Polynomial regression with normal covariate measurement error

Andrew Chesher¤
Department of Economics
University College London

5th November 1999

Abstract. This paper derives the exact functional form of an error contaminated regression function when the error free regression is a polynomial function of error free covariates (discrete or continuous) which are contaminated by normally distributed measurement error, with coe¢cients which may be arbitrary functions of error free covariates. The form of higher order central moment error contaminated regressions is examined and by way of example the form of normal measurement error induced heteroskedasticity when the error free regression is linear and homoskedastic is derived.

The results of this paper may provide at least a partial explanation of mild non-linearity and heteroskedasticity found in applied econometric work with survey data when error contamination, e.g. of income and expenditure data, is likely. The error contaminated regression function is completely determined by the coe¢cients in the error free regression, the measurement error variance and the density of the observed covariates. This density can be estimated, opening the way to estimation of error free regression functions using only data on the response and the error contaminated covariate, to investigation of the potential impact of measurement error on structural error free regressions and to the development of speci…cation tests sensitive to unmodelled measurement error.

Keywords: measurement error, attenuation, non-linear regression, non-parametric regression, heteroskedasticity.

1. Introduction

Measurement error is pervasive in micro-econometric work and it can have damaging e¤ects on inference. It is well known that measurement error can bring attenuation e¤ects causing behavioural responses to appear less sharp than they actually are. Less well known is that covariate measurement error can change the shapes of regression

¤Corresponding address: Department of Economics, University College London, Gower Street, London WC1E 6BT, UK. Email: andrew.chesher@ucl.ac.uk.

I am grateful to Hans Schneeweiss for pointing out an error in an earlier version of this paper. Support provided by ESRC grant number R 008237386 is gratefully ackowledged.

functions, although this phenomenon has been understood for over 50 years - see for example Lindley (1947) and Cochran (1972). If the regression of Y on an error free covariate, X, (i.e. $E_{Y|X}[Y|X = x]$) is linear then in general the regression of Y on Z, an error contaminated version of X, is non-linear and when the regression of Y on X is non-linear the regression of Y on Z generally has a different non-linear form. Since there is increasing econometric interest in non-linear modelling and increasing use of non-parametric regression estimation methods, recognition and understanding of this effect of measurement error is important.

Econometric applications of nonparametric regression often reveal mild nonlinearity, and in parametric regression estimation it is often found that including squared and higher powers of covariates improves fit. An old example appears in the production function literature where developments such as the translog production function were motivated partly by perceived failings in simpler functional forms. Recent examples are provided by Banks, Blundell and Lewbel (1997) and Attfield and Bhalotra (1998) who, providing their own evidence contained in data from respectively Great Britain and rural Pakistan, argue for the addition to the conventional Working-Leser Engel curve of a quadratic term in log total expenditure. Banks et al survey many other studies which come to a similar conclusion. Since covariate measurement error generally distorts the shapes of regression functions, it is possible that such observed non-linearities are in part due to covariate measurement error, the error free regression perhaps taking a relatively simple form.

This suggests an econometric strategy in which relatively complex relationships among observed variates are modelled using simple structural models combined with realistic models of the observation process. This strategy has been widely used in modelling duration data where simple hazard function models combined with simple models of uncontrolled across individual heterogeneity can explain complex forms of observed hazard function duration dependence[1]. In situations where measurement error may be significant the same sort of strategy can be employed. To do this it is necessary to understand how measurement error distorts perceptions of underlying structural relationships.

The relationship between error contaminated and error free regressions is studied in general continuously distributed measurement error models in Chesher (1991, 1997) using small parameter asymptotic approximations. The present paper considers a leading special case, namely that in which measurement error is normally distributed (the response and error free covariates can be non-normal, continuous or discrete). It provides the exact form of the error contaminated regression function when the error free regression is a polynomial function of a scalar X_2 that is observed with error, where the coefficients in the polynomial are arbitrary functions of an error free vector X_1. The tools to extend the result to multivariate polynomial functions of

[1] The impact of response measurement error on duration models, the relationship between response measurement error and across individual heterogeneity and specification tests sensitive to measurement error are studied in Chesher, Dumangane and Smith (1998).

vector X_2 observed with error are provided. Higher order moment regression functions are also studied and, as an example, the exact form of measurement error induced heteroskedasticity is obtained for the case in which the error free regression is linear and homoskedastic. In all these cases the distortion introduced by measurement error is determined entirely upon the parameters of the error free regression, the measurement error variance and the conditional density of the error contaminated covariates given the error free covariates, $f(z_2|z_1)$.

