favoring a 1% Trimming rule when $\gamma > 1$, but as the variance of the measurement error increases, this result can no longer be justified on a RMSE criteria. Interestingly, Table 3C also documents another problem with cleaning procedures: *while a certain cleaning rule may improve matters over not cleaning the data for a single coefficient, the same rule can exacerbate the bias for another coefficient.* This problem is clearly seen in Table 3C, in the cases where [$\gamma = 0.9$, $var(e) = 1$], when [$\gamma = 0.9$, $var(e) = 1$] and when [$\gamma = 1.1$, var(e)=1]. In all these cases, the 1% trimming rule improves the estimate of either the coefficient on x or that on x-squared. Simultaneously, note that the corresponding coefficient on x-squared (or x) has worsened on a RMSE criteria. Together the results presented in this section demonstrate that it is impossible to agree on a universal cleaning procedure– there are occasions when a certain rule appears to work, but this finding depends on the distribution of the covariate, the degree of cleaning, the presence of another covariate, and the variance of the measurement error. In other words, when we do find a rule that works, it appears to do so only serendipidously.

5.1.3 Comparing rescaling approaches to trimming approaches.

As noted in previous sections, another identification approach is to rescale the estimates. It was noted above in section 4 that in order to arrive at an optimal trimming rule, one needed both information about distributions and information about γ. The Monte Carlo results highlight this aspect. If we know that $\gamma = 1.1$, a strong assumption indeed, then scanning across the different tables, no particular trimming or winsorizing rule will work in every case: however rescaling will always work. Examination of the "Nothing" column shows that no matter what value γ may take, knowledge of γ alone will be sufficient to arrive at consistent estimates. It is clear in examining the results across these tables that knowledge of γ is not sufficient to determine how much trimming is necessary. Indeed even when γ is know, trimming cannot always be used to obtain consistent estimates of all slope coefficients.

Even if γ is not known, one can use the rescaling result to perform sensitivity analysis: how sensitive to different values of γ are the conclusions we draw from our OLS slopes? The robustness of our conclusions can 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 (forthcoming), 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 Monte Carlo Evidence from U.S. Decennial Census Data

To shed light on the more general case we present evidence from quasi-simulated data drawn from the PUMS samples of the 1990 US Decennial Census. We study the conventional problem of assessing 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. The PUMS data are treated as having no measurement error: we are using them as the true population for y_i^* and x. This allows us to examine a case where the underlying joint distribution of (y_i^*, x) is a more complicated multivariate distribution. We use the full PUMS data and delete all observations that had a value of minimum wage under \$3.35 in 1989. To keep the analysis tractable, we delete all non-workers from the sample and restrict our analysis to prime aged men (those between 25-55). These sample-selection criteria were imposed in order to abstract from simultaneously having to model labor force participation or enrollment in college. We artificially add measurement error to reported wages (based on the models above), while retaining the distribution of the covariates. We are treating the PUMS sample of 346,900 men as a population from which we will draw random samples of size 1,000. Table 4 presents mean and standard deviation parameters for the population. Black men comprise 8.3 percent of the sample and the average years of potential experience is 17.67. The men in our sample had on average completed 13.37 years of schooling. The table reports "population parameters" in column 3. These were generated by estimating the Micerian wage equation on the sample of 346,900 men. In Tables 4 and 5, we use 1000 replications of samples of size 1,000.

Table 4 is generated in the spirit of Table 1– no measurement error has been added to the PUMS data. In this situation, the cleaning procedures do not generally do 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 also found to be simultaneously inferior to not cleaning the data in regards to the estimation of the coefficients on schooling and race. Together these results confirm the results from the univariate case– no cleaning procedure is neutral when applied to already clean data.

Measurement error is added to the data in Table 5; we select two values for the variance of this measure-

ment error using the results of Bound and Kruger (1991), who note that $Var(lnY) = 0.458$ and 0.529 with corresponding error variances are .083 and .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 the case of additive white noise we find that a 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 with the censoring rate is increased to 5%. Least-squares is found to be always 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. Trimming procedures dominate not cleaning the data on a RMSE criteria, but can be worse in terms of the bias component.

6 Conclusions

The common practice of cleaning data by removing observations where the dependent variable is larger or smaller than some threshold has often been justified by claiming that it reduced the impact of measurement error. For example, observations where the wage falls below the minimum wage, or is so high as to be "incredible" are often discarded as being "obvious measurement error." While this appears to be a sensible argument, the problem is that the observations are not removed randomly, but rather systematically. This induces bias in the estimates. Under certain circumstances, as discussed in section 4, it may be possible to achieve an optimal cleaning strategy, but if the information necessary for that result were available, a simpler approach works too. Moreover, the optimal solution is highly dependent upon the underlying distributions and cannot be generalized: it must be treated on a case by case basis. Yet the simpler solution is available under much weaker conditions.

The analytic results demonstrate that only the case where measurement error in the dependent variable results in an upward bias of the magnitude of coefficients can cleaning strategies work. This case is not supported empirically by investigations into the structure of response error, at least for earnings data. Typically, rather than actually improve estimates, the cleaning strategies further exacerbate bias due to measurement error. Following medical terminology, we term this source of bias "Iatrogenic" (or econometrician induced) specification error.

