Errors in Variables and Lending Discrimination

Jed L. DeVaro and Jeffrey M. Lacker

Do banks discriminate against minority loan applicants? One approach to answering this question is to estimate a model of bank lending decisions in which the probability of being denied a loan is a function of a set of creditworthiness variables and a dummy variable for the applicant's race ($z = 1$ for minorities, $z = 0$ for whites). A positive coefficient on the race dummy is taken as evidence that minority applicants are less likely to be granted loans than white applicants with similar qualifications. This approach is employed in many empirical studies of lending discrimination (Schill and Wachter 1994; Munnell et al. 1992), in U.S. Department of Justice lending discrimination suits (Seiberg 1994), and in regulatory examination procedures (Bauer and Cromwell 1994; Cummins 1994).

One weakness of this approach is that an estimate of the discrimination coefficient may be biased when measures of creditworthiness are fallible. In such situations, distinguishing racial discrimination from unmeasured racial disparities in creditworthiness can be difficult. If true creditworthiness is lower on average for minority applicants, the model may indicate that race adversely affects the probability of denial, even if race plays no direct causal role.

There are good reasons to believe that measures of creditworthiness are fallible. First, regulatory field examiners report difficulty finding matched pairs of loan files to corroborate discrimination identified by regression models. An applicant's file often yields a picture of creditworthiness different from the one given by model variables. Second, including more borrower financial characteristics generally reduces discrimination estimates, sometimes to zero (Schill and Wachter 1994). Third, studies of default data find that minority borrowers are more likely than white borrowers to default, even after controlling for income,

The authors thank Mary Finn and Peter Ireland for helpful conversations during this research and Tom Humphrey, Tony Kuprianov, and Stacey Schreft for helpful comments on an earlier draft. The authors are solely responsible for the contents of this article. The views expressed do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

wealth, and other borrower characteristics related to creditworthiness (Berkovec et al. 1994). This finding suggests that there are race-related discrepancies between the true determinants of creditworthiness and the measures available to econometricians.

Our objective is to develop a method for assessing the sensitivity of lending discrimination estimates to measurement error. In particular, we study the classical errors-in-variables model, in which the components of a vector x of observed measures of creditworthiness are, one for one, fallible measures of those in a vector of true qualifications x^*.[1] The implications of errors in variables in the standard linear regression model are well known (Klepper and Leamer 1984; Goldberger 1984).[2] We briefly review these implications in Section 1. Models of lending discrimination generally specify a nonlinear regression model, such as the logit model, because the dependent variable is dichotomous ($y = 1$ if the loan application is denied; $y = 0$ if it is accepted). In this article we extend the results for the linear case to cover the nonlinear logit regression model widely used in lending discrimination studies.

Linear errors-in-variables models are underidentified because variation in true qualifications cannot be distinguished from error variance. Assuming that the errors are normally distributed with known parameters, however, the linear model is just-identified, allowing estimation of model parameters depending on the assumed error-variance parameters. Assuming zero error variance yields the standard linear regression model as a special case. By estimating under a range of error-variance assumptions, one can trace out the potential effect of measurement error on model parameter estimates. Note that since the error-variance assumptions make the model just-identified, no one assumption about the error-variance parameters is more likely than any other; that is, estimates of model parameters under alternative error-variance assumptions are all equally consistent with the data. Also note that in the case of normally distributed regressors in the linear model, parameter estimates for alternative error-variance

[1] The classical errors-in-variables model is not the only one in which observed variables, taken together, are fallible measures of true creditworthiness. Alternatives include “multiple-indicator” models in which observed variables are fallible measures of a single index of creditworthiness, and “omitted-variable” models in which some determinants of creditworthiness are unobservable. All are alike in that a component of the true model is unobserved by the econometrician; thus, all are latent-variable models. Because errors in variables is one of the simplest and most widely studied models of fallible regressors, it is a useful starting point in examining fallibility in empirical models of lending discrimination.

[2] Interest in the errors-in-variables problem has surged since 1970. As Hausman and colleagues (1995) stated, “During the formative period of econometrics in the 1930’s, considerable attention was given to the errors-in-variable[s] problem. However, with the subsequent emphasis on aggregate time series research, the errors-in-variables problem decreased in importance in most econometric research. In the past decade as econometric research on micro data has increased dramatically, the errors-in-variables problem has once again moved to the forefront of econometric research” (p. 206).

assumptions can be obtained through an algebraic correction to the ordinary least squares estimates.

