

Is the Magic Still There? The Use of the Heckman Two-Step Correction for Selection Bias in Criminology

Shawn Bushway & Brian D. Johnson & Lee Ann Slocum

Springer Science+Business Media, LLC 2007

Abstract Issues of selection bias pervade criminological research. Despite their ubiquity, considerable confusion surrounds various approaches for addressing sample selection. The most common approach for dealing with selection bias in criminology remains Heckman's [(1976) *Ann Econ Social Measure* 5:475 – 492] two-step correction. This technique has often been misapplied in criminological research. This paper highlights some common problems with its application, including its use with dichotomous dependent variables, difficulties with calculating the hazard rate, misestimated standard error estimates, and collinearity between the correction term and other regressors in the substantive model of interest. We also discuss the fundamental importance of exclusion restrictions, or theoretically determined variables that affect selection but not the substantive problem of interest. Standard statistical software can readily address some of these common errors, but the real problem with selection bias is substantive, not technical. Any correction for selection bias requires that the researcher understand the source and magnitude of the bias. To illustrate this, we apply a diagnostic technique by Stolzenberg and Relles [(1997) *Am Sociol Rev* 62:494 – 507] to help develop intuition about selection bias in the context of criminal sentencing research. Our investigation suggests that while Heckman's two-step correction can be an appropriate technique for addressing this bias, it is not a *magic* solution to the problem. Thoughtful consideration is therefore needed before employing this common but overused technique.

Keywords Sample selection, Selection bias, Heckman correction, Two-step estimator

S. Bushway
School of Criminal Justice, State University of New York at Albany, Albany, NY, USA

B. D. Johnson (&) & L. A. Slocum
Department of Criminology and Criminal Justice, University of Maryland, 2220 LeFrak Hall,
College Park, MD 20742, USA
e-mail: bjohanson@crim.umd.edu

Any sufficiently advanced technology is indistinguishable from magic.
 -Arthur C. Clarke

Introduction

For several decades criminologists have recognized the widespread threat of sample selection bias in criminological research.¹ Sample selection issues arise when a researcher is limited to information on a non-random sub-sample of the population of interest. Specifically, when observations are selected in a process that is not independent of the outcome of interest, selection effects may lead to biased inferences regarding a variety of different criminological outcomes. In criminology, one common approach to this problem is Heckman's (1976) two-step estimator, also known simply as the Heckman. This approach involves estimation of a probit model for selection, followed by the insertion of a correction factor—the inverse Mills ratio, calculated from the probit model—into the second OLS model of interest.

After Berk's (1983) seminal paper introduced the approach to the social sciences, the Heckman two-step estimator was initially used by criminologists studying sentencing, where a series of formal selection processes results in a non-random sub-sample at later stages of the criminal justice process. Hagan and colleagues applied it to the study of sentencing in a series of papers that demonstrated the prominence of selection effects (Hagan and Palloni 1986; Hagan and Parker 1985; Peterson and Hagan 1984; Zatz and Hagan 1985). Since that time, the Heckman two-step estimator has been prominently featured in research examining various stages of criminal justice case processing, such as studies of police arrests (D'Alessio and Stolzenberg 2003; Kingsnorth et al. 1999), prosecutorial charging decisions (Kingsnorth et al. 2002), bail amounts (Demuth 2003) and criminal sentencing outcomes (e.g. Nobiling et al. 1998; Steffensmeier and Demuth 2001; Ulmer and Johnson 2004). An increasing number of studies have also begun to use the Heckman approach to examine outcomes across multiple stages of criminal case processing (e.g. Kingsnorth et al. 1998; Leiber and Mack 2003; Spohn and Horney 1996; Wooldredge and Thistlewaite 2004). Despite the continued prominence of sample selection issues in criminology, however, much confusion still surrounds the appropriate application and effectiveness of Heckman's two-step procedure. This paper presents a systematic review of common problems associated with the Heckman estimator and offers directions for future research regarding issues of sample selection. To keep the scope manageable, we focus our specific comments about the use and misuse of the Heckman on the 25 articles published in *Criminology* over the last two decades which utilized some form of the Heckman technique.²

Our review finds that the Heckman estimator tends to be utilized mechanistically, without sufficient attention devoted to the particular circumstances that give rise to sample selection. This contention is not new, nor is it a problem limited to

¹ These topics range from those concerned with sample attrition or non-response (e.g. Gondolf 2000; Maxwell et al. 2002; Robertson et al. 2002; Worrall 2002) to studies of racial profiling in police stops (Lundman and Kaufman 2003), to research examining the effects of race on criminal justice outcomes (Klepper et al. 1983; Wooldredge 1998).