The results here only begin to shed light on this subject. Other questions remain: one important extension may be to examine cleaning rules based on X variable values. While in general, truncation and censoring along the X axis does not bias slope coefficients (provided the linear model applies), however, it may be that an interaction with certain measurement error processes can result in either bias, or bias correction. Another extension may be to examine more complex measurement error processes: where the error process for Y differs across values of X's. Finally, in the case where the researcher has access to panel data (and multiple outcomes on the dependent variable), it may be possible to derive complicated cleaning procedures that exploit the signal to noise ratio in that data.

We conclude by noting that for any approach to correcting measurement error the key information is some knowledge of the structure of the error process. Our results don't contradict those of Stigler (1977), but rather consider a broader class. Stigler (1977) only considers estimation of the mean, the results here demonstrate that the problem is more complicated in a regression setting. Stigler's (1977) results are shown for a set of lab experiments where the assumption of additive white noise measurement error is quite plausible. In social sciences that assumption is more difficult to maintain. Echoing the conclusions of Hyslop and Imbens (2001), we note that information about the measurement error structure cannot be inferred from primary data and so we highlight the need for studies like Bollinger (1998), Bound and Kreuger (1991), Bound, Brown, Duncan and Rodgers (1994), Mellow, and Sider (1983), or Rodgers,. Brown and Duncan (1993). These type of studies provide the important information that can allow researchers to address problems of measurement error. Ad hoc approaches, such as trimming or winsorizing are unlikely to cure the disease and may even further bias conclusions.

7 Appendix

7.0.1 Derivation of equation 17

To solve for the expression in 17, 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

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

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

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:

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

Substitute for θ and solve

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

7.0.2 Proof of Proposition 3

To show that the bias gets worse with more variance in ε, consider the derivative of the bias term:

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

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

This term is negative iff the numerator is negative, considering only the numerator and grouping like 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 (see Goldberger) the ratio of the variance of the truncated standard normal to the variance of the untruncated standard normal, increasing c and decreasing C will result in a lower variance for the truncated distribution and thus a lower θ. Hence, by inspection, $\frac{\partial \theta_{LI}}{\partial \sigma_\varepsilon^2} < 0$ also. Combined with the bounds on ρ^2 and θ, the result is established.

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 Nonclassical Measurement Error," *Journal of the American Statistical Association* 95: 739-48.

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

[4] Bollinger, Christopher R., forthcoming "Measurement Error in Human Capital and the Black -White Wage Differential," *Review of Economics and Statistics.*

[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] 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.

[8] 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.

[9] 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.

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

[11] 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.

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

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

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

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

[16] Hyslop, Dean R. and Guido W. Imbens, 2001. "Bias From Classical and Other Forms of Measurement Error," Journal of Business and Economic Statistics 19(2): 475-481.

[17] 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.

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

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

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

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

[22] 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.

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

Table 1: Monte Carlo Simulations of the Effect of Cleaning Procedures on Uncorrupted Data, by Distribution of Covariate; True Model is y* = 1 + x + u; u~N(0,1)

	Do Nothing	Trim 1%	Trim 5%	Wins 1%	Wins 5%	Median
NORMAL X						
Beta	1.0000	0.9323	0.7683	0.9800	0.9002	1.0002
se (b)	0.0312	0.0316	0.0328	0.0310	0.0305	0.0400
RMSE (b)	0.0312	0.0747	0.2340	0.0369	0.1043	0.0400
Cons	1.0000	1.0000	0.9999	0.9999	0.9999	1.0010
se (cons)	0.0325	0.0326	0.0338	0.0325	0.0330	0.0413
UNIFORM X						
Beta	1.0037	0.8916	0.6505	0.9862	0.9091	1.0062
se (b)	0.1066	0.1017	0.0937	0.1047	0.0987	0.1380
RMSE (b)	0.1067	0.1487	0.3618	0.1056	0.1342	0.1381
Cons	0.9979	1.0540	1.1745	1.0066	1.0451	0.9960
se (cons)	0.0634	0.0612	0.0574	0.0625	0.0598	0.0818
EXPONENTIAL X						
Beta	1.0011	0.9437	0.7693	0.9510	0.8137	1.0006
se (b)	0.0316	0.0344	0.0383	0.0374	0.0409	0.0402
RMSE (b)	0.0316	0.0660	0.2339	0.0616	0.1907	0.0402
Cons	0.9997	1.0702	1.2556	1.0439	1.1605	1.0003
se (cons)	0.0447	0.0463	0.0480	0.0483	0.0502	0.0570
LOGNORMAL X						
Beta	1.0000	0.9798	0.8718	0.7964	0.5517	0.9994
se (b)	0.0147	0.0201	0.0311	0.1083	0.0927	0.0193
RMSE (b)	0.0147	0.0285	0.1319	0.2306	0.4578	0.0193
Cons	0.9994	1.0518	1.2477	1.2931	1.5984	1.0011
se (cons)	0.0393	0.0441	0.052	0.1693	0.1444	0.0505

Reported estimates are empirical sample moments from 1,000 replications each with a sample size of 1,000. Each replication resampled both the measurement error and the regression error. The regression error and measurement error are uncorrelated with each other.