In Section 2 we examine the logit model under errors in variables and show how estimators depend on assumptions about error variance. Adjusting estimators for error variance is no longer an algebraic correction as it is in the linear setup; the model must be reestimated for each error-variance assumption. For the case in which the independent variables are continuous-valued, we show how to estimate the logit model under various assumptions about error variance. Because of the nonlinearity, the logit model is in some cases identified without error-variance assumptions. In practice, however, the logit model is quite close to underidentified, and little information can be obtained from the data about error-variance parameters. Therefore, we advocate estimating models under a range of error-variance assumptions to check the sensitivity of estimates to measurement error.

In Section 3 we demonstrate our method using artificial data. We show how estimates of a discrimination parameter can be biased when a relatively modest amount of measurement error is present. The magnitude of the bias depends on the model's fundamental parameters. By estimating the model under different assumptions about measurement error variance, we can gauge the sensitivity of the estimators to errors in variables. Section 4 concludes and offers directions for further research.

Bauer and Cromwell (1994) have also studied the properties of logit regression models of lending discrimination, focusing on the small-sample properties of a misspecified model using simulated data. They found that tests for lending discrimination were sensitive to sample size. Our work focuses on the effect of errors in variables on the large-sample properties of otherwise correctly specified logit models of lending discrimination.

1. ERRORS IN VARIABLES

The implications of errors in variables are easiest to see in a linear setup such as the following simple model of salary discrimination.[3] Suppose that an earnings variable (y) is determined according to the following equations:

$$y = \beta x^* + \alpha z + v, \tag{1a}$$

$$x^* = x_0 + \mu z + u, \tag{1b}$$

[3] The exposition in this section is based on Goldberger (1984). This model of salary discrimination has a close parallel in the permanent income theory. Friedman (1957) discusses how racial differences in unobserved permanent income (the counterpart of qualifications in the salary model and creditworthiness in the lending model) bias estimates of racial differences in the consumption function intercept.

$$x = x^* + e, \tag{1c}$$

where the scalar x^* = true qualification, x = measured qualification, and z is a race dummy ($z = 1$ for minorities, $z = 0$ for whites). We take v, u, and e to be mutually independent random variables with zero means and variances σ_v^2, σ_u^2, and σ_e^2, all independent of z. The earnings variable in (1a) is a stochastic function of the true qualifications and race. The parameter α represents the independent effect of race on salary, and $\alpha < 0$ represents discrimination against minorities. If better-qualified applicants obtain higher salaries, then $\beta > 0$. In (1b) qualification is allowed to be correlated with race; the expectation of x^* is x_0 for whites and $x_0 + \mu$ for minorities. The empirically relevant case has $\mu < 0$. Observed qualification in (1c) is contaminated by measurement error e.

Consider a regression of y on the observed variables x and z. This estimates

$$E[y \mid x, z] = bx + az.$$

Since the variances and covariances are the same for both white and minority applicants, we can use conditional covariances to calculate the regression slopes. We focus on relationships in a population and thus ignore sampling variability. The least squares estimators are

$$b = \text{cov}(x, y \mid z)/v(x \mid z) = \text{cov}(x^*, y \mid z)/v(x \mid z) = (1 - \delta)\beta$$

and

$$\begin{aligned} a &= E[y \mid z = 1] - E[y \mid z = 0] - b\{E[x \mid z = 1] - E[x \mid z = 0]\} \\ &= \alpha + \beta\mu - b\mu \\ &= \alpha + \delta\beta\mu, \end{aligned}$$

where

$$\delta \equiv \sigma_e^2/(\sigma_u^2 + \sigma_e^2).$$

When there is measurement error ($\sigma_e^2 > 0$), the regression estimator of β is biased toward zero. To see why, substitute for x^* in (1a) using (1c) to obtain $y = \beta x + \alpha z + (v - \beta e)$. The "error" $v - \beta e$ in the regression of y on x and z is correlated with x via (1c). Thus a key assumption of the classical linear regression model is violated, and the coefficients are no longer unbiased.