The assumption of normal measurement error is crucial in obtaining the exact results reported here. However that assumption is often a quite realistic one since in many applications a central limit theorem argument can be invoked in its favour. I have in mind here the many situations in econometric work in which unobserved covariates are replaced by estimates obtained from a …rst pass through the data. Examples include the use of estimated Mills ratio terms in sample selection models and the use of estimated regional prices in demand models. In macro-economic contexts an aggregation argument will often lead to the same conclusion.

The results given here have a number of potential uses. First they allow exploration of the potential impact of measurement error contamination of data on particular covariates. With an estimate of, or even a guess at, the variance of measurement error and with an estimate of $f(z_2|z_1)$, which could be obtained by parametric or nonparametric methods, it is possible to gain a view of the impact of measurement error on the shape of any posited structural error free regression function. For example, the results imply that with one covariate and a linear error free regression, $E[Y|X = x] = a_0 + a_1 x$, the concavity of the error contaminated regression is at each value of z precisely that of $a_1 \, r_z \log f(z)$ where $f(z)$ is the density of the error contaminated covariate, $Z = X + V$ and $V \sim N(0; \sigma^2)$ independently of Y and X. Second, and as elaborated[2] in Chesher (1998), armed with an estimate of $f(z_2|z_1)$ it is possible to estimate the structural coe¢cients in the error free regression using only data on the response and the error contaminated covariate data[3] and if instrumental variables are available the information they contain can be incorporated to gain improved e¢ciency. Third, the results allow construction of test statistics for the hypothesis of no covariate measurement error. Finally, they suggest caution in giving behavioural interpretations to all observed non-linearities in econometric relationships if there is the possibility that covariates are measured with error, as of course is often the case in microeconometric work.

[2] The procedure described in Chesher (1998) produces approximately consistent (in the sense that the inconsistency is of order $o(\sigma^2)$) estimators of coe¢cients in error free regressions without assuming normality of measurement errors for a wide class of error free regressions..

[3] This is in contrast to the estimation procedure of Wolter and Fuller (1982) for the quadratic error free regression model which has similar data requirements but rests on an assumption of joint normality of the covariate measurement error and the response conditional on the error free covariates and knowledge of their joint covariance matrix, and to the estimation procedure of Hausman, Newey and Powell (1995) which requires instrumental variables or replicated measurements.

2. Error free and error contaminated regressions

2.1. The exact error contaminated regression function. Let $X^{\prime} = [X_1^{\prime} : X_2]$ with X_2 scalar, let $Z_1 = X_1$, $Z_2 = X_2 + \sigma U$ where U is distributed independently of X and Y, each of which may be continuous, discrete or mixed. The variates $V = \sigma U$ and Z_2 are to be interpreted as respectively measurement error and an error contaminated measure of X_2. Let $f(z_2|z_1)$ be the conditional density of Z_2 given Z_1. It is shown in Section 3 that, for Gaussian U and an error free regression which is a Jth order polynomial in x_2 with coefficients that may depend upon the error free x_1:

$$E[Y|X = x] = \sum_{j=0}^{J} a_j(x_1) x_2^j$$

the error contaminated regression of Y on Z is

$$E[Y|Z = z] = \sum_{j=0}^{J} a_j(z_1) \frac{1}{f(z_2|z_1)} \sum_{i=0}^{j} \binom{j}{i} (-1)^i G_i(z; \sigma^2) z_2^{j-i} \qquad (1)$$

where the functions $G_i(z; \sigma^2)$ satisfy the recurrence relation

$$\begin{aligned} G_{i+1}(z; \sigma^2) &= -\tfrac{1}{2}\sigma^2 \, r_{z_2} G_i(z; \sigma^2) + 1_{[i \geq 1]} i \sigma^2 G_{i-1}(z; \sigma^2) \\ G_0(z; \sigma^2) &= f(z_2|z_1): \end{aligned}$$

2.2. Discussion. The error contaminated regression function can be expressed as a polynomial of degree J in z_2 with coefficients depending on the functions $a_j(z_1)$, the measurement error variance σ^2, the conditional density $f(z_2|z_1)$ and its derivatives with respect to z_2 of order up to J.

As a simple example, when the error free regression is a quadratic function of scalar x_2 measured with error and of error free vector x_1 (each of which can be continuous or discrete) with

$$E[Y|X = x] = x_1^{\prime}\gamma + a_0 + a_1 x_2 + a_2 x_2^2$$

the error contaminated regression function is

$$E[Y|Z = z] = z_1^{\prime}\gamma + (a_0 + a_2\sigma^2) + a_1 z_2 + a_2 z_2^2 + \sigma^2(a_1 + 2a_2 z_2) g^{(1)}(z) + a_2 \sigma^4 g^{(2)}(z) \quad (2)$$

where

$$g^{(i)}(z) = \frac{f^{(i)}(z_2|z_1)}{f(z_2|z_1)}$$

$f(z_2|z_1)$ is the conditional density of Z_2 given $Z_1 = z_1$ and superscript $^{(i)}$ indicates differentiation i times with respect to z_2.

Consider the case of linear error free regression. Setting a_2 to zero in (2) gives the following simple result.

$$E[Y|Z = z] = z_1^{\beta} + a_0 + a_1 z_2 + \sigma^2 a_1 g^{(1)}(z) \tag{3}$$

When $g^{(1)}(z) = \partial \log f(z_2|z_1)/\partial z_2$ is non-linear in z_2, measurement error destroys the linearity of the error free regression. This log density derivative is linear when Z_2 given Z_1 is itself normal and then this shape deformation is absent confirming a well known property of the fully Gaussian measurement error model. Note that when $f(z_2|z_1)$ is not normal the $O(\sigma^2)$ term in (3) can bring non-linear functions of z_1 and z_2 into the error contaminated regression function.

The conditional density $f(z_2|z_1)$ can be estimated by parametric methods, for example as a mixture of normal distributions, or by nonparametric methods, and then estimates of its derivatives can be obtained. It is then possible to compute a score test of the hypothesis $H_0 : \sigma^2 = 0$ which in this linear error free regression case will examine the sample covariance of residuals around the fitted error free regression, $\hat{\epsilon}_i$, and the estimated log density derivatives $\hat{g}^{(1)}(z)$.

The measurement error variance and the coefficients in the error free regression can be estimated by a two step procedure, plugging the log density derivatives estimated in the first step into (3) and applying some suitable M-estimation procedure. This is investigated further in Chesher (1998).

The next Section formally states and proves the main result concerning the exact error contaminated regression and compares it with the small variance approximation given in Chesher (1991). Section 4 studies two extensions; first the case in which many covariates are measured with error and then the exact form of higher moment error contaminated regressions. Section 5 concludes.

3. The exact form of normal measurement error contaminated polynomial regressions

In this Section the exact form of the error contaminated regression is derived for the case in which the error free regression is a polynomial function of a scalar X_2 which is observed with additive normal measurement error. The coefficients in this polynomial are arbitrary functions of a vector X_1 observed without error. The case in which the error free regression is a multivariate polynomial function of vector X_2 observed with error differs only in notational complexity and the tools required to produce the error contaminated regression for this case are given in Section 4.1.

The form of the error contaminated regression is given in the following Theorem.

Theorem 1. Let Y be a scalar random variable and let $X^0 = [X_1^0 : X_2]$ be a vector random variable with X_2 scalar. Let $Z_1 = X_1$, $Z_2 = X_2 + \gamma U_2$ where $U_2 \sim N(0;1)$ independently of X and Y. Let the conditional density of Z_2 given Z_1 be denoted by $f(z_2|z_1)$. Assume that there exist functions of x_1, $a_j(x_1)$, $j = 1;:::;J$ such that the regression of Y on X exists and takes the following form.

$$E[Y|X = x] = \sum_{j=0}^{\infty} a_j(x_1)x_2^j$$

Then the regression of Y on Z is

$$E[Y|Z = z] = \sum_{j=0}^{\infty} a_j(z_1)\frac{1}{f(z_2|z_1)} \sum_{i=0}^{\infty} \binom{j}{i} (-1)^i G_i(z;\gamma^2) z_2^{j-i}$$

where the functions $G_i(z;\gamma^2)$ satisfy the recurrence relation

$$\begin{aligned} G_{i+1}(z;\gamma^2) &= \gamma^2 \left[1_{[i>1]} i G_{i-1}(z;\gamma^2) - r_{z_2} G_i(z;\gamma^2) \right] \\ G_0(z;\gamma^2) &= f(z_2|z_1) \end{aligned} \tag{4}$$

The proof, which is given below, involves expressing $E[Y|Z = z]$ as a linear function (given z_1) of $E[X_2^j|Z = z]$, obtaining the latter expectations by exploiting the properties of the normal density under differentiation. First note the following points about the conditions of the Theorem and the result.