Table 2: Monte Carlo Simulations of the Effect of Cleaning Procedures on Corrupted Data, for Normally Distributed Covariate, by variance of measurement error; True Model is y* = 1 + x + u; u~N(0,1) and x ~ N(0,1)

	Variance (e) = 0.25						Variance (e) = 1						Variance (e) = 3					
	Nothing	Trim 1%	Trim 5%	Wins 1%	Wins 5%	Median	Nothing	Trim 1%	Trim 5%	Wins 1%	Wins 5%	Median	Nothing	Trim 1%	Trim 5%	Wins 1%	Wins 5%	Median
Error Model: y = y* + e																		
Beta	0.9991	0.9305	0.7630	0.9801	0.9001	1.0006	1.0020	0.9150	0.7151	0.9830	0.9027	1.0032	1.0053	0.8911	0.6515	0.9862	0.9064	1.0081
SE (b)	0.0321	0.0325	0.0335	0.0320	0.0313	0.0405	0.0452	0.0448	0.0443	0.0446	0.0429	0.0561	0.0978	0.0943	0.0857	0.0961	0.0898	0.1226
RMSE (b)	0.0321	0.0768	0.2394	0.0377	0.1047	0.0405	0.0452	0.0961	0.2883	0.0477	0.1063	0.0562	0.0979	0.1440	0.3589	0.0971	0.1298	0.1229
Constant	1.0004	1.0004	1.0003	1.0003	1.0002	1.0010	0.9999	1.0000	1.0000	0.9999	1.0001	1.0009	1.0019	1.0021	1.0014	1.0021	1.0018	1.0032
se (cons)	0.0335	0.0336	0.0347	0.0335	0.0338	0.0412	0.0454	0.0456	0.0467	0.0455	0.0460	0.0557	0.1055	0.1060	0.1074	0.1058	0.1062	0.1301
Error Model: y = 0.9y* + e																		
Beta	0.8989	0.8369	0.6855	0.8818	0.8099	0.8995	0.9023	0.8216	0.6378	0.8852	0.8127	0.9005	0.9031	0.7980	0.5788	0.8859	0.8131	0.9051
SE (b)	0.0292	0.0296	0.0307	0.0290	0.0287	0.0370	0.0434	0.0428	0.0427	0.0428	0.0409	0.0546	0.0966	0.0904	0.0843	0.0944	0.0875	0.1240
RMSE (b)	0.1053	0.1658	0.3160	0.1217	0.1923	0.1070	0.1069	0.1835	0.3647	0.1226	0.1917	0.1135	0.1368	0.2213	0.4296	0.1481	0.2063	0.1561
Constant	0.9003	0.9003	0.9003	0.9002	0.9003	0.9008	0.9008	0.9010	0.9009	0.9009	0.9009	0.8986	0.9006	0.9003	0.8999	0.9006	0.8999	0.9035
se (cons)	0.0302	0.0303	0.0313	0.0301	0.0306	0.0381	0.0412	0.0415	0.0432	0.0413	0.0421	0.0542	0.0981	0.0980	0.0992	0.0980	0.0987	0.1250
Error Model: y = 1.1y* +e																		
Beta	1.0986	1.0235	0.8399	1.0777	0.9896	1.0997	1.1031	1.0101	0.7946	1.0820	0.9934	1.1047	1.0985	0.9746	0.7163	1.0779	0.9899	1.0968
SE (b)	0.0351	0.0356	0.0369	0.0349	0.0343	0.0453	0.0487	0.0482	0.0472	0.0480	0.0454	0.0616	0.1006	0.0958	0.0879	0.0991	0.0922	0.1242
RMSE (b)	0.1047	0.0427	0.1643	0.0852	0.0359	0.1095	0.1140	0.0492	0.2107	0.0950	0.0459	0.1214	0.1408	0.0991	0.2970	0.1261	0.0928	0.1575
Constant	1.1001	1.1001	1.1000	1.1001	1.1000	1.1005	1.1022	1.1024	1.1026	1.1022	1.1022	1.1012	1.0965	1.0967	1.0971	1.0966	1.0963	1.1002
se (cons)	0.0367	0.0369	0.0381	0.0367	0.0372	0.0450	0.0450	0.0456	0.0474	0.0451	0.0456	0.0590	0.1042	0.1041	0.1049	0.1042	0.1040	0.1289

Reported estimates are empirical sample moments from 1,000 replications each with a sample size of 1,000. Each replication resampled both the measurement error and the regression error. The regression error and measurement error are uncorrelated with each other.

Table 3A: Monte Carlo Simulations of the Effect of Cleaning Procedures on Corrupted Data with Multivariate Covariates, by variance of measurement error; True Model is $y^* = 1 + x + x^2 + u$; u~N(0,1) and x ~ N(0,1)