In our case ($\beta > 0$, $\mu < 0$), the estimator of α is biased downward as well. Bias creeps in because z is informative about x^*, given x;

$$E[x^* \mid x, z] = (1 - \delta)x + \delta(x_0 + \mu z).$$

Given observed qualification x, race can help "predict" true qualification x^*. Race can then help "explain" earnings, even in the absence of discrimination ($\alpha = 0$), because race is correlated with true qualifications.

The model (1) is underidentified (Kapteyn and Wansbeek 1983). A regression of x on z recovers the nuisance parameters x_0 and μ, along with

$v(x \mid z) = \sigma_u^2 + \sigma_e^2$. Other population moments provide us with a and b, but these are not sufficient to identify α, β, and δ. No sample can provide us with enough information to divide $v(x \mid z)$ between the variance in true qualifications σ_u^2 and the variance in measurement error σ_e^2. Under the assumptions $\beta > 0$ and $\mu < 0$, any value of $\alpha > a$, including the no-discrimination case $\alpha = 0$, is consistent with the data for some β and σ_e^2.

If σ_e^2 were known independently, then we would know $\delta = \sigma_e^2/(\sigma_u^2 + \sigma_e^2)$ and could calculate the unbiased estimators $\hat{\alpha}$ and $\hat{\beta}$ by correcting the ordinary least squares estimators as follows:

$$\hat{\beta} = b/(1-\delta) \tag{2a}$$

$$\hat{\alpha} = a - \delta b\mu/(1-\delta). \tag{2b}$$

One could use (2) to study the implications of alternative assumptions about the variance of measurement error; different values of σ_e^2 would trace out different estimates of α.

In (1) the direction of bias in a is known when the sign of $\beta\mu$ is known. Matters are different when x is a vector of characteristics affecting qualifications. Consider a multivariate model:

$$y = \beta' x^* + \alpha z + v, \tag{3a}$$

$$x^* = x_0 + \mu z + u, \tag{3b}$$

$$x = x^* + e, \tag{3c}$$

where x^* and x are now $k \times 1$ random vectors and β, μ, and x_0 are $k \times 1$ parameter vectors. We take u and e to be normally distributed random vectors, independent of v, z, and each other, with zero means and covariance matrices Σ^* and D. The classical assumption is that measurement errors are mutually independent, so D is diagonal.

The least squares estimators are now

$$b = (\Sigma^* + D)^{-1}\Sigma^*\beta \tag{4a}$$

and

$$a = \alpha + (\beta - b)'\mu. \tag{4b}$$

The direction of bias is now uncertain, even under the usual assumption that measurement errors are independent (D is diagonal). To see why, suppose that $k = 2$, Σ^* has ρ as the off-diagonal element, and $\Sigma^* + D$ has ones on the diagonal (a normalization of units). Then (4b) becomes

$$a = \alpha + [(D_{11}\beta_1 - \rho D_{22}\beta_2)\mu_1 + (D_{22}\beta_2 - \rho D_{11}\beta_1)\mu_2]/(1-\rho^2).$$

The bias in a could be positive or negative, depending on parameter values. For example, suppose only one component of x is subject to measurement error, say, x_1 ($D_{11} > 0$ and $D_{22} = 0$). By itself this would bias b_1 downward, resulting in an upward bias in a. But $b_2 = \rho\beta_1 D_{11}/(1-\rho^2) + \beta_2$ is now biased as well, and this would induce downward bias in a if $\rho\beta\mu > 0$. The overall direction of bias is indeterminate (Rao 1973; Hashimoto and Kochin 1980). But again, if the measurement error parameters D were known, then the least squares estimators a and b could be corrected by a simple transformation of (4) (using $\Sigma^* = \Sigma - D$, where $\Sigma = v(x \mid z)$). Each alternative measurement error assumption would imply a different estimator.[4]

2. ERRORS IN VARIABLES IN A LOGIT MODEL OF DISCRIMINATION

In model (3) the dependent variable is a linear function of the explanatory variables. In models of lending decisions the dependent variable is dichotomous: $y = 1$ if the applicant is denied a loan, and $y = 0$ if the applicant is accepted. In this case the linear formulation in (3) is unattractive (Maddala 1983). A common alternative is the logit model, shown here without errors in variables:

$$\Pr[y = 1 \mid x, z] = G(\beta' x + \alpha z), \tag{5a}$$

$$G(t) = \frac{1}{1 + e^{-t}}, \tag{5b}$$

where x is a vector of characteristics influencing creditworthiness. The empirically relevant case has $\beta < 0$, so applicants who are more creditworthy are less likely to be denied loans. A value of $\alpha > 0$ would indicate discrimination against minorities: a minority applicant is approximately $\alpha(1 - G)$ times more likely than an identical white applicant to be denied a loan.[5]

