# Misclassification in Binary Choice Models

Bruce D. Meyer*
University of Chicago and NBER

Nikolas Mittag*
CERGE-EI

March 8, 2016

**Abstract:** While measurement error in the dependent variable of a regression model does not lead to bias in some cases, with a binary dependent variable the bias can be pronounced. We examine what can still be learned from such contaminated data. First, we derive the asymptotic bias in parametric models allowing measurement errors to be correlated with both observables and unobservables. Using simulations and administrative records on food stamp receipt linked to survey data as validation data, we show that the bias formulas are accurate in finite samples and imply a tendency to attenuation. Second, we examine the bias in a prototypical application (receipt of food stamps) using two validation datasets and several methods to account for misclassification. Estimators that are consistent when misclassification is independent of the covariates (conditional on the true outcome) aggravate the bias if this assumption is invalid in all cases we examine. Estimators that relax this assumption perform well if a true model of misclassification or validation data are available.

**Keywords:** measurement error; binary choice models; program take-up; food stamps;
**JEL Classification Numbers:** C35, C81, H53

*Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. We would like to thank Frank Limehouse for assistance with the data and participants at presentations at the Census Research Data Center Conference, UNCE at Charles University and Xiamen University for their comments. Mittag would like to thank the Czech Science Foundation for financial support from grant no. 16-07603Y and the Czech Academy of Sciences for funding during early stages of this project through institutional support RVO 67985998. Meyer: Harris School of Public Policy, University of Chicago, 1155 E. 60th Street, Chicago, IL 60637, bdmeyer@uchicago.edu; Mittag: CERGE-EI, a joint workplace of Charles University in Prague and the Economics Institute of the Czech Academy of Sciences, Politických vězňů 7, Prague 1, 111 21, Czech Republic, nikolas.mittag@cerge-ei.cz

# 1 Introduction

Many important outcomes are binary such as program receipt, labor market status, and educational attainment. These outcomes are frequently misclassified due to interviewer or respondent error or for other reasons. It is a common misconception that measurement error in a dependent variable does not lead to bias, but this requires classical measurement error. Misclassification of a binary variable is necessarily non-classical measurement error, and thus leads to bias. However, there are few general results on bias in binary choice models. Yet, given the pervasiveness of misclassification in common data sources, it is important to know whether we can still learn from contaminated data and which methods allow it. To address this issue, we first examine the properties of binary choice models with measurement error in the dependent variable. We then discuss the performance of several estimators designed to account for misclassification. We rely on a combination of analytical results, simulations, and results from an application to the Food Stamp Program.

Several papers that examine misreporting in surveys have found high rates of misclassification in binary variables such as participation in welfare programs (Marquis and Moore, 1990; Meyer, Mok and Sullivan, 2009; Meyer, Goerge and Mittag, 2015), Medicaid enrollment (Call et al., 2008; Davern et al., 2009) and education (Black, Sanders and Taylor, 2003). Bound, Brown and Mathiowetz (2001) provide an overview. In the case of program reporting, false negatives, i.e. recipients who fail to report receiving program benefits, are the main problem, with rates of underreporting sometimes exceeding 50%. Measurement error can badly bias substantive studies with binary outcomes such as those examining take-up of government programs (e.g. Bitler, Currie and Scholz, 2003; Haider, Jacknowitz and Schoeni, 2003), labor market status (e.g. Poterba and Summers, 1995) or educational attainment (e.g. Eckstein and Wolpin, 1999; Cameron and Heckman, 2001). A frequent cause of error besides misreporting is subjective classification of a dependent variable, for example whether there is a recession or not (e.g. Estrella and Mishkin, 1998), the presence of an armed civil conflict (e.g Fearon and Laitin, 2003) or whether an individual is disabled or not (e.g. Benítez-Silva

et al., 2004; Kreider and Pepper, 2008). Similarly, a proxy variable is often used instead of the true variable of interest, such as arrests or incarcerations instead of crimes (e.g. Lochner and Moretti, 2004). Another reason for misclassification is prediction error, for example, if some observations of a variable are missing and imputed values are substituted. Moreover, there is little ex ante reason to believe that misclassification is independent of the covariates, not even conditional on the true value of the binary variable, though it could occur if misclassification stems from coding errors or failure to link some records.

A few papers have analyzed the consequences of misclassification for econometric estimates. For example, Bollinger and David (1997, 2001) and Meyer, Goerge and Mittag (2015) examine how misclassification affects estimates of food stamp participation and Davern et al. (2009) analyze Medicaid enrollment. They show that misclassification affects the estimates of common econometric models and distorts conclusions in meaningful ways. From these studies we know that misclassification can seriously alter estimates from binary choice models, but we know very little about the ways it affects estimates in general. This situation is aggravated by the scarcity of analytic results on bias in binary choice models. Carroll et al. (2006) and Chen, Hong and Nekipelov (2011) provide overviews of the literature on measurement error in non-linear models and there is a small body of literature on misspecification in binary choice models (e.g. Yatchew and Griliches, 1985; Ruud, 1983, 1986), but general results or formulas for biases are scarce.

Thus, the literature has established that misclassification is pervasive and affects estimates, but not how it affects them or what can still be done with contaminated data. This paper characterizes the consequences of misclassification of the dependent variable in binary choice models and assesses whether substantive conclusions can still be drawn from the observed data and if so, which methods to do so work well. We first present a closed form solution for the bias in the linear probability model that allows for simple corrections. For non-linear binary choice models such as the Probit model, we decompose the asymptotic bias into four components. We derive closed form expressions for three bias components and an

equation that determines the fourth component. The formulas imply that if misclassification is conditionally random, only the probabilities of misclassification are required to obtain the exact bias in the linear probability model and an approximation in the Probit model. If misclassification is related to the covariates, additional information on this relation is required to assess the (asymptotic) bias, but the results still imply a tendency for the bias to be in the opposite direction of the sign of the coefficient.

Next, we examine whether our analytic results help to explain the bias found in empirical applications and whether one can use them to interpret coefficient estimates from contaminated data. We conduct simulations and estimate models of food stamp receipt in two unique data sets that include "true" food stamp receipt from administrative data along with the survey reports. Overall, the results underscore that our formulas can be used to judge whether substantive conclusions obtained from misclassified data are likely to be valid. They show that the signs of coefficients are robust to a wide range of misclassification mechanisms and that there is a tendency for the coefficients to be attenuated, though this result is not expected to hold in all cases. The results suggest that in some cases one can learn about coefficient signs from the contaminated data.

Finally, we evaluate six estimators that take misclassification into account to examine whether it is feasible to obtain consistent parameter estimates from data with misclassification. Several such estimators have been introduced in the literature (e.g. Bollinger and David, 1997; Hausman, Abrevaya and Scott-Morton, 1998), but unless the true parameters are known, it is impossible to know whether these estimators improve parameter estimates or only change them. We use the same data with a measure of "truth", and model food stamp participation as above. If misclassification is conditionally random, it is feasible to obtain consistent parameter estimates from the observed data. We find that incorporating information on the misclassification rates greatly improves the robustness of the estimates to misspecification. However, applying an estimator that assumes misclassification to be conditionally random can make estimates substantively worse when this assumption is false.

In such cases, one can still obtain consistent estimates if validation data or a model of misclassification is available. Our results suggest that such additional information is still likely to improve parameter estimates even if it is not completely accurate.

In summary, our results underline that misclassification can lead to severe bias. Nonetheless, the observed data remain informative about true parameters, particularly when misclassification is conditionally random or additional information on the misclassification model is available. The next section introduces the models and discusses the bias in theory, while section 3 examines it in practice using the matched survey data and simulations. Section 4 introduces consistent Probit estimators and evaluates their performance when misclassification is unrelated to the covariates (4.1) and when it is related to them (4.2). Section 5 concludes.

# 2 Bias Due to Misclassification of a Binary Dependent Variable

We are concerned with a situation in which a binary outcome $y$ is related to observed characteristics $X$, but the outcome indicator is subject to misclassification. Let $y_i^T$ be the true indicator for the outcome of individual $i$ and $y_i$ be the observed indicator that is subject to misclassification. The sample size is $N$ and $N_{MC}$ observations are misclassified, $N_{FP}$ of which are false positives and $N_{FN}$ are false negatives. We define the probabilities of false positives and false negatives conditional on the true response as

$$\Pr(y_i = 1 | y_i^T = 0) = \alpha_{0i}$$
$$\Pr(y_i = 0 | y_i^T = 1) = \alpha_{1i}$$

We refer to them as the conditional probabilities of misclassification. Additionally, we define a binary random variable $M$ that equals one if the outcome of individual $i$ is misclassified

$$m_i = \begin{cases} 0 & \text{if } y_i^T = y_i \\ 1 & \text{if } y_i^T \neq y_i \end{cases}$$

We consider two cases, the linear probability model and the Probit model. For the linear probability model, $E(y^T|X) = X\beta$, so one would like to run the following OLS regression

$$y_i^T = x_i'\beta^{LPM} + \varepsilon_i^{LPM}$$

to obtain the $K$-by-1 vector $\hat{\beta}^{LPM}$. Using only the observed data yields

$$y_i = x_i'\tilde{\beta}^{LPM} + \tilde{\varepsilon}_i^{LPM}$$

The Probit model can be motivated by a latent variable $y_i^{T*}$ such that

$$y_i^T = 1\{y_i^{T*} = x_i'\beta + \varepsilon_i \geq 0\} \tag{1}$$

where $\varepsilon_i$ is drawn independently from a standard normal distribution and $\beta$ is the K-by-1 coefficient vector of interest. Extending our results to other binary choice models in which $\varepsilon_i$ is drawn from a different distribution is straightforward. Estimating a Probit model using the observed indicator $y_i$ instead of $y_i^T$ yields $\hat{\tilde{\beta}}$, which is potentially inconsistent. Little is known about the effects of measurement error in non-linear models (see Carroll et al., 2006). While some of the papers mentioned above propose consistent estimation strategies and show that ignoring the problem leads to inconsistent estimates, they do not discuss the nature of this inconsistency. Hausman, Abrevaya and Scott-Morton (1998) assume that the probabilities of false negatives and false positives conditional on the true response are constants for all individuals, i.e. $\alpha_{0i} = \alpha_0$ and $\alpha_{1i} = \alpha_1$ for all $i$. We refer to this kind of misclassification

as "conditionally random", because conditional on the true value, $y_i^T$, misclassification is independent of the covariates $X$. Hausman, Abrevaya and Scott-Morton (1998) show that under this assumption the marginal effects in the observed data are proportional to the true marginal effects

$$\frac{\partial \Pr(y=1|x)}{\partial x} = (1-\alpha_0-\alpha_1) f(x'\beta)\beta \tag{2}$$

where $f()$ is the derivative of the link function (e.g. the normal cdf in the Probit model), so that $f(x'\beta)\beta$ are the true marginal effects. As long as $\alpha_0 + \alpha_1 < 1$, the marginal effects are attenuated: they are smaller in absolute value, but retain the correct signs.

If one has consistent estimates of the marginal effects in the observed data, equation (2) implies that they are all attenuated proportionally, which suggests that inference based on coefficient ratios may be valid. If one also has consistent estimates of the probabilities of misclassification, $\hat{\alpha}_0$ and $\hat{\alpha}_1$, one can use $(1-\hat{\alpha}_0-\hat{\alpha}_1)^{-1}\widehat{\frac{\partial \Pr(y=1|x)}{\partial x}}$ to consistently estimate the true marginal effects. However, while estimating the marginal effects in the observed data is often possible using semi- or non-parametric estimators, we show below that using the Probit marginal effects from the observed data in (2) usually yields inconsistent estimates. Section 3 discusses this further and examines the extent to which equation (2) is useful in practice.

## 2.1 Bias in the Linear Probability Model

Measurement error in binary variables is a form of non-classical measurement error (Aigner, 1973; Bollinger, 1996). The bias in OLS coefficients when the dependent variable is subject to non-classical measurement error is the coefficient in the (usually infeasible) regression of the measurement error on the covariates (Bound, Brown and Mathiowetz, 2001). Our dependent variable is binary, so the measurement error takes the following simple form:

$$u_i = y_i - y_i^T = \begin{cases} -1 & \text{if } i \text{ is a false negative} \\ 0 & \text{if } i \text{ reported correctly} \\ 1 & \text{if } i \text{ is a false positive} \end{cases}$$

Consequently, the coefficient in this OLS regression $X$ (if it were feasible) would be:

$$\hat{\delta} = (X'X)^{-1}X'u \tag{3}$$

$\hat{\delta}$ can only be zero if the measurement error is uncorrelated with $X$, which is unlikely: If a variable is a relevant regressor, $\Pr(y = 1)$ is a function of $X$. Since $u = -1$ can only occur if $y = 1$, this creates a dependence between $u$ and $X$.[1] Equation (3) implies that the coefficient in an OLS regression of the misclassified indicator on $X$, $\hat{\tilde{\beta}}^{LPM}$, is $\hat{\beta}^{LPM} + \hat{\delta}$. So the bias is

$$\mathbb{E}(\hat{\tilde{\beta}}^{LPM}) - \beta^{LPM} = \mathbb{E}(\hat{\delta}) \tag{4}$$

The measurement error only takes on three values, so equation (3) simplifies to

$$\hat{\delta} = (X'X)^{-1}(N_{FP}\bar{x}_{FP} - N_{FN}\bar{x}_{FN}) = N(X'X)^{-1}\left(\frac{N_{FP}}{N}\bar{x}_{FP} - \frac{N_{FN}}{N}\bar{x}_{FN}\right) \tag{5}$$

where $\bar{x}_{FP}$ and $\bar{x}_{FN}$ are the means of $X$ among the false positives and false negatives. Appendix A provides more detail on the derivation. Consequently in expectation[2]

$$\begin{aligned}\mathbb{E}(\hat{\delta}) = N(X'X)^{-1}[&\Pr(y = 1, y^T = 0)\mathbb{E}(X|y = 1, y^T = 0) \\ &- \Pr(y = 0, y^T = 1)\mathbb{E}(X|y = 0, y^T = 1)]\end{aligned} \tag{6}$$

That is, the bias in $\hat{\tilde{\beta}}^{LPM}$ depends on the difference between the conditional means of $X$ among false positives and false negatives where these conditional means are weighted by the probability of observing a false negative or positive. The bias is this vector of differences pre-multiplied by the inverse of the covariance matrix of the data.

