Stacy, Brian

Preprint

Left with Bias? Quantile Regression with Measurement Error in Left Hand Side Variables

Left with Bias? Quantile Regression with Measurement Error in Left Hand Side Variables

Brian Stacy

Michigan State University

December 8, 2014

Abstract

This paper examines the effect of measurement error in the dependent variable on quantile regression, because unlike OLS regression, even classical measurement error can generate bias. I examine the pattern and size of the bias using both simulation and an empirical example. The simulations indicate that classical error can cause bias and that non-classical measurement error, particularly heteroskedastic measurement error, has the potential to produce substantial bias. Also, the size and direction of the bias depends on the amount of heterogeneity in the effects across quantiles and the regression error distribution. Using restricted access Health and Retirement Study data containing matched IRS W-2 earnings records, I examine whether estimates of the returns to education statistically differ using a precisely measured and mismeasured earnings variable. I find that returns to education are over-stated by roughly 1 percentage point at the median and 75th percentile using earnings reported by survey respondents.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education through Grant R305B090011 to Michigan State University. This grant supports MSU's doctoral training program in the economics of education. The opinions expressed are those of the author and do not represent the views of the Institute or the U.S. Department of Education. I would like to thank Steven Haider, Scott Imberman, Jeffrey Wooldridge, Mark Reckase, Christina Plerhoples, Quentin Brummet, Steven Dieterle, and Otavio Bartalotti for helpful comments. I would also like to thank the Michigan Center on the Demography of Aging for their permission and help using their restricted access HRS data.

1 Introduction

Quantile regression, which allows a researcher to examine the effects of covariates on different points of the conditional distribution of the outcome variable, is an important tool for empirical research. For instance, such methods have been used to examine the returns to schooling (Buchinsky (1994)), inter-generational earnings (Eide & Showalter (1999)), birth weight (Abrevaya & Dahl (2008)), and empirical finance (Chernozhukov & Umantsev (2001)). See Koenker & Hallock (2001) for a review.

Despite it's popularity as an empirical tool, a relatively small literature exists on the effects of measurement error on quantile regression estimation, and within this literature, most of the work has been concentrated on measurement error in independent variables.[1] Almost no research has been done on the issue of bias in quantile regression estimation caused by measurement error in the dependent variable, except for a brief discussion in a footnote in Hausman (2001) and in Chen, Hong, & Tamer (2005), who only examine the issue in the context of censored quantile regression at the median.

This lack of research is surprising because, unlike OLS, even classical measurement error in the dependent variable can cause quantile coefficient estimates to be biased.[2] Moreover, many other realistic types of measurement error, such as mean-reverting and heteroskedastic measurement error, complicate matters quickly.[3]

[1] See Angrist, Chernozhukov, & Fernandez-Val (2006) and Wei & Carroll (2009) for examples.

[2] Hausman (2001) mentions this fact and that the bias tends to be in the direction of the median coefficient estimate.

[3] Bound & Krueger (1991), Bound, Brown, Duncan, & Rodgers (1994), and Pischke (1995)

In this paper, I examine bias in the quantile regression estimator caused by measurement error in the dependent variable using simulation and an empirical example. In the simulations, I examine the cases of classical measurement error, mean-reverting measurement error, and heteroskedastic measurement error. My results confirm that the introduction of classical measurement error when the underlying error term is symmetrically distributed can bias the quantile regression estimator towards the coefficient at the median, and that the estimator at the median is largely unbiased, a finding reported in Hausman (2001).[4] However, my results further show that this finding is not generalizable to the case in which the regression error is asymmetric. In this case, the estimator at the median can be biased as well, and no clear pattern emerges for the bias at the other quantiles. The simulations also show that mean reverting and heteroskedastic measurement error can potentially cause substantial bias.

In the empirical application, I examine quantile regression estimates of the returns to education using both reported earnings from the Health and Retirement Study and matched IRS W-2 records, which I assume to be accurate. I find that estimates of the returns to education at the median and 75th percentile are overstated by around 1 percentage point (a bias of around 12-15%) using reported earnings instead of the more accurate W-2 records. These differences are statistically significant at the 5% level. For context, this bias is similar in magnitude to the upward bias caused by omitted ability in the OLS estimator that has been

have found evidence that the measurement error is mean reverting, and Hausman (2001) reports that heteroskedastic measurement error may exacerbate bias.

[4]No simulation results are presented in the paper, although the issue is raised in a footnote.

found by others.[5] Also, the pattern of the estimates suggests that the returns to education are less heterogeneous than previously thought.

2 Model and Estimator

This section will provide a brief overview of quantile regression. For more details, one can read Koenker & Bassett (1978), Koenker (2005), or Wooldridge (2010) among many other sources.

The goal of quantile regression is typically to examine the effects of covariates on different points of the conditional distribution of the outcome variable. It is common to model conditional quantiles using a model that is linear in parameters. In which case, we can express the τth conditional quantile of y_i as

$$Q_\tau(y_i|\mathbf{x_i}) = \mathbf{x_i}\boldsymbol{\beta}_0(\tau), \tag{1}$$

where $\mathbf{x_i}$ is a vector of covariates, and $\boldsymbol{\beta}_0$ is a vector of population parameters.

It can be shown that $\boldsymbol{\beta}_0(\tau)$ satisfies the condition that

$$\min_{\boldsymbol{\beta}\in\Re^K} E[(\tau - \mathbf{1}[y_i - \mathbf{x_i}\boldsymbol{\beta} < 0])(y_i - \mathbf{x_i}\boldsymbol{\beta})], \tag{2}$$

where $\mathbf{1}[\cdot]$ is the indicator function. Assuming that $\boldsymbol{\beta}_0(\tau)$ uniquely satisfies Equation (2), the parameters can be consistently estimated under some weak regularity conditions by finding values that satisfy the sample analog.

[5] Upward ability bias in the OLS estimator of the return to education is also around 10-15%, as reported in Card (1999).

In many cases, instead of observing the dependent variable y_i, the researcher observes the variable measured with error, call it Y_i. As is well-known, such measurement error causes no bias in the OLS estimator if it follows the classical assumptions.[6] Heuristically, this is the case, because the measurement error is simply absorbed into a composite error term.

This fortunate outcome is not generally the case for the quantile regression estimator for the following reason. Let $u_i = y_i - \mathbf{x_i}\boldsymbol{\beta}$ be the quantile regression error term. In the case of no measurement error, it can be shown that the first order conditions for (2) are:

$$E(\mathbf{x_i'}(\mathbf{1}[u_i < 0] - \tau)) = 0. \tag{3}$$

When measurement error in the dependent variable is introduced, the first order conditions are:

$$E(\mathbf{x_i'}(\mathbf{1}[u_i + e_i < 0] - \tau)) = 0. \tag{4}$$

Because the expected value operator does not pass through the indicator function, the first order conditions are not the same, so there is no guarantee that the parameters that solve (4) also solve (3) even under classical measurement error.[7]

[6] I define classic measurement error as measurement error that is independent of the true value of the dependent variable and the covariates.

[7] Note that in the OLS case with classical measurement error, the first order conditions with and without classical measurement error are the same. This is true since the expected value operator does pass through linear functions.

3 Simulation Evidence of Bias in Quantile Regression

Since no closed form solution exists for the quantile regression estimator, it is difficult to examine bias caused by measurement error in the dependent variable analytically. In order to study the issue further, I produce simulation evidence on how various forms of measurement error affect the quantile coefficient estimates.

My data generating processes consist of a dependent variable with a single explanatory variable. In order to generate different parameters at different quantiles, a random coefficients model is used. My baseline data generating process is meant to be a very simple model of returns to schooling and takes the following form:

$$y_i = \alpha_o + x_i\beta_0 + \gamma x_i\eta_i + \omega_i, \tag{5}$$

where η_i and ω_i are independent of one another and x_i and have a standard normal distribution. Note that throughout the discussion of the simulations, I will refer to ω_i as the regression error and e_i, which I will define below, as the measurement error. The regressor x_i, which can be thought of as years of schooling, has a binomial distribution with $n = 16$ and $p = .75$ producing a distribution with a mean of 12 and standard deviation of 3.[8] In my simulation, β is set to .075, α is

[8]These parameters were chosen to give a very basic approximation to the distribution of number of years of schooling. I have also produced simulation results where x_i has a uniform, a normal,

set to 5, and γ is set to .04. The parameter choices are meant to roughly mimic what is found in previous literature and in my HRS/W-2 data.[9]

In addition to the baseline, I also examine the performance of the estimator under a number of alternate data generating processes, which can help inform on which data characteristics can make biases larger or smaller and which findings seem to be general and which are specific to a particular data generating process. In the second data generating process examined, I make the effect of x_i negative rather than positive. In the third, I report simulation results in which the effect at the 10th percentile is the largest. In the fourth, I examine the bias caused by measurement error when the effect of x_i is much more heterogeneous. In the fifth, I examine the case where the distribution of ω_i, in equation (5), is the Student's t distribution with 3 degrees of freedom instead of the standard normal distribution. This distribution is a symmetric distribution that has a thicker tail than the standard normal distribution. In the sixth, I examine the results when the distribution of ω_i has an asymmetric, lognormal distribution. In this case, $\omega_i = exp(Z_i)$, where Z_i has a standard normal distribution. For this case, the effect at the mean and median will no longer be identical.

I examine three cases of measurement error: classical measurement error, mean-reverting measurement error, and heteroskedastic measurement error.[10] In

and a Poisson distribution. The general patterns are the same as described below.

[9]The choice of .075 is meant to be reflective of the estimates of the mean return to education found previously by other authors, which typically are in the .07-.10 range. For an overview of the mean returns to education literature, see Card (1999).

[10]Classical measurement error is a case in which the measurement error is independent of the covariates. Mean-reverting measurement error is a case where there is a negative correlation between ω_i and e_i. As discussed in Kim & Solon (2005), one way to interpret mean-reversion found

each case, the measurement error is additive with the form:

$$Y_i = y_i + e_i. \tag{6}$$

Simulations were done using Stata. 1,000 simulation repetitions were performed. Each repetition contained 10,000 simulated observations. In the tables, I report the quantile regression estimates at the .10, .25, .50, .75, and .90 quantiles. For purposes of comparison, I also report OLS estimates.

3.1 Simulation Results Under Classical Measurement Error

For the simulations with classical measurement error, I generate e_i to be independent of y_i and x_i. Also, e_i is normally distributed with a mean of zero. The variance is chosen so that the reliability ratio is

$$\frac{Var(y_i)}{Var(y_i) + Var(e_i)} = .8. \tag{7}$$

This reliability ratio is approximately the value calculated by Bound & Krueger (1991) for men's reported income in the CPS data. In addition, I examine classical measurement error when the reliability is .6 as a more extreme case.

In row (1) of Table 1a, I report the results for the baseline specification with normally distributed regression error and classical measurement error with a reli-

in the measurement error in earnings records is that when workers are asked to report their earnings for the year, the workers under report transitory earnings and shade toward their usual earnings. Heteroskedastic measurement error is a case where the variance of the measurement error depends on the covariates in x_i.

ability of .8. In column (2), the quantile regression estimator at the .10 quantile is shown to be biased towards the median coefficient in this simulation. The estimate is .056, while the true value of the parameter is .053 (a bias of roughly 6%) and the median coefficient is .075. The estimator at the .25 quantile is also biased towards the median, but to a lesser degree. The median estimator is unbiased. The estimator at the .75 quantile is slightly biased again towards the median coefficient, and the estimator at the .90 quantile is also biased toward the median coefficient, by an amount nearly symmetric with estimator at the .10 quantile. This pattern is consistent with the pattern reported in the footnote in Hausman (2001) that quantile regression estimators at the tails of the distribution are biased towards the true parameter at the median.

In row (2), I report the estimates when the reliability is .6. The estimates follow the same pattern as those in row (1), but the results show a stronger bias towards the true parameter at the median.

In row (3), I report results in which the coefficient on x_i is negative. In row (4), I report results in which the effect at the 10th percentile is the largest. In both of these cases, the finding that under classical error the estimator at the tails are biased toward the median coefficient holds.

Next, I examine the bias caused by measurement error when the effect of x_i is much more heterogeneous. In this simulation, the bias at the tails is much larger in magnitude than the baseline simulation. As shown in Table 1b, the bias at the 10th and 90th percentiles is still towards the median, but the magnitude of the bias is around .06 instead of .003 in the baseline simulation (a bias of around 14%

instead of 6%). The simulations do not prove this, but they may hint that as the effects at different quantiles become more heterogeneous, the bias becomes larger with classical measurement error. With larger differences between the effect in the tails and the effect at the median, there may be more room for bias towards the median.

In rows (6) and (7), I change the distribution of the regression error. In row (6), I report results were the distribution of ω_i, in equation (5), is the Student's t distribution with 3 degrees of freedom instead of the standard normal distribution. Despite this difference, the results look very similar to the orginal simulation in row (1). The results still display bias towards the median in the case of classical measurement error.

Finally, I examine the results when the distribution of ω_i is asymmetric with a lognormal distribution in row (7). An important thing to note is that in this case the coefficients at the tails of the distribution are not necessarily biased towards the median coefficient and there is no symmetric bias at either end of the conditional distribution. While the estimator at the 10th percentile is biased towards the median coefficient by around .014 (a bias of around 32%), the estimator at the 90th percentile is slightly biased away from the median by around .001. Additionally, the estimator at the median is biased in this case by around 1 percentage point, which was not the case with the symmetric distributions. This shows that the finding reported in Hausman (2001) are not generalizable to the case where the regression error has an asymmetric distribution.

A few key points emerge from this set of simulation results. First, under some

alternate simulation parameters and distributions, when the error term, ω_i, is symmetrically distributed, the quantile regression estimator at the tails tend to be biased towards the median coefficient when there is classical measurment error in my simulations. I conjecture that this is true generally for symmetric distributions, but the simulations do not prove this. Second, when the effects across the conditional distribution are relatively more heterogeneous using my data generating process and normally distributed errors, the bias at the tails can be larger. Third, when the error term, ω_i, is asymmetrically distributed, bias still exists and the direction is less clear when there is classical error. The estimator for the coefficient at the median may also be biased.

3.2 Simulation Results Under Mean-Reverting Measurement Error

In Table 2, I report estimates when mean-reverting measurement error is added to the dependent variable. I omit the results for the case with a negative coefficient on x_i, the case where the effect at the 10th percentile is the largest, and the case where the distribution of ω_i follows the Student's t distribution with 3 degrees of freedom, because these results closely parallel the baseline results with normally distributed regression error. In the cases presented below, the measurement error has the following form:[11]

[11] As discussed in Kim & Solon (2005), one way to interpret mean reversion in the measurement error for earnings is that when workers are asked to report their earnings for the year, the workers under report transitory earnings and shade toward their usual earnings. In my simulation, this is reflected with a negative correlation between ω_i and e_i.

$$E(e_i|\omega_i) = -.3\omega_i. \tag{8}$$

The parameters are chosen to match what is found in Bound & Krueger (1991) for measurement error in log earnings and matches what is found in my HRS data discussed below. As a more extreme case, I also examine, in row (2), mean-reverting measurement error of the form:

$$E(e_i|\omega_i) = -.45\omega_i. \tag{9}$$

In row (1) and (2), results are reported for the baseline specification with normally distributed regression error. In these cases, the estimators at the tails of the distribution are biased away from the true parameter at the median in this simulation. In row (1), the bias is -.002 at the .10 quantile (a bias of roughly 4%) and the bias is .001 at the .90 quantile. In row (2) the bias is more pronounced, with a bias of -.004 at the .10 quantile (a bias of roughly 7.5%) and a bias of .004 at the .90 quantile. The OLS estimator and the estimator at the median are unbiased by this form of mean reverting measurement error in this simulation. In row (3) the results are reported for the simulations with a more heterogeneous effect. In this case, the bias is towards the median, meaning that there appears to be no general result for mean reverting error regarding bias towards or away from the median coefficient. Also, the magnitude of the bias for the estimators at the tails of the distribution is again much larger in the case with more heterogeneous effects than for the baseline case in row (1) (a bias of around 13.2% versus 4% in the case

of the .10 quantile). Finally in row (4), I present the results for the case in which the regression error has an asymmetric lognormal distribution. Again, there is no clear pattern to the bias when the regression error is asymmetric. The estimator at the 10th percentile is biased towards the median by around .014 (a bias of 32%), and the estimator at the 90th percentile is biased away from the median by .005 (a bias of around 6%). Also, the estimator at the median is again biased by around 1 percentage point.

3.3 Simulation Results Under Heteroskedastic Measurement Error

In the cases of heteroskedastic measurement error, reported in Table 3, the measurement error has the following form:[12]

$$Var(e_i|x_i) = .25exp(-.1x_i + .01x_i^2). \tag{10}$$

The parameters are chosen to match what is found empirically in my HRS/W-2 earnings data.[13] In row (2), I again examine a more extreme case that takes the form:

[12] I again omit the case with a negative coefficient on x_i, the case where the effect at the 10th percentile is the largest, and the case where the distribution of ω_i follows the Student's t distribution with 3 degrees of freedom. These results look generally similar to the baseline results presented in row (1).

[13] More details on the data can be found in section 4.1. More details on the approach to estimating the parameters can be found in section 4.2.

$$Var(e_i|x_i) = .25exp(-.1x_i + .02x_i^2). \tag{11}$$

Heteroskedastic measurement error has the potential to produce bias at the tails that is considerably larger than previous cases. The results in row (1) show heteroskedasticity producing a bias of -.019 (a bias of 36% compared to 6% in the baseline case with classical measurement error) in the case of the estimator at the .10 quantile and a bias of .019 in the case of the .90 quantile. The estimators at the .25 and .75 quantiles are also biased, but to a lesser degree. The median estimator does not appear to be biased by heteroskedasticity in the case with the normally distributed regression error. In row (2), which are based on added measurement error with a more extreme form of heteroskedasticity, we see severe bias at the tails of the distribution. The bias at the .10 quantile is -.15 (a bias of 283%), and the bias at the .90 quantile is .186. The estimators at the .25 and .75 quantiles are also substantially biased, while the estimator at the median is largely unbiased. In row (3) with more heterogeneous effects, the bias is actually smaller in magnitude than the bias in the baseline case in row (1) of the table (a bias of with a magnitude of roughly 8.7% versus 36%). This could be a case of the strong bias towards the median exhibited with classical measurement error partially cancelling out the bias generated by the heteroskedasticity. Finally, in row (4) with the lognormal regression error, we see a bias with a magnitude of around 125% at the 10th percentile, a bias of around 2.3% at the median, and a bias with a magnitude of around 68.6% at the 90th percentile.

Overall, the simulation evidence suggests that the quantile regression estimator can be biased by classical measurement error under a variety of distributions and data generating processes. The bias can potentially be made worse when there is non-classical measurement error, particularly in the case of heteroskedastic measurement error. The size of the bias also depends on the underlying true data generating process and the distribution of the regression error. In this next section, I offer an empirical example showing bias.

4 Quantile Returns to Education as an Application

In my empirical example, I use reported earnings in data from the Health and Retirment study benchmarked against what I maintain are more reliable IRS W-2 records data. Bound & Krueger (1991) find that reported earnings in Current Population Study data contains substantial measurement error when bench-marked against more reliable Social Security earnings records data. Since quantile regression is often applied to income data, the effect of measurement error in these income variables on quantile coefficient estimates is important to understand. I am following a convention in the literature, for instance Chen et al. (2005), maintaining that the administrative earnings records are more reliable.[14]

[14]There is good reason to think that the actual dependent variable of interest is permanent income, since the income in any one year may not be an accurate reflection of the return to an additional year of education (see Haider & Solon (2006)). Constructing a measure of permanent income and examining how estimates using this measure compare to using reported annual earnings may be a topic of future research.

Buchinsky (1994) has an excellent paper examining the returns to education using quantile regression. I will closely follow the specification in that paper. The regressions are based on the familiar Mincer (1974) equation.

$$log(y_i) = \beta_0 + S_i\beta_1 + E_i\beta_2 + E_i^2\beta_3 + B_i\beta_4 + \epsilon_i \tag{12}$$

where $log(y_i)$ is the log of annual earnings, S_i is years of schooling, E_i is experience, and B_i is an indicator variable for being African American.

I will follow Buchinsky (1994) and estimate parameters for a reduced form equation that does not factor in omitted ability. In addition, I will not address the issue of measurement error in reported years of schooling. The focus of this analysis will be on measurement error in earnings.

4.1 Data

The Health and Retirement Study is a survey of over 26,000 Americans (and their spouses) over the age of 50. The purpose of the study was to examine the transition of individuals from the labor force into retirement. The study collects information on income, employment, demographics, as well as on the participants health, retirement assets, and health care expenditures.

Participants are asked to report their total wage earnings, labor force status, age, experience and education level.[15] Importantly for my analysis, many HRS

[15]In order to keep things as similar as possible to Buchinsky (1994) I use potential experience, defined as age minus years of education minus, as my measure of experience instead of years reported working.

respondents also consented to having their survey records matched with their W-2 earnings records, which allows me to match reported earnings with the respondent's W-2 records. Haider & Solon (2000) show that the respondents who consented have observable characteristics which are similar overall to the complete sample. The total wage earnings from the W-2 data comes from the box described as, 'Wages, tips, and other compensation'. Income from self employment or income contributed to 401(k) pensions is not included. Income above $250,000 is top coded.

I make a number of sample restrictions in the analysis. I use only the first wave of the study, which took place in 1992-93, since many of the workers, particularly in later waves of the survey, are not prime working age. In the 1992-93 survey, workers are surveyed about earnings in 1991. I exclude women from the analysis in order to avoid sample selection issues with female participation in the labor force. My main set of results includes all workers that have at least $2500 in self-reported and W-2 earnings in 1991 dollars. Summary statistics of the final sample are reported in Table 4.

4.2 Characteristics of Measurement Error in Log Earnings

I define the measurement error as the difference between log reported earnings and the more accurately measured log W-2 earnings.[16] In this section, I provide an

[16]To be more clear, the measurement error for observation i, e_i is constructed as:

$$e_i = log(sv\_earn_i) - log(irs\_earn_i),$$

overview of measurment error in my self reported earnings data.[17] Given that non-classical measurement error, and in particular heteroskedastic measurement error, has the potential to exacerbate bias caused by measurement error in the dependent variable, I also examine the relationship between the measurement error and the true earnings variable and covariates.

The raw summary statistics of the measurement error are reported in Table 5. The mean, standard deviation, and 10th, 25th, 50th, 75th, and 90th percentiles of the measurement error are included in the table. A kernel estimate of the density of the measurement error is included in Figure 1. The measurement error in log reported earnings has a mean close to zero and the standard deviation is .486. The measurement error also shows some rightward skewness, with the mean larger than the median.

I examine the degree of mean reversion in the measuremt error in my data by an OLS regression of the measurement error on the log of the true W-2 earnings. As discussed in Kim & Solon (2005), a coefficient of zero for the log of true earnings indicates no mean reversion in the measurement error, and a negative coefficient indicates mean reversion. Results are reported in column (1) of Table 6. The coefficient on the log true earnings variable is -.234, which is statistically significant at the 1% level, and is similar to the degree of mean reversion detected

where $log(sv\_earn_i)$ is the log of survey earnings, and $log(irs\_earn_i)$ is the log of IRS W-2 earnings.

[17] Bricker & Engelhardt (2008) also study measurement error in HRS earnings data using the HRS/IRS W-2 matched earnings records. They find evidence of a negative correlation between the measurement error and the true earnings variable. They also find a positive correlation between the measurement error and the education level of the respondent.

in Bound & Krueger (1991), Bound et al. (1994), and Pischke (1995).

Next, I examine the relationship between the measurement error and the covariates. I assume the following functional form for the conditional expectation and variance:

$$E(e_i|S_i, E_i, B_i) = \gamma_0 + S_i\gamma_1 + E_i\gamma_2 + E_i^2\gamma_3 + B_i\gamma_4, \tag{13}$$

$$Var(e_i|S_i, E_i, B_i) = \sigma^2 exp(S_i\delta_0 + S_i^2\delta_1 + E_i\delta_2 + E_i^2\delta_3 + B_i\delta_4). \tag{14}$$

I estimate the parameters in Equation (13) by an OLS regression of e_i on years of education, experience, experience squared, and the indicator for being black. I estimate the parameters in (14), which will tell us whether the measurement error is conditionally heteroskedastic, by non-linear least squares of the squared residuals, wich come from the OLS regression to estimate Equation (13), on the same covariates.[18]

The results for the conditional mean are reported in column (2) of Table 6. The estimated coefficients are insignificant when experience, experience squared, and the indicator for being black are included in column (3). Overall, the estimates suggest a small or negligible effect of the covariates on the conditional mean of

[18]This produces consistent estimates of the parameters in the conditional variance, because $Var(e_i|S_i, E_i, B_i) = E(v_i^2|S_i, E_i, B_i)$ by definition, where $v_i = e_i - E(e_i|S_i, E_i, B_i)$, and because the OLS residuals converge in distribution to v_i, as noted in Harvey (1976).

the measurement error.[19]

The estimates of the coefficients in Equation (14) are reported in column (3). The coefficient on education squared is statistically significant at the 5% level, suggesting that the measurement error is conditionally heteroskedastic. The point estimates suggest the degree of heteroskedasticity with respect to education in the measurement error is more similar to the baseline heteroskedastic measurement error examined in row (1) of Table 3 for the simulation rather than the more extreme case examined in row (2). The other estimated coefficients, such as for experience, are not statistically significant. If the degree of heteroskedasticity is different with respect to say experience than it is for education, then we may see diffferences in the degree of bias produced by the measurement error.

Overall, the measurement error displays mean-reversion and heteroskedasticity. The heteroskedasticity is particularly a cause for concern, because it had such a strong effect in the simulations. In the next section, I report estimates of the returns to education and experience using both the log of reported earnings as the dependent variable and the log of true earnings using W-2 earnings records and test for differences.

4.3 Estimates of the Returns to Education and Experience

In order to test whether estimates based on IRS W-2 records statistically differ from estimates based on the reported earnings, I perform the following procedure:

[19] These results are consistent with Bound & Krueger (1991), who do a similar analysis in Table 3 of their paper.

1. Estimate the Mincer equation in (12) using reported earnings and again using the (true) W-2 records at the .1, .25, .50, .75, and .9 quantiles.

2. Form the difference between the estimates using the W-2 records and the estimate using reported earnings for each quantile.

3. Repeat the procedure 1000 times sampling with replacement to produce bootstrapped standard errors for the differences between the estimates using reported earnings and (true) W-2 earnings records.

In Table 7, I show estimates of the returns to education and experience using true W-2 earnings in row (1) and estimates using the reported earnings records in row (2).[20] Stars in row (2) signifiy that the difference between the estimates using W-2 earnings and reported earnings are statistically different from zero. For comparison, the first column shows estimates for the mean from an OLS regression of log annual earnings on years of education, experience, experience squared, and an indicator for whether the respondent is black. Columns (2) through (6) show estimates of the quantile coefficients for the .10, .25, .50, .75, and .90 quantiles.

The return to a year of education at the 10th conditional percentile is estimated to be .046 using log reported earnings and .056 using log W-2 earnings. The return at the 25th percentile is .078 using log reported earnings and .079 for true earnings. However, neither of these differences are statistically significant. At the 50th percentile, the estimated return is .086 for reported earnings and .075 for

[20] I also have examined the returns using only workers with more than 7 years of education and also only workers who report working full time. The patterns are very similar to those reported below.

true earnings. This difference of .0113 is statistically significant at the 5% level. Interestingly, this difference is very similar to the difference found by Chen et al. (2005), who find that using the mismeasured earnings variable biases the censored quantile regression estimate of the return to education at the median by around .014.[21] The estimate at the 75th percentile is .083 and .074 for true earnings, and this difference is also statistically significant at the 5% level. The estimates at the 90th percentile are .090 for reported and .083 for true earnings, but this difference is not statistically significant.

In the lower two panels, I show returns to experience. Since the Mincer equation in (12) includes a quadratic in experience, the return depends on the level of experience of the individual. I report the returns at 10 years of experience in the middle panel and 25 years of experience in the lower panel of Table 7. Overall, the estimates of the return to experience tend to be low compared to estimates found in the literature. This may be because the Health and Retirement study participants are older, with an average age around 55. At this age, experience may have only a small return. The mean return to experience estimated using OLS and reported earnings is statistically different from the return estimated using true earnings at the 5% level. However, none of the quantile regression estimates statistically differ using the different earnings measures.

[21] The authors use the 1978 CPS-SSR match file, which combines reported earnings with social security earnings records. The authors do not report estimates at quantiles other than the median. They also do not report uncensored quantile regression results because of severe top coding in the social security earnings records.

4.4 Discussion

To summarize, I find strong evidence that the estimators of the return to education at the median and 75th percentiles are biased by measurement error in the dependent variable. The point estimates suggest the estimator at the median is overstated by 1.13 percentage points and that the estimator of the effect at the 75th percentile is overstated by .91 percentage points. To put this into context, the roughly 12-15% bias produced is similar in magnitude to the upward bias caused by omitted ability in the OLS estimator of the return to education.[22] The fact that I find the estimator at the median to be biased may suggest that the conditional distribution of true log earnings or the measurement error is asymmetric, given my simulation results. While the differences are not statistically significant at the other quantiles, the point estimates of the bias are also important in magnitude. The point estimate at the 90th percentile suggest a bias with a magnitude of around 8.4%, and the point estimate at the 10th percentile suggest a bias with a magnitude of around 14.8%. Finally, when looking at the point estimates, the returns to education appear less heterogeneous using the true earnings variable compared to using the mismeasured earnings.[23]

[22] Upward ability bias in the OLS estimator of the return to education is reported to be around 10-15% by Card (1999).

[23] This may have implications for explanations of the increase in income inequality over the past few decades. Machado & Mata (2005) cite heterogeneity in returns to education combined with a trend of higher educational attainment as an important factor explaining increases in income inequality. The argument is as follows. Previous research, for instance Buchinsky (1994) and Machado & Mata (2005), has found that there is a higher return to education at the 90th percentile than the 10th percentile. This implies that the distribution of earnings has a higher dispersion at higher levels of education, because the right tail increases faster than the left tail as education increases. Now because education levels have generally increases over time, Machado & Mata (2005) cite this as one reason why income inequality has increased.

The estimates of the effects of experience do not statistically differ. One potential explanation for not detecting bias for experience is that the quantile effects of experience are estimated imprecisely. This is true because there is not much variation in experience in my sample. The lack of precision in the estimates may explain the lack of statistically significant differences. Another possibility is that the measurement error is less heteroskedastic with respect to experience than years of education, something that is hinted at in section 4.2. The smaller degree of heteroskedasticity with respect to experience may result in less bias.

5 Conclusions

This paper makes several important contributions. I add to a small literature on how quantile regression estimates are affected by measurement error in the dependent variable. I show in my simulations that even under classical measurement error the quantile regression estimator may be biased by measurement error in the dependent variable. If one assumes classical measurement error and that the conditional distribution of the true dependent variable is symmetrically distributed, then the simulation evidence suggests that the estimator at the tails of the distribution may be biased towards the median coefficient, although a rigorous proof of this could be a useful topic of future research. However, this result does not hold for the case where the regression error is asymmetric. In this case, the median is also biased and the bias at the other quantiles is not always towards the median coefficient. Also, the size of the bias depends on the amount of hetero-

geneity in the effects across the distribution and the amount of mean reversion and heteroskedasticity in the measurement error.

Empirically, I show that quantile regression estimator of the returns to education may be biased by measurement error in log reported earnings when compared to the more accurate W-2 earnings records. I find evidence that returns to education estimated at the median and 75th percentile are moderately over stated using reported earnings.

This paper can serve as a cautionary note to researchers using quantile regression techniques with possible mismeasured dependent variables. A bright side is that even though the estimates appear biased, the bias is not overwhelmingly large in the context of the returns to education. The largest bias seen is around 1.13 percentage points. In other contexts, however, the bias could be larger. Future research in other contexts could be useful. Also, finding a solution to the problem may be another important topic for future research.

References

Abrevaya, Jason, & Dahl, Christian M. 2008. The effects of birth inputs on birthweight: evidence from quantile estimation on panel data. *Journal of Business & Economic Statistics*, *26*(4): 379–397.

Angrist, Joshua, Chernozhukov, Victor, & Fernandez-Val, Ivan. 2006. Quantile regression under mispecification, with an application to the u.s. wage structure. *Econometrica*, *74*(2): 539–563.

Bound, John, Brown, Charles, Duncan, Greg J, & Rodgers, Willard L. 1994. Evidence on the validity of cross-sectional and longitudinal labor market data. *Journal of Labor Economics*, *12*(3): 345–368.

Bound, John, & Krueger, Alan B.. 1991. The extent of measurement error in longitudinal earnings data: Do two wrongs make a right? *Journal of Labor Economics*, *9*(1): 1–24.

Bricker, Jesse, & Engelhardt, Gary V. 2008. Measurement error in earnings data in the health and retirement study. *Journal of Economic and Social Measurement*, *33*(1): 39–61.

Buchinsky, Moshe. 1994. Changes in the u.s. wage structure 1963-1987: Application of quantile regression. *Econometrica*, *62*(2): 405–458.

Buchinsky, Moshe. 1998. Recent advances in quantile regression models: A practical guideline for empirical research. *The Journal of Human Resources*, *33*(1): 88–126.

Card, David. 1999. The causal effect of education on earnings. *Handbook of labor economics*, *3* 1801–1863.

Chen, Xiaohong, Hong, Han, & Tamer, Elie. 2005. Measurement error models with auxiliary data. *The Review of Economic Studies*, *72*(2): 343–366.

Chernozhukov, Victor, & Umantsev, Len. 2001. Conditional value-at-risk: Aspects of modeling and estimation. *Empirical Economics*, *26*(1): 271–292.

Eide, Eric R, & Showalter, Mark H. 1999. Factors affecting the transmission of earnings across generations: A quantile regression approach. *Journal of Human Resources*, (pp. 253–267).

Haider, Steven, & Solon, Gary. 2000. Non random selection in the hrs social security earnings sample. *RAND, Labor and Population Program Working Paper Series*, (No 00-01):. Http://www.rand.org/pubs/drafts/DRU2254.

Haider, Steven, & Solon, Gary. 2006. Life-cycle variation in the association between current and lifetime earnings. *The American Economic Review*, *96*(4): 1308–1320.

Harvey, Andrew C. 1976. Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, (pp. 461–465).

Hausman, Jerry. 2001. Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *The Journal of Economic Perspectives*, *15*(4): 57–67.

Kim, Bonggeun, & Solon, Gary. 2005. Implications of mean-reverting measurement error for longitudinal studies of wages and employment. *Review of Economics and Statistics*, *87*(1): 193–196.

Koenker, Roger. 2005. *Quantile Regression*. New York: Cambridge University Press.

Koenker, Roger, & Bassett, Gilbert. 1978. Regression quantiles. *Econometrica*, *46*(1): 33–50.

Koenker, Roger, & Hallock, Kevin F.. 2001. Quantile regression. *The Journal of Economic Perspectives*, *15*(4): 143–156.

Machado, José AF, & Mata, José. 2005. Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of applied Econometrics*, *20*(4): 445–465.

Mincer, Jacob A. 1974. Schooling and earnings. In *Schooling, experience, and earnings*, (pp. 41–63). Columbia University Press.

Pischke, Jörn-Steffen. 1995. Measurement error and earnings dynamics: Some estimates from the psid validation study. *Journal of Business & Economic Statistics*, *13*(3): 305–314.

Wei, Ying, & Carroll, Raymond J. 2009. Quantile regression with measurement error. *Journal of the American Statistical Association*, *104*(487): 1129–1143.

Wooldridge, Jeffrey M.. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2nd ed.

Tables and Figures

Table 1a: Simulation Results for OLS/Quantile Regression Estimates with Classical Measurement Error in Dependent Variable.

	OLS	.10	.25	.50	.75	.90
Baseline Spec: Normally Distributed Regression Error						
True Parameter	.075	.053	.063	.075	.087	.097
(1) Estimator w/ Rel .8	.075 (.00023)	.056 (.00037)	.065 (.00031)	.075 (.00028)	.085 (.00032)	.095 (.00039)
(2) Estimator w/ Rel .6	.075 (.00027)	.058 (.00044)	.066 (.00035)	.075 (.00033)	.084 (.00037)	.092 (.00046)
Negative Effect: Normal Regression Error						
True Parameter	-.075	-.097	-.087	-.075	-.064	-.053
(3) Estimator w/ Rel .8	-.075 (.00023)	-.094 (.00037)	-.085 (.00031)	-.075 (.00028)	-.065 (.00032)	-.055 (.00039)
Largest Effect at .10 Quantile: Normal Regression Error						
True Parameter	.075	.088	.082	.075	.068	.061
(4) Estimator w/ Rel .8	.075 (.00022)	.087 (.00036)	.082 (.0003)	.075 (.00027)	.069 (.0003)	.063 (.00037)

The first row of each panel reports the true value of the parameter. The second row reports the estimated coefficient. Rel .8 refers classical measurement error with a reliability ratio of .8. Rel .6 refers classical measurement error with a reliability ratio of .6. The lognormal distribution in row (7) is such that $log(\omega_i)$ has standard normal distribution. Standard errors for the mean across the 1000 reps in parenthesis.

Table 1b: Simulation Results for OLS/Quantile Regression Estimates with Classical Measurement Error in Dependent Variable.

	OLS	.10	.25	.50	.75	.90
More Heterogeneous Effects: Normal Regression Error						
True Parameter	.077	-.424	-.187	.077	.341	.576
(5) Estimator w/ Rel .8	.076 (.001)	-.366 (.00164)	-.156 (.00132)	.077 (.00125)	.31 (.00136)	.516 (.00169)
Baseline Spec: Student's t w/ 3 d.f. Regression Error						
True Parameter	.075	.055	.062	.075	.088	.094
(6) Estimator w/ Rel .8	.074 (.00037)	.06 (.00059)	.066 (.00041)	.075 (.00035)	.084 (.00042)	.09 (.0006)
Baseline Spec: Lognormal Regression Error						
True Parameter	.075	.044	.067	.087	.091	.086
(7) Estimator w/ Rel .8	.075 (.00045)	.058 (.00043)	.068 (.00035)	.078 (.00038)	.086 (.00053)	.087 (.00106)

The first row of each panel reports the true value of the parameter. The second row reports the estimated coefficient. Rel .8 refers classical measurement error with a reliability ratio of .8. Rel .6 refers classical measurement error with a reliability ratio of .6. The lognormal distribution in row (7) is such that $log(\omega_i)$ has standard normal distribution. Standard errors for the mean across the 1000 reps in parenthesis.

Table 2: Simulation Results for OLS/Quantile Regression Estimates with Mean-Reverting Measurement Error in Dependent Variable.

	OLS	.10	.25	.50	.75	.90
Baseline Spec: Normally Distributed Regression Error						
True Parameter	.075	.053	.063	.075	.087	.097
(1) Estimator w/ Mean Reverting	.075 (.00018)	.051 (.00031)	.062 (.00025)	.075 (.00022)	.087 (.00026)	.098 (.00031)
(2) Estimator w/ Larger Mean Reverting	.075 (.00016)	.049 (.00028)	.061 (.00022)	.075 (.0002)	.089 (.00023)	.101 (.00029)
More Heterogeneous Effects: Normal Regression Error						
True Parameter	.077	-.424	-.187	.077	.341	.576
(3) Estimator w/ Mean Reverting	.077 (.00098)	-.368 (.00162)	-.159 (.00126)	.078 (.00119)	.312 (.00133)	.521 (.0016)
Baseline Spec: Lognormal Regression Error						
True Parameter	.075	.044	.067	.087	.091	.086
(4) Estimator w/ Mean Reverting	.075 (.00034)	.058 (.00042)	.067 (.00034)	.077 (.00034)	.086 (.00042)	.091 (.00072)

The first row of each panel reports the true value of the parameter. The second row reports the estimated coefficient. Mean-Reverting refers to mean-reverting measurement error with the following form: $E(e_i|\omega_i) = -.3\omega_i$. Larger Mean-Reverting refers to measurement error with the following form: $E(e_i|\omega_i) = -.45\omega_i$. The lognormal distribution in row (4) is such that $log(\omega_i)$ has standard normal distribution. Standard errors for the mean across the 1000 reps in parenthesis.

Table 3: Simulation Results for OLS/Quantile Regression Estimates with Heteroskedastic Measurement Error in Dependent Variable.

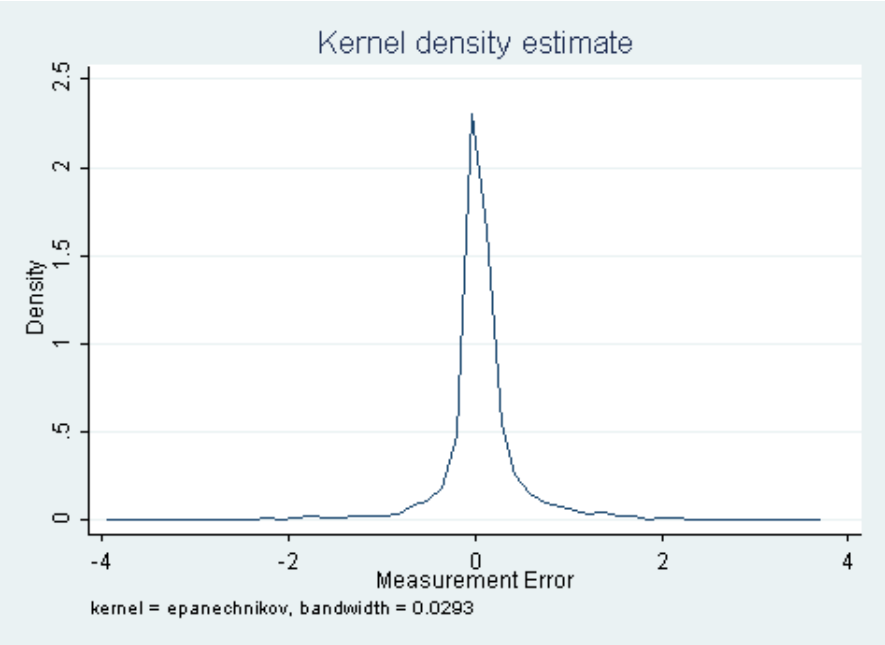
	OLS	.10	.25	.50	.75	.90
Baseline Spec: Normally Distributed Regression Error						
True Parameter	.075	.053	.063	.075	.087	.097
(1) Estimator w/ Heteroskedastic	.075 (.00023)	.034 (.00038)	.053 (.00031)	.075 (.00028)	.097 (.00033)	.116 (.0004)
(2) Estimator w/ Larger Heteroskedastic	.075 (.00034)	-.133 (.00049)	-.035 (.00041)	.075 (.00039)	.184 (.00042)	.282 (.00052)
More Heterogeneous Effects: Normal Regression Error						
True Parameter	.077	-.424	-.187	.077	.341	.576
(3) Estimator w/ Heteroskedastic	.077 (.001)	-.461 (.00159)	-.205 (.00125)	.078 (.00119)	.36 (.00128)	.614 (.00162)
Baseline Spec: Lognormal Regression Error						
True Parameter	.075	.044	.067	.087	.091	.086
(4) Estimator w/ Heteroskedastic	.075 (.00045)	-.011 (.00044)	.037 (.00038)	.089 (.00039)	.134 (.00053)	.145 (.00106)

The first row of each panel reports the true value of the parameter. The second row reports the estimated coefficient. Heteroskedastic refers to heteroskedastic measurement error with the following form: $Var(e_i|x_i) = .25exp(-.1x_i + .01x_i^2)$. Larger Heteroskedastic refers to measurement error with the following form: $Var(e_i|x_i) = .25exp(-.1x_i + .02x_i^2)$. The lognormal distribution in row (4) is such that $log(\omega_i)$ has standard normal distribution. Standard errors for the mean across the 1000 reps in parenthesis.

Table 4: Summary Statistics, Wave 1 (1992) Male Workers with Positive Earnings

Variable	Mean	Std. Dev.	Min.	Max.
Total Reported Annual Earnings	36052.17	28173.51	2800	410000
Total Annual Earnings W-2	33157.61	25492.41	2600	245000
Hours Worked/Week Main Job	43.69	10.52	1	95
Weeks Worked/Year Main Job	50.43	5.47	1	52
Hourly Wage Rate	27.74	486.15	0.96	24000
Years of Tenure Current Job	15.46	11.78	0	55.8
Total Years Worked	37.46	5.92	3	65
Total Years of Education	12.7	3.29	0	17
Age	55.87	4.61	23	77
Black	0.129	0.335	0	1
Hispanic	.091	0.288	0	1
Number of Observations				2975

Figure 1: Kernel Estimate of the Density of Measurement Error in Log Earnings



Measurement error defined as difference between log reported earnings and log W-2 earnings.

Table 5: Measurement Error Descriptive Statistics

			Quantiles				
Variable	**Mean**	**Std. Dev.**	**.10**	**.25**	**.50**	**.75**	**.90**
Measurement Error	.060	.486	-.268	-.051	.032	.166	.443
Number of Observations						2975	

Measurement error defined as difference between log reported earnings and log W-2 earnings.

Table 6: Estimates of Conditional Distribution of Measurement Error

	Mean		**Variance**
VARIABLES	(1)	(2)	(3)
Log W-2 Earnings	-.234***		
	(.018)		
Education		.001	-.103
		(.004)	(.078)
Education Squared			.008**
			(.004)
Experience		-.019	-.079
		(.014)	(.076)
Experience Squared		.0002	.001
		(.0002)	(.001)
Black		-.019	.069
		(.026)	(.190)
Observations	2,975	2,975	2,975

Estimates in column (1) come from OLS regression of measurement error on log true earnings. Estimates in column (2) come from an OLS regression of the measurement error on the covariates. Estimates in column (3) come from an NLS regression of the squared residuals from the OLS regression in column (3) on the covariates. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 7: Estimates of Mincer Equation: Male Workers with Positive Earnings

	OLS	.10	.25	.50	.75	.90
Returns to Education						
Using (True) W-2 Earnings	.073	.054	.079	.075	.074	.083
	(.005)	(.014)	(.007)	(.005)	(.005)	(.006)
Using Reported Earnings	.075	.046	.078	.086**	.083**	.090
	(.005)	(.014)	(.006)	(.006)	(.005)	(.007)
Returns to Experience at 10 Years						
Using (True) W-2 Earnings	.016	.065	.016	.002	.003	.001
	(.011)	(.041)	(.016)	(.013)	(.012)	(.019)
Using Reported Earnings	.002*	.032	.012	-.003	-.007	.001
	(.011)	(.023)	(.021)	(.013)	(.010)	(.026)
Returns to Experience at 25 Years						
Using (True) W-2 Earnings	.004	.020	.002	-.002	-.000	.002
	(.005)	(.022)	(.008)	(.006)	(.001)	(.001)
Using Reported Earnings	-.005**	-.000	-.002	-.005	-.004	.001
	(.006)	(.011)	(.010)	(.006)	(.005)	(.012)
Number of Observations				2975		

All regressions include years of education, experience, experience squared, and an indicator for whether black. All workers have at least $2500 in reported and W-2 earnings in 1991 dollars. Bootstrapped standard errors in parenthesis. 1000 bootstrap replications performed. *** Difference between estimates using W-2 and reported earnings statistically significant at 1% level. ** Difference statistically significant at 5% level. * Difference statistically significant at 10% level