Remark 1. Leading cases of interest are those in which

1. the functions $a_j(x_1)$ are all independent of x_1 so that x_1 is absent from the error free regression and the error free regression is a pure polynomial model. In this case the conditional density $f(z_2|z_1)$ can be replaced by the marginal density of Z_2, $f(z_2)$, and,
2. only $a_0(x_1)$ depends on x_1 in which case the error free regression is a partially polynomial model, the sum of a polynomial function of x_2 and an arbitrary function of x_1.

Remark 2. The function $G_i(z;\gamma^2)$ is a polynomial of degree i in γ^2 with coefficients involving derivatives of order up to i with respect to z_2 of $f(z_2|z_1)$. The first four of these functions are as follows.

$$\begin{aligned} G_1(z;\gamma^2) &= -\gamma^2 f^{(1)}(z_2|z_1) \\ G_2(z;\gamma^2) &= \gamma^4 f^{(2)}(z_2|z_1) + \gamma^2 f_Z(z_2|z_1) \\ G_3(z;\gamma^2) &= -\gamma^6 f^{(3)}(z_2|z_1) - 3\gamma^4 f^{(1)}(z_2|z_1) \\ G_4(z;\gamma^2) &= \gamma^8 f^{(4)}(z_2|z_1) + 6\gamma^6 f^{(2)}(z_2|z_1) + 3\gamma^4 f(z_2|z_1) \end{aligned}$$

where $f^{(i)}(z_2|z_1) = @^i f(z_2|z_1)/@z_2^i$.

Remark 3. Both Y and X can be continuous, discrete or mixed discrete continuous. Clearly only the regression of Y on X is relevant and so the nature of the distribution of Y is irrelevant, except of course that its regression on X must exist. Since U_2 is normal, Z_2 given $Z_1 = z_1$ has a continuous density with …nite derivatives of all orders, as is shown in the proof which follows.

Proof of Theorem 1

The possibility that the distribution of X_2 given X_1 is mixed discrete continuous is allowed. Let the conditional distribution function of X_2 given X_1 be

$$P[X_2 \cdot x_2|X_1 = x_1] = \sum_{j \in J(x_2)} p(x_{2j}|x_1) + \int_{i\ 1}^{x_2} p^{\alpha}(w|x_1)dw$$

where $p^{\alpha}(w|x_1)$ is a nonnegative function, $p(x_{2j}|x_1)$ is the positive probability mass located at the mass point x_{2j}, $J(w) = \{j : x_{2j} \cdot w\}$, and

$$\sum_{j \in J(1)} p(x_{2j}|x_1) + \int_{i\ 1}^{1} p^{\alpha}(w|x_1)dw = 1.$$

The location and number of mass points may depend upon x_1 but this is not made explicit in the notation[4].

Let $Á(¢)$ denote the standard normal density function and de…ne

$$G_i(z;¾^2) = E[(Z_2 ¡ X_2)^i|Z = z] £ f(z_2|z_1) \tag{5}$$

where

$$f(z_2|z_1) = \sum_{j \in J(1)} p(x_{2j}|z_1)\frac{1}{¾}Á\left(\frac{z_2 ¡ x_{2j}}{¾}\right) + \int_{i\ 1}^{1} p^{\alpha}(x_2|z_1)\frac{1}{¾}Á\left(\frac{z_2 ¡ x_2}{¾}\right)dx_2$$

is the conditional density of Z_2 given $Z_1 = z_1$ which, note, is continuous with bounded derivatives with respect to z_2 of all orders. Then

$$G_i(z;¾^2) = \sum_{j \in J(1)} (z_2 ¡ x_{2j})^i p(x_{2j}|z_1)\frac{1}{¾}Á\left(\frac{z_2 ¡ x_{2j}}{¾}\right) + \int_{i\ 1}^{1} (z_2 ¡ x_2)^i p^{\alpha}(x_2|z_1)\frac{1}{¾}Á\left(\frac{z_2 ¡ x_2}{¾}\right)dx_2$$

and di¤erentiating with respect to z_2 gives

$$G_{i+1}(z;¾^2) = ¾^2 \{1_{[i\ ,\ 1]}\, i\, G_{i\ 1}(z;¾^2) ¡ r_z G_i(z;¾^2)¢$$

[4] In the purely discrete case $p^{\alpha}(w|x_1) = 0$ for all w. In the purely continuous case the index set $J(1)$ is empty.

which with $G_0(z;\sigma^2) = f(z_2|z_1)$, which is implied by the definition of $G_i(z;\sigma^2)$ in (5), gives the expectation of $(Z_2 - X_2)^i$ given $Z = z$ for all $i \geq 0$.

Expanding $(Z_2 - X_2)^i$ in (5) gives

$$\sum_{j=0}^{\infty} \binom{i}{j} (-1)^j z_2^{i-j} E[X_2^j | Z = z] = G_i(z;\sigma^2)/f(z_2|z_1)$$

and upon inverting

$$E[X_2^j | Z = z] = \frac{1}{f(z_2|z_1)} \sum_{i=0}^{\infty} \binom{j}{i} (-1)^i G_i(z;\sigma^2) z_2^{j-i}. \tag{6}$$

Finally, because U is independent of Y,

$$E[Y|Z = z] = \sum_{j=0}^{\infty} a_j(z_1) E[X_2^j | Z = z]$$

and substituting (6) the proof is complete. □

The conditional expectations of the first four powers of X_2 which appear in the error contaminated regression, expressed in terms of the functions $g^{(i)}(z) = f^{(i)}(z_2|z_1)/f(z_2|z_1)$ are[5] as follows.

$$E[X_2|Z = z] = z_2 + \sigma^2 g^{(1)}(z)$$

$$E[X_2^2|Z = z] = z_2^2 + 2z_2\sigma^2 g^{(1)}(z) + \sigma^4 g^{(2)}(z) + \sigma^2$$

$$E[X_2^3|Z = z] = z_2^3 + 3z_2^2\sigma^2 g^{(1)}(z) + 3z_2(\sigma^4 g^{(2)}(z) + \sigma^2) + \sigma^6 g^{(3)}(z) + 3\sigma^4 g^{(1)}(z)$$

$$E[X_2^4|Z = z] = z_2^4 + 4z_2^3\sigma^2 g^{(1)}(z) + 6z_2^2(\sigma^4 g^{(2)}(z) + \sigma^2) + 4z(\sigma^6 g^{(3)}(z) + 3\sigma^4 g^{(1)}(z)) + \sigma^8 g^{(4)}(z) + 6\sigma^6 g^{(2)}(z) + 3\sigma^4$$

The error contaminated regression can be re-expressed as

$$E[Y|Z = z] = \frac{1}{f(z_2|z_1)} \sum_{j=0}^{\infty} z_2^j \left(\sum_{s=0}^{J-j} a_{s+j}(z_1) \binom{s+j}{s} (-1)^s G_s(z;\sigma^2) \right)$$

which is a polynomial function of z_2 of the same degree, J, as the polynomial error free regression with coefficients depending on σ^2, and the functions $a_j(z_1)$ and $g^{(j)}(z)$, $j = 0,\ldots,J$. If $f(z_2|z_1)$, and hence the $g^{(j)}(z)$, were known then the parameters of

[5] For the case in which X_1 is absent from the problem the first of these results is given in Das and Mulder (1983).

the error free regression could, when identi…cation permits, be estimated directly by non-linear least squares using only realisations of the response and the error contaminated covariates. Since realisations of Z are available the functions $g^{(j)}(z)$ can be estimated and the estimates "plugged in" prior to calculation of non-linear least squares estimates. Of course this sort of estimation procedure cannot be implemented when the structural parameters of the error free regression cannot be identi…ed from knowledge of the function (1). Lack of identi…ability arises for example when the functions $g^{(i)}(z)$ are polynomials in z_2 ruling out the case in which Z (and therefore X) is itself normally distributed[6].

For the case in which X is continuously distributed and for a wide class of continuous measurement error distributions, Chesher (1991) gives small measurement error variance approximations to error contaminated regressions which, when there is just one covariate measured with error, specialise to

$$r_Z^a(z) = r_X(z) + \tfrac{1}{4}\sigma^2 r_X^{(1)}(z)g^{(1)}(z) + \frac{\tfrac{1}{4}\sigma^2}{2} r_X^{(2)}(z)$$

where $r_Z(z)$ and $r_X(z)$ are respectively the error contaminated and error free regression functions expressed as functions of z, $r_Z(z) = r_Z^a(z) + o(\sigma^2)$ and $r_X^{(i)}(z)$ is the ith derivative with respect to z_2 of $r_X(z)$.

Theorem 1 shows that with normally distributed measurement error the remainder terms for degree $J = 1; 2; 3$ polynomial error free regressions $r_X(z) = \sum_{j=0}^{J} a_j x^j$ are as follows.

$$\begin{aligned}
J = 1 \;&:\; r_Z(z) - r_Z^a(z) = 0 \\
J = 2 \;&:\; r_Z(z) - r_Z^a(z) = a_2\tfrac{1}{4}\sigma^4 g^{(2)}(z) \\
J = 3 \;&:\; r_Z(z) - r_Z^a(z) = \tfrac{1}{4}\sigma^4\left[a_2 g^{(2)}(z) + 3a_3 z g^{(2)}(z) + 3a_3 g^{(1)}(z)\right] + \tfrac{1}{4}\sigma^6 a_3 g^{(3)}(z)
\end{aligned}$$

The small variance approximation is exact when the error free regression is linear. When it is quadratic the remainder term is larger the further from zero is

$$g^{(2)}(z) = \left[r_{z_2}^{(2)} \log f(z_2|z_1) + \left(r_{z_2}^{(1)} \log f(z_2|z_1)\right)^2\right].$$

In the case in which Z_2 given Z_1 is normally distributed with mean $\mu(z_1)$ and variance ω^2 the remainder term for this quadratic case is

$$r_Z(z) - r_Z^a(z) = a_2 \left[\frac{\tfrac{1}{4}\sigma^4}{\omega^4}\left((z_2 - \mu(z_1))^2 - \omega^2\right)\right]$$

which quickly becomes negligible as the variance of the measurement error becomes small relative to the conditional variance of Z_2 given Z_1.

[6] Unless $\tfrac{1}{4}\sigma^2$ is known or a consistent estimate of it is available.

4. Extensions

4.1. Many covariates measured with error. Now suppose that the error free regression is a multivariate polynomial function of a P element vector X with constant coefficients

$$E[Y|X = x] = \sum_{m=1}^{M} a_m \prod_{p=1}^{P} x_p^{J_{mp}}$$

where M is the number of terms (not necessarily distinct) in the polynomial regression, a_m is the coefficient associated with the mth term, and J is a $M \times P$ array containing in its (m,p) position the power to which x_p is raised in the mth term. Suppose that X is observed with error, which potentially affects all of its components. Let the error contaminated Z be

$$Z = X + U$$

where U is $N_P(0,\Sigma)$, distributed independently of Y and X. The regression of Y on Z is then

$$E[Y|Z = z] = \sum_{m=1}^{M} a_m E\left[\prod_{p=1}^{P} X_p^{J_{mp}} | Z = z\right]. \tag{7}$$

The conditional expectations in this expression can be obtained using the following result, a multivariate extension of (4). Let

$$G_{i_1,i_2,\ldots,i_P} = E\left[\prod_{p=1}^{P} (z_p - x_p)^{i_p} | Z = z\right] f(z)$$

where $f(z)$ is the density function of Z. These functions obey the recursion

$$\begin{bmatrix} G_{i_1+1,i_2,\ldots,i_P} \\ G_{i_1,i_2+1,\ldots,i_P} \\ \vdots \\ G_{i_1,i_2,\ldots,i_P+1} \end{bmatrix} = \Sigma\, \mathrm{diag}(1_{[i_p>1]} i_p) \begin{bmatrix} G_{i_1-1,i_2,\ldots,i_P} \\ G_{i_1,i_2-1,\ldots,i_P} \\ \vdots \\ G_{i_1,i_2,\ldots,i_P-1} \end{bmatrix} - \Sigma \begin{bmatrix} r_{z_1} G_{i_1,i_2,\ldots,i_P} \\ r_{z_2} G_{i_1,i_2,\ldots,i_P} \\ \vdots \\ r_{z_P} G_{i_1,i_2,\ldots,i_P} \end{bmatrix}$$