The parameters α and β can be estimated by the method of maximum likelihood. The log likelihood function for a sample of n observations $\{y_i, x_i, z_i, i = 1, \ldots, n\}$ is

$$\log L = \sum_{i=1}^{n} \log \Pr(y_i, x_i, z_i) = \sum_{i=1}^{n} \log \Pr(y_i \mid x_i, z_i) + \sum_{i=1}^{n} \log \Pr(x_i, z_i), \tag{6}$$

where

$$\Pr(y_i \mid x_i, z_i) = G(\beta' x_i + \alpha z_i)^{y_i} [1 - G(\beta' x_i + \alpha z_i)]^{(1-y_i)}.$$

Estimators are found by choosing parameter values that maximize log L. The likelihood depends on the parameters of the conditional distribution in (5) as

[4] Klepper and Leamer (1984) and Klepper (1988b) show how to find bounds and other diagnostics for the linear errors-in-variables model.

[5] The elasticity of G with respect to z is $\alpha G'/G = \alpha(1 + e^{-t})e^{-t}/(1 + e^{-t})^2 = \alpha e^{-t}/(1 + e^{-t}) = \alpha(1 - G)$, where G is evaluated at $\beta' x + \alpha z$.

well as on the “nuisance parameters” governing the unconditional distribution of (x, z). Since the nuisance parameters appear only in the second sum in (6), while α and β appear only in the first sum, α and β can be estimated in this case without estimating the nuisance parameters.

Under errors in variables, (5a) is replaced with

$$\Pr[y = 1 \mid x^*, z] = G(\beta' x^* + \alpha z), \tag{7}$$

where x^* is the vector of true characteristics. The resulting log likelihood function is

$$\begin{aligned} \log L &= \sum_{i=1}^{n} \log \Pr(y_i, x_i, z_i) \\ &= \sum_{i=1}^{n} \log \int \Pr(y_i \mid x_i^*, z_i) \Pr(x_i \mid x_i^*) \Pr(x_i^*, z_i) dx_i^*. \end{aligned} \tag{8}$$

The likelihood function now depends on $\Pr(x \mid x^*)$, the probability that x is observed if the vector of true characteristics is x^*. Since $x - x^*$ is the vector of measurement errors, $\Pr(x \mid x^*)$ is the probability distribution governing the measurement error. In the linear model (3) the least squares estimators could be corrected algebraically for measurement error of known variance. In the logit model, however, there is no simple way to adjust maximum likelihood estimators for errors in variables, since the regression function is nonlinear. Instead, we must estimate α and β for each distinct assumption about $\Pr(x \mid x^*)$.

Unlike the one in (6), the log likelihood function in (8) is not separable in the nuisance parameters of the distribution $\Pr(x^*, z)$. Even if we posit an error distribution $\Pr(x \mid x^*)$, estimating α and β requires estimating the parameters of $\Pr(x^*, z)$ as well. The estimation of these nuisance parameters will be sidestepped here by maximizing the conditional likelihood function

$$\begin{aligned} \log \tilde{L} &= \sum_{i=1}^{n} \log \Pr(y_i \mid x_i, z_i) \\ &= \sum_{i=1}^{n} \log \int \Pr(y_i \mid x_i^*, z_i) \Pr(x_i^* \mid x_i, z_i) dx_i^*. \end{aligned} \tag{9}$$

We will assume that $\Pr(x^* \mid x, z)$, the distribution of true characteristics conditional on observed characteristics and race, is known.