	Variance (e) = 0.25						Variance (e) = 1						Variance (e) = 3					
	Nothing	Trim 1%	Trim 5%	Wins 1%	Wins 5%	Median	Nothing	Trim 1%	Trim 5%	Wins 1%	Wins 5%	Median	Nothing	Trim 1%	Trim 5%	Wins 1%	Wins 5%	Median
Error Model: y = y* + e																		
Beta 1	1.0026	0.9763	0.8677	0.9468	0.7814	1.0015	0.9999	0.9595	0.8054	0.9452	0.7842	1.0002	0.9998	0.8938	0.6624	0.9534	0.8294	0.9963
SE (b)	0.0327	0.0344	0.0394	0.0381	0.0447	0.0411	0.0445	0.0453	0.0501	0.0492	0.0512	0.0558	0.0972	0.0973	0.0924	0.0954	0.0881	0.1243
RMSE (b)	0.0328	0.0418	0.1380	0.0654	0.2231	0.0412	0.0445	0.0608	0.2009	0.0737	0.2218	0.0558	0.0972	0.1440	0.3500	0.1062	0.1920	0.1244
Beta 2	1.0000	0.9721	0.8626	0.9183	0.7222	0.9996	0.9989	0.9553	0.7987	0.9203	0.7314	0.9978	1.0026	0.8882	0.6598	0.9397	0.8004	1.0025
SE (b)	0.0237	0.0274	0.0339	0.0398	0.0479	0.0293	0.0313	0.0354	0.0436	0.0450	0.0529	0.0400	0.0689	0.0711	0.0696	0.0697	0.0661	0.0892
RMSE (b)	0.0237	0.0391	0.1415	0.0908	0.2819	0.0293	0.0313	0.0571	0.2059	0.0915	0.2738	0.0400	0.0689	0.1325	0.3472	0.0921	0.2103	0.0892
Constant	1.0020	1.0480	1.1825	1.0670	1.2013	1.0028	1.0026	1.0689	1.2484	1.0660	1.2009	1.0020	0.9971	1.1367	1.3718	1.0509	1.1665	0.9935
se (cons)	0.0406	0.0421	0.0439	0.0485	0.0516	0.0511	0.0544	0.0554	0.0576	0.0605	0.0636	0.0696	0.1249	0.1251	0.1205	0.1257	0.1232	0.1567
Error Model: y = 0.9y* +e																		
Beta 1	0.9020	0.8782	0.7795	0.8520	0.7030	0.9025	0.8976	0.8585	0.7108	0.8488	0.7043	0.8967	0.9003	0.7997	0.5904	0.8609	0.7538	0.9042
SE (b)	0.0300	0.0309	0.0353	0.0344	0.0404	0.0365	0.0423	0.0432	0.0467	0.0457	0.0480	0.0526	0.0984	0.0974	0.0917	0.0971	0.0901	0.1276
RMSE (b)	0.1025	0.1257	0.2233	0.1520	0.2997	0.1041	0.1108	0.1479	0.2930	0.1580	0.2996	0.1159	0.1401	0.2227	0.4197	0.1697	0.2621	0.1595
Beta 2	0.9001	0.8749	0.7753	0.8268	0.6500	0.9001	0.9000	0.8577	0.7091	0.8303	0.6607	0.8991	0.8987	0.7861	0.5786	0.8449	0.7249	0.9011
SE (b)	0.0213	0.0246	0.0305	0.0357	0.0432	0.0267	0.0296	0.0339	0.0409	0.0422	0.0479	0.0369	0.0706	0.0712	0.0671	0.0706	0.0643	0.0879
RMSE (b)	0.1021	0.1275	0.2268	0.1768	0.3527	0.1034	0.1042	0.1463	0.2937	0.1749	0.3427	0.1075	0.1235	0.2255	0.4267	0.1704	0.2825	0.1323
Constant	0.9017	0.9433	1.0653	0.9600	1.0812	0.9021	0.9008	0.9641	1.1324	0.9572	1.0791	0.9011	0.8981	1.0320	1.2438	0.9448	1.0459	0.8983
se (cons)	0.0361	0.0374	0.0392	0.0433	0.0462	0.0454	0.0512	0.0528	0.0545	0.0571	0.0599	0.0653	0.1190	0.1196	0.1183	0.1189	0.1163	0.1525
Error Model: y = 1.1y* +e																		
Beta 1	1.1025	1.0738	0.9548	1.0413	0.8593	1.1020	1.1007	1.0594	0.8966	1.0405	0.8623	1.1019	1.1003	0.9923	0.7408	1.0482	0.9063	1.1006
SE (b)	0.0362	0.0376	0.0436	0.0418	0.0487	0.0444	0.0465	0.0482	0.0540	0.0519	0.0564	0.0578	0.0988	0.0978	0.0952	0.0988	0.0924	0.1232
RMSE (b)	0.1087	0.0828	0.0628	0.0588	0.1489	0.1112	0.1109	0.0764	0.1167	0.0658	0.1488	0.1171	0.1407	0.0981	0.2761	0.1099	0.1316	0.1590
Beta 2	1.0999	1.0694	0.9499	1.0103	0.7942	1.1001	1.1004	1.0562	0.8914	1.0139	0.8045	1.0989	1.1001	0.9821	0.7308	1.0284	0.8669	1.1022
SE (b)	0.0261	0.0301	0.0374	0.0435	0.0526	0.0320	0.0339	0.0388	0.0472	0.0506	0.0591	0.0434	0.0759	0.0785	0.0758	0.0782	0.0737	0.0938
RMSE (b)	0.1033	0.0757	0.0625	0.0447	0.2124	0.1051	0.1060	0.0683	0.1184	0.0525	0.2042	0.1080	0.1257	0.0805	0.2797	0.0832	0.1522	0.1387
Constant	1.1020	1.1522	1.2994	1.1733	1.3212	1.1024	1.1010	1.1694	1.3600	1.1706	1.3192	1.1006	1.0999	1.2469	1.5069	1.1610	1.2924	1.0983
se (cons)	0.0449	0.0463	0.0483	0.0533	0.0569	0.0554	0.0573	0.0588	0.0602	0.0648	0.0679	0.0712	0.1323	0.1323	0.1295	0.1320	0.1290	0.1603

Reported estimates are empirical sample moments from 1,000 replications each with a sample size of 1,000. Each replication resampled both the measurement error and the regression error. The regression error and measurement error are uncorrelated with each other.

Table 3B: Monte Carlo Simulations of the Effect of Cleaning Procedures on Corrupted Data with Multivariate Covariates, by variance of measurement error; True Model is $y^* = 1 + x + x^2 + u$; u~N(0,1) and x ~ Exponential

	Variance (e) = 0.25						Variance (e) = 1						Variance (e) = 3					
	Nothing	Trim 1%	Trim 5%	Wins 1%	Wins 5%	Median	Nothing	Trim 1%	Trim 5%	Wins 1%	Wins 5%	Median	Nothing	Trim 1%	Trim 5%	Wins 1%	Wins 5%	Median
Error Model: y= y* + e																		
Beta 1	0.9987	0.9393	0.7461	2.7357	3.7847	0.9987	0.9983	0.9389	0.7780	2.7657	3.7715	0.9987	0.9951	1.0048	1.0972	2.6833	3.6101	0.9895
SE (b)	0.0755	0.1049	0.1464	0.6129	0.2437	0.0929	0.1095	0.1529	0.2042	0.6270	0.2486	0.1383	0.2411	0.3127	0.3920	0.6666	0.3308	0.3032
RMSE (b)	0.0755	0.1212	0.2931	1.8407	2.7953	0.0929	0.1095	0.1646	0.3016	1.8737	2.7826	0.1384	0.2412	0.3127	0.4039	1.8104	2.6310	0.3034
Beta 2	1.0000	1.0098	1.0392	0.4075	-0.1674	1.0000	1.0003	1.0062	0.9981	0.3972	-0.1644	1.0004	1.0003	0.9632	0.7289	0.4215	-0.1117	1.0011
SE (b)	0.0181	0.0316	0.0584	0.1906	0.0780	0.0217	0.0259	0.0453	0.0809	0.1954	0.0778	0.0320	0.0574	0.0934	0.1504	0.2014	0.0905	0.0707
RMSE (b)	0.0181	0.0331	0.0704	0.6224	1.1700	0.0217	0.0259	0.0457	0.0809	0.6337	1.1670	0.0320	0.0574	0.1004	0.3100	0.6125	1.1154	0.0707
Constant	1.0015	1.0653	1.2620	0.3401	0.1302	1.0025	1.0000	1.0794	1.3182	0.3279	0.1445	1.0011	1.0060	1.1302	1.5179	0.3742	0.2531	1.0127
se (cons)	0.0568	0.0636	0.0723	0.2610	0.1163	0.0715	0.0802	0.0913	0.1002	0.2704	0.1279	0.1026	0.1730	0.1872	0.2011	0.3163	0.2040	0.2242
Error Model: y= 0.9y* +e																		
Beta 1	0.8993	0.8467	0.6744	2.4629	3.4058	0.8973	0.8978	0.8456	0.7171	2.4874	3.3888	0.8965	0.8941	0.9188	1.0390	2.3924	3.2241	0.8947
SE (b)	0.0682	0.0943	0.1313	0.5512	0.2197	0.0834	0.1005	0.1402	0.1895	0.5636	0.2225	0.1225	0.2286	0.3159	0.3796	0.6003	0.3113	0.2934
RMSE (b)	0.1217	0.1800	0.3510	1.5633	2.4158	0.1323	0.1433	0.2085	0.3405	1.5906	2.3991	0.1604	0.2520	0.3262	0.3816	1.5163	2.2458	0.3117
Beta 2	0.8998	0.9082	0.9338	0.3664	-0.1502	0.9002	0.8999	0.9039	0.8848	0.3574	-0.1475	0.9003	0.9011	0.8591	0.6150	0.3872	-0.0914	0.9013
SE (b)	0.0163	0.0284	0.0529	0.1714	0.0707	0.0197	0.0236	0.0424	0.0742	0.1757	0.0706	0.0284	0.0526	0.0945	0.1442	0.1795	0.0826	0.0676
RMSE (b)	0.1015	0.0961	0.0848	0.6563	1.1524	0.1017	0.1028	0.1051	0.1370	0.6662	1.1497	0.1037	0.1121	0.1697	0.4111	0.6385	1.0945	0.1196
Constant	0.9015	0.9589	1.1362	0.3061	0.1176	0.9040	0.9021	0.9762	1.1961	0.2981	0.1364	0.9043	0.9035	1.0200	1.3889	0.3440	0.2398	0.9011
se (cons)	0.0513	0.0571	0.0647	0.2349	0.1048	0.0630	0.0767	0.0852	0.0949	0.2431	0.1151	0.0964	0.1773	0.1956	0.2016	0.2980	0.2008	0.2192
Error Model: y= 1.1y* +e																		
Beta 1	1.0976	1.0320	0.8234	3.0077	4.1624	1.0955	1.0981	1.0307	0.8515	3.0432	4.1538	1.0978	1.1032	1.0956	1.1856	2.9627	4.0152	1.1062
SE (b)	0.0821	0.1146	0.1620	0.6744	0.2689	0.1020	0.1113	0.1583	0.2199	0.6901	0.2717	0.1394	0.2401	0.3200	0.4156	0.7363	0.3500	0.3024
RMSE (b)	0.1276	0.1190	0.2397	2.1179	3.1738	0.1398	0.1484	0.1613	0.2654	2.1566	3.1655	0.1702	0.2614	0.3340	0.4551	2.0963	3.0355	0.3205
Beta 2	1.1002	1.1111	1.1420	0.4487	-0.1840	1.1007	1.1004	1.1083	1.1045	0.4367	-0.1829	1.1006	1.1017	1.0706	0.8381	0.4646	-0.1356	1.1007
SE (b)	0.0198	0.0347	0.0650	0.2097	0.0866	0.0243	0.0268	0.0473	0.0876	0.2155	0.0861	0.0323	0.0561	0.0960	0.1592	0.2244	0.0985	0.0702
RMSE (b)	0.1022	0.1164	0.1561	0.5898	1.1872	0.1036	0.1040	0.1182	0.1364	0.6031	1.1860	0.1057	0.1161	0.1192	0.2271	0.5805	1.1399	0.1228
Constant	1.1020	1.1720	1.3865	0.3746	0.1432	1.1034	1.1018	1.1871	1.4408	0.3621	0.1569	1.1026	1.0924	1.2240	1.6189	0.3948	0.2454	1.0944
se (cons)	0.0609	0.0686	0.0785	0.2870	0.1272	0.0773	0.0807	0.0919	0.1031	0.2936	0.1352	0.1060	0.1785	0.1959	0.2119	0.3421	0.2114	0.2315