Consequently, misclassifying the dependent variable from 1 to 0 at higher values of a particular variable, while holding everything else fixed, decreases the estimated coefficient

[1] However, if misclassification conditional on $y^T$ depends on $X$ in a peculiar way, $X'u$ can still be 0.
[2] This assumes that $X$ is non-stochastic. The extension to the stochastic case is straightforward.

on that particular variable, while misclassifying it from 0 to 1 increases it. The opposite effect occurs at lower values of the variable. Misclassifying more observations amplifies this effect. The bias can only be zero in knife-edge cases in which the expression in brackets is 0. Neither equal probabilities of misclassification nor (conditional) independence of $X$ and misclassification are sufficient for the bias to be zero. Equation (6) only depends on the probabilities of misclassification and the conditional means of the covariates, so one only needs these quantities (or an estimate of the expectation of $\hat{\delta}$) to correct the bias or assess its likely direction and magnitude.

If the conditional probabilities of misclassification are constants as in Hausman, Abrevaya and Scott-Morton (1998), the results above simplify to $\mathbb{E}(\hat{\beta}_k^{LPM}) = (1-\alpha_0-\alpha_1)\beta_k^{LPM}$ for the slope coefficients.[3] So if the true model is a linear probability model and misclassification is not correlated with $X$, knowing the conditional probabilities of misclassification is enough to correct both coefficients and marginal effects.

## 2.2 Asymptotic Bias in the Probit Model

No general result on the consequences of measurement error in the dependent variable exists for non-linear models. We first show that misclassification of the dependent variable is equivalent to a specific form of omitted variable bias. We then use results on the effect of omitting variables from Yatchew and Griliches (1984, 1985) to decompose the asymptotic bias due to misclassification. The results below are for the Probit model, but the extension to other binary choice models such as the Logit is straightforward. The true data generating process without misclassification is assumed to be given by equation (1). Thus, with $m_i$ the indicator of misclassification, the data generating process with misclassification is

$$y_i = \begin{cases} 1\{x_i'\beta + \varepsilon_i \geq 0\} & \text{if } m_i = 0 \\ 1\{x_i'\beta + \varepsilon_i \leq 0\} & \text{if } m_i = 1 \end{cases} \qquad \Leftrightarrow$$

[3]For the intercept: $\mathbb{E}(\hat{\beta}_0^{LPM}) = \alpha_0 + (1-\alpha_0-\alpha_1)\beta_0^{LPM}$. See appendix A for proof.

$$y_i = \begin{cases} 1\{x_i'\beta + \varepsilon_i \geq 0\} & \text{if } m_i = 0 \\ 1\{-x_i'\beta - \varepsilon_i \geq 0\} & \text{if } m_i = 1 \end{cases} \tag{7}$$

Therefore, the true data generating process has the following latent variable representation:

$$\begin{aligned} y_i^* &= (1 - m_i)(x_i'\beta + \varepsilon_i) + m_i(-x_i'\beta - \varepsilon_i) \qquad \Leftrightarrow \\ y_i^* &= \underbrace{x_i'\beta + \varepsilon_i}_{\text{Well-Specified Probit Model}} \quad \underbrace{-2m_i x_i'\beta - 2m_i\varepsilon_i}_{\text{Omitted Variable}} \end{aligned} \tag{8}$$

The first two terms form a well specified Probit, because $\varepsilon_i$ is not affected by misclassification, so it is still a standard normal variable. We can decompose each of the omitted variable terms into its linear projection on $X$ and deviations from it:

$$\begin{aligned} -2m_i x_i'\beta &= x_i'\lambda + \nu_i \\ -2m_i\varepsilon_i &= x_i'\gamma + \eta_i \end{aligned} \tag{9}$$

Substituting this back into the equation (8) gives:

$$y_i^* = x_i' \underbrace{(\beta + \lambda + \gamma)}_{\text{biased coefficient}} + \underbrace{\varepsilon_i + \nu_i + \eta_i}_{\text{misspecified error term}} = x_i'\tilde{\beta} + \tilde{\varepsilon}_i \tag{10}$$

Equation (10) implies that the observed data do not conform to the assumptions of a Probit model unless $\tilde{\varepsilon}_i$ is drawn independently from a normal distribution that is identical for all $i$. While $\tilde{\varepsilon}$ is uncorrelated with $X$ and has a mean of zero by construction, it is unlikely to have constant variance and cannot come from a normal distribution.[4] Consequently, running a Probit on the observed data does not yield consistent estimates of the marginal effects in the observed data, so that using equation (2) to obtain estimates of the true marginal effects is inconsistent.

In summary, equation (10) highlights three violations of the assumptions of the original Probit model: First, the linear projection of the latent variable is $X\tilde{\beta}$ instead of $X\beta$. Second,

[4]That they cannot be normal can be seen from the fact that the omitted variables have point mass at 0.

the variance of the misspecified error term $\tilde{\varepsilon}$ is different from the variance of the true error term $\varepsilon$. Finally, $\tilde{\varepsilon}$ is not drawn from a normal distribution that is identical for all observations. The next sections discuss the implications of the violation of these assumptions for maximum likelihood estimates of $\beta$. We start by deriving an expression for the estimate of $\tilde{\beta}$ that one would obtain if $\tilde{\varepsilon}$ were a regular iid normal error term. We then apply results from the literature on functional form misspecification in binary choice models (Yatchew and Griliches, 1985) to derive how parameter estimates of a Probit model differ from $\tilde{\beta}$ in a second step.

### 2.2.1 Bias in the Linear Projection

The first component of the asymptotic bias is the result of the coefficients on $X$ incorporating the linear projection of the omitted terms. The linear projection has two parts that are analogous to the two bias terms Bound et al. (1994) derive for linear models. The first term arises from a relation between misclassification and the covariates $X$. The second part stems from a relation of misclassification and the error term $\varepsilon$. The familiar linear projection formula gives

$$\hat{\lambda} = -2(X'X)^{-1}X'SX\beta \tag{11}$$

where $S$ is an $N$-by-$N$ matrix with indicators for misclassification on the diagonal. Equation (11) shows that $\hat{\lambda}$ can be interpreted as minus twice the coefficient on $X$ when regressing a variable that equals the linear index $X\beta$ for misclassified observations and 0 for correctly reported observations on $X$. Under the usual Probit assumptions, $N^{-1}X'X$ converges to the uncentered covariance matrix of $X$. Let $\operatorname{plim}_{N\to\infty} N^{-1}X'X = Q$. Additionally, we define the probability limit of the uncentered covariance matrix of $X$ among the misclassified observations as $\operatorname{plim}_{N\to\infty} N_{MC}^{-1}(X'X|M=1) = Q_{MC}$. A typical element $(r,c)$ of $X'SX$ is $\sum_{i=1}^{N} x_{ri}m_i x_{ci}$, whereas a typical element $(r,c)$ of $X'X$ is $\sum_{i=1}^{N} x_{ri}x_{ci}$. From the sums in $X'X$, $S$ selects only

the $x_i$ that belong to misclassified observations, so that

$$\operatorname*{plim}_{N\to\infty}\frac{1}{N}X'SX = \operatorname*{plim}_{N\to\infty}\frac{N_{MC}}{N}\operatorname*{plim}_{N\to\infty}N_{MC}^{-1}\left(X'X|M=1\right) = \Pr(M=1)Q_{MC}$$

i.e. it converges to the uncentered covariance matrix of $X$ among those that are misclassified multiplied by the probability of misclassification. Thus, the probability limit of $\hat{\lambda}$ is

$$\operatorname*{plim}_{N\to\infty}\hat{\lambda} = -2\Pr(M=1)Q^{-1}Q_{MC}\beta \tag{12}$$

Equation (12) shows that the asymptotic bias from this source cannot be zero for all coefficients if there is any misclassification, i.e. if $\Pr(M=1)\neq 0$. Otherwise, both right hand side matrices are positive definite, so lambda has positive rank, i.e. it contains non-zero elements. Thus, while some elements of $\lambda$ can be 0 in special cases, misclassification always induces bias in some coefficients. Multiplication by $-2\Pr(M=1)$ creates a tendency for the asymptotic bias to be in the direction opposite from the sign of the coefficient, which reduces to the rescaling effect if misclassification is not related to $X$. This effect can be amplified or reduced by $Q^{-1}Q_{MC}$, which is due to the relation of misclassification to $X$. Both matrices are positive definite, so the diagonal elements are positive, which creates a tendency for $\lambda$ and $\beta$ to have different signs, causing the asymptotic bias to be in the opposite direction from the sign of the coefficient. However, unless the off-diagonal elements are zero, bias from other coefficients "spreads" and may reverse this tendency.

In summary, the magnitude of the first component of the asymptotic bias depends on three things: all else equal, it is larger if the probability of misclassification is larger, if misclassification comes from a wider range of $X$, or is more frequent among extreme values of $X$. The second point follows from the fact that in such cases the conditional covariance matrix is large relative to the full covariance matrix. The third effect is due to the covariance matrices being uncentered, so if the mean of $X$ among the misclassified observations differs much from that in the general sample, the asymptotic bias will be larger.

The second component of the bias in the linear projection stems from the fact that misclassification may create a relation between $X$ and the error term. Using the standard formula for linear projections, equation (9) and the definition of $S$ from above, it is:

$$\hat{\gamma} = -2(X'X)^{-1}X'S\varepsilon \tag{13}$$

$\hat{\gamma}$ can also be interpreted as minus twice the regression coefficient on $X$ when regressing a vector that contains $\varepsilon_i$ for misclassified observations and zeros for all other observations on $X$. Using exactly the same arguments as above yields

$$\operatorname*{plim}_{N\to\infty}\hat{\gamma} = -2\Pr(M=1)Q^{-1}\operatorname*{plim}_{N\to\infty}N^{-1}(X'\varepsilon|M=1) \tag{14}$$

While $\operatorname*{plim}_{N\to\infty}N^{-1}X'\varepsilon = 0$ by assumption, this restriction does not determine the *conditional* covariance between $X$ and the error term, $\operatorname*{plim}_{N\to\infty}N^{-1}(X'\varepsilon|M=1)$. The conditional covariance and thus $\operatorname*{plim}_{N\to\infty}\hat{\gamma}$ are 0 if besides the assumed independence between $X$ and $\varepsilon$ it is also true that $\varepsilon$ and $M$ are independent. If $X$ is independent of $\varepsilon$ and $M$ and the model includes an intercept, this bias component does not affect the slope coefficients. However, if the probability of misclassification depends on the true value $y^T$, because the determinants of false positives and false negatives differ, the bias is unlikely to be 0. In this case, $M$ can only be independent of $X$ *or* $\varepsilon$ and is likely to depend on both.

### 2.2.2 Rescaling Bias

The second effect of misclassification is a rescaling effect that always occurs when misspecification affects the variance of the error term in Probit models. The coefficients of the latent variable model are only identified up to scale, so one normalizes the variance of the error term to one, which normalizes the coefficients to $\beta/\sigma_\varepsilon$. Consequently, misspecification that affects the variance of the error term leads to coefficients with the wrong scale. In the absence of the additional asymptotic bias discussed below (i.e. if $\tilde{\varepsilon}$ were iid normal), estimating (10)

by a Probit model gives

$$\operatorname*{plim}_{N\to\infty}\hat{\tilde{\beta}} = \frac{\tilde{\beta}}{SD(\tilde{\varepsilon})} = \frac{\beta+\lambda+\gamma}{SD(\varepsilon+\nu+\eta)} \equiv \bar{\beta} \tag{15}$$

One may expect the error components due to misclassification to increase the variance of the error term, i.e. $SD(\tilde{\varepsilon}) > SD(\varepsilon)$, so the rescaling will tend to result in an asymptotic bias towards zero. However, the variance can decrease if misclassification depends on $\varepsilon$. The rescaling factor is the same for all coefficients, so it does not affect their relative magnitudes.

### 2.2.3 Bias Due to Misspecification of the Error Distribution

If $\tilde{\varepsilon}$ were iid normal with constant variance, estimating equation (10) by a Probit model would yield a consistent estimate of $\bar{\beta}$ as given by (15). However, additional asymptotic bias may arise from heteroskedasticity or differences between the higher order moments of the distribution of the misspecified error term and those of the normal distribution. Ruud (1983, 1986) characterizes this bias, but closed form solutions do not exist. Adapting a result from Yatchew and Griliches (1985) provides an implied formula for the exact asymptotic bias. Taking the probability limit of $N^{-1}$ times the first order conditions of the log-likelihood function, the parameter estimate converges to the vector b that solves

$$\int f_X(x_j)\frac{x_j'\phi(x_j'b)}{\Phi(x_j'b)(1-\Phi(x_j'b))}\left[F_{\tilde{\varepsilon}|X=x_j}(-x_j'\bar{\beta}) - \Phi(-x_j'b)\right]\,\mathrm{d}x_j = 0 \tag{16}$$

where $F_{\tilde{\varepsilon}|X=x_j}$ is the conditional cumulative distribution function of $\tilde{\varepsilon}/Var(\tilde{\varepsilon})$, i.e. the misspecified error term normalized to have (unconditional) variance 1, evaluated at $X = x_j$. Consequently, $F_{\tilde{\varepsilon}|X=x_j}(-x_j'\bar{\beta})$ provides the probability that $\tilde{\varepsilon}_i/Var(\tilde{\varepsilon})$ is smaller than $-x_j'\bar{\beta}$ in the sub-population with a specific value of the covariates ($X = x_j$). Thus, it provides the probability of observing $y = 1$ when drawing from the sub-populations with covariates equal to $x_j$. Note that the left hand side of (16) is the first derivative of a concave function, so the equation has a unique solution.

If $F_{\tilde{\varepsilon}|X}$ is a normal cdf with the same variance for all values of $X$, $b = \bar{\beta}$ solves (16) so that (15) gives the exact probability limit of the coefficient estimate. Unfortunately, (16) has no closed form solution and can only be solved numerically for specific cases of $F_{\tilde{\varepsilon}|X}$ which is usually unknown. Note, however, that the unconditional distributions, $F_{\tilde{\varepsilon}}$ and $\Phi$ have the same first and second moments by construction. Consequently, (asymptotic) deviations of the parameter estimates from $\bar{\beta}$ only occur due to a dependence of the first two moments of $F_{\tilde{\varepsilon}}$ on $X$ (e.g. heteroskedasticity) and differences in higher order moments of the two distributions (so we refer to this bias component as the "higher order bias").

If one has reasons to believe that the bias due to functional form misspecification is small (such as in Ruud, 1983, 1986), equation (15) provides a tractable approximation to the inconsistent coefficients and thereby makes the asymptotic bias easier to analyze in practice. We find the effect of misspecification of the error term to be small in Monte Carlo simulations, but the bias can become large if misclassification depends heavily on $\varepsilon$ or the true value of $y$. The web appendix discusses how to further assess the likely direction and severity of this component of the asymptotic bias and conditions under which one can expect (15) to provide a good approximation.

# 3 The Bias in Practice

We use administrative data matched to two surveys to illustrate the applicability of the results from the previous section and examine what can still be learned from the observed data. Misclassification is correlated with the covariates in the matched data, but we also conduct a Monte Carlo study in which we induce conditionally random misclassification in our matched sample. Finally, we use simulated data to assess the size of the components of the bias in the Probit model in more detail.

## 3.1 Bias from Misclassification in Models of Food Stamp Take-up

We use the data employed in Meyer, Goerge and Mittag (2015): the 2001 American Community Survey (ACS) and the 2002-2005 Current Population Survey Annual Social and Economic Supplement (CPS ASEC) matched to administrative food stamp data from Illinois and Maryland. We successfully link more than 90 percent of households in the ACS and more than 70 percent of households in the CPS to the administrative data. We estimate the probability of a household being linked and use inverse probability weighting (Wooldridge, 2007) to adjust the household weights in order to keep the sample representative of the population in the two states. We weight to make the analysis comparable to common applications and allow the reader to interpret the estimates and biases as population summary statistics. We treat the administrative food stamp indicator as truth, even though it may contain errors, e.g. due to an imperfect match to the survey data. Given that there should be few mistakes in the administrative records and the match rate is high, this assumption seems plausible. Additionally, most of the analysis below does not require this assumption - the survey data can be considered a version of the administrative data with misclassification even if neither of them represent "truth". We restrict the sample to matched households with income less than 200% of the federal poverty line and estimate simple Probit and linear probability models of food stamp take-up with three covariates: a continuous poverty index (income/poverty line) as well as dummies for whether the householder is 50 or older and whether the household is in Maryland.[5] Misreporting is related to all three variables in both surveys. False negatives are more likely for higher income households, while both households in Maryland and those with a head over the age of 50 are less likely to be false positives and more likely to be false negatives. See Meyer, Goerge and Mittag (2015) for further details on the data, the matching process and the determinants of misreporting.

The closed form solutions for the bias in the linear probability model are straightforward

[5]Income is also known to be measured with error, but we focus on the consequences of measurement error in the dependent variable here.

Table 1: Estimated Bias in the Linear Probability Model, ACS

| | $\hat{\beta}^{LPM}$ Matched Data | $\hat{\beta}^{LPM}$ Survey Data | Bias in MC Study (random MR) | Bias in Survey Data | (4)-(3): Bias due to correlation |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Poverty index | -0.0018 | -0.0019 | -27.94% | 5.56% | 33.50% |
| | (0.0001) | (0.0001) | | | |
| Age≥50 | -0.1166 | -0.1046 | -28.93% | -10.29% | 18.64% |
| | (0.0145) | (0.0133) | | | |
| Maryland | -0.0034 | -0.0217 | -10.03% | 538.24% | 548.26% |
| | (0.0157) | (0.0141) | | | |

*Note:* Sample size: 5945 matched households from IL and MD with income less than 200% of the federal poverty line. All analyses include a constant term (not reported) and are conducted using household weights adjusted for match probability. All estimated biases are in % of the coefficient from the matched data. The first two columns report OLS coefficients using FS receipt according to the administrative data (column 1) and according to survey reports (column 2) as the dependent variable. In the MC design in column 3, the dependent variable is administrative FS receipt with misreporting induced with the misreporting probabilities observed in the actual sample (Pr(FP)=0.02374 and Pr(FN)=0.2596), but independent of the covariates conditional on true receipt. 500 replications are performed and we report the average bias in %.

to analyze. The results in table 1 for the ACS and table 2 for the CPS conform to the expectations from section 2.1. The first column presents the "true" coefficient estimate from the matched data and the second column contains the inconsistent estimates from the survey. In the simulations in column 3, the dependent variable is administrative food stamp receipt with misclassification induced with the probabilities observed in the actual samples, but independent of the covariates conditional on true receipt. This exercise leaves the remainder of the data unchanged, so it allows us to focus on conditionally random misclassification. We perform 500 replications. The factor of attenuation in column 3 is very similar across slopes. In both surveys, it is close to $\alpha_0+\alpha_1$ as expected. The coefficient on the Maryland dummy in the ACS is an exception due to the fact that it is imprecisely estimated and indistinguishable from 0. Column 4 shows that in the survey data, where misclassification is related to the covariates, the factor differs substantially between coefficients and is different from $\alpha_0+\alpha_1$. Since the only difference in the data used for column 3 and 4 is the correlation between

Table 2: Estimated Bias in the Linear Probability Model, CPS

| | $\hat{\beta}^{LPM}$ Matched Data | $\hat{\beta}^{LPM}$ Survey Data | Bias in MC Study (random MR) | Bias in Survey Data | (4)-(3): Bias due to correlation |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Poverty index | -0.0023 | -0.0021 | -42.18% | -8.70% | 33.48% |
| | (0.0002) | (0.0001) | | | |
| Age≥50 | -0.1264 | -0.0985 | -41.79% | -22.07% | 19.72% |
| | (0.0174) | (0.0151) | | | |
| Maryland | -0.0937 | -0.0706 | -42.60% | -24.65% | 17.95% |
| | (0.0184) | (0.0156) | | | |

*Note:* Sample size: 2791 matched households from IL and MD with income less than 200% of the federal poverty line. All analyses include a constant term (not reported) and are conducted using household weights adjusted for match probability. All estimated biases are in % of the coefficient from the matched data. The first two columns report OLS coefficients using FS receipt according to the administrative data (column 1) and according to survey reports (column 2) as the dependent variable. In the MC design in column 3, the dependent variable is administrative FS receipt with misreporting induced with the misreporting probabilities observed in the actual sample (Pr(FP)=0.03271 and Pr(FN)=0.3907), but independent of the covariates conditional on true receipt. 500 replications are performed and we report the average bias in %.

misclassification and the covariates, the difference between the columns is an estimate of the effect of the correlation of misclassification with the covariates. This difference is presented in column 5 and in our case biases all coefficients away from 0. Thus, it is in the opposite direction of the estimated bias in the random case, i.e. the two effects partly offset each other. Consequently, the estimated bias in the correlated case is smaller for all coefficients except for the imprecise Maryland dummy in the ACS. In both the random and the correlated case, the estimated bias is always equal to $\hat{\delta}$ as defined by equation (3) (results not presented).

Table 3 examines to what extent equation (2) allows us to learn something from the biased coefficients. The first three columns present coefficient ratios, since equation (2) suggests that if misclassification is conditionally random, the constant of proportionality may cancel. Coefficient ratios are informative about the relative magnitude of the coefficients and, if the sign of one coefficient is known, their direction. The table contains estimates of the "true" coefficient ratios from the matched data (column 1), the average estimated bias in percent

of the true ratio from simulating conditionally random misclassification as described above (column 2) as well as the difference to the ratios from the survey data (column 3). If the conditional probabilities of misclassification are known, one can also multiply the estimated coefficients by the constant of proportionality, $(1 - \alpha_0 - \alpha_1)^{-1}$ to obtain estimates of the true coefficients. The last two columns examine whether this reduces the bias in simulations where misclassification is conditionally random (column 4) and in the estimates from the survey data, where misclassification is related to the covariates (column 5). Columns 2 and 5 show that if misclassification is indeed not related to the covariates, ratios and scaled up coefficients indicate the right relative magnitudes and signs of the true parameters. However, a small difference remains for all coefficients, which becomes sizable for imprecisely estimated coefficients. As one would expect, the results in columns 3 and 5 show that both approaches perform poorly if the assumption that misclassification is conditionally random fails. As we have seen above, the correlation of misclassification with the covariates partly offsets the attenuation effect in our application, so that the rescaling factor under the assumption of no correlation induces an upward bias.

Table 3: What Can Be Learned From Survey Coefficients, LPM

| | Coefficient Ratios (relative to age coefficient) | | | Bias Rescaled Marginal Effects | |
|---|---|---|---|---|---|
| | Matched Data | Bias in MC Study | Bias in Survey | in MC Study | in Survey |
| | (1) | (2) | (3) | (4) | (5) |
| | | | **ACS** | | |
| Poverty index | 0.0154 | 1.41% | 17.67% | 0.37% | 36.84% |
| Age≥50 | - | - | - | -0.85% | 39.48% |
| Maryland | 0.0292 | 12.82% | 611.46% | 25.95% | 39.63% |
| | | | **CPS** | | |
| Poverty index | 0.0182 | 3.51% | 17.17% | -0.29% | 71.43% |
| Age≥50 | - | - | - | 0.97% | 73.50% |
| Maryland | 0.7413 | 5.56% | -3.31% | -0.44% | 73.51% |

*Note:* See note Table 1 and 2

We perform the same analyses for the Probit models and present the results in tables

4-6. The third column of table 4 and 5 shows that, as in the linear probability model, coefficients are attenuated by a similar factor if misclassification is conditionally random and the coefficient is reasonably precisely estimated. The factor of attenuation is different from $\alpha_0 + \alpha_1$, because coefficients and marginal effects are not equal in the Probit model. As discussed above, equation (2) does not apply to Probit estimates due to the higher order bias from misspecification of the distribution of the error term. So, contrary to the linear probability model, this proportionality may not generalize to other applications.

Table 4: Estimated Bias in the Probit Model, ACS

| | $\hat{\beta}$ Matched Data | $\hat{\beta}$ Survey Data | Bias in MC Study (random MR) | Bias in Survey Data | (4)-(3): Bias due to Correlation |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Poverty index | -0.0060 | -0.0071 | -20.51% | 18.33% | 38.84% |
| | (0.0004) | (0.0005) | | | |
| Age≥50 | -0.4062 | -0.4167 | -22.15% | 2.58% | 24.74% |
| | (0.0512) | (0.0543) | | | |
| Maryland | -0.0187 | -0.0978 | -9.08% | 422.99% | 432.07% |
| | (0.0548) | (0.0582) | | | |

*Note:* Sample size: 5945 matched households from IL and MD with income less than 200% of the federal poverty line. All analyses include a constant term (not reported) and are conducted using household weights adjusted for match probability. All biases are in % of the coefficient from the matched data. The first two columns report Probit coefficients using FS receipt according to the administrative data (column 1) and according to survey reports (column 2) as the dependent variable. In the MC design in column 3, the dependent variable is administrative FS receipt with misreporting induced with the misreporting probabilities observed in the actual sample (Pr(FP)=0.02374 and Pr(FN)=0.2596), but independent of the covariates conditional on true receipt. 500 replications are performed and we report the average bias in %.

Column 4 underlines that both the proportionality and the attenuation only apply if misclassification is conditionally random: The difference between survey estimates and those in column 1 differs between coefficients and is sometimes positive, indicating a bias away from zero. It is again smaller in absolute value than the estimated bias without correlation. Column 5 confirms that the effect of correlation is in the opposite direction of the estimated bias in the conditionally random case, so the two effects partly offset each other. This

Table 5: Estimated Bias in the Probit Model, CPS

| | $\hat{\beta}$ Matched Data | $\hat{\beta}$ Survey Data | Bias in MC Study (random MR) | Bias in Survey Data | (4)-(3): Bias due to Correlation |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Poverty index | -0.0074 | -0.0083 | -32.47% | 12.16% | 44.63% |
| | (0.0005) | (0.0006) | | | |
| Age≥50 | -0.4297 | -0.3992 | -32.17% | -7.10% | 25.07% |
| | (0.0614) | (0.0682) | | | |
| Maryland | -0.3338 | -0.3189 | -33.15% | -4.46% | 28.69% |
| | (0.0736) | (0.0808) | | | |

*Note:* Sample size: 2791 matched households from IL and MD with income less than 200% of the federal poverty line. All analyses include a constant term (not reported) and are conducted using household weights adjusted for match probability. All biases are in % of the coefficient from the matched data. The first two columns report Probit coefficients using FS receipt according to the administrative data (column 1) and according to survey reports (column 2) as the dependent variable. In the MC design in column 3, the dependent variable is administrative FS receipt with misreporting induced with the misreporting probabilities observed in the actual sample (Pr(FP)=0.03271 and Pr(FN)=0.3907), but independent of the covariates conditional on true receipt. 500 replications are performed and we report the average bias in %.

explains why Meyer, Goerge and Mittag (2015) find that the bias in models of program take-up is relatively small given the extent of misreporting in the data.

Table 6: What Can Be Learned From Survey Coefficients, Probit

| | Coefficient Ratios (relative to age coefficient) | | | Observed Marginal Effects | | Bias in Rescaled Marginal Effects | |
|---|---|---|---|---|---|---|---|
| | Matched Data | Bias in MC Study | Bias in Survey | Matched Data | Survey Data | MC Study | Survey Data |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| | | | | **ACS** | | | |
| Poverty index | 0.0148 | 2.19% | 15.35% | -0.0018 | -0.0017 | 1.39% | 38.89% |
| Age≥50 | - | - | - | -0.1033 | -0.1164 | -0.58% | 39.50% |
| Maryland | 0.0460 | 9.40% | 409.82% | -0.0242 | -0.0053 | 3.66% | 39.67% |
| | | | | **CPS** | | | |
| Poverty index | 0.0172 | 4.84% | 20.73% | -0.0019 | -0.0021 | 1.98% | 73.68% |
| Age≥50 | - | - | - | -0.0902 | -0.1214 | 2.76% | 73.50% |
| Maryland | 0.7768 | 6.32% | 2.84% | -0.0721 | -0.0943 | 0.95% | 73.37% |

*Note:* See note Table 1 and 2

Table 6 examines the implications of equation (2) for Probit models. Columns 1-3 and 6-7 are the equivalent of table 3, i.e. they present coefficient ratios and rescaled marginal effects. Columns 4 and 5 contain Probit estimates of the observed marginal effects in the matched data and the survey data, since, contrary to the linear probability model, coefficients and marginal effects differ. Contrary to the linear probability model, using a Probit on the survey data does not yield consistent estimates of the observed marginal effects in the presence of misclassification due to the higher order bias. Nonetheless, the results in table 6 are qualitatively similar to the results for the linear probability model in table 3: only a small difference remains in the conditionally random cases in columns 2 and 6, suggesting that the bias from functional form misspecification is small here. However, both ratios and rescaled coefficients are substantively misleading if misclassification is related to $X$.

Overall, the results in this section show that the results from section 2 describe the (asymptotic) bias well in practice, which makes the formulas useful to interpret estimates obtained from contaminated data. The results in tables 3 and 6 show that examining coefficient ratios and scaling-up marginal effects can be very useful to learn something from data that is subject to misclassification if it is conditionally random. However, one should be cautious with this assumption, as using results for the conditionally random case when this assumption fails can be severely misleading and make matters worse than if the problem of misclassification were ignored.

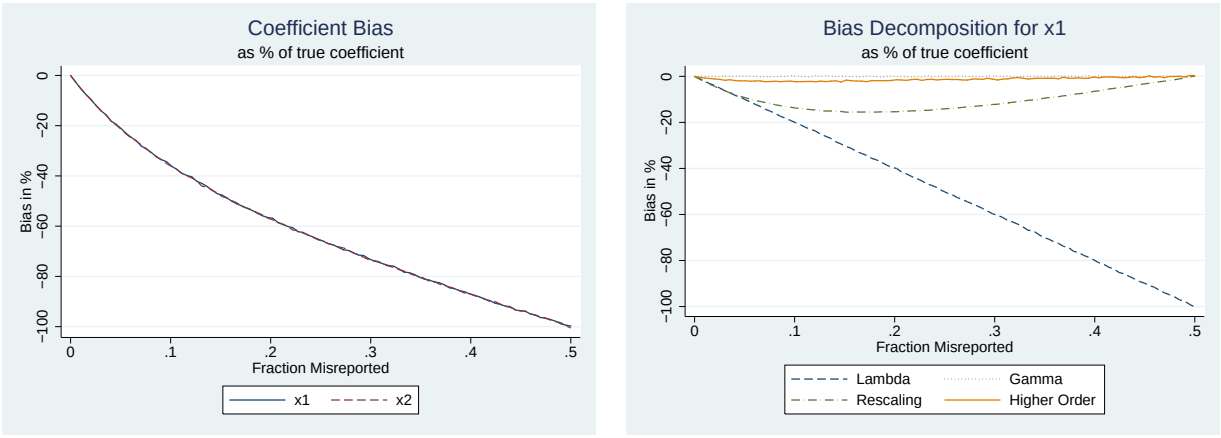
## 3.2 Assessing the Bias and its Components in Simulations

In order to obtain more evidence on the determinants and relative sizes of the bias components derived in part 2.2, we perform three simulation studies that allow us to change the factors that cause the bias components. We generate data with a similar structure as the observed data. In particular, we generate two covariates ($x_1$ and $x_2$) from normal distributions with variance 1. The mean is 0 for $x_1$ and $0.1x_1$ for $x_2$, so they are mildly correlated. The dependent variable is generated according to the Probit model

$y^T = 1\{a + 0.5x_1 + 0.5x_2 + e \geq 0\}$ where $e$ is drawn from a standard normal distribution. The intercept $a$ is chosen such that the mean of $y^T$ is always 0.25. We generate misclassification according to $m = 1\{c + bx_1 + e_{MC} \geq 0\}$ where $e_{MC}$ is another standard normal variable. Note that misclassification is related to $x_2$ only through the correlation between $x_1$ and $x_2$.

In the first two setups, we increase the level of misclassification from 0% to 50% holding everything else constant. In the first setup, misclassification is not related to $x_1$, while in the second setup $b$ is constant at 0.25, creating a modest correlation. In the third setup, we hold the level of misclassification fixed at 30% and increase $b$ from -1 to 1, so that the correlation between $x_1$ and $m$ increases from roughly -0.5 to 0.5. In all cases we take 100 equally spaced grid points over the parameter space. At each point, we draw 100 samples of 10,000 observations and run a Probit on $y^T$ to obtain $\hat{\beta}$ and on the data with misclassification (yielding $\hat{\beta}^{MC}$). We record the difference between the two estimated coefficient vectors $(\hat{\beta} - \hat{\beta}^{MC})$ as well as the bias due to $\hat{\lambda}$ as given by equation (11), $\hat{\gamma}$ as given by (13) and the rescaling bias implied by (15): $(\hat{\beta}^{MC} + \hat{\lambda} + \hat{\gamma})/SD(\hat{\varepsilon} + \hat{\nu} + \hat{\eta}) - (\hat{\beta}^{MC} + \hat{\lambda} + \hat{\gamma})$. We calculate the higher order bias as the "residual" difference, i.e. $(\hat{\beta}^{MC} - \hat{\lambda} - \hat{\gamma}) \cdot SD(\hat{\varepsilon} + \hat{\nu} + \hat{\eta}) - \hat{\beta}$.

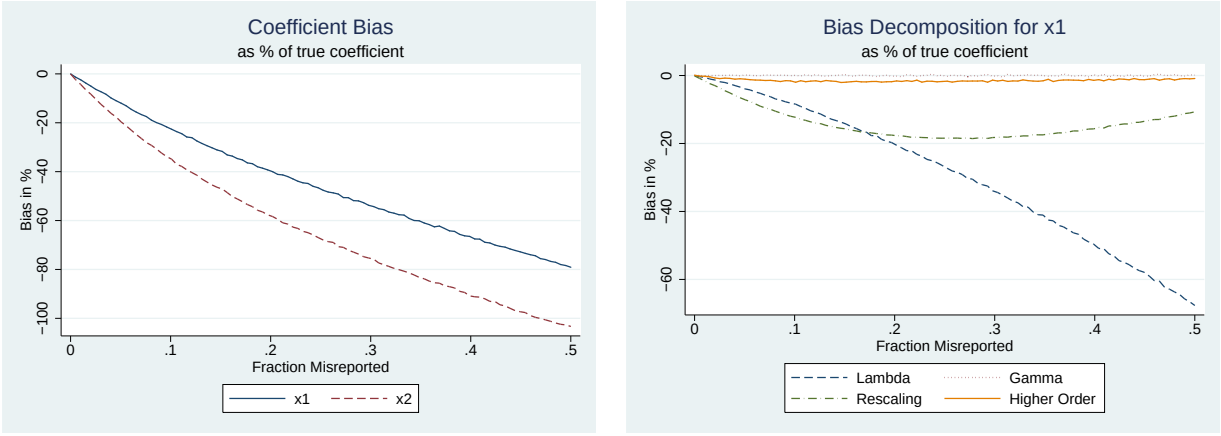
Figure 1: MC Design 1 - Uncorrelated Misclassification (b=0)




Figures 1-3 show the bias of the two coefficients in the left panel and the decomposition for the coefficient on $x_1$ into the four components in the right panel. Overall, the results are as expected based on our analytic results. The left panel of figure 1 shows that in the

uncorrelated case, the bias is the same for both coefficients (in relative terms) and always between 0% and -100%, i.e. both coefficients are always attenuated. It also shows that the bias is not linear in $\alpha_0 + \alpha_1$, which underlines that equation (2) is a relation between true values and not between Probit estimates. The bias decomposition in the right panel shows why this is the case: In addition to $\hat{\lambda}$, which is linear and has a slope of -2 as predicted by (12), there is a rescaling bias that is non-linear in the level of misclassification.

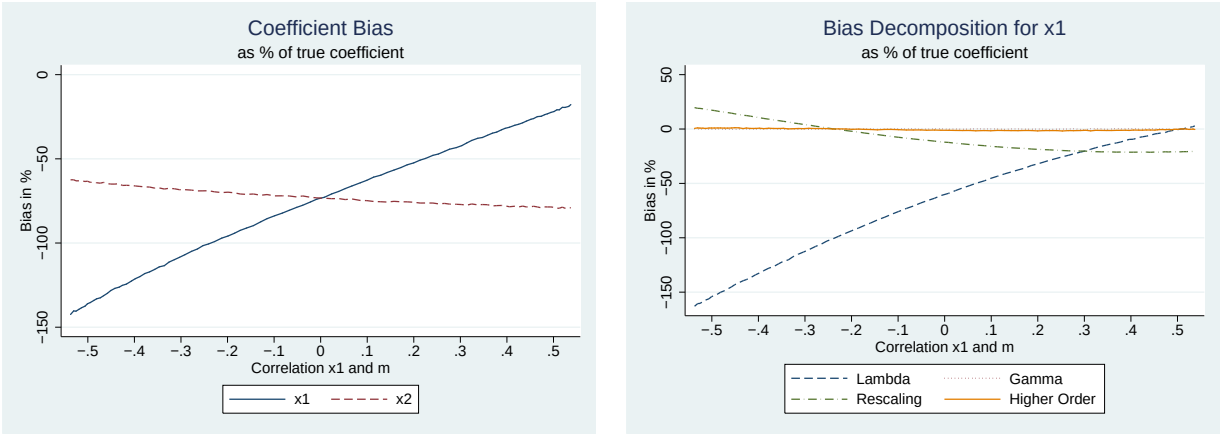
Figure 2: MC Design 2 - Correlated Misclassification (b=0.25)




The main difference in figure 2, where misclassification is correlated with $x_1$, is that the coefficient on $x_1$ is less severely biased. As in the food stamp data, the bias due to correlation reduces the bias from $\hat{\lambda}$, but is not strong enough to override it. This finding does not generalize. If the (negative) correlation is stronger, the coeffcient can be biased away from zero. If the correlation is positive, the bias is more severe than in figure 1, so that the coeffcient can change its sign. The bias decomposition is similar to the previous case, but $\hat{\lambda}$ is not linear in the fraction of misclassified observations. The difference in the rescaling bias is mainly due to its numerator (which is less attenuated and thus bigger in absolute value).

The key insight from the results in figure 3 is that most of the regularities we have stressed can be overturned if there is a strong relation between the covariates and misclassification. With extreme negative correlation, the bias exceeds -100%, so that the coefficient changes sign. For high positive values of the correlation the bias from $\hat{\lambda}$ is away from zero. If the

Figure 3: MC Design 3 - 30% Misclassification, Increasing Correlation With $x_1$




off-diagonal elements in (12) are negative and large relative to the diagonal elements, some components of $\hat{\lambda}$ can become positive and cause bias away from 0. Thus, coefficients are not always attenuated. However, while not always true, coefficients tend to retain their sign and are attenuated for a range of reasonable correlations even at a relatively high level of misclassification (30%).

The results from the last simulation raise the question in how far the patterns in these simulations are artifacts of the simulation setups. In order to assess this question, we conducted several additional simulation designs, the results are available upon request. Our misclassification model generates misclassification that is independent of $\varepsilon$, so $\hat{\gamma}$ is zero in all three cases. However, if misclassification depends on both $X$ and $\varepsilon$, $\hat{\gamma}$ becomes non-zero and behaves as one would expect based on equation (13). Similarly, the higher order bias is small in all three cases here, but can become noticeable, for example when the models for false positives and false negatives differ. It becomes sizable in our simulations when misclassification has an asymmetric effect on the distribution of $\varepsilon$, i.e. if the probability of misclassification for large values of the error term in the outcome equation is higher or lower than for small values. This is in line with the symmetric weighting function in equation (16), which suggests that symmetric deviations of $F_{\varepsilon|X}$ and $\Phi$ may average out over the sample.

In summary, the results in this section underline several useful conjectures suggested by the analytic results, which can be used to interpret estimates from data with misclassification

and assess the robustness of substantive conclusions. However, they also verify that there are exceptions to the regularities we find. While the usual caution regarding generalizations from simulations and particular cases applies, a robust finding is that slope coefficients only change sign if misclassification is strongly related to the covariates or $\varepsilon$ in particular ways: none of the coefficients in any of our applications changes sign, and it only happens for some extreme cases in the Monte Carlo studies. In addition, one of the components of the bias may sometimes bias the coefficients away from zero, but an overall bias away from zero only seems to arise in cases where $M$ is highly correlated with $X$ or $\varepsilon$. Thus, the coefficient estimates tend to be attenuated, i.e. lie between 0 and the true coefficient. This is always the case when misclassification is conditionally random, but can be overturned if it is strongly related to $X$. If one can rule out such cases, the estimates can be interpreted as lower bounds for the true coefficients and one may be able to infer the sign of the coefficients, which is often of key interest.

# 4 Consistent Estimators

This section evaluates estimators for the Probit model that are consistent under different assumptions. What is unique about our analysis is that we use the actual data for a common use of binary choice models, and, more importantly, we know the true value of the dependent variable from administrative data. We focus on the Probit model, because it is the most common parametric model and other maximum likelihood estimators can be corrected in similar ways.[6] We use six different estimators: three estimators are only consistent under conditionally random misclassification, and three are still consistent if misclassification is related to $X$. After describing the estimators, which are all variants of estimators that have been proposed elsewhere, section 4.1 evaluates their performance under conditionally random misclassification and section 4.2 allows misclassification to be related to $X$.

[6]Section 3 shows that corrections for the linear probability model work well if an estimate of $\hat{\delta}$ is available, so we do not examine corrections for the linear probability model.

All estimators can be derived from equation (7), which implies that the probability distribution of the observed outcome $y_i$ can be written as:

$$\Pr(y_i) = [\alpha_{0i} + (1 - \alpha_{0i} - \alpha_{1i})\Phi(x_i'\beta)]^{y_i} + [1 - \alpha_{0i} - (1 - \alpha_{0i} - \alpha_{1i})\Phi(x_i'\beta)]^{1-y_i}$$

This immediately implies the log-likelihood function of the observed data:

$$\begin{aligned}\ell(\alpha, \beta) = &\sum_{i=1}^{N} y_i \ln\left(\alpha_{0i} + (1 - \alpha_{0i} - \alpha_{1i})\Phi(x_i'\beta)\right) + \\ &(1 - y_i)\ln\left(1 - \alpha_{0i} - (1 - \alpha_{0i} - \alpha_{1i})\Phi(x_i'\beta)\right)\end{aligned} \tag{17}$$

The parameters of this likelihood are not identified, as there are $2N + K$ parameters (two $\alpha$s for each observation plus the $K$-by-1 vector $\beta$). Hausman, Abrevaya and Scott-Morton (1998) assume that the conditional probabilities of misclassification ($\alpha_{0i}$ and $\alpha_{1i}$) are constants, which reduces the number of parameters to $K + 2$: $\alpha_0$, $\alpha_1$ and $\beta$. They show that these parameters are identified due to the non-linearity of the normal cdf as long as $\alpha_0 + \alpha_1 < 1$. We refer to this estimator as the HAS-Probit. Their assumption that the probabilities of misclassification are constants implies that $\alpha_0$ is the population probability of a false positive and $\alpha_1$ is the population probability of a false negative. These probabilities or estimates of them may be known from validation data or other out of sample information. Let $\tilde{\alpha}_0$ and $\tilde{\alpha}_1$ denote such estimates. Our second estimator uses these estimates in (17) and maximizes it with respect to $\beta$ only. Poterba and Summers (1995) take a similar approach to a model of labor market transitions. In our application, the probabilities can be considered known, because they are either calculated within sample or are known from data on the whole population. If this is not the case, one should estimate standard errors as in Imbens and Lancaster (1994), because the usual standard errors are inconsistent (Hausman, Abrevaya and Scott-Morton, 1998).

Meyer, Mok and Sullivan (2009) assume that $\tilde{\alpha}_0$ is small enough to be ignored and cal-

culate $\tilde{\alpha}_1$ as the ratio of the population weighted number of people who report receipt of a program to the number of people who receive it according to administrative totals. While assuming $\tilde{\alpha}_0 = 0$ is a likely misspecification, estimates of the needed ratio are often available when separate estimates of $\alpha_0$ and $\alpha_1$ are not. Therefore, we also examine the performance of the estimator that maximizes (17) with $\alpha_0$ constrained to 0 and $\alpha_1$ constrained to $\tilde{\alpha}_1$ estimated as in Meyer, Mok and Sullivan (2009). All three estimators assume that, conditional on truth, misclassification is independent of the covariates. The unconstrained HAS-Probit teases out $\hat{\alpha}_0$ and $\hat{\alpha}_1$ from the observed binary responses, while the other two estimators constrain these parameters based on outside information.

In many cases, the assumption that misclassification is conditionally random does not hold. One can allow the misclassification probabilities to depend on $X$ if one can predict $\alpha_{0i}$ and $\alpha_{1i}$. For example, such predictions could be obtained by using the parameters from models of misclassification that use validation data (e.g. Meyer, Goerge and Mittag, 2015; Marquis and Moore, 1990). As Bollinger and David (1997) show, using such predicted probabilities in equation (17) and maximizing the resulting pseudo-likelihood with respect to $\beta$ yields consistent estimates. We refer to this estimator as the "predicted probabilities estimator" and bootstrap standard errors to account for the estimation of first stage parameters.

The predicted probabilities estimator does not require access to the validation data used to estimate the probabilities of misclassification. If both the validation data and the data used to estimate the outcome model are available, one could estimate the misclassification model and the outcome model jointly. Assuming that misclassification can be described by single index models, the two models imply a system of 3 equations: One for the model of false positives, a second equation for the model of false negatives and a third equation for the true outcome of interest. We assume that the misclassification models are Probit models, which yields a fully specified parametric system of equations that can be estimated by maximum likelihood. The likelihood function is derived in appendix B and depends on three components. Which components an observation contributes to depends on whether it contains

$y_i^T, x_i$ or both. The observations with $y_i^T$, but not $x_i$ identify the misclassification models, while those with $x_i$, but not $y_i^T$ identify the outcome equation in the predicted probabilities estimator. The observations with both $y_i^T$ and $x_i$ identify both the misclassification model and the outcome model, so in principle they could be used to estimate the outcome model directly. One may still want to estimate the full model, either because one is interested in the misclassification model or because one considers the observations in the intersection to be insufficient to estimate the parameters of interest (e.g. for reasons of efficiency or sample selection). Such cases often arise if a small subset of the observations has been validated: the validated observations allow estimation of the true outcome model and the misclassification model while those that were not validated only identify the observed outcome model. We examine an estimator for this setting that we refer to as the joint estimator with common observations. In other cases, such as those discussed by Bollinger and David (1997, 2001), observations that identify the true outcome model by themselves are not available, so we also consider an estimator in which there are no observations with both $y_i^T$ and $x_i$: Some observations identify the misclassification model and others the observed outcome model, but none identify both. We refer to this estimator as the "joint estimator without common observations".

Several other estimators for misclassified binary dependent variables have been proposed and examining their performance would be an interesting extension to the evidence presented below. Some papers consider point identification and estimation with panel data (Feng and Hu, 2013) or in the presence of instruments (e.g. Hu, 2008; Lewbel, 2000). However, we do not evaluate these estimators since we have neither panel data nor sufficiently credible instruments. Similarly, some semi-parametric estimators have been proposed (e.g Hausman, Abrevaya and Scott-Morton, 1998), but misspecification and misclassification are different problems and we focus on the latter here. Another related line of literature builds on Horowitz and Manski (1995) to examine bounds in the presence of contaminated sampling. Most closely related to our approach is Molinari (2008), who derives tight bounds for models

with misclassified discrete variables for a broad class of restrictions on the misclassification process (see also Dominitz and Sherman, 2004; Kreider and Pepper, 2008, 2011). Evaluating these approaches, or partial identification in general, is beyond the scope of this paper. However, bounds are an attractive alternative in cases where one does not have enough prior information to point identify the parameters of the model.

## 4.1 Performance When Misclassification is Conditionally Random

We apply the estimators to the matched survey data used in part 3. We begin by examining their performance when misclassification is conditionally random by conducting Monte Carlo simulations. In these simulations, we induce false positives and false negatives at the rate actually observed in the real data, but misclassification is unrelated to $X$ conditional on $y_i^T$. We run 500 replications. For the Predicted Probabilities estimator, we obtain estimates of the probabilities of misclassification (the first stage) from the same sample that we use for the outcome model (the second stage). The joint estimators allow estimation with two different samples with different information, so we split the sample in half randomly and use the two halves for the two samples required for the joint estimators. With the exception of the estimator that fixes $\alpha_0$ at 0, all estimators are consistent in this setting if the Probit assumption holds in the matched data.

The results in table 7 show that the HAS-Probit greatly improves upon the uncorrected Probit estimator in terms of bias (in columns 2 and 6) if one has estimates of $\alpha_0$ and $\alpha_1$. Columns 4 and 8 use the "true" probabilities of misreporting which results in estimates that have little bias (with the exception of the imprecise Maryland coefficient in the ACS). Sometimes only inaccurate estimates may be available, such as the net underreporting rate for $\alpha_1$ and 0 for $\alpha_0$. This means that the estimates of $\alpha_0$ and $\alpha_1$ are biased, but not too far off, so it is useful to know whether using them still improves the results. The results in columns 3 and 7 give an example where biased estimates of the probabilities of misclassification affect the slope coefficients.

In our case, choosing $\alpha_0$ and $\alpha_1$ to be lower than the true probabilities leads to a partial correction: the corrected slope estimates are between the estimates from the survey and the matched data. They are still substantially closer to the estimates using well-measured administrative data than with the error-ridden survey data, but it is unclear to what extent this holds more generally. In our application, the estimated bias is small when the error in the assumed $\alpha_0$ and $\alpha_1$ is small and is approximately proportional to this error. This suggests that choosing $\alpha_0$ and $\alpha_1$ between 0 and their true values may lead to bias between that of the naive Probit and the HAS Probit using the correct values of $\alpha_0$ and $\alpha_1$. However, it is unclear whether this finding generalizes. Our estimates from the ACS and CPS by and large support this conjecture, but also show that it is not always the case as it does not hold for the imprecisely estimated Maryland coefficient.

Table 7: Comparing Estimators - Conditional Random Misreporting

| | **ACS** | | | | **CPS** | | | |
|---|---|---|---|---|---|---|---|---|
| | $\hat{\beta}$ | Percent Bias | | | $\hat{\beta}$ | Percent Bias | | |
| | Matched Data | Standard Probit | $\tilde{\alpha}_0 = 0$ $\tilde{\alpha}_1 = .19$ | $\tilde{\alpha}_0 = .024$ $\tilde{\alpha}_1 = .26$ | Matched Data | Standard Probit | $\tilde{\alpha}_0 = 0$ $\tilde{\alpha}_1 = .29$ | $\tilde{\alpha}_0 = .033$ $\tilde{\alpha}_1 = .39$ |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Poverty index | -0.0060 | -20.51% | -11.12% | 1.34% | -0.0074 | -32.47% | -18.43% | 3.30% |
| Age≥50 | -0.4062 | -22.15% | -14.45% | -2.30% | -0.4297 | -32.17% | -20.20% | 0.00% |
| Maryland | -0.0187 | -9.08% | 15.22% | 13.20% | -0.3338 | -33.15% | -24.90% | 5.84% |

*Note:* Sample sizes: 5945 (ACS), 2791 (CPS) matched households, based on 500 replications. See notes Table 1 and 2 for further details on the samples and MC design. All analyses include a constant term (not reported). The conditional probabilities of misreporting in column 4 and 8 are based on the actual probabilities; column 3 and 7 use the (expected) net under count as the probability of false negatives.

The results in table 7 underline that the HAS-Probit can yield substantial improvements if misclassification is conditionally random and one constrains the probabilities of misclassification based on outside information. However, we find that the HAS-Probit performs poorly in our application to real data when these probabilities are left unconstrained. We do not report the results from the linked data, since they are too far off to be informative, but we examined the problem further using simulated data and the ACS public use files.[7]

[7]All results are available upon request.

In summary, we find that the estimator works well when the model is correctly specified and does not seem to be particularly sensitive to misspecification of the functional form of $\varepsilon$, such as skewness and heavy or light tails. However, we find that the estimator is sensitive to violations of the assumption that $\varepsilon$ is independent of the covariates. A possible explanation is that if $X$ and $\varepsilon$ are related, there are surprisingly many or few ones at certain values of $X$. The estimator explains this as misclassification. Thus, it is likely to pick up spurious misclassification, which leads to inconsistent estimates of $\alpha_0$ and $\alpha_1$. If the probabilities of misclassification are poorly estimated, the estimates of the slope coefficients can be severely off. Knowing these probabilities from outside information fixes this fragility in our application, and greatly reduces the bias even if the estimates of $\alpha_0$ and $\alpha_1$ are biased as in the MC study using the linked data.

Results from the estimators that remain consistent under correlated misclassification in the same MC setup are as expected, so we do not present them. In conclusion, our results suggest that if misclassification is conditionally random, estimators that are able to account for misclassification can greatly improve the estimates in terms of bias. However, unless one has great faith that the underlying binary choice model is correctly specified, it is useful to have external estimates of the probabilities of misclassification (even if they are slightly inaccurate). Otherwise, since the error rates are hard to estimate, their bias and imprecision can lead to severe bias in the slope coefficients.

## 4.2 Performance When Misclassification is Not Conditionally Random

The assumption that misclassification is conditionally random clearly fails in our data, which raises the question of how the estimators perform when this key assumption is violated and whether one can still obtain consistent estimates from the observed data using other methods. Our matched data contains both the "true" dependent variable $y^T$ as well as the misreported indicator $y$. This enables us to compare estimates of the true coefficients using

the administrative dependent variable to inconsistent coefficient estimates from the survey reports that suffer from misclassification.

Table 8: Comparing Estimators - Correlated Misreporting

| | **ACS** | | | | | **CPS** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Matched Data | Standard Probit | HAS Probit | $\tilde{\alpha}_0 = 0$ $\tilde{\alpha}_1 = .19$ | $\tilde{\alpha}_0 = .024$ $\tilde{\alpha}_1 = .26$ | Matched Data | Standard Probit | HAS Probit | $\tilde{\alpha}_0 = 0$ $\tilde{\alpha}_1 = .29$ | $\tilde{\alpha}_0 = .033$ $\tilde{\alpha}_1 = .39$ |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Poverty index | -0.0060 | -0.0071 | -0.0188 | -0.0082 | -0.0094 | -0.0074 | -0.0083 | -0.0082 | -0.0107 | -0.0146 |
| | (0.0004) | (0.0005) | (0.0019) | (0.0006) | (0.0007) | (0.0005) | (0.0006) | (0.0006) | (0.0008) | (0.0015) |
| Age≥50 | -0.4062 | -0.4167 | -0.8267 | -0.4622 | -0.5287 | -0.4297 | -0.3992 | -0.3879 | -0.4739 | -0.6270 |
| | (0.0512) | (0.0543) | (0.1097) | (0.0599) | (0.0713) | (0.0614) | (0.0682) | (0.0679) | (0.0812) | (0.1170) |
| Maryland | -0.0187 | -0.0978 | -0.3187 | -0.1140 | -0.1238 | -0.3338 | -0.3189 | -0.3058 | -0.3604 | -0.5760 |
| | (0.0548) | (0.0582) | (0.1137) | (0.0640) | (0.0757) | (0.0736) | (0.0808) | (0.0805) | (0.0977) | (0.1655) |
| $\hat{\alpha}_0$ | | | 0.0000 | | | | | 0.0011 | | |
| | | | (0.0000) | | | | | (0.0000) | | |
| $\hat{\alpha}_1$ | | | 0.6114 | | | | | 0.0000 | | |
| | | | (0.0233) | | | | | (0.0000) | | |

*Note:* Sample sizes: 5945 (ACS), 2791 (CPS), SEs in parentheses. All analyses conducted using household weights adjusted for PIK probability. All analyses include a constant term (not reported). The conditional probabilities of misreporting in columns 5 and 10 are based on the actual probabilities and column 4 and 9 use the (expected) net undercount as the false negative probability.

We find that all estimators that are consistent only under conditionally random misclassification fare poorly if misclassification is related to the covariates. Table 8 shows that in both surveys, all estimates are further from the true coefficients than the naive estimates. This strongly suggests that one should be cautious with the assumption that misclassification is conditionally independent of the covariates.

Tables 9 and 10 present the results from the estimators that are consistent if misclassification is related to the covariates: the predicted probabilities estimator from Bollinger and David (1997) and the two joint estimators. The last two rows contain two measures to evaluate their performance. "Weighted Distance" gives the average distance to the coefficients from the matched data weighted by the inverse of the variance matrix of the estimates from the matched data. We only use the variance matrix from the matched data in order to avoid dependence on differences in the efficiency of the estimators. The number in the last row is the F-Statistic of the coefficients from the matched data using the variance matrix of the estimator in that column. This can be interpreted as a measure of efficiency with

Table 9: Comparing Estimators - Correlated Misreporting, ACS

| | Survey Data | Matched Data | Pred. Prob. | JE 1: no common observations | JE 2: common observations |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Poverty index | -0.0071 | -0.0060 | -0.0059 | -0.0076 | -0.0063 |
| | (0.0005) | (0.0004) | (0.0006) | (0.0013) | (0.0005) |
| Age≥50 | -0.4167 | -0.4062 | -0.3950 | -0.5086 | -0.3660 |
| | (0.0543) | (0.0512) | (0.0530) | (0.1229) | (0.0615) |
| Maryland | -0.0978 | -0.0187 | 0.0050 | 0.1322 | -0.0366 |
| | (0.0582) | (0.0548) | (0.0574) | (0.1387) | (0.0662) |
| Constant | 0.0686 | 0.0987 | 0.1199 | 0.3088 | 0.1568 |
| | (0.0605) | (0.0583) | (0.0721) | (0.1489) | (0.0701) |
| Weighted Distance | 15.672 | 0 | 0.694 | | 0.779 |
| Precision | 230.824 | 272.011 | 220.563 | 66.650 | 209.915 |

*Note:* Sample size 5945 matched households. The first stage model for (3)-(5) includes age≥50, a MD dummy, the poverty index and its square. The model for false negatives also includes a cubic term in poverty. SEs for (3) are bootstrapped to account for the estimated first stage parameters. All analyses conducted using household weights adjusted for match probability. A mistake prevented the distance statistic for column (4) from being disclosed.

higher values being better. We use the coefficient from the matched data rather than the estimates in each column in order to avoid confounding efficiency with estimates that are larger in absolute value.[8] The measures of efficiency of the joint estimators are not directly comparable to the other estimators, since the sample definitions differ.

The results show that all three estimators work well and thereby underline that a model of misclassification can serve as a substitute for "clean" data. The joint estimator without common observations is less efficient than the joint estimator with common observations, but we cannot reject the hypothesis that it is consistent. Its lack of precision suggests that it may only be an attractive option in large data sets. Both the predicted probabilities estimator and the joint estimator with common observations work extremely well. The predicted probabilities estimator fares a little better in our applications, but at least in

[8]As any summary measure, these two statistics measure a particular aspect of the performance and may not capture other aspects well. For example, the first statistic is not a test of equality, but is the $\chi^2_4$ statistic of a test that the coefficients from the matched data are equal to the values in a given column. The test statistic from a test of equality may rank the estimators differently.

Table 10: Comparing Estimators - Correlated Misreporting, CPS

| | Survey Data | Matched Data | Pred. Prob. | JE 1: no common observations | JE 2: common observations |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Poverty index | -0.0083 | -0.0074 | -0.0070 | -0.0044 | -0.0070 |
| | (0.0006) | (0.0005) | (0.0007) | (0.0028) | (0.0008) |
| Age≥50 | -0.3992 | -0.4297 | -0.4109 | -0.3368 | -0.3693 |
| | (0.0682) | (0.0614) | (0.0616) | (0.1775) | (0.0777) |
| Maryland | -0.3189 | -0.3338 | -0.3733 | -0.4076 | -0.3347 |
| | (0.0808) | (0.0736) | (0.0819) | (0.1958) | (0.0931) |
| Constant | 0.1892 | 0.4108 | 0.3688 | 0.1741 | 0.3454 |
| | (0.0735) | (0.0723) | (0.0836) | (0.2462) | (0.0998) |
| Weighted Distance | 27.197 | 0 | 0.318 | | 0.481 |
| Precision | 148.198 | 189.154 | 153.357 | 37.682 | 131.888 |

*Note:* Sample size 2791 matched households. The first stage model for (3)-(5) includes age≥50, a MD dummy and the poverty index. SEs for (3) are bootstrapped to account for the estimated first stage parameters. All analyses conducted using household weights adjusted for match probability. A mistake prevented the distance statistic for column (4) from being disclosed.

terms of efficiency, we have stacked the deck in its favor given that we split the sample for the joint estimator. One would expect the joint estimator to be more efficient, as it is the maximum likelihood estimator.[9] The main drawback of the joint estimator with common observations is that it requires observations that identify the outcome model because they contain both $y_i^T$ and $x_i$. Such observations are rarely available. When they are available, one may prefer to use *only* these validated observations to obtain consistent, but inefficient estimates of the outcome model rather than also using the unvalidated observations in the efficient, but potentially misspecified full maximum likelihood joint estimator. On the other hand, the predicted probabilities estimator does not require the linked data to be available. It only requires consistent estimates of the parameters of the misclassification model, which can often be obtained from other studies, as in Bollinger and David (1997).

[9]The results for the joint estimator with common observations support this, as the SEs are only slightly larger than those of the predicted probabilities estimator, despite the fact that the joint estimator only uses half of the sample to estimate the outcome model. The SEs from the joint estimator without common observations are surprisingly large compared to the predicted probabilities estimator.

An important concern for these estimators besides bias and efficiency is their robustness to misspecification. One will usually not be able to assess whether one has actually reduced bias by using a correction for misclassification since validation data are not generally available. The results from applying estimators that are only consistent if misclassification is conditionally random indicate that increased bias is certainly possible if the model of misclassification is wrong. Informal evidence from using subsamples (such as one of the two states) to identify the misclassification model suggest that neither of the estimators is particularly sensitive to minor misspecification, but the joint estimator with common observations seems more robust than the predicted probabilities estimator. In the MC study where misclassification is conditionally random, both joint estimators produced estimates that were closer to the "true" estimates from the matched data than the uncorrected Probit estimates. The predicted probabilities estimator, on the other hand, produced estimates that were further off than the survey estimates and fared worse in terms of mean squared error than the joint estimator with common observations. This suggests that the predicted probabilities estimator may be sensitive to the inclusion of irrelevant variables in the first stage, which can usually be avoided by standard t- or F-tests.

The results above show that the information obtained from validation studies can improve survey based estimates considerably, but validation studies are costly. This raises the question of how much one loses from correcting estimates based on validation data from previous years, a subset of the population or even a different survey. Nothing is lost if the misclassification models are the same in the two data sets, but if they are slightly different the loss depends on the robustness to misspecification of the misclassification model. We examine this issue by estimating the misclassification model on the IL sample of the ACS and using it to see how well it corrects food stamp take up in MD. The misclassification models are statistically different in the two states, but qualitatively similar, so one may be tempted to use them to correct estimates if validation data are only available for some states. The results are in line with our previous findings: The joint estimator with common

observations performs best according to our measure of both bias and efficiency. The joint estimator without common observations still suffers from a lack of precision. The predicted probabilities estimator works well in terms of estimated bias (as measured by the distance metric defined above) compared to an uncorrected Probit, but contrary to the previous case it performs worse than the joint estimator with common observations. All estimators work better than the naive survey estimates in terms of the distance metric used above, so if similar data have been validated or parameter estimates from similar data are available, using them to correct the survey coefficients may be worth trying.[10]

# 5 Conclusion

In order to assess what can be learned from data with misclassification, we analyze common binary choice estimators when the dependent variable is subject to misclassification and evaluate the performance of several estimators that account for misclassification. In the first part of the paper, we derive analytic results for the (asymptotic) bias due to misclassification when misclassification depends on the covariates in arbitrary ways. We do so for the linear probability model and the Probit model. We show that there is always bias and describe the determinants of its size and direction. For the linear probability model, the bias formula is tractable and allows for simple corrections if the mean of the covariates among false positives and false negatives is available. For the Probit model, the asymptotic bias consists of four components. We derive formulas for three components, but there is no closed form solution for the last component, which is the effect of misclassification on the higher order moments of the error distribution. Nonetheless, the results imply a tendency for the asymptotic bias to be in the opposite direction of the sign of the coefficient. With additional information, the formulas allow an assessment of whether this tendency is likely to hold in the case at

[10] We also correct estimates of food stamp take-up in the ACS using the misclassification model we observe in the CPS and vice versa to see how extrapolating from a different survey works. The results, which are available upon request, underline that the joint estimator with common observations is more efficient than the other two estimators, but do not provide conclusive evidence on their performance.

hand. For example, the covariance of $X$ among the misclassified observations plays a key role in the asymptotic bias, so additional information on it helps to assess the size of the bias and thereby the robustness of substantive conclusions. If misclassification is conditionally random, only the probabilities of misclassification are required to obtain the exact bias in the linear probability model and an approximation to the asymptotic bias in the Probit model.

We then show that the bias can be substantial and that our analytic results are useful in practice by using simulations and models of food stamp receipt using two unique validation data sets. The formulas describe the estimated bias in the survey data accurately for the linear probability model and approximately for the Probit model. For example, in our application, the correlation with covariates reduces the bias, which explains why Meyer, Goerge and Mittag (2015) find relatively small bias. If misclassification is conditionally random, the inconsistent estimates are attenuated and coefficient ratios as well as scaled up marginal effects are informative about the true coefficients. However, the estimates can be misleading if misclassification is related to the covariates. If misclassification is not conditionally random, there still is a robust tendency for the estimates to be attenuated, but it can be overturned if misclassification is strongly related to the covariates. Thus, if one can rule out these extreme cases, one can still infer the signs of the coefficients from the data with misclassification.

Finally, we examine the performance of six estimators that account for misclassification. If misclassification is conditionally random, the HAS Probit with the probabilities of misclassification estimated from outside sources works well. Estimates of the error rates may be available from validation data or comparisons to administrative aggregates and they improve estimates in our application even when they are estimated with error. Without such estimates, the HAS Probit does not work well in our application to real data, likely because it is sensitive to violations of independence between $X$ and the error term. The estimators that assume conditionally random misclassification can easily be further from the true coefficients than the naive estimator when this assumption fails. However, we show that even if misclas-

sification is not conditionally random, it is still possible to obtain consistent estimates from the observed data. This situation requires additional information on the misclassification model, such as parameter estimates from which one can predict the probabilities of misclassification. Among the estimators we evaluate, the joint estimator with common observations is not only the efficient estimator, but it also performs best in our applications. However, it is often infeasible in practice and our results suggest that the predicted probabilities estimator is an attractive alternative. Neither estimator seems particularly sensitive to modest misspecification of the misclassification model. This result suggests that if validation data are not available, using approximate models may still improve the estimates or provide a robustness check. One may often be able to obtain an approximate model of misclassification from validation studies in other samples or time periods or based on more detailed aggregate tabulations.

Overall, our results show that misclassification of the dependent variable can lead to severe bias and standard results from linear models do not carry over. However, if misclassification is conditionally random or additional information on the misclassification model is available, the formulas and estimators in this paper can still be used to draw conclusions from data that are subject to misclassification, or even to obtain consistent estimates. This result underlines the value of validation data to understand the nature of misclassification and its effects, particularly since we find that corrections based on incorrect assumptions can also increase bias. Consequently, when little is known about the misclassification process, one may prefer to use our results to assess the robustness of substantive conclusions.

# Appendix A: Derivation of Bias in the Linear Probability Model

Equation (5) follows from equation (3) and the fact that the measurement error, $u$, only takes three values (-1,0,1):

$$\begin{aligned}
\hat{\delta} &= (X'X)^{-1}X'u = (X'X)^{-1}\sum_{i=1}^{N} x_i u_i \\
&= (X'X)^{-1}\left(\sum_{\substack{i \text{ s.t. } y_i=1 \\ \& y_i^T=0}} x_i \cdot 1 + \sum_{i \text{ s.t. } y_i=y_i^T} x_i \cdot 0 + \sum_{\substack{i \text{ s.t. } y_i=0 \\ \& y_i^T=1}} x_i \cdot (-1)\right) \\
&= (X'X)^{-1}(N_{FP}\bar{x}_{FP} - N_{FN}\bar{x}_{FN}) \\
&= N(X'X)^{-1}\left(\frac{N_{FP}}{N}\bar{x}_{FP} - \frac{N_{FN}}{N}\bar{x}_{FN}\right)
\end{aligned}$$

We also consider the special case when the conditional probabilities of misclassification are constants as in Hausman, Abrevaya and Scott-Morton (1998), i.e. when

$$\begin{aligned}
&\Pr(y_i = 1|y_i^T = 0) = \alpha_{0i} = \alpha_0 \\
&\Pr(y_i = 0|y_i^T = 1) = \alpha_{1i} = \alpha_1
\end{aligned} \qquad \forall i$$

By the assumptions of the linear probability model $\Pr(y_i^T = 1|X) = x_i'\beta^{LPM}$ and $\Pr(y_i^T = 0|X) = 1 - x_i'\beta^{LPM}$, so that the probability mass function of the measurement error, $U$, conditional on $X$ is

$$\Pr(U = u_i|X = x_i) = \begin{cases} \Pr(y_i = 1|y_i^T = 0) \cdot \Pr(y_i^T = 0|X) = \alpha_0(1 - x_i'\beta^{LPM}) & \text{if } u_i = 1 \\ 1 - \alpha_0 + (\alpha_0 - \alpha_1)x_i'\beta^{LPM} & \text{if } u_i = 0 \\ \Pr(y_i = 0|y_i^T = 1) \cdot \Pr(y_i^T = 1|X) = \alpha_1 x_i'\beta^{LPM} & \text{if } u_i = -1 \end{cases}$$

where the probability for $u_i = 0$ follows immediately from the fact that the probabilities

have to sum to 1. Consequently, the conditional expectation of the measurement error is

$$\begin{aligned}\mathbb{E}(u_i|X = x_i) &= 1 \cdot [\alpha_0(1 - x_i'\beta^{LPM})] + 0 \cdot [1 - \alpha_0 + (\alpha_0 - \alpha_1)x_i'\beta^{LPM}] - 1 \cdot [\alpha_1 x_i'\beta^{LPM}] \\ &= \alpha_0 - (\alpha_0 + \alpha_1)x_i'\beta^{LPM}\end{aligned}$$

Assuming that $X$ is non-stochastic,[11] this implies that the bias, $\mathbb{E}(\hat{\delta})$, is

$$\begin{aligned}\mathbb{E}(\hat{\delta}) &= \mathbb{E}(X'X)^{-1}(X'u) \\ &= (X'X)^{-1}\mathbb{E}(X'u) \\ &= (X'X)^{-1}\sum_i x_i'\mathbb{E}(u_i|x_i) \\ &= (X'X)^{-1}\sum_i x_i'(\alpha_0 - (\alpha_0 + \alpha_1)x_i'\beta^{LPM}) \\ &= (X'X)^{-1}\sum_i x_i'\alpha_0 - (\alpha_0 + \alpha_1)(X'X)^{-1}\sum_i x_i x_i'\beta^{LPM} \\ &= (X'X)^{-1}X'\alpha_0 - (\alpha_0 + \alpha_1)(X'X)^{-1}(X'X)\beta^{LPM} \\ &= (\alpha_0, 0, \ldots, 0)' - (\alpha_0 + \alpha_1)\beta^{LPM}\end{aligned}$$

The first term is the coefficient vector from a regression of a constant, $\alpha_0$, on $X$, so the intercept will be equal to $\alpha_0$ and all other coefficients will be zero. The expectation of the biased coefficient vector is $\mathbb{E}(\hat{\tilde{\beta}}^{LPM}) = \beta^{LPM} + \mathbb{E}(\hat{\delta})$. Consequently, the expectation of the (biased) intercept is

$$\mathbb{E}(\hat{\tilde{\beta}}_0^{LPM}) = \alpha_0 + (1 - \alpha_0 - \alpha_1)\beta_0^{LPM}$$

and the expectation of the (biased) slope parameters is

$$\mathbb{E}(\hat{\tilde{\beta}}_{1\ldots k}^{LPM}) = (1 - \alpha_0 - \alpha_1)\beta_{1\ldots k}^{LPM}$$

[11] The extension to stochastic $X$ follows from the law of iterated expectation.

# Appendix B: Likelihood of the Joint Estimators

The setup in section 4 implies a system of 3 equations, two that describe misclassification and an outcome equation:

$$\begin{aligned}
\Pr(y_i|y_i^T = 0, x_i^{FP}) &= \left[F^{FP}(x_i^{FP\prime}\gamma^{FP})\right]^{y_i} + \left[1 - F^{FP}(x_i^{FP\prime}\gamma^{FP})\right]^{1-y_i} \\
\Pr(y_i|y_i^T = 1, x_i^{FN}) &= \left[F^{FN}(x_i^{FN\prime}\gamma^{FN})\right]^{y_i} + \left[1 - F^{FN}(x_i^{FN\prime}\gamma^{FN})\right]^{1-y_i} \\
\Pr(y_i^T|x_i) &= \Phi(x_i'\beta)^{y_i^T} + [1 - \Phi(x_i'\beta)]^{1-y_i^T}
\end{aligned} \tag{18}$$

The first equation gives the model for false positives, which depend on covariates $X^{FP}$ through the parameters $\gamma^{FP}$ and the link function $F^{FP}$. Similarly, the second equation gives the model for false negatives and the third equation the model for the true outcome of interest, which depends on the parameters of interest, $\beta$. We assume that the misclassification models are Probit models, i.e. $F^{FP}$ and $F^{FN}$ are standard normal cumulative distribution functions. This assumption yields a fully specified parametric system of equations that can be estimated by maximum likelihood.

As discussed in section 4, the sample can be divided into three disjoint subsamples according to whether an observation contains $\left[y_i, y_i^T, x_i^{FP}, x_i^{FN}\right]$, $\left[y_i, x_i, x_i^{FP}, x_i^{FN}\right]$ or both. The first subsample, $S_1$ contains observations that are in the validation data, but not in the data used to identify the outcome model and thus contains $(y_i, y_i^T, x_i^{FP}, x_i^{FN})$. This sample identifies the parameters of the misclassification equations, but not the outcome equation. The second subsample, $S_2$, contains all observations that have been validated and includes all variables in the outcome model, so that it contains $(y_i, y_i^T, x_i, x_i^{FP}, x_i^{FN})$. This sample identifies all parameters. The third subsample, $S_3$, contains the observations used to estimate the outcome model that have not been validated. Thus, it contains $(y_i, x_i, x_i^{FP}, x_i^{FN})$, which by itself identifies none of the parameters. Frequently, one of the first two samples will be empty. For the joint estimator with common observations, $S_1$ is empty; for the joint estimator with common observations, $S_2$ is empty, as in Bollinger and David (1997).

Since the subsamples are disjoint and independent, the log-likelihood of the entire sample is the sum of the three log-likelihoods of the subsamples. Thus, it depends on three components:

$$\begin{aligned}\ell(y, y^T, x, x^{FP}, x^{FN}; \beta, \gamma^{FP}, \gamma^{FN}) = &\sum_{i \epsilon S_1} \ell_i^{S_1}(y_i, y_i^T, x_i^{FP}, x_i^{FN}; \gamma^{FP}, \gamma^{FN}) + \\ &\sum_{i \epsilon S_2} \ell_i^{S_2}(y_i, y_i^T, x_i, x_i^{FP}, x_i^{FN}; \beta, \gamma^{FP}, \gamma^{FN}) + \\ &\sum_{i \epsilon S_3} \ell_i^{S_3}(y_i, x_i, x_i^{FP}, x_i^{FN}; \beta, \gamma^{FP}, \gamma^{FN})\end{aligned} \tag{19}$$

Equation (18) and the assumption that the error terms in all three equations are independent draws from standard normal distributions imply the following (conditional) probabilities

$$\begin{aligned}\Pr(y_i | y_i^T = 0, x_i^{FP}) &= \left[\Phi(x_i^{FP\prime}\gamma^{FP})\right]^{y_i} + \left[1 - \Phi(x_i^{FP\prime}\gamma^{FP})\right]^{1-y_i} \\ \Pr(y_i | y_i^T = 1, x_i^{FN}) &= \left[\Phi(x_i^{FN\prime}\gamma^{FN})\right]^{y_i} + \left[1 - \Phi(x_i^{FN\prime}\gamma^{FN})\right]^{1-y_i} \\ \Pr(y_i^T | x_i) &= \Phi(x_i'\beta)^{y_i^T} + \left[1 - \Phi(x_i'\beta)\right]^{1-y_i^T}\end{aligned} \tag{20}$$

The first probability is the likelihood contribution of an observation in $S_1$ with $y_i^T = 0$, while the second probability is the likelihood contribution of an observation in $S_1$ with $y_i^T = 1$. Consequently, the likelihood contribution of an observation from $S_1$ is

$$\begin{aligned}\ell_i^{S_1}(y_i, y_i^T, x_i^{FP}, x_i^{FN}; \gamma^{FP}, \gamma^{FN}) = &(1 - y_i^T)[y_i \ln \Phi(x_i^{FP\prime}\gamma^{FP}) + (1 - y_i) \ln \Phi(-x_i^{FP\prime}\gamma^{FP})] + \\ &y_i^T[(1 - y_i) \ln \Phi(x_i^{FN\prime}\gamma^{FN}) + y_i \ln \Phi(-x_i^{FN\prime}\gamma^{FN})]\end{aligned}$$

This is the sum of the likelihoods of Probit models for false positives and false negatives.

The likelihood contribution of an observation in sample $S_2$ is the probability of the observed combination of $y_i$ and $y_i^T$, which is

$$\Pr(y_i, y_i^T | x_i, x_i^{FP}, x_i^{FN}) = \begin{cases} \Pr(y_i | y_i^T = 0, x_i^{FP}) \Pr(y_i^T = 0 | x_i) & \text{if } y_i^T = 0 \\ \Pr(y_i | y_i^T = 1, x_i^{FN}) \Pr(y_i^T = 1 | x_i) & \text{if } y_i^T = 1 \end{cases}$$

Using the probabilities from (20) yields the likelihood contribution of an observation from sample $S_2$:

$$\ell_i^{S_2}(y_i, y_i^T, x, x_i^{FP}, x_i^{FN}; \beta, \gamma^{FP}, \gamma^{FN}) =$$
$$(1 - y_i^T) \ln \left( [y_i \Phi(x_i^{FP\prime} \gamma^{FP}) + (1 - y_i) \Phi(-x_i^{FP\prime} \gamma^{FP})] \Phi(-x_i' \beta) \right) +$$
$$y_i^T \ln \left( [(1 - y_i) \Phi(x_i^{FN\prime} \gamma^{FN}) + y_i \Phi(-x_i^{FN\prime} \gamma^{FN})] \Phi(x_i' \beta) \right)$$

Equation (20) defines the probabilities of false positives as $\alpha_{0i} = \Phi(x_i^{FP\prime} \gamma^{FP})$ and of false negatives as $\alpha_{1i} = \Phi(x_i^{FN\prime} \gamma^{FN})$. Using these probabilities of misclassification in equation (17) yields the contribution of an observation from $S_3$:

$$\ell_i^{S_3}(y_i, x_i, x_i^{FP}, x_i^{FN}; \beta, \gamma^{FP}, \gamma^{FN}) =$$
$$y_i \ln[\Phi(x_i^{FP\prime} \gamma^{FP}) + (1 - \Phi(x_i^{FP\prime} \gamma^{FP}) - \Phi(x_i^{FN\prime} \gamma^{FN})) \Phi(x_i' \beta)] +$$
$$(1 - y_i) \ln[1 - \Phi(x_i^{FP\prime} \gamma^{FP}) - (1 - \Phi(x_i^{FP\prime} \gamma^{FP}) - \Phi(x_i^{FN\prime} \gamma^{FN})) \Phi(x_i' \beta)]$$

Using the three likelihood contributions $\ell_i^{S_1}, \ell_i^{S_2}$ and $\ell_i^{S_3}$ in equation (19) and maximizing it with respect to $(\beta, \gamma^{FP}, \gamma^{FN})$ yields consistent estimates of the three parameter vectors by the standard arguments for the consistency of maximum likelihood. Standard errors of all parameters can be obtained as usual. The joint estimator with common observations used above assumes that $S_1$ is empty, so the log-likelihood reduces to

$$\ell^{JE1}(y, y^T, x, x^{FP}, x^{FN}; \beta, \gamma^{FP}, \gamma^{FN}) =$$
$$\sum_{i \epsilon S_2} \Big[ (1 - y_i^T) \ln \left( [y_i \Phi(x_i^{FP\prime} \gamma^{FP}) + (1 - y_i) \Phi(-x_i^{FP\prime} \gamma^{FP})] \Phi(-x_i' \beta) \right) +$$
$$y_i^T \ln \left( [(1 - y_i) \Phi(x_i^{FN\prime} \gamma^{FN}) + y_i \Phi(-x_i^{FN\prime} \gamma^{FN})] \Phi(x_i' \beta) \right) \Big] +$$
$$\sum_{i \epsilon S_3} \Big[ y_i \ln[\Phi(x_i^{FP\prime} \gamma^{FP}) + (1 - \Phi(x_i^{FP\prime} \gamma^{FP}) - \Phi(x_i^{FN\prime} \gamma^{FN})) \Phi(x_i' \beta)] +$$
$$(1 - y_i) \ln[1 - \Phi(x_i^{FP\prime} \gamma^{FP}) - (1 - \Phi(x_i^{FP\prime} \gamma^{FP}) - \Phi(x_i^{FN\prime} \gamma^{FN})) \Phi(x_i' \beta)] \Big]$$

The second joint estimator we use above, the joint estimator without common observa-

tions, assumes that $S_2$ is empty, so the log-likelihood reduces to

$$\begin{aligned}
\ell^{JE2}(y, y^T, x, x^{FP}, x^{FN}; \beta, \gamma^{FP}, \gamma^{FN}) = \\
&\sum_{i \epsilon S_1} \big[(1-y_i^T)[y_i \ln \Phi(x_i^{FP\prime}\gamma^{FP}) + (1-y_i)\ln\Phi(-x_i^{FP\prime}\gamma^{FP})] + \\
&\qquad y_i^T[(1-y_i)\ln\Phi(x_i^{FN\prime}\gamma^{FN}) + y_i \ln\Phi(-x_i^{FN\prime}\gamma^{FN})]\big] + \\
&\sum_{i \epsilon S_3} \big[y_i \ln[\Phi(x_i^{FP\prime}\gamma^{FP}) + (1-\Phi(x_i^{FP\prime}\gamma^{FP}) - \Phi(x_i^{FN\prime}\gamma^{FN}))\Phi(x_i'\beta)] + \\
&\qquad (1-y_i)\ln[1-\Phi(x_i^{FP\prime}\gamma^{FP}) - (1-\Phi(x_i^{FP\prime}\gamma^{FP}) - \Phi(x_i^{FN\prime}\gamma^{FN}))\Phi(x_i'\beta)]\big]
\end{aligned}$$

Note that the likelihood contribution of $S_2$ can be re-written as the sum of the likelihood contribution if the observation were in sample $S_1$ and the likelihood contribution to a standard probit likelihood:

$$\begin{aligned}
\ell_i^{S_2}(y_i, y_i^T, x, x_i^{FP}, x_i^{FN}; \beta, \gamma^{FP}, \gamma^{FN}) = &\ell_i^{S_1}(y_i, y_i^T, x_i^{FP}, x_i^{FN}; \gamma^{FP}, \gamma^{FN}) + \\
&y_i^T \ln\Phi(x_i'\beta) + (1-y_i^T)\ln\Phi(-x_i'\beta)
\end{aligned}$$

This can be used to re-write the log-likelihood function as

$$\begin{aligned}
\ell(y, y^T, x, x^{FP}, x^{FN}; \beta, \gamma^{FP}, \gamma^{FN}) = &\sum_{i \epsilon S_1 \cup S_2} \ell_i^{S_1}(y_i, y_i^T, x_i^{FP}, x_i^{FN}; \gamma^{FP}, \gamma^{FN}) + \\
&\sum_{i \epsilon S_2} \ell_i^P(y_i^T, x_i; \beta) + \sum_{i \epsilon S_3} \ell_i^{S_3}(y_i, x_i, x_i^{FP}, x_i^{FN}; \beta, \gamma^{FP}, \gamma^{FN})
\end{aligned}$$

where $\ell^P$ is the log-likelihood function of a standard Probit model. This shows more clearly which observations contribute to the identification of the parameters. In particular, it shows the value of observations in $S_2$, because in addition to the likelihood contribution of an observation in $S_1$, they add a term that directly identifies the parameters of the outcome model, $\beta$. It also shows that observations in $S_3$ contribute to the identification of the parameters of false positives and false negatives even though these observations only contain information on the observed dependent variable.

# References

**Aigner, Dennis J.** 1973. "Regression with a binary independent variable subject to errors of observation." *Journal of Econometrics*, 1(1): 49–59.

**Benítez-Silva, Hugo, Moshe Buchinsky, Hiu Man Chan, Sofia Cheidvasser, and John Rust.** 2004. "How large is the bias in self-reported disability?" *Journal of Applied Econometrics*, 19(6): 649–670.

**Bitler, Marianne P., Janet Currie, and John Karl Scholz.** 2003. "WIC Eligibility and Participation." *The Journal of Human Resources*, 38: 1139–1179.

**Black, Dan A., Seth Sanders, and Lowell Taylor.** 2003. "Measurement of Higher Education in the Census and Current Population Survey." *Journal of the American Statistical Association*, 98: 545–554.

**Bollinger, Christopher R.** 1996. "Bounding mean regressions when a binary regressor is mismeasured." *Journal of Econometrics*, 73(2): 387–399.

**Bollinger, Christopher R., and Martin H. David.** 1997. "Modeling Discrete Choice With Response Error: Food Stamp Participation." *Journal of the American Statistical Association*, 92(439): 827–835.

**Bollinger, Christopher R., and Martin H. David.** 2001. "Estimation with Response Error and Nonresponse: Food-Stamp Participation in the SIPP." *Journal of Business & Economic Statistics*, 19(2): 129–141.

**Bound, John, Charles Brown, and Nancy Mathiowetz.** 2001. "Measurement error in survey data." In *Handbook of Econometrics*. Vol. 5, , ed. James J. Heckman and Edward Leamer, Chapter 59, 3705–3843. Amsterdam:Elsevier.

**Bound, John, Charles Brown, Greg J. Duncan, and Willard L. Rodgers.** 1994. "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data." *Journal of Labor Economics*, 12(3): 345–368.

**Call, Kathleen T., Gestur Davidson, Michael Davern, and Rebecca Nyman.** 2008. "Medicaid undercount and bias to estimates of uninsurance: new estimates and existing evidence." *Health Services Research*, 43(3): 901–914.

**Cameron, Stephen V., and James J. Heckman.** 2001. "The Dynamics of Educational Attainment for Black, Hispanic, and White Males." *Journal of Political Economy*, 109(3): 455–499.

**Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu.** 2006. *Measurement Error in Nonlinear Models: A Modern Perspective. Monographs on Statistics and Applied Probability.* 2nd ed., Boca Raton:Chapman & Hall/CRC.

**Chen, Xiaohong, Han Hong, and Denis Nekipelov.** 2011. "Nonlinear Models of Measurement Errors." *The Journal of Economic Literature*, 49(4): 901–937.

**Davern, Michael, Jacob A. Klerman, Jeanette Ziegenfuss, Victoria Lynch, and George Greenberg.** 2009. "A partially corrected estimate of medicaid enrollment and uninsurance: Results from an imputational model developed off linked survey and administrative data." *Journal of Economic & Social Measurement*, 34(4): 219–240.

**Dominitz, Jeff, and Robert P. Sherman.** 2004. "Sharp bounds under contaminated or corrupted sampling with verification, with an application to environmental pollutant data." *Journal of Agricultural, Biological, and Environmental Statistics*, 9(3): 319–338.

**Eckstein, Zvi, and Kenneth I. Wolpin.** 1999. "Why Youths Drop out of High School: The Impact of Preferences, Opportunities, and Abilities." *Econometrica*, 67(6): 1295–1339.

**Estrella, Arturo, and Frederic S. Mishkin.** 1998. "Predicting U.S. Recessions: Financial Variables as Leading Indicators." *Review of Economics and Statistics*, 80(1): 45–61.

**Fearon, James D., and David D. Laitin.** 2003. "Ethnicity, insurgency, and civil war." *American Political Science Review*, 97(1): 75–90.

**Feng, Shuaizhang, and Yingyao Hu.** 2013. "Misclassification Errors and the Underestimation of the US Unemployment Rate." *American Economic Review*, 103(2): 1054–1070.

**Haider, Steven J., Alison Jacknowitz, and Robert F. Schoeni.** 2003. "Food Stamps and the Elderly: Why Is Participation so Low?" *The Journal of Human Resources*, 38: 1080–1111.

**Hausman, Jerry A., Jason Abrevaya, and Fiona M. Scott-Morton.** 1998. "Misclassification of the dependent variable in a discrete-response setting." *Journal of Econometrics*, 87(2): 239–269.

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

**Hu, Yingyao.** 2008. "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution." *Journal of Econometrics*, 144(1): 27–61.

**Imbens, Guido W., and Tony Lancaster.** 1994. "Combining Micro and Macro Data in Microeconometric Models." *The Review of Economic Studies*, 61(4): 655–680.

**Kreider, Brent, and John Pepper.** 2008. "Inferring disability status from corrupt data." *Journal of Applied Econometrics*, 23(3): 329–349.

**Kreider, Brent, and John V. Pepper.** 2011. "Identification of Expected Outcomes in a Data Error Mixing Model With Multiplicative Mean Independence." *Journal of Business & Economic Statistics*, 29(1): 49–60.

**Lewbel, Arthur.** 2000. "Identification of the Binary Choice Model With Misclassification." *Econometric Theory*, 16(4): 603–609.

**Lochner, Lance, and Enrico Moretti.** 2004. "The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports." *The American Economic Review*, 94(1): 155–189.

**Marquis, Kent H., and Jeffrey C. Moore.** 1990. "Measurement Errors in SIPP Program Reports." In *Proceedings of the 1990 Annual Research Conference*. 721–745. Washington, D.C.:U.S. Bureau of the Census.

**Meyer, Bruce D., Robert Goerge, and Nikolas Mittag.** 2015. "Errors in Survey Reporting and Imputation and Their Effects on Estimates of Food Stamp Program Participation." Unpublished Manuscript.

**Meyer, Bruce D., Wallace K. C. Mok, and James X. Sullivan.** 2009. "The Under-Reporting of Transfers in Household Surveys: Its Nature and Consequences." Harris School of Public Policy Studies, University of Chicago Working Paper 0903.

**Molinari, Francesca.** 2008. "Partial identification of probability distributions with misclassified data." *Journal of Econometrics*, 144(1): 81–117.

**Poterba, James M., and Lawrence H. Summers.** 1995. "Unemployment Benefits and Labor Market Transitions: A Multinomial Logit Model with Errors in Classification." *The Review of Economics and Statistics*, 77(2): 207–216.

**Ruud, Paul A.** 1983. "Sufficient Conditions for the Consistency of Maximum Likelihood Estimation Despite Misspecification of Distribution in Multinomial Discrete Choice Models." *Econometrica*, 51(1): 225–228.

**Ruud, Paul A.** 1986. "Consistent estimation of limited dependent variable models despite misspecification of distribution." *Journal of Econometrics*, 32(1): 157–187.

**Wooldridge, Jeffrey M.** 2007. "Inverse probability weighted estimation for general missing data problems." *Journal of Econometrics*, 141(2): 1281–1301.

**Yatchew, Adonis, and Zvi Griliches.** 1984. "Specification Error in Probit Models." Institute for Policy Analysis, University of Toronto Working Paper 8429.

**Yatchew, Adonis, and Zvi Griliches.** 1985. "Specification Error in Probit Models." *The Review of Economics and Statistics*, 67(1): 134–139.

# Web Appendix: Further Analysis of the Higher Order Bias

We have shown that the bias in the Probit model depends on four components: two components due to the linear projection, a rescaling bias, and bias due to misspecification of the higher order moments of the distribution function. The bias due to misspecification of the error distribution is harder to assess than the other components. If one has enough information about $F_{\tilde{\varepsilon}|X}$ to take random draws from it, one can simulate the exact bias or even obtain an exact solution of (16). In practice, however, such detailed information will rarely be available, so we discuss some factors that influence the size and direction of the bias. They may allow the researcher to informally assess whether this bias is likely to be small, which justifies considering $\bar{\beta} - \beta$ a good approximation to the full bias.

Given that the left hand side of equation (16) is the derivative of a Probit likelihood, which is globally concave in b, the left hand side of each equation in (16) crosses 0 only once and does so from above. In the absence of bias due to functional form misspecification, it does so at $b = \bar{\beta}$. In the univariate case, this implies that if the left hand side of (16) is positive at this point, the additional bias will be positive, while the additional bias will be negative if the left hand side is negative at $b = \bar{\beta}$. In the multivariate case, this can in principle be offset for some, but not all coefficients by the fact that the bias "spreads" between coefficients. For the multivariate case, note that

$$f_X(x_i)\frac{\phi(x_i'b)}{\Phi(x_i'b)(1-\Phi(x_i'b))} > 0 \tag{21}$$

so (16) can be interpreted as a weighted average of $x_i' \left[F_{\tilde{\varepsilon}|X=x_i}(x_i'\bar{\beta}) - \Phi(x_i'\bar{\beta})\right]$ with the weights given by (21). Consequently, observations for which $sign(x_i) = sign(F_{\tilde{\varepsilon}|X=x_i}(x_i'\bar{\beta}) - \Phi(x_i'\bar{\beta}))$ tend to cause a positive bias in the coefficient on $x$, while observations with opposing signs tend to cause a negative bias. The weight function has a minimum at 0 and increases

in either direction, so differences at more extreme values of $x'b$ have a larger impact. Larger values of $x$ also tend to make $x_i' \left[F_{\tilde{\varepsilon}|X=x_i}(x_i'\bar{\beta}) - \Phi(x_i'\bar{\beta})\right]$ larger, because $x$ enters it multiplicatively. The expression at each value of $x$ is weighted by its density $f_X(x)$, so differences at frequent values of $x$ have a larger impact. In summary, one can get an idea of the direction of the bias if one knows how $F_{\tilde{\varepsilon}_i}$ and $\Phi$ differ. If the former is larger in regions where the sample density of $x$ is high, $|x'b|$ is high or $|x|$ is large, the bias will tend to be positive if $x$ is positive in this region and negative if $x$ is negative in this region.

However, knowing more about the conditions under which the higher order bias is small could allow researchers to characterize the bias using the more tractable formulas and scaling up marginal effects if misclassification is conditionally random. Thus, we additionally conducted several simulation exercises to further examine conditions under which the higher order bias is small. Results are available upon request. While the regularities we find are in line with our analytic results, the simulations are just case studies, so they may not generalize. We find that the higher order bias tends to be small, but can become large if misclassification depends heavily on $\varepsilon$ or the true value of $y$, for example because the models for false positives and false negatives differ. Both are likely to induce asymetries (such as skewness) in $F_{\tilde{\varepsilon}}$, which creates differences between $F_{\tilde{\varepsilon}|X}$ and $\Phi$ that are unlikely to even out over the sample.