Our model is completed by adding specific assumptions about the distributions $\Pr(x \mid x^*)$ and $\Pr(x^* \mid z)$, which will allow us to derive $\Pr(x^* \mid x, z)$. We will maintain the assumptions embodied in (3b) and (3c):

$$x^* = x_0 + \mu z + u, \tag{10a}$$

$$x = x^* + e, \tag{10b}$$

where β, μ, and x_0 are $k \times 1$ parameter vectors and where u and e are normally distributed random vectors, independent of v, z, and each other, with zero means and covariance matrices Σ^* and D. Given x and z, x^* is then normally distributed with mean vector m^* and covariance matrix S^*, where

$$m^* = D\Sigma^{-1}\mu z + (I - D\Sigma^{-1})x, \tag{11a}$$

$$S^* = (I - D\Sigma^{-1})D. \tag{11b}$$

With this result in hand, we find that, conditional on x and z, the argument of G is normally distributed with mean $\beta' m^* + \alpha z$ and variance $\beta' S^* \beta$. Therefore, the likelihood in (9) can be written as

$$\Pr(y \mid x, z) = \int G(m + \sigma s)(2\pi)^{-1/2} \exp(-s^2/2) ds, \tag{12}$$

where

$$m = \beta'(I - D\Sigma^{-1})x + (\alpha + \beta' D\Sigma^{-1}\mu)z,$$

$$\sigma = [\beta'(I - D\Sigma^{-1})D\beta]^{1/2}.$$

When $D = 0$, m collapses to $\beta' x + \alpha z$ and $\sigma = 0$, which is the error-free model.[6]

Because of the nonlinearity of G, the logit model can potentially be identified without error-variance assumptions, unlike the linear model in Section 1. Thus, in principle, the error-variance parameters could be estimated rather than imposed. In practice, however, the model is so close to linear that the error-variance parameters cannot be estimated; even large samples are uninformative about D. We therefore recommend estimating the model under a range of alternative error-variance assumptions.

To summarize the procedure, first calculate least squares estimators for the parameters x_0, μ, and Σ. These parameters are treated as fixed and combined with an assumed D to obtain the distribution $\Pr(x^* \mid x, z)$, which is used in (12) and (9) to obtain maximum likelihood estimates of α and β. This procedure treats the error variance D as known, just as the error-free model treats D

[6] The joint normality of x and x^* given z implies that given x and z, x^* is normal with parameters that can be derived algebraically from the parameters of $\Pr(x \mid x^*)$ and $\Pr(x^* \mid z)$. Other distributional assumptions on x and x^* are far less convenient. For example, when x^* takes on discrete values, a more general approach is required to derive $\Pr(x^* \mid x, z)$. Given a distribution of the observables $\Pr(x, z)$, recover $\Pr(x^* \mid z)$ using $\Pr(x \mid z) = \int \Pr(x \mid x^*)\Pr(x^* \mid z)dx^*$, and then use Bayes's rule to obtain $\Pr(x^* \mid x, z) = \Pr(x \mid x^*)\Pr(x^* \mid z)/\Pr(x \mid z)$. The first of these steps involves inverting a very large matrix.

as identically zero. Estimates of α can then be traced out under alternative assumptions on D.[7]

Our procedure will misstate the uncertainty about parameter estimates, even conditioning on D. By implicitly assuming that the estimated parameters x_0, μ, and Σ are known, we are neglecting their sampling variability. These parameters appear in (12) and thus influence estimates of α and β. Our procedure therefore misstates their sampling variability as well. When $D = 0$, the nuisance parameters disappear from (12), and this problem does not arise.[8]

3. EXAMPLES

In the examples in this section, we apply our procedure in a logit model of discrimination to show how the technique is capable of detecting the sensitivity of parameter estimates to errors in variables. We find it convenient to use artificially generated data sets to illustrate our results. Artificial data allow us to isolate important features of the errors-in-variables model for a wide array of cases. Observations are randomly generated under a given, true error variance, and the model is then estimated under various hypothesized error variances.

In the simplest case there is only one explanatory variable besides race ($k = 1$). We assume $\alpha = 0$, $\beta = -1$, $\mu = -2$, and $\Sigma = 1$. (We focus on the no-discrimination case, $\alpha = 0$, solely for convenience.) In this case, if a is significantly different from zero, then it is also significantly greater than α, and the usual t-statistic on a will also show whether a is significantly biased. The sample was assumed to be half white ($z = 0$) and half minority ($z = 1$). Using these values and an assumed true error variance D, we generated 10,000 random observations on x^*, x, and y using equations (7) and (10). We then estimated the model using maximum likelihood, assuming that the true values of μ and Σ were known and making an assumption about $\tilde{D}$ (not necessarily the same as D). The results are displayed in Table 1. The sample size of 10,000 was chosen to reduce sampling variance.

For the estimates shown in Panel A of Table 1, the true variance of the measurement error is $D = 0.1$. This represents one-tenth of the total variance in observed x, a relatively modest amount. The first line reports estimation under the (incorrect) assumption that the error variance is zero. As expected, the estimate b is biased toward zero. Consequently, a is biased upward, toward showing discrimination, and is significant.

[7] In related work, Klepper (1988a) extended the diagnostic results of Klepper and Leamer (1984) and Klepper (1988b) to a linear regression model with dichotomous independent variables. These earlier approaches attempted to characterize the set of parameters that maximize the likelihood function. Levine (1986) extended the results of Klepper and Leamer (1984) to the probit model.

[8] Specifically, the hessian of the log likelihood function is then block diagonal across (α, β) and (x_0, μ, Σ).

Table 1 Coefficient Estimates for Alternative Error-Variance Assumptions, $k = 1$
$\mu = -2$, $\Sigma = 1$, $n = 10,000$.

	a	***b***
A. True parameters $\alpha = 0$, $\beta = -1$, and $D = 0.1$:		
Assumed $\tilde{D}$		
0.0	0.1446	−0.9208
	(2.4380)	(−32.3477)
0.05	0.0482	−0.9775
	(0.7780)	(−31.8322)
0.1	−0.0607	−1.0418
	(−0.9308)	(−31.2626)
B. True parameters $\alpha = 0.1$, $\beta = -0.9$, and $D = 0.0$:		
Assumed $\tilde{D}$		
0.0	0.1609	−0.9260
	(2.7101)	(−32.4378)
0.05	0.0640	−0.9832
	(1.0315)	(−31.9159)
0.1	−0.0456	−1.0480
	(−0.6986)	(−31.3393)

Notes: t-statistics are shown in parentheses beneath the coefficient estimates. For each panel, we drew a set of 10,000 random realizations for (y, x): 5,000 with $z = 0$ and 5,000 with $z = 1$. Within each panel, estimation was performed on the same data set with different assumptions about the error variance $\tilde{D}$.

The last two lines in Panel A show estimates assuming positive error variance. For larger values of $\tilde{D}$, b is closer to one and a is closer to zero, the true value. The discrimination parameter is not significantly different from zero when estimated assuming D is 0.05 or 0.1. In this case, then, our procedure successfully detects the sensitivity of parameter estimates to errors in variables.

In Panel B we examine the case in which no measurement error is present and the true discrimination parameter is positive. The (correct) assumption of no measurement error now yields estimates that are unbiased; they differ from the true parameters only because of sampling error. Imposing the (incorrect) assumption of positive measurement error variance "undoes" a nonexistant bias, resulting in a near zero and a larger negative b.

Table 2 shows how the magnitude of the bias varies with the correlation between components of x when $k = 2$. Σ has diagonal elements equal to one and off-diagonal elements equal to a scalar ρ, where $-1 < \rho < 1$. D has diagonal elements all equal to 0.1; the independent variables other than race suffer from measurement error of the same variance. We maintain $\alpha = 0$,

Table 2 Coefficient Estimates for Alternative Correlation and Error-Variance Assumptions, $k = 2$

$\alpha = 0$, $\beta = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $\mu = \begin{bmatrix} -2 \\ -2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ and $D = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}$, $\tilde{D} = \begin{bmatrix} \tilde{d} & 0 \\ 0 & \tilde{d} \end{bmatrix}$, $n = 10,000$.

	$\boldsymbol{a}$	$\boldsymbol{b_1}$	$\boldsymbol{b_2}$
A. $\rho = 0$:			
Assumed $\tilde{d}$			
0.0	0.4299	−0.8340	−0.8663
	(4.2028)	(−25.2057)	(−25.7966)
0.1	0.0394	−0.9557	−0.9924
	(0.3483)	(−24.3926)	(−24.9389)
B. $\rho = 0.5$:			
Assumed $\tilde{d}$			
0.0	0.2975	−0.8797	−0.8705
	(3.4439)	(−22.8422)	(−22.5769)
0.1	0.0419	−0.9726	−0.9597
	(0.4506)	(−20.2974)	(−20.0418)
C. $\rho = -0.5$:			
Assumed $\tilde{d}$			
0.0	0.7672	−0.7816	−0.7714
	(5.5997)	(−21.8801)	(−21.5103)
0.1	−0.0457	−1.0084	−0.9969
	(−0.2720)	(−21.4374)	(−21.1531)

Notes: t-statistics are shown in parentheses beneath the coefficient estimates. For each panel, we drew a set of 10,000 random realizations for (y, x): 5,000 with $z = 0$ and 5,000 with $z = 1$. Within each panel, estimation was performed on the same data set.

$\beta = (-1, -1)$, and $\mu = (-2, -2)$. Panel A shows that when the components of x are uncorrelated, the bias is larger than in the comparable $k = 1$ model: 0.43 versus 0.14. When the components of x are positively correlated ($\rho = 0.5$), the bias is smaller by almost a third but is still significant. When the components of x are negatively correlated ($\rho = -0.5$), the bias is substantially larger. Thus the bias in a varies negatively with ρ, just as the linear case suggested. A positive value of ρ implies that measurement error in x_1 biases the coefficient on x_2 away from zero, counteracting the effect of measurement error in x_2. Although b_i is biased toward zero by measurement error in x_i, the bias is somewhat offset by the effects of measurement error in other components of x.

When $k = 1$, the direction of bias is determined entirely by the sign of $\beta\mu$. When $k > 1$, the direction of bias depends on Σ and D, even when $\beta'\mu$ can be signed. Table 3 illustrates this fact for $k = 2$, showing a set of parameters for which a is biased against finding discrimination. Both x_1 and x_2 are plagued by measurement error, but with a strong positive correlation between the two,

Table 3 Coefficient Estimates for Alternative Error-Variance Assumptions, $k = 2$

$\alpha = 0, \beta = \begin{bmatrix} -0.1 \\ -1 \end{bmatrix}, \mu = \begin{bmatrix} -2 \\ -0.1 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.75 \\ 0.75 & 1 \end{bmatrix}$ and

$D = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}, \tilde{D} = \begin{bmatrix} \tilde{d} & 0 \\ 0 & \tilde{d} \end{bmatrix}, n = 10,000.$

	a	b_1	b_2
Assumed $\tilde{d}$			
0.0	−0.2445	−0.2352	−0.7703
	(−3.4442)	(−7.0602)	(−21.9616)
0.1	0.0312	−0.0887	−0.9962
	(0.2888)	(−1.6430)	(−17.3444)

Notes: t-statistics are shown in parentheses beneath the coefficient estimates. For each panel, we drew a set of 10,000 random realizations for (y, x): 5,000 with $z = 0$ and 5,000 with $z = 1$. Within each panel, estimation was performed on the same data set.

each has a dampening effect on the bias in the coefficient of the other variable. The net bias in b_2 is toward zero, but b_1 is biased away from zero. Since x_1 is more strongly correlated with z, the net effect is a negative bias in a. With the correct error-variance assumption, the model detects the lack of discrimination.

In Table 4 we display results for a model with $k = 10$, a size that is more like that of the data sets encountered in actual practice. With $\rho = 0$, we see in Panel A that with more correlates plagued by measurement error, the bias in a is larger. With $\rho = 0.5$, the various measurement errors partially offset each other, but a remains significantly biased. Once again, our technique faithfully compensates for known measurement error.

4. SUMMARY

We have described a method for estimating logit models of discrimination under a range of assumptions about the magnitude of errors in variables. Using artificially generated data, we showed how the bias in the discrimination coefficient varies with measurement error and other basic model parameters. Our method successfully corrects for known measurement error, and can gauge the sensitivity of parameter estimates to errors in variables. Our method can be applied to the studies of lending discrimination cited in the introduction. It can also be applied to the empirical models employed in lending discrimination suits and regulatory examinations. Since the stakes are high in such applications, the models ought to be routinely tested for sensitivity to errors in variables.

Further extensions of our method would be worthwhile. Although we allow for errors only in continuous-valued independent variables, studies of lending

Table 4 Race Coefficient Estimates for Alternative Correlation and Error-Variance Assumptions, $k = 10$

$\alpha = 0$, β is a $k \times 1$ vector of -1s, μ is a $k \times 1$ vector of 1s, Σ is a $k \times k$ matrix with 1s on the diagonal and off-diagonal elements equal to ρ, D is a $k \times k$ matrix with 0.1s on the diagonal and off-diagonal elements equal to 0, $\tilde{D}$ is a $k \times k$ matrix with elements $\tilde{d}$ on the diagonal and off-diagonal elements equal to 0, and $n = 10,000$.

	a
A. True parameter $\rho = 0$:	
Assumed $\tilde{d}$	
0.0	1.0033
	(3.3154)
0.1	−0.0339
	(−0.1006)
B. True parameter $\rho = 0.5$:	
Assumed $\tilde{d}$	
0.0	0.2266
	(3.4658)
0.1	0.0645
	(0.5988)

Notes: t-statistics are shown in parentheses beneath the coefficient estimate. For each panel, we drew a set of 10,000 random realizations for (y, x): 5,000 with $z = 0$ and 5,000 with $z = 1$. Within each panel, estimation was performed on the same data set.

discrimination often include discrete variables that are likely to be fallible as well. It would be worthwhile to allow for errors in the discrete variables, as Klepper (1988a) does for the linear regression model. In addition, it would be useful to allow for uncertainty about the nuisance distributional parameters that our method treats as known.

REFERENCES

Bauer, Paul W., and Brian A. Cromwell. "A Monte Carlo Examination of Bias Tests in Mortgage Lending," Federal Reserve Bank of Cleveland *Economic Review*, vol. 30 (July/August/September 1994), pp. 27–44.

Berkovec, James, Glenn Canner, Stuart Gabriel, and Timothy Hannan. "Race, Redlining, and Residential Mortgage Loan Performance," *Journal of Real Estate Finance and Economics*, vol. 9 (November 1994), pp. 263–94.

Cummins, Claudia. "Fed Using New Statistical Tool to Detect Bias," *American Banker,* June 8, 1994.

Friedman, Milton. *A Theory of the Consumption Function*. Princeton, N.J.: Princeton University Press, 1957.

Goldberger, Arthur S. "Reverse Regression and Salary Discrimination," *Journal of Human Resources,* vol. 19 (Summer 1984), pp. 293–318.

Hashimoto, Masanori, and Levis Kochin. "A Bias in the Statistical Estimation of the Effects of Discrimination," *Economic Inquiry,* vol. 18 (July 1980), pp. 478–86.

Hausman, J. A., W. K. Newey, and J. L. Powell. "Nonlinear Errors in Variables: Estimation of Some Engel Curves," *Journal of Econometrics,* vol. 65 (January 1995), pp. 205–33.

Kapteyn, Arie, and Tom Wansbeek. "Identification in the Linear Errors in Variables Model," *Econometrica,* vol. 51 (November 1983), pp. 1847–49.

Klepper, Steven. "Bounding the Effects of Measurement Error in Regressions Involving Dichotomous Variables," *Journal of Econometrics,* vol. 37 (March 1988a), pp. 343–59.

__________. "Regressor Diagnostics for the Classical Errors-in-Variables Model," *Journal of Econometrics,* vol. 37 (February 1988b), pp. 225–50.

__________, and Edward E. Leamer. "Consistent Sets of Estimates for Regressions with Errors in All Variables," *Econometrica,* vol. 52 (January 1984), pp. 163–83.

Levine, David K. "Reverse Regressions for Latent-Variable Models," *Journal of Econometrics*, vol. 32 (July 1986), pp. 291–92.

Maddala, G. S. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press, 1983.

Munnell, Alicia H., Lynn E. Browne, James McEneaney, and Geoffrey M. B. Tootell. "Mortgage Lending in Boston: Interpreting the HMDA Data," Working Paper Series No. 92. Boston: Federal Reserve Bank of Boston, 1992.

Rao, Potluri. "Some Notes on the Errors-in-Variables Model," *American Statistician,* vol. 27 (December 1973), pp. 217–28.

Schill, Michael H., and Susan M. Wachter. "Borrower and Neighborhood Racial and Income Characteristics and Financial Institution Mortgage Application Screening," *Journal of Real Estate Finance and Economics,* vol. 9 (November 1994), pp. 223–39.

Seiberg, Jaret. "When Justice Department Fights Bias by the Numbers, They're His Numbers," *American Banker,* September 14, 1994.