² We identified articles by electronically searching for papers in *Criminology* that cite the seminal article by Berk (1983) and/or included the word Heckman. This process resulted in 25 articles that use the Heckman in some shape or form (7 other studies cite Heckman or Berk but were not directly relevant).

criminology. As one leading econometrician poignantly stated not long after Heckman introduced the approach:

It is tempting to apply the ‘Heckman correction’ for selection bias in every situation involving selectivity. This type of analysis, popular because it is easy to use, should be treated only as a preliminary step, but not as a final analysis, as is often done. The results are sensitive to distributional assumptions and are also often uninformative about the basic economic decisions that produce the selectivity bias. One should think more about these basic decisions and attempt to formulate the selection criterion on its structural form before the Heckman correction is even applied (Maddala 1985, p. 16).

As Maddala suggests, the Heckman estimator is only appropriate for estimating a theoretical model of a particular kind of selection; different selection processes necessitate different modeling approaches, which require different estimators.

We also find important problems with the way the Heckman estimator has been applied in the vast majority of the studies we reviewed. These errors include use of the logit rather than probit in the first stage, use of models other than OLS in the second stage, failure to correct heteroskedastic errors, and improper calculation of the inverse Mills ratio. Although some of these errors are more problematic than others, collectively they raise serious questions about the validity of prior results derived from miscalculated or misapplied versions of Heckman’s correction for selection bias.

Even when the correction has been properly implemented, however, research evidence demonstrates that the Heckman approach can seriously inflate standard errors due to collinearity between the correction term and the included regressors (Moffitt 1999; Stolzenberg and Relles 1990). This problem is exacerbated in the absence of *exclusion restrictions*. Exclusion restrictions, like instrumental variables, are variables that affect the selection process but not the substantive equation of interest. Models with exclusion restrictions are superior to models without exclusion restrictions because they lend themselves to a more explicitly causal approach to the problem of selection bias. They also reduce the problematic correlation introduced by Heckman’s correction factor. Our review of the literature identified only four cases in *Criminology* in which potentially valid exclusion restrictions were implemented. Unlike other approaches to selection bias (e.g. the bivariate probit), the Heckman model does not require exclusion restrictions to be estimated, but it is imperative that it is utilized cautiously and not simply as a matter of habit or precedent. In this paper, we discuss the controversy surrounding the use of Heckman without exclusion restrictions and provide some guidance about the costs and benefits of this approach.

We conclude the paper by highlighting two approaches which provide important intuition regarding the severity of sample selection bias and the potential benefits of correcting it with Heckman’s technique. The first approach, by Stolzenberg and Relles (1997), highlights a little-used technique which provides important intuition regarding the severity of bias in one’s data, and the second, suggested by Leung and Yu (1996), assesses the extent to which collinearity introduced by the correction factor is likely to be problematic. We apply these techniques to the study of criminal sentencing using a well-known dataset on criminal convictions in Pennsylvania. Our investigation suggests there is relatively little bias in key race effects when compared to other extralegal considerations, such as gender and ethnicity, which demonstrate the potential to be biased in substantively important ways. We begin with a

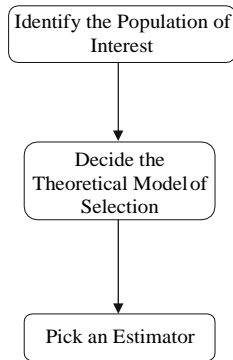


Fig. 1 The process of picking a good statistical estimator

conceptual discussion of the selection problem before proceeding to a more detailed discussion of the Heckman technique and its use in criminology.

The Conceptual Sample Selection Framework

Criminological research questions often involve non-random sub-samples of observations from a population of interest. Criminal sentencing research, for instance, is often concerned with the effects of race (or other status characteristics) on sentence lengths for convicted offenders. The problem is that sentence lengths are only observed for offenders who are incarcerated. This distinction is analogous to what economists label the ‘wage’ equation, where one is interested in returns to schooling on wages but only observes wages for employed individuals (Moffitt 1999). An important but seldom appreciated distinction is between this ‘wage’ equation and the ‘treatment’ equation, which involves identifying the impact of selection into a treatment on a given outcome. Spohn and Holleran’s (2002) recent study of recidivism among offenders sentenced to prison versus probation is an example of the treatment case. Selection into prison is potentially non-random, so a control for selection or unobserved heterogeneity is needed before the causal model is identified. Although the Heckman two-step method can be applied to both the wage and treatment equations, its use in criminology is dominated by the former.³ We therefore focus our discussion on its common application to the wage equation.

When faced with a sample selection problem, the process of selecting a good estimator involves three steps (see Fig. 1). The first step is to identify the relevant population based on the question of interest. Statistical models allow the researcher to make inferences from a sample to a population, so it almost goes without saying that we need to know our population of interest; however, this somewhat obvious issue becomes less trivial in models of selection. Sample selection occurs when a researcher is working with a non-random sub-sample from a larger population of interest. The estimate of the coefficient of interest, *conditional* on membership in the selected sub-group, will be a biased estimate of the parameter in the larger

³ Part of the reason for the limited application of Heckman in the treatment equation approach is the availability of a number of other approaches for dealing with selection, including experiments, instrumental variable estimation and panel models. For useful in-depth discussions of alternative approaches to the treatment effects model, see Halaby (2004), Angrist (2001), and Heckman et al. (1999).

population of interest.⁴ The estimate for this larger population is sometimes called the *unconditional* estimate. Researchers modeling sentence length, for example, are often interested in drawing inferences to the entire population of convicted offenders, despite only having data on sentence lengths for incarcerated offenders. The distinction between the conditional and unconditional populations is non-trivial because it has consequences for subsequent modeling decisions.

As Berk (1983) has shown, both the conditional and unconditional estimates are subject to potential selection bias. Claiming an interest in only the observed subsample of offenders, therefore, does not alleviate selection bias concerns. Even the estimates for the marginal impact of an effect on the observed population (i.e. the conditional estimate) can be subject to selection-induced biases, depending on the type of bias that exists. The more common interest, however, is in the larger, unconditional population of interest. Whereas the conditional estimates refer to the *actual* population, the unconditional estimates refer to the *potential* population, which consists of everyone who could have possibly been selected. Despite its obvious importance, rarely is this distinction explicitly recognized.

The second step involves defining the proper theoretical selection model. There are a number of theoretical models of selection that can be used to describe the substantive selection process, but in what follows we focus on the two main types—explicit and incidental. After choosing an appropriate model, the final step is to choose a suitable statistical estimator.

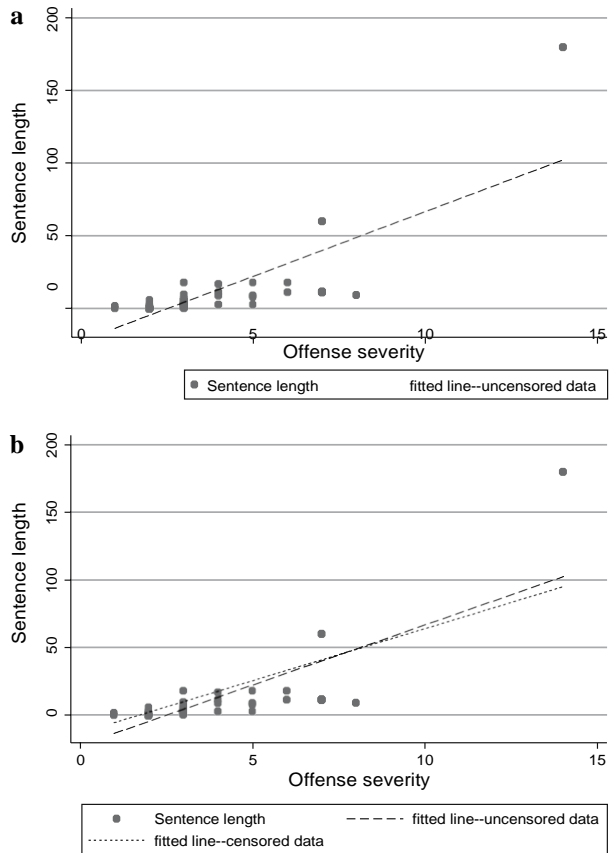
Estimators are just different ways to generate estimates of the parameters in the models under study. Each estimator carries its own set of assumptions that are used to identify an estimation strategy. Estimators can be chosen based on a number of criteria including consistency, which deals with the amount of bias in the estimate, and efficiency, which concerns the size of the standard error. In addition, the ease of estimation and unrestrictiveness of underlying assumptions represent additional qualities of desirable estimators. Econometricians often speak about *identification* of these estimators. In this context, identification is the key feature(s) of the model which allows or facilitates estimation. Although there are several different classes of available estimators (e.g. semi-parametric and non-parametric estimators, see Vella 1999 and Moffitt 1999), we focus here on the common class of bivariate normal estimators, which assume normally distributed errors.

Explicit Selection

As the above discussion suggests, different types of selection necessitate different model specifications and statistical estimators. A number of classic econometric treatments of these issues already exist (Amemiya 1985; Heckman 1979; Maddala 1985, Vella 1999), along with versions written specifically for sociologists (Berk 1983; Berk and Ray 1982; Stolzenberg and Relles 1997; Winship and Mare 1992). We will not completely replicate those discussions here, but for illustrative purposes we begin by providing a brief mathematical representation of the basic selection model.

⁴ It is important to distinguish between the terms conditional and marginal. A slope coefficient is an estimate of the marginal impact of x on y . We can usually estimate two different terms—the marginal effect of x in the conditional and unconditional models. As Greene (1993) makes clear, the estimate of interest will vary by case.

Fig. 2 (a) Fitted regression line for a random sample of 54 cases. (b) Fitted regression line with observations censored below 9



We then utilize this model to discuss the appropriateness of different estimators for different types of selection.

We begin by replicating Berk's (1983) Figure 1, a simple bivariate scatterplot of X and Y (see Fig. 2a). To make the conversation concrete, we will use sentencing as the example. Y will be sentence length and X will be crime severity. To begin, sentence length will be a relative scale that represents the judgment of the actor(s) in the system about the severity of the punishment. This scale will be translated into real values at a later time. Crime severity can be imagined to be a score such as a typical guideline score of the penalty. As with Berk, we have 54 observations in our population of interest. The simple scatterplot and fitted regression line make it clear that there is a strong linear relationship between crime severity and punishment in fitting a strict retributive model of punishment. We represent the simple model in (Eq. 1), where Y_1^* is the dependent variable sentence length, X_1 is the independent variable crime severity and e is a normally distributed error term:

$$\delta Y_1^* = \beta_0 + \beta_1 X_1 + e \tag{1}$$

The premise of the sample selection problem is that we will not actually capture all 54 observations, but rather a subset. If this subset is just a random subset, clearly

we would expect no selection bias. Simple OLS theory says that a random sample of a population will allow us to make unbiased inference about the coefficients of interest. In this case, the simple *two part model* (TPM), or the ‘ ‘uncorrected’ ’ model, will be the appropriate modeling strategy (Duan et al. 1983, 1984).⁵ In the TPM, selection into the subset is modeled as a dichotomous dependent variable using probit.⁶ Then a second outcome of interest is modeled for only the selected subset of observations using OLS. In the context of our sentencing example, the likelihood of incarceration would be modeled followed by a separate analysis of sentence length for only those offenders who were incarcerated. Results from the sentence length regression would therefore represent the influence of offense severity on sentence length, *conditional* on being selected into the incarcerated population. These simple models are presented in Eqs. 2.1 and 2.2 below where P_1 is a dummy variable equal to 1 if the person is selected and 0 otherwise:

$$\begin{aligned} \text{Probit } \delta P_1 &= \beta_0 + \beta_1 Z_1 + \epsilon & \delta 2:1P \\ \delta Y_1 &= \beta_0 + \beta_1 X_1 + \epsilon & \delta 2:2P \end{aligned}$$

There are many ways that non-random selection can arise, but as discussed above they generally fall into two basic categories, explicit and incidental selection. Explicit selection is the simplest, and continues to involve only one dependent variable, Y_1^* . In this case, we will not observe the true value of Y_1^* for certain values of Y_1^* . This is the simplest case because the selection is based explicitly on the outcome of interest. But, to complicate the example, the sample may be either truncated or censored. A *truncated* sample occurs if we have unobserved data for both sentence length and crime severity. A sample is considered *censored*, however, if we have information on our independent variable of interest, i.e. severity, but lack information on the outcome of interest, i.e. sentence length.⁷ In (Eq. 2.2), Y_1 is what we observe, and we can write down the cutoff values for censoring/truncation. These cutoff values can take different forms in different theoretical frameworks. There can be multiple cutoff values on either end of the distribution, they can vary by individual, and they can even be unknown. In Fig. 2b, we have demonstrated the case of censoring where all values of Y_1^* less than 9 have been set equal to 9. In this case,

⁵ An additional distinction can be made between the simple Two Part Model (TPM) and the censored two stage model (CTSM). Whereas the substantive regression in the TPM is based on the sub-sample of selected cases, for the CTSM, it includes all observations (including zeros for unobserved values). The TPM provides the *conditional* estimate given selection into the sub-sample of the population whereas the CTSM provides the *unconditional* estimate based on all observations. This distinction can be useful in certain applications, but it is often confusing because the CTSM requires that censored observations have meaningful 0 values (e.g. sentence lengths of 0 months of incarceration). This distinction is additionally muddled by the common convention of using the TPM to obtain estimates that ostensibly represent the unconditional population of interest (Puhani 2000). We constrain our discussion to the TPM given its focus in the literature, but further comparison of these two estimators offers an additional line of potentially interesting future research.

⁶ While it is possible to estimate the TPM with alternative specifications, such as the logit instead of the probit, it is important to recognize that the estimator then assumes a different error structure (i.e. errors are distributed log normal).

⁷ This distinction is relevant to the extent that slightly different modeling procedures are appropriate for the two types of samples. Truncated regression models are more appropriate for situations with truncated samples whereas limited-dependent-variable models or Tobit models are generally preferred for censored samples (see Fig. 3).

the estimated regression line will be fatter than the real line, leading to an underestimate of the coefficient of interest, and the more censoring/truncation, the more bias. The most common estimator of the explicit selection model presented in (Eq. 3) is the Tobit model (Tobin 1958).

$$Y_1 = \begin{cases} Y_1^* & \text{if } Y_1^* \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{Eq. 3}$$

Tobit regression has been increasingly applied in the field of sentencing research (Albonetti 1997; Bushway and Piehl 2001; Kurlychek and Johnson 2004; Rhodes 1991), as well as in criminology more broadly (see Osgood et al. 2002).⁸

Incidental Selection

A slightly more complicated type of selection is incidental selection. Here, we have another endogenous variable, Y_2 , which determines the selection process. We refer to the equation for Y_2 as the selection equation, and the equation for Y_1^* as the substantive equation. In the case of incidental selection, the researcher specifically models the factors that influence the selection process. Again, there can be censoring and truncation, and the actual form of Y_2 and the cutoff values will depend on the specific model. Several other things also can vary in an incidental selection model, including whether we believe that the error terms for the selection and substantive equations are correlated. If the error terms are entirely uncorrelated, then by definition selection bias is not a problem. In the case where the error terms from the selection and substantive equations are truly independent, the correct estimator is simply the TPM. The first step would be estimated with a probit and the second with a simple OLS. No additional technique is required.

Under most circumstances, however, the assumption of independent error terms will not be met because of specification problems. If any factors that affect both the selection and substantive equations are omitted from the model, these factors will enter both error terms and induce correlation between them. For instance, recent reviews of the sentencing literature have highlighted the fact that offender socioeconomic status is often an omitted variable (Zatz 2000). If class status does indeed affect the likelihood of being incarcerated and the length of incarceration, these two decisions are no longer statistically independent. As a result of omitted variables, most models cannot assume independent error terms, leading to biased estimates if the TPM is used for model estimation. When selection bias is present, the size of the bias is driven primarily by the correlation between the errors in the two models, the amount of error in the substantive regression (i.e. the model fit), and the degree of censoring (Stolzenberg and Relles 1990).

Importantly, the independence of error terms cannot be directly tested. Therefore, a common response in the literature is to compare the estimates from the TPM

⁸ The Tobit model also has some limitations. These limitations include restrictive normality assumptions regarding the dependent variable (Chay and Powell 2001; Chesher and Irish 1987) restrictive homoskeasticity assumptions regarding error terms (Wooldridge 2005), and the key assumption that the effects of independent variables are constant for the selection process and the outcome of interest (Smith and Brame 2003). See Osgood et al. (2002) for a recent overview of Tobit models.

to the estimates from one of several models that attempt to correct for selection bias. Researchers then argue that if their results remain the same, selection is not a problem; however, as we discuss in detail below, this approach assumes that the model correction was valid, an assumption that may not be appropriate.

The Heckman Estimator and its Problematic Use

The most common approach to estimate incidental selection models involves a class of estimators known generically as bivariate normal selection models.⁹ The key feature of these models is that the error terms in the two equations are distributed bivariate normal. When the dependent variable of interest is continuous, there are two basic choices of estimators—a maximum likelihood model sometimes called the Full Information Maximum Likelihood (FIML) and the Heckman two-step estimator.¹⁰ The FIML is a straightforward maximum likelihood model, like a probit or logit, that maximizes a specified likelihood function. By definition, when the error assumptions are met the FIML will always be more efficient than the Heