DRAFT, January 3, 2001
Please do not quote without permission
Forthcoming: Journal of Economic Perspectives

Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left

Jerry Hausman[1]

The effect of mismeasured variables in statistical and econometric analysis is one of the oldest known problems, dating from the 1880's in Adcock (1887). In the most straightforward regression analysis with a single regressor variable under classical mismeasurement assumptions the least squares estimate is downward biased in magnitude towards zero.[2] At MIT I have called this the "Iron law of econometrics"—the magnitude of the estimate is usually smaller than expected. While a mismeasured right hand side variable creates this problem, a mismeasured left hand side variable under classical assumptions does not lead to bias. The only result is less precision in the estimated coefficient and a lower t-statistic.

In this paper I consider three recent developments for mismeasurement econometric models that may be less familiar to readers. The typical solution to the right hand side mismeasurement problem is to use instrumental variables to achieve a consistent estimate.[3] However, in the last decade a literature has arisen regarding the use of "weak instruments" that can result in significant finite sample bias.[4] Thus, reliance on an instrumental variable estimator to solve themismeasurement problem may be misplaced in a particular application. Here, I discuss a specification test of Hahn and Hausman (1999) that permits a test of the weak instrument hypothesis.

Next, for non-linear models with mismeasured variables, application of instrumental variables leads to inconsistent results. At MIT I call this the "forbidden regression" where improper use of instrumental variables leads to inconsistency. A paper

---

[1] Department of Economics, MIT, jhausman@mit.edu. This paper is given in memory of Zvi Griliches. Jason Abrevaya provided helpful comments.
[2] Surveys of these classical results are found in Aigner, Hsiao, Kapteyn and Wansbeek, T. (1984), Fuller (1987), and Hausman, Newey, and Powell (1995). Henceforth, I will assume that the "true" parameter is positive so that I will not repeat the "in magnitude" qualifier.
[3] Of course, a solution may not exist if suitable instruments cannot be found.
[4] See e.g. Bound, Jaeger, and Baker (1997).

by Hausman, Ichimura, Newey and Powell (HINP) (1991) proposes a consistent estimator for the polynomial regression specification in the presence of mismeasurement. I discuss that estimator here and a recent extension of the approach to general non-linear specifications by Schennach (2000). Thus, consistent estimators now are known for mismeasured non-linear regression models.

Lastly, I return to mismeasured left hand side variables. Many limited dependent variable models in econometrics are inconsistently estimated if the left hand side variables have measurement error.[5] I discuss probit/logit type models of binary outcomes or discrete choice and demonstrate that inconsistency occurs. I also discuss consistent estimators of these models with measurement errors and also specification tests. Interestingly, these models do not require instrumental variables for consistent estimation so long as a monotonicity condition is satisfied. I next consider duration models and demonstrate that under measurement errors on the left hand side variables, inconsistency will again result for general models. But so long as a stochastic dominance condition is satisfied, these models can again be estimated consistently without the use of an instrument. The last model I consider with a mismeasured left hand side variable is the quantile regression (QR) model. Again inconsistent estimates result, but I do not currently have a solution to this problem.

I. Linear Models With Mismeasured Variables

I begin with the classic linear regression mismeasurement model. I assume a linear specification after means of the variables have been removed with the right hand side variable mismeasured with uncorrelated measurement error where z_i is the true unobserved regressor and x_i is the observed variable that contains measurement error:

$$\begin{aligned} y_i &= \beta z_i + e_i = \beta x_i + e_i - \beta h_i = \beta x_i + t_i \qquad i = 1,\ldots,n \\ x_i &= z_i + h_i \end{aligned} \tag{1}$$

[5] All the models discussed here will also be inconsistently estimated if the right hand side variable is mismeasured. However, I do not discuss that topic specifically in the paper since the specifications can be considered as nonlinear specifications.

I will assume that the sample is independent and identically distributed (i.i.d.) and that z_i is the unobserved true variable and that the error in observation h_i is mean zero and is uncorrelated with z_i and with e_i.[6] Other right hand side variables, assumed to be measured without error, have been "partialled out" of the model. I will assume that $b > 0$.

The classic result is that usually the least squares (OLS) estimator $b < b$.[7] This result causes the "iron law of econometrics" that I suggested above, and it is also called "attenuation" in the statistics literature. In terms of a large sample result, $p\lim b = ab < b$ where $a = s^2_z / s^2_x = s^2_z / (s^2_z + s^2_h) < 1$. Thus, the amount of large sample bias depends on the ratio of the variance of the "signal" (true variable) to the sum of the variance of the signal and the variance of the "noise" (error in measurement).[8]

Another well-known classic result arises if the left hand side and right hand side variables are interchanged in the regression specification. The inverse of the OLS estimator of the coefficient in the reverse regression, g, gives an upward biased estimate of the true coefficient b. In large samples we have the bounds, $p\lim b < b < p\lim g$. A less well known result tells the width of the bounds since $b / g = R^2$ where the R^2 arises from the regression equation. Thus, for cross section econometrics where R^2 is often about 0.3 the bounds can be quite wide while in a times series application if the R^2 is quite high, mismeasurement will not have a large effect on the least squares estimate.[9]

As a last result, note that if the left hand side variable is measured with error but not the right hand side variable, the OLS estimator b would be unbiased under the usual Gauss-Markov assumptions.[10] Indeed, the new measurement error term, w_i, assumed to

---

[6] The i.i.d. assumption will be maintained throughout the paper.

[7] This result does not hold in general if the measurement error is correlated with z or with e, although the downward bias still often occurs. A straightforward calculation demonstrates the effect of correlated measurement error on the estimator b.

[8] For the effect on the estimated coefficients of the other variables in the regression specification that have been "partialled out" see Meijer and Wansbeek (2000).

[9] In times series, the appropriate R^2 is after partialling out other right hand side variables, which can reduce the original R^2 by quite a lot. Also, note that this result demonstrates that the times series version of the "permanent income hypothesis", in its most simple form of equation (1) with z and x permanent and measured income, does not lead to a significantly greater measured MPC because the R^2 in time series data is quite high.

[10] By Gauss-Markov assumptions I mean the assumptions that lead to OLS being the best linear unbiased estimator.

be uncorrelated with e_i would not be separately identified because the measured variable $q_i = y_i + w_i$ would be equivalent to replacing y_i by q_i in the original regression specification and denoting the residual as $n_i = e_i + w_i$. The only result would be reduced precision in the estimate b, a lower t-statistic, and a reduced R^2. However, if both left and right hand side variables are mismeasured with errors of measurement uncorrelated with each other and with e_i, the b remains the same as above but g increases because the R^2 has decreased since n_i has replaced e_i and $s^2_n = s^2_e + s^2_w$. Thus, the bounds for the true coefficient increase when measurement error occurs in both the left and right hand side variables.

Most solutions to the mismeasurement problem for the linear specification in econometrics depend on the use of instrumental variables.[11] Instrumental variables are assumed to be correlated with the true z_i but uncorrelated with either e_i and h_i. Let the vector of instrumental variables be w_i and the instrumental variable (IV) estimator will use a linear combination of the w_i to achieve a consistent estimator b_{IV} of the true coefficient b. In the usual case under Gauss-Markov assumptions 2SLS will give the efficient instrumental variable estimator, while in more general situations of conditional heteroscedasticity the White (1984) efficient IV estimator should be used.[12]

However, a significant understanding has emerged over the past few years that IV estimation of the errors-in-variables model can lead to problems of inference in the situation of "weak instruments," which can arise when the instruments do not have a high degree of explanatory power for the mismeasured variable(s) or when the number of instruments becomes large. The situation of "weak instruments" causes conventional (first order) asymptotictheory to provide a poor guide to finite sample inference. These problems of inference in the weak instrument situation can arise when conventional (first order) asymptotic inference techniques are used. In particular, conventional first order

[11] Of course, it has long been known that consistent estimators exist in the non-normal case, which do not require instrumental variables; see e.g. Aigner et. al. (1984). However, these estimators have not been used very much in econometrics.

[12] Henceforth, I will confine the choice of estimators and tests to the Gauss-Markov setup of no heteroscedasticity or serial correlation.

asymptotics can lead to a lack of indication of a problem even though significant (large sample) bias is present because estimated standard errors are not very accurate.

Hahn-Hausman (HH 1999) take a new approach to identifying the potential problem and use higher order asymptotic distribution theory to determine if the conventional first order IV asymptotics are reliable in a particular situation. HH demonstrate that the finite sample bias depends on 3 factors: (1) bias is a monotonically increasing function of $b^2 s^2_h$ the variance of the measurement error multiplied by the true coefficient; (2) bias is a monotonically increasing function of K the number of instrumental variables; and (3) bias is a monotonically decreasing function of the R^2 of the 'reduced form" specification of x_i regressed on the instrumental variables w_i. The new specification test takes the general approach of Hausman (1978) and estimates the same parameter(s) in two different ways. In particular, we compare the difference of the forward (conventional) 2SLS estimator of the coefficient of the right hand side mismeasured variable with the reverse regression 2SLS estimator of the same unknown parameter when the normalization is changed.[13] Under the null hypothesis that conventional first order asymptotics provides a reliable guide, the two estimates should be very similar. Indeed, they have unitary correlation according to first order asymptotic distribution theory. If the IV estimator is working well, the forward and reverse estimators take the form under conventional (first order) asymptotics, $p\lim\sqrt{n}\cdot(b_{IV} - g_{IV}) = 0,$ so that the forward and reverse estimators for b should be quite close. However, if they are not close where the measure of closeness is determined by second order asymptotics in HH, then problems of inferences are likely to be present.

When second order asymptotic distribution theory is used, the two estimators will differ due to second order bias terms. The HH test subtracts off these bias terms and then sees whether the resulting difference in the two estimates satisfies the results of second order asymptotic theory. If it does and the second order bias term is small, the HH test does not reject the use of first order asymptotic theory. If the new specification test rejects HH then consider estimation of the equation by second-order unbiased estimators of the type first proposed by Nagar (1959). A second specification test is then used to see

[13] Thus, 2SLS is applied to the forward and reverse regression, which was discussed above.

if these second-order unbiased estimator provides a suitable basis for inference. The HH specification test may be quite useful in cross section situations with a mismeasured right had side variable because of the significant finite sample bias that often exists along with instruments that do not have high explanatory power for the mismeasured variable.

II. Non-Linear Models With Mismeasured Variables

The linear model with measurement error is equivalent to a linear simultaneous equation model, so that two stage least squares or a closely related estimator is used. However, this relationship no longer holds in the nonlinear regression framework as noted by Y. Amemiya (1985). The reason that 2SLS no longer leads to a consistent estimator in the nonlinear errors in variables problem is because the error of measurement is no longer additively separable from the true variable in the nonlinear regression model. Application of an IV estimator such as 2SLS or nonlinear 2SLS (N2SLS) leads to inconsistent estimates. A straightforward way in which to see why 2SLS or N2SLS does not yield consistent estimators in the nonlinear errors in variable model is to consider the linear in parameters and nonlinear in variables specification for equation (1) where z_i is replaced by the nonlinear function $g(z_i)$, which is assumed to be a sufficiently smooth function to do Taylor approximations. Replacing the unobservable z_i with the observed variable x_i leads to higher order terms of the Taylor expansion which enter the regression error term but contain the true variable z_i. Thus, as demonstrated by Hausman-Newey-Powell (HNP, 1995), IV estimation will be inconsistent because the instruments will necessarily be correlated with both the mismeasured right hand side and the error term in the econometric specification.

Hausman-Ichimura-Newey-Powell (HINP, 1991) demonstrate how to solve the problem and achieve consistent estimates in the case of a polynomial specification:

$$y_i = \sum_{j=0}^{K} b_j (z_i)^j + e_i \qquad i = 1,\ldots,n \qquad (2)$$

where the unobserved variable z_i is replaced by x_i, which is observed with measurement error h_i as in equation (1). HINP develop a consistent estimator under either of two sets of assumptions. The first is a "repeated measurement" specification where $w_i = z_i + v_i$ and v_i is independent of z_i, where the stronger (than no correlation) assumption is required because of the nonlinear specification. The second specification is the instrumental variables specification $z_i = w_i d + v_i$ where v_i is independent of q_i so that $x_i = w_i d + v_i + h_i$. The unknown parameter vector d is estimated using OLS.

The HINP (1991) approach uses estimates of higher order moments to derive consistent estimators. The usual IV approach multiplies through equation (1) by $w_i d$, where d is the estimate of d. Then taking probability limits allows a solution of the normal equations for b. In the polynomial specification HINP use higher order moments of the indicator variable $w^p{}_i$ or the instrument variable $(w_i d)^p$ to multiply both sides of the equation. HINP(1991) derive recursive relationships, which permit calculation of the estimates of the elements of the unknown parameter vector and solution for the vector b which is a consistent estimate of the unknown parameters b_j in equation (2). HINP also derive asymptotic variance estimators for the b vector.

HNP(1995) use this approach to estimate Engel curves for family expenditure in the U.S. Estimation of Engel curves has long been an area of interest among econometricians where the "Leser (1963)-Working" form of Engel curve in which budget shares are regressed on the log of income or expenditure has been widely adopted in recent research. However, in a notable paper Gorman (1981) considered Engel curves in which either expenditure or budget shares are specified as polynomials in functions of expenditure, e.g. log of expenditure. Given an "exactly aggregable function", Gorman demonstrates that the rank of the matrix of coefficients for the polynomial terms in income is at most three. HNP investigate Engel curve specifications of the Gorman form and provide tests of his rank three restriction.

Few studies of Engel curves have used estimators other than least squares or nonlinear least squares. HNP investigate two alternative sets of instruments: expenditure in other periods which follows from a life cycle model approach to consumption or

determinants of income and expenditure such as education and age. HNP use the 1982 U.S. Consumer Expenditure Survey (CES). The CES collects data from families over four quarters so that HNP can apply the repeated measurement technique. HNP use budget share and total expenditure for each family from 1982:1 and for the repeated measurement total expenditure from 1982:2. They estimate Engel curves on five commodity groups: food, clothing, recreation, health care, and transportation. HNP find that the usual assumption of constant budget share elasticities, which the Leser-Working specification imposes, appears inconsistent with the 1982 CES data. HNP also find that a Hausman (1978) type specification test of the IV estimates versus the OLS estimates strongly rejects the OLS estimates, indicating the importance of the measurement error specification. Thus, HNP find strong evidence that use of current expenditure in estimation of Engel curves on micro data leads to errors in variables problems. Lastly, HNP explore the Gorman results. For the polynomial specification of equation (2), the rank restriction takes the form that the ratio of coefficients of the cubic terms to the coefficients of the quadratic terms will be constant across budget share equations. HNP estimate the "Gorman statistic" to see whether the coefficients in the cubic specific of equation (2) have rank three. HNP find a rather remarkable result. HNP find the ratios of the coefficients to be extremely close in actual values and estimated precisely. The ratios of the coefficients for the different budget share equations are –25.0, -25.2, -25.1, -23.3, and –25.6.[14] The near equality of the ratios provides strong support for the Gorman hypothesis. Thus, use of consistent estimates of a mismeasurement model for Engel curve analysis provides an interesting result that a restriction from economic theory applies to the data.

In a recent doctoral thesis at MIT, Schennach (2000) has extended the HINP (1991) results from the polynomial specification of equation (2) to the consistent estimation of nonlinear models with measurement errors in the explanatory variables when one repeated observation exists. Thus, regression models with nonlinear functions such as $g(z_i)$, where z_i is an unobserved variable can be estimated consistently. Schennach uses the Fourier transform to convert the integral equations that relate the

[14] The HINP consistent estimates provide stronger support for the Gorman hypothesis than do the OLS estimates.

distribution of the unobserved true variables to the mismeasured observed variables into algebraic equations. She then solves these equations to identify moments of the true unobserved variables. These moments are used to construct a traditional nonlinear least squares estimator. While HINP restricted their estimator to polynomial moments of the unobserved variables and its product with the left hand side variable, Schennach allows for arbitrary functions of the mismeasured right hand side variable x_i.[15] Thus, the approach is considerably more flexible than the HINP approach, although it does involve calculation of the empirical Fourier transforms. Schennach estimates nonlinear Engel curves using CES data from 1984:1 using the next quarter expenditure as the repeated measurement. For both of her nonlinear specification she finds that for 4 out or the 5 budget categories that she estimates, a Hausman (1978) test rejects the traditional NLS estimator in favor of her estimator that allows for mismeasured right hand side variables. Thus, both HNP (1995) and Schennach (2000) find that in estimation of Engel curves, a least squares approach which uses current expenditure instead of a permanent income approach leads to inconsistent estimates where the usual downward bias (attenuation) is present in the least squares estimates.

III. Measurement Error in the Left Hand Side Variables: Probit and Logit

I now consider the consequences of mismeasurement in the left hand side variable in binary outcome models where the observed left hand side variable is a function of an unobserved (latent) dependent variable. This specification often arises in limited dependent variable models. The outcome is different from the usual regression specification where a mismeasured left hand side variable does not lead to problems, as discussed above. In the limited dependent variable specification, misclassification of the left hand side variable leads to biased and inconsistent estimators.

The usual latent variable specification has an observed latent variable y^*_i and an observed variable $y_i = 1$ if $y^*_i \geq 0$ and $y_i = 0$ if $y^*_i < 0$. The regression specification that allows for misclassification is:

[15] In particular, consistent estimation for probit and logit models with mismeasured right hand side (explanatory) variables follows.

$$q_i = a_0 + (1 - a_0 - a_1)F(z_i b) + e_i \qquad (3)$$

where a_0 is the probability that $q_i = 1$ although the true $y_i = 0$, a_1 is the probability that $q_i = 0$ although the true $y_i = 1$, and F is the cumulative distribution for the probit model or logit model. With no mismeasurement $a_0 = a_1 = 0$ and nonlinear least squares (NLS) or maximum likelihood estimation (MLE) of equation (3), imposing the no misclassification assumption, leads to consistent estimates. However, if misclassification is present, both NLS and MLE lead to biased and inconsistent estimation. However, if equation (3) is estimate by NLS allowing for misclassification, Hausman, Abrevaya, and Scott-Morton (HAS 1998) demonstrate that consistent estimation results. Thus, the estimated coefficients of α_0 and α_1 provide a specification test for mismeasurement. Similarly, HAS demonstrate that MLE, which permits for misclassification, is straightforward and consistent estimates result so long as a monotonicity condition holds: $a_0 + a_1 < 1$. This monotonicity condition is relatively weak since it says that the combined probability of misclassification is not so high that on average you cannot tell which result actually occurred. Thus, two results arise in the binary limited dependent variable cases which are different from the classical regression specification with measurement error in the left hand side variable: (1) inconsistent estimation results from the mismeasurement and (2) consistent estimation does not require instrumental variables. The analysis extends to discrete response models with more than two categories, and MLE estimation is consistent. HAS find that relatively small amounts of misclassification, as little as 2%, can lead to significant amounts of large sample bias in Monte Carlo experiments.

The assumption of normally distributed disturbances required by the probit specification or extreme value disturbances for the logit specification, used in the specification of F in equation (3), is not necessary for consistent estimation. Thus, HAS also demonstrate that the monotonicity condition allows for semiparametric estimation where no cumulative distribution needs to be assumed. HAS demonstrate that the maximum rank correlation (MRC) estimator of Han (1987) provides a consistent

estimator for b. A further advantage of the MRC estimator is that a more flexible model of misclassification is sufficient for consistent estimation as the exact form of misclassification need not be specified or estimated.

HAS then demonstrate how to non-parametrically estimate the response function F as a function of $z_i b$, the consistent estimate, using an isotonic regression (IR) method. The IR method works in the current situation since F is monotonic in $z_i b$ since it is a distribution function. The IR technique imposes a monotonicity condition on the left-hand side variable in a regression specification. The IR estimator is pointwise consistent, and HAS develop asymptotic standard errors and confidence intervals for the consistent estimator of F.

As an empirical example HAS specify and estimate a model of job change using both the CPS (Current Population Survey) and PSID (Panel Study of Income Dynamics) data sets. The results from the CPS data set, discussed here, demonstrate strong evidence of misclassification. Using the MLE of a probit specification which allows for misclassification, HAS find that a_0, the probability of misclassification for non-job changers is estimated to be 6% and is very precisely estimated, and a_1, the probability of misclassification for job changers is estimated to be 31% and is also very precisely estimated. The MLE estimates of many of the right hand side variables change by large amounts when misclassification is permitted. HAS then use the MRC and IR estimators that do not impose functional form restrictions. They find that a_0, the probability of misclassification for non-job changers is estimated to be 4% and is very precisely estimated, and a_1, the probability of misclassification for job changers is estimated to be 40% and is also very precisely estimated. Thus, HAS find that the semiparametric estimates are quite similar to the MLEestimates that allow for misclassification. Also, the MRC estimates of most of the right hand side variables change by only relatively small amounts compared to the parametric MLE estimates that allow for misclassification. Thus, misclassification appears to be a potentially serious problem in micro data where inconsistent results may arise. While the combination MRC/IR approach is quite flexible, the limited empirical experience of HAS appears to

demonstrate that the MLE approach to probit or logit allowing for misclassification may give reasonable results in many actual empirical situations.

IV. Measurement Error in the Left Hand Side Variables: More General Models

The model in the last section is a particular example of a class of models that take the form:

$$y^*_i = f(z_i b) + e_i \qquad (4)$$

where f is strictly increasing and y^*_i is an unobserved (latent) variable. To allow for mismeasurement in this more general situation Abrevaya-Hausman (AH 1999) model the observed left hand side variable q_i as a stochastic function of the underlying y^*_i. AH demonstration that this model allows for binary choice with misclassification as in HAS (1998), mismeasured discrete dependent variables, and mismeasured continuous dependent variables. In this latter situation the observed left hand side variable is continuous and is a function of the unobserved (latent) variable and a random disturbance so that $q_i = h(y^*_i, w_i)$. This last model specification includes the increasingly used duration and hazard models.

AH consider semi-parametric estimation of these models using either the MRC estimators or the more general monotone rank estimator (MRE) developed by Cavanagh and Sherman (1998). To achieve identification and allow consistent estimation using the MRC and MRE estimators AH develop a sufficient condition on the mismeasurement process. The sufficient condition is that the distribution for the observed variable q_i for a higher latent y^*_i stochastically dominates the distribution of q_j for a lower latent y^*_j. Thus, the effect of the mismeasurement cannot on average permute the ordering of the observed left hand side variables with respect to the ordering of the unobserved latent variables. So the effect of the mismeasurement cannot be “too large” on average. The use of the notion of first order stochastic dominance is familiar from microeconomics

where it is used to order portfolios or risky distributions.[16] In terms of an actual problem with mismeasurement in the left hand side variable the appropriate question to ask to satisfy the sufficient condition is: "Are observational units with larger "true" values for their left hand side variable more likely to report larger values than observations units with smaller "true" values?" If the answer is "yes", the AH techniques can be used. Again instrumental variables are not required for consistent estimation.

AH consider their approach for the duration models such as the well-known Cox (1972) partial likelihood model, the Han-Hausman (1990) and Meyer (1990) specifications, and more generally the proportional hazard model. A hazard model, in the context of the return to employment for an unemployed person, answers the question of what is the probability of becoming employed in the next time period conditional on being unemployed up to the previous time period. Thus, the hazard model has a similarity to discrete response models, but it allows for the effect of past outcomes on the probability of the return to employment to affect the outcome in the current period.

AH demonstrate that conventional MLE estimation of commonly used duration models and hazard models is inconsistent when the left hand side variable is mismeasured, with the single exception of the highly restrictive Weibull duration model.[17] So if the person were actually unemployed for say 24 weeks and answered 26 weeks in the survey, mismeasurement willexist and would, in general, lead to inconsistent estimation. AH find in Monte Carlo experiments that both the Cox model and the Han-Hausman-Meyer (HHM) models have coefficient estimates that are attenuated, e.g. biased towards zero. Thus, duration model specification that do not allow for mismeasurement in the left hand side variable have results similar to the classical regression model with mismeasurement in the right hand side variable. AH demonstrate that MRC and MRE estimators do well in the Monte Carlo experiments in taking account of the mismeasurement and estimating the unknown parameters.

AH discuss the finding of mismeasured duration data in survey or unemployment duration data. Mismeasured durations are common to all data sets that have survey

[16] See e.g. Mas-Colell, Whinston, and Green (1995), p. 195.
[17] The Weibull specification requires a monotonic (baseline) hazard, which makes it too restrictive for most problems. Also, the particular application cannot allow for censoring (check) which is typically present in applications of duration models.

responses, and AH concentrate on the Survey of Income and Program participation (SIPP) data set. Some common findings are that a significant number of reporting error are found—in the CPS about 37% of unemployed workers overstated unemployment durations; longer spells have a higher proportion of reporting errors, and responses tend to be "focal responses" for instance at a number of weeks that corresponds to an integer month amount, e.g. 4 or 8 weeks. AH conduct an empirical analysis on unemployment duration in the SIPP data and use the Cox and HHM models, which do not allow for mismeasurement as well as a semi-parametric MRE model that does permit mismeasurement. While many of the MRE parameter estimates for the demographic variables are similar to the Cox and HHM parameter estimates, one potentially important difference is found. Allowing for mismeasurement finds a much smaller and statistically insignificant effect on the UI benefit levels on unemployment duration. The previous wage interacted with receiving UI continues to be significant, but the actual size of the UI benefit ceases to be significant. In particular the commonly used "replacement-rate" specification is rejected when mismeasurement is permitted in the unemployment durations. AH conclude that mismeasurement in the left hand side variable can have an important effect in duration models.

V. An Unsolved Problem: Measurement Error in the Left Hand Side Variable of Quantile Regression Models

Koenker and Bassett (KB 1978) introduced the quantile regression (QR) estimator into econometrics. The QR estimator gives a view of the entire conditional distribution rather than just, say, the conditional mean. The advantage of the QR estimator is that it is a "robust" estimator of the regression specification so that it is not as sensitive to "outliers", i.e. extreme observations.[18] Alternatively, the QR estimator performs well on efficiency grounds when the stochastic disturbance has "thick tails" that depart from the normal (Gaussian) assumption. Lastly, the QR estimator does well in the heteroscedastic situation. The QR estimator estimates a regression specification at a number of q th regression quantiles for $0 < q < 1$ in the regression model specification $e_i(q) = y_i - z_i b(q)$. The QR estimator thus generalizes the regression median estimator

for the regression quantile of $q = 1/2$. Interesting applications of the QR estimator are increasingly common. Buchinsky (1994) used the QR technique to explore changes in the wage distribution in the CPS over a 25 year period. Buchinsky focuses on changes in the returns to education and experience at different points of the wage distribution, given that he believes that heteroscedasticity is likely to be an important factor. He finds that the returns to education and experience are different at the different quantiles, $q = \{0.1, 0.25, 0.5, 0.75, 0.9\}$.

To the best of my knowledge, no one has explored the effect of mismeasured left hand side variables for the QR approach. As before, assume that the measured left hand side variable $q_i = y_i + w_i$. Mismeasurement in the left hand side variable might be expected in some applications. For instance, in the Buchnisky paper the main left hand side variable is the weekly wage, defined in the CPS as the "total income from wages and salaries last year" divided by the "number of weeks worked last year." Both numerator and denominator arise from survey responses in the CPS (check). However, it is not difficult to demonstrate that a mismeasured left hand side variable will lead to inconsistent QR estimation because the stochastic disturbance is now $e_i + w_i$. The presence of the error in measurement "attenuates" the quantiles coefficient estimates, leading to estimates that are linear combination of the actual quantile coefficients $b(q)$ for different q's.

As an empirical example I took a QR specification with a constant and one right hand variable in the presence of heteroscedasticity and then added a (standard) normal measurement error to the left hand side variable. The results for estimates of the slope coefficient are in Table 1:

[18] See the paper by Professor Koenker in this issue.

Table 1: Estimated QR Slope Coefficients with and without Measurement Error (ME)[19]

Quantile	No ME	Normal ME
0.10	.010	.278
0.25	.064	.234
0.50	.254	.303
0.75	.549	.329
0.90	.795	.368

As expected, the QR slope estimates with normal measurement error are attenuated towards the median coefficient because of the mixture of the regression error e_i and the measurement error w_i. While the QR estimator is robust to certain data features that create problems for the classical linear regression model, it is not robust to a mismeasured left hand side variable, which does not create problems for the classical linear regression model. I have not yet found a solution to this problem, similar to the solution for the limited dependent variable problems and duration model problems. Finding a solution seems a good topic for future research.

[19] I used 10,000 observations in the Monte Carlo runs. The right hand side variable is drawn from a uniform distribution. The specification contains significant conditional heteroscedasticity.

References

Abrevaya J. and J. Hausman (1999), “Semiparametric Estimation with Mismeasured Dependent Variables: An Application to Duration Models for Unemployment Spells”, Annales D’Economie et de Statistique, 55-56, 243-275.

Bound, J., D. A. Jaeger, and R. M. Baker (1997), “Problems with Instrumental Variables Estimation when the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak,” Journal of the American Statistical Association, 90, 443-450.

Adcock, R. (1887), "A Problem in Least Squares", The Analyst, 5, 53-54

Aigner, D.J., Hsiao, C., Kapteyn A., Wansbeek, T. (1984), "Latent Variable Models in Econometrics", in Z. Griliches and M. D. Intriligator, Handbook of Econometrics, vol. II, 1323-1393.

Amemiya, Y. (1985), "Instrumental Variable Estimator for the Nonlinear Errors-in-Variables Model," Journal of Econometrics, 28, 273-289.

Buchinsky, M. “Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression”, Econometrica, 62, 405-458.

Cavanagh C., and R. Sherman (1998), “Rank Estimators for Monotonic Index Models”, Journal of Econometrics, 84, 351-381.

Cox D.R. (1972), “Regression Models and Life Tables (with discussion)”, Journal of the Royal Statistical Society, B, 34: 187-220.

Gorman, W.M. (1981), "Some Engel Curves", in A. Deaton ed., Essays in the Theory and Measurement of Consumer Behaviour in Honor of Sir Richard Stone, Cambridge: Cambridge Univ. Press.

Hahn J. and J. Hausman (1999), “A New Specification Test for the Validity of Instrumental Variables”, forthcoming Econometrica.

Han A. (1987), “Non-Parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator”, Journal of Econometrics, 35: 303-316.

Han A. and J. Hausman (1990), “Flexible Parametric Estimation of duration and Competing Risk Models”, Journal of Applied Econometrics, 5, 1-28.

Hausman, J. (1978), "Specification Tests in Econometrics", Econometrica, 46, 1251-1271.

Hausman J., J. Abrevaya, and F. Scott-Morton (1998), “Misclassification of an Dependent Variable in a Discrete-Response Setting”, Journal of Econometrics, 87, 239-269.

Hausman, J., H. Ichimura, W. Newey, and J. Powell (1991), "Measurement Errors in Polynomial Regression Models," Journal of Econometrics, 50, 271-295.

Hausman, J., W. Newey, and J. Powell (1995), "Measurement Errors in Polynomial Regression Models," Journal of Econometrics, 65, 205-233.

Koenker R. and G. Bassett, “Regression Quantiles”, Econometrica, 46, 33-50.

Leser, C.E.V. (1963), "Forms of Engel Functions", Econometrica, 31, 694-703

Livitan N. (1961), "Errors in Variables and Engel Curve Analysis", Econometrica, 29, 336-362.

Mas-Colell A., M. Whinston, and J. Green, Microeconomic Theory, Oxford Univ. Press, 1995.

Meijer E. and T. Wansbeek, “Measurement Error in a Single Regressor”, Economic Letters, 69, 277-284.

Meyer B. (1990), “Unemployment Insurance and Unemployment Spells”, Econometrica, 58, 757-782.

Nagar, A.L. (1959): “The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations,” Econometrica, 27, 575 – 595.

Schennach S. (2000), Estimation of Nonlinear Models with Measurement Error, unpublished MIT Ph.D. thesis.

White H. (1984), Asymptotic Theory for Econometricians, New York: Academic Press.