which, together with the initial condition $G_{0,0,\ldots,0} = f(z)\iota_P$ where ι_P is a P-element vector of ones, generates $E[\prod_{p=1}^{P}(z_p - x_p)^{i_p}|Z = z]$ for all combinations of nonnegative $i_1,\ldots,i_P$. From this the expectations in (7) can be obtained. The result for the case in which some components of X are observed without error are obtained by setting the appropriate elements of Σ to zero and the coefficients, a_m, can be functions of error free covariates. As earlier it is then possible to express the error contaminated regression in terms of the conditional density of the error contaminated covariates given the error free covariates.

For low order polynomial regression a direct attack is straight forward as the following example shows.

Example 1. Let the error free regression be the following quadratic function of x.

$$E[Y|X = x] = \beta + x'\gamma + x'Ax$$

The marginal density function of Z is

$$\begin{aligned} f(z) &= \int f_{ZX}(z;x)dx \\ &= \int |\Sigma|^{-1/2} f_X(x)(2\pi)^{-p/2}\exp(-\frac{1}{2}(z-x)\Sigma^{-1}(z-x))dx \end{aligned}$$

which has (vector) first and (matrix) second derivatives with respect to z

$$f^{(1)}(z) = -\int \Sigma^{-1}(z-x)f_{ZX}(z;x)dx$$

$$f^{(2)}(z) = -\int \left[\Sigma^{-1} - \Sigma^{-1}(z-x)(z-x)'\Sigma^{-1}\right] f_{ZX}(z;x)dx$$

from which it follows that

$$E[X|Z = z] = z + \Sigma\frac{f^{(1)}(z)}{f(z)}$$

$$E[XX'|Z = z] = zz' + \Sigma + \Sigma\frac{f^{(1)}(z)}{f(z)}z' + z\frac{f^{(1)}(z)'}{f(z)}\Sigma + \Sigma\frac{f^{(2)}(z)}{f(z)}\Sigma.$$

Therefore the error contaminated regression is as follows.

$$\begin{aligned} E[Y|Z = z] = \beta + z'\gamma + z'Az + \frac{f^{(1)}(z)'}{f(z)}\Sigma\gamma + \mathrm{trace}(A\Sigma) + z'A\Sigma\frac{f^{(1)}(z)}{f(z)} \\ + \frac{f^{(1)}(z)'}{f(z)}\Sigma Az + \mathrm{trace}(\Sigma A\Sigma\frac{f^{(2)}(z)}{f(z)}) \end{aligned}$$

4.2. Measurement error induced heteroskedasticity. When error free covariates are non-normal there is generally heteroskedastic variation around the error contaminated regression even though the variation around the error free regression may be homoskedastic. The exact form of this variation can be obtained when measurement error is normal. Returning to the case in which one covariate is measured with error, let Y have a polynomial regression on potentially error contaminated scalar X_2 with coefficients that are functions of error free X_1 and suppose the second and higher central moments,

$$E[(Y - \sum_{j=0}^{k} a_j(X_1)X_2^j)^s|X = x] = \mu_s;$$

(where $\gamma_0 = 1$, $\gamma_1 = 0$) are independent[7] of X. It follows that

$$E[Y^s|X = x] = \sum_{k=0}^{s} \binom{s}{k} \gamma_k \left(\sum_{j=0} a_j(x_1) x_2^j \right)^{s-k} :$$

This is a polynomial function of x_2 and applying Theorem 1 gives the conditional moments of Y given $Z = z$ in terms of powers of z_2, the coefficients $a_j(z_1)$, the conditional density function of Z_2 given Z_1 and its derivatives with respect to z_2.