Reported estimates are empirical sample moments from 1,000 replications each with a sample size of 1,000. Each replication resampled both the measurement error and the regression error. The regression error and measurement error are uncorrelated with each other.

Table 3C: Monte Carlo Simulations of the Effect of Cleaning Procedures on Corrupted Data with Multivariate Covariates, by variance of measurement error; True Model is $y^* = 1 + x + x^2 + u$; u~N(0,1) and x ~ Log Normal

	Variance (e) = 0.25						Variance (e) = 1						Variance (e) = 3					
	Nothing	Trim 1%	Trim 5%	Wins 1%	Wins 5%	Median	Nothing	Trim 1%	Trim 5%	Wins 1%	Wins 5%	Median	Nothing	Trim 1%	Trim 5%	Wins 1%	Wins 5%	Median
Error Model: y= y* + e																		
Beta 1	1.0016	0.9634	0.7431	7.6474	5.7768	1.0012	1.0002	0.9492	0.6931	7.4649	5.8290	1.0009	1.0027	0.9144	0.6399	7.4817	5.7951	1.0029
SE (b)	0.0309	0.0563	0.1010	1.4233	0.6132	0.0395	0.0426	0.0728	0.1374	1.5670	0.5893	0.0550	0.0962	0.1696	0.2919	1.5160	0.5780	0.1185
RMSE (b)	0.0309	0.0672	0.2760	6.7980	4.8160	0.0395	0.0426	0.0888	0.3363	6.6521	4.8648	0.0550	0.0962	0.1900	0.4635	6.6566	4.8298	0.1185
Beta 2	0.9998	1.0038	1.0428	0.0096	-0.1825	0.9999	1.0001	1.0053	1.0463	0.0359	-0.1885	1.0001	1.0000	1.0079	1.0219	0.0323	-0.1854	0.9998
SE (b)	0.0023	0.0081	0.0243	0.1578	0.0583	0.0029	0.0034	0.0107	0.0332	0.1769	0.0573	0.0044	0.0077	0.0249	0.0690	0.1722	0.0559	0.0096
RMSE (b)	0.0024	0.0090	0.0492	1.0029	1.1839	0.0029	0.0034	0.0119	0.0570	0.9802	1.1899	0.0044	0.0077	0.0261	0.0724	0.9829	1.1867	0.0096
Constant	0.9969	1.0594	1.3048	-4.2895	-1.6872	0.9976	1.0018	1.0876	1.4022	-4.1384	-1.7213	0.9995	1.0010	1.1752	1.7180	-4.1416	-1.6464	1.0028
se (cons)	0.0506	0.0627	0.0801	1.3870	0.5919	0.0664	0.0673	0.0814	0.1056	1.5093	0.5682	0.0859	0.1535	0.1856	0.2251	1.4671	0.5644	0.1914
Error Model: y= 0.9y* +e																		
Beta 1	0.9013	0.8664	0.6680	6.8825	5.1990	0.9014	0.8998	0.8504	0.6141	6.7178	5.2434	0.8992	0.8990	0.8212	0.5789	6.7271	5.2116	0.8995
SE (b)	0.0279	0.0500	0.0922	1.2813	0.5512	0.0355	0.0420	0.0730	0.1271	1.4111	0.5294	0.0534	0.0941	0.1697	0.2869	1.3668	0.5212	0.1165
RMSE (b)	0.1025	0.1426	0.3445	6.0204	4.2350	0.1048	0.1086	0.1665	0.4063	5.8893	4.2763	0.1141	0.1381	0.2465	0.5095	5.8879	4.2437	0.1539
Beta 2	0.8998	0.9035	0.9385	0.0087	-0.1642	0.8999	0.9000	0.9051	0.9424	0.0323	-0.1695	0.9000	0.9001	0.9059	0.9116	0.0291	-0.1666	0.9000
SE (b)	0.0021	0.0073	0.0224	0.1420	0.0523	0.0027	0.0035	0.0105	0.0308	0.1591	0.0513	0.0042	0.0075	0.0242	0.0678	0.1555	0.0501	0.0093
RMSE (b)	0.1002	0.0968	0.0655	1.0014	1.1654	0.1002	0.1001	0.0954	0.0653	0.9806	1.1706	0.1001	0.1002	0.0971	0.1114	0.9833	1.1677	0.1004
Constant	0.8974	0.9543	1.1761	-3.8604	-1.5180	0.8978	0.9017	0.9840	1.2793	-3.7243	-1.5451	0.9028	0.8979	1.0647	1.5862	-3.7253	-1.4771	0.8982
se (cons)	0.0458	0.0554	0.0728	1.2488	0.5322	0.0585	0.0639	0.0786	0.0982	1.3596	0.5083	0.0812	0.1518	0.1844	0.2266	1.3214	0.5138	0.1889
Error Model: y= 1.1y* +e																		
Beta 1	1.1018	1.0592	0.8185	8.4121	6.3537	1.1020	1.1019	1.0472	0.7653	8.2125	6.4144	1.1021	1.1015	1.0117	0.6943	8.2253	6.3745	1.1037
SE (b)	0.0325	0.0611	0.1104	1.5660	0.6743	0.0408	0.0460	0.0812	0.1472	1.7238	0.6471	0.0574	0.0972	0.1768	0.2957	1.6680	0.6334	0.1239
RMSE (b)	0.1069	0.0851	0.2125	7.5757	5.3960	0.1098	0.1118	0.0939	0.2770	7.4156	5.4530	0.1171	0.1405	0.1772	0.4253	7.4154	5.4117	0.1616
Beta 2	1.0998	1.1043	1.1468	0.0106	-0.2007	1.0998	1.0999	1.1056	1.1523	0.0395	-0.2075	1.0999	1.0999	1.1078	1.1326	0.0354	-0.2043	1.0998
SE (b)	0.0025	0.0089	0.0269	0.1736	0.0641	0.0031	0.0036	0.0117	0.0353	0.1945	0.0631	0.0045	0.0078	0.0260	0.0691	0.1899	0.0614	0.0099
RMSE (b)	0.0998	0.1047	0.1493	1.0045	1.2024	0.0999	0.1000	0.1063	0.1563	0.9800	1.2092	0.1000	0.1002	0.1109	0.1495	0.9831	1.2058	0.1003
Constant	1.0964	1.1655	1.4337	-4.7187	-1.8553	1.0964	1.0992	1.1898	1.5269	-4.5552	-1.8990	1.1001	1.0957	1.2731	1.8432	-4.5578	-1.8231	1.0931
se (cons)	0.0543	0.0675	0.0869	1.5258	0.6506	0.0697	0.0758	0.0935	0.1175	1.6642	0.6213	0.0951	0.1550	0.1931	0.2381	1.6122	0.6157	0.1966

Reported estimates are empirical sample moments from 1,000 replications each with a sample size of 1,000. Each replication resampled both the measurement error and the regression error. The regression error and measurement error are uncorrelated with each other.

Table 4: Monte Carlo Simulations of the Effect of Cleaning Procedures on Uncorrupted Data, Evidence from the Returns to Schooling in 1990 PUMS Data

	Mean	Population β	Do Nothing	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 $3.35 in 1989 (the nominal value of the minimum wage) are deleted from the analysis. Column (1) reports means and standard deviations for this sample of 346,900 individuals, and column 2 reports the parameters from estimating the model: ln wage=$\beta_0 + \beta_1$ Schooling $+\beta_2$ Exp$+\beta_3$ Exp^2 $+\beta_4$ 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 5: Monte Carlo Simulations of the Effect of Cleaning Procedures on Corrupted Data, Evidence from the Returns to Schooling in 1990 PUMS Data

	Var (e) = 0.1 x Var (wage); Var (wage) = 0.3144						Var (e) = 0.3 x Var (wage); Var (wage) = 0.3144					
	Nothing	**Trim 1%**	**Trim 5%**	**Wins 1%**	**Wins 5%**	**Median**	**Nothing**	**Trim 1%**	**Trim 5%**	**Wins 1%**	**Wins 5%**	**Median**
Error Model: y = y* + e												
Schooling	0.0918	0.0880	0.0706	0.0910	0.0856	0.1000	0.0925	0.0883	0.0704	0.0916	0.0859	0.1001
SE (Schooling)	0.0078	0.0072	0.0069	0.0075	0.0070	0.0086	0.0081	0.0074	0.0068	0.0078	0.0072	0.0091
RMSE	0.0078	0.0083	0.0226	0.0076	0.0095	0.0117	0.0082	0.0084	0.0227	0.0078	0.0095	0.0121
Pot. Exp	0.0375	0.0362	0.0295	0.0372	0.0353	0.0401	0.0377	0.0363	0.0294	0.0374	0.0354	0.0402
SE (Pot Exp)	0.0084	0.0079	0.0068	0.0081	0.0076	0.0095	0.0084	0.0078	0.0070	0.0081	0.0075	0.0096
RMSE	0.0084	0.0080	0.0105	0.0081	0.0079	0.0098	0.0084	0.0079	0.0107	0.0081	0.0078	0.0100
Pot. Exp. Sq /100	-0.0538	-0.0522	-0.0420	-0.0535	-0.0508	-0.0562	-0.0539	-0.0524	-0.0419	-0.0537	-0.0510	-0.0566
SE (Pot. Exp Sq)	0.0218	0.0203	0.0172	0.0211	0.0196	0.0244	0.0215	0.0201	0.0178	0.0209	0.0193	0.0248
RMSE	0.0218	0.0203	0.0208	0.0211	0.0198	0.0246	0.0215	0.0202	0.0213	0.0209	0.0195	0.0250
Black (1= yes)	-0.1417	-0.1383	-0.1083	-0.1418	-0.1339	-0.1637	-0.1426	-0.1394	-0.1097	-0.1426	-0.1346	-0.1616
SE (Black)	0.0596	0.0545	0.0501	0.0579	0.0532	0.0702	0.0639	0.0578	0.0513	0.0616	0.0564	0.0730
RMSE	0.0596	0.0546	0.0604	0.0579	0.0538	0.0734	0.0639	0.0578	0.0606	0.0616	0.0569	0.0756
Constant	0.8640	0.9263	1.2354	0.8748	0.9657	0.7296	0.8535	0.9222	1.2396	0.8667	0.9621	0.7281
SE	0.1273	0.1186	0.1116	0.1227	0.1146	0.1413	0.1345	0.1237	0.1147	0.1289	0.1203	0.1502
Error Model: y= 0.9y* + e												
Schooling	0.0826	0.0792	0.0635	0.0819	0.0770	0.0903	0.0833	0.0795	0.0634	0.0825	0.0773	0.0901