Example 2. Let the regression of Y on scalar X be linear ($E[Y|X = x] = a_0 + a_1 x$) with $Var[Y|X = x] = \gamma_2$, independent of x. With $Z = X + \varepsilon U$ and $U \sim N(0;1)$ independent of Y and X, the skedastic function of Y given Z is as follows.

$$V[Y|Z = z] = \gamma_2 + a_1^2 \varepsilon^2 + a_1^2 \varepsilon^4 \left[g_Z^{(2)}(z) - g_Z^{(1)}(z)^2 \right] \tag{8}$$

Note that

$$g_Z^{(2)}(z) - g_Z^{(1)}(z)^2 = r_z^{(2)} \log f_Z(z):$$

This is constant when X (and therefore Z) is normally distributed so that this result reproduces the well known result for the fully Gaussian model, namely that in that case measurement error does not cause heteroskedasticity. Equation (8) shows that with non-normal error free covariates, normal measurement error and homoskedastic linear regression on error free covariates there is always heteroskedasticity around the error contaminated regression function, which, as shown in the small variance approximations in Chesher (1997, 1998), is of order $o(\varepsilon^2)$.

5. Concluding remarks

This paper has derived the exact functional form of an error contaminated regression function when the error free regression is a polynomial function of an error free covariate (discrete or continuous), subject to normally distributed measurement error, with coefficients which may be arbitrary functions of error free covariates. The extension to the case in which the error free regression is a polynomial function of a vector of potentially error contaminated covariates has been indicated. Higher order, central moment, error contaminated regressions are easily derived and, as an example, the exact form of heteroskedasticity induced by normal measurement error in a linear, homoskedastic error free regression has been derived.

[7] The method used here could also be employed if these conditional moments were polynomial functions of x.

These results may provide a partial explanation of mild non-linearity and heteroskedasticity found in applied econometric work with survey data when error contamination, e.g. of income and expenditure data, is likely. The error contaminated regressions depend upon the density of the observed covariates which can be estimated, opening the way to estimation of error free regression functions using only data on the response and the error contaminated covariate. This provides an alternative to the instrumental variables procedure with similar data requirements for linear error free regression models proposed by Lewbel (1997) whose procedure, unlike that proposed here, does not readily extend to non-linear error free regressions.

Hausman, Newey and Powell (1995), in a situation in which there are replicated measurements of error contaminated covariates, propose consistent estimation of coefficients of general non-linear regressions, $E[Y|X = x] = h(x;\theta)$, via estimation of the coefficients of the linear projection of Y onto polynomial functions of x, minimum distance estimation producing estimators of θ from these estimated coefficients. For the normal measurement error case, employing an estimate of the density of the error contaminated covariates as proposed in Chesher (1998), the result of this paper suggests that it is possible to dispense with the requirement of replicated measurements.

References

Attfield, C.L.F., and S. Bhalotra, (1998) "Intrahousehold Resource Allocation in Rural Pakistan: a Semiparametric Approach, " forthcoming Journal of Econometrics.

Banks, J., R. Blundell and A. Lewbel, (1997) "Quadratic Engel curves and consumer demand," Review of Economics and Statistics, 179, 527-539.

Chesher, A.D., (1991) "The effect of measurement error," Biometrika, 78, 451-462.

Chesher, A.D., (1997) "Non-normal variation and regression to the mean," Statistical Methods in Medical Research, 6, 147-166.

Chesher, A.D., (1998) "Structural estimation in the presence of covariate measurement error," revised version of: "Measurement error bias reduction," (1998) University of Bristol Department of Economics Discussion Paper No. 98/449, presented at the 1998 European Meeting of the Econometric Society, Berlin.

Chesher, A.D., Dumangane, M.B.G., and R.J. Smith, (1998) "Duration response measurement error," University of Bristol Department of Economics Discussion Paper No. 98/451.

Cochran, W.G., (1972) "Some Effects of Measurement Error on Linear Regression," in Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and

Probability, Volume 1, ed., Le Cam, L.M,. Neyman, J. and E.L. Scott, Berkeley CA: University of California Press.

Das, P., and P.G.H. Mulder, (1983) "Regression to the mode," Statistica Neerlandica, 37, 15-21.

Hausman, J.A., W.K. Newey and J.L. Powell, (1995) "Nonlinear errors in variables: estimation of some Engel curves," Journal of Econometrics, 65, 205-233.

Lewbel, A., (1997) "Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R&D," Econometrica, 65, 1201-1214.

Lindley, D.V., (1947) "Regression lines and the linear functional relationship," Journal of the Royal Statistical Society, B, 6, 218-244.

Wolter, K.M., and W.A. Fuller (1982) "Estimation of the quadratic errors-in-variables model," Biometrika, 69, 175-182.