SE (Schooling)	0.0070	0.0065	0.0062	0.0068	0.0063	0.0078	0.0074	0.0067	0.0062	0.0071	0.0066	0.0082
RMSE	0.0117	0.0144	0.0292	0.0122	0.0163	0.0080	0.0115	0.0142	0.0294	0.0119	0.0161	0.0084
Pot. Exp	0.0338	0.0326	0.0265	0.0335	0.0318	0.0363	0.0339	0.0327	0.0264	0.0337	0.0318	0.0360
SE (Pot Exp)	0.0076	0.0071	0.0062	0.0074	0.0069	0.0085	0.0075	0.0071	0.0061	0.0073	0.0068	0.0087
RMSE	0.0085	0.0086	0.0125	0.0084	0.0089	0.0086	0.0083	0.0086	0.0127	0.0083	0.0088	0.0088
Pot. Exp. Sq /100	-0.0484	-0.0470	-0.0378	-0.0481	-0.0457	-0.0511	-0.0485	-0.0471	-0.0375	-0.0483	-0.0459	-0.0502
SE (Pot. Exp Sq)	0.0197	0.0183	0.0158	0.0190	0.0176	0.0220	0.0195	0.0182	0.0158	0.0190	0.0175	0.0225
RMSE	0.0203	0.0194	0.0223	0.0198	0.0193	0.0221	0.0201	0.0193	0.0225	0.0197	0.0191	0.0227
Black (1= yes)	-0.1276	-0.1244	-0.0979	-0.1277	-0.1206	-0.1474	-0.1276	-0.1250	-0.0976	-0.1276	-0.1205	-0.1439
SE (Black)	0.0539	0.0494	0.0452	0.0522	0.0480	0.0640	0.0576	0.0518	0.0458	0.0554	0.0507	0.0641
RMSE	0.0558	0.0525	0.0631	0.0542	0.0525	0.0642	0.0594	0.0545	0.0637	0.0572	0.0551	0.0641
Constant	0.7772	0.8337	1.1111	0.7870	0.8688	0.6511	0.7672	0.8295	1.1165	0.7789	0.8651	0.6558
SE	0.1154	0.1077	0.1008	0.1112	0.1039	0.1291	0.1208	0.1108	0.1020	0.1158	0.1075	0.1351
Error Model: y= 1.1y* + e												
Schooling	0.1010	0.0968	0.0777	0.1001	0.0942	0.1102	0.1018	0.0972	0.0775	0.1008	0.0945	0.1104
SE (Schooling)	0.0086	0.0079	0.0075	0.0082	0.0077	0.0095	0.0090	0.0082	0.0075	0.0087	0.0080	0.0100
RMSE	0.0124	0.0092	0.0162	0.0115	0.0079	0.0205	0.0132	0.0097	0.0164	0.0123	0.0084	0.0208
Pot. Exp	0.0413	0.0397	0.0324	0.0410	0.0388	0.0442	0.0415	0.0400	0.0325	0.0412	0.0390	0.0441
SE (Pot Exp)	0.0093	0.0086	0.0076	0.0090	0.0084	0.0105	0.0092	0.0087	0.0075	0.0090	0.0083	0.0106
RMSE	0.0100	0.0089	0.0091	0.0096	0.0085	0.0125	0.0101	0.0090	0.0089	0.0097	0.0085	0.0125
Pot. Exp. Sq /100	-0.0591	-0.0573	-0.0461	-0.0588	-0.0558	-0.0620	-0.0594	-0.0577	-0.0466	-0.0591	-0.0562	-0.0616
SE (Pot. Exp Sq)	0.0241	0.0222	0.0194	0.0233	0.0216	0.0272	0.0239	0.0224	0.0192	0.0232	0.0214	0.0275
RMSE	0.0247	0.0225	0.0208	0.0239	0.0217	0.0284	0.0246	0.0227	0.0205	0.0238	0.0216	0.0286
Black (1= yes)	-0.1557	-0.1514	-0.1192	-0.1558	-0.1471	-0.1801	-0.1560	-0.1527	-0.1204	-0.1562	-0.1478	-0.1789
SE (Black)	0.0656	0.0603	0.0543	0.0636	0.0584	0.0752	0.0702	0.0635	0.0554	0.0675	0.0618	0.0803
RMSE	0.0670	0.0611	0.0589	0.0651	0.0586	0.0843	0.0716	0.0644	0.0594	0.0690	0.0621	0.0884
Constant	0.9502	1.0195	1.3577	0.9621	1.0621	0.8004	0.9380	1.0126	1.3612	0.9522	1.0567	0.7982
SE	0.1404	0.1302	0.1218	0.1352	0.1262	0.1551	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. 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 are empirical sample moments from 1,000 replications each with a sample size of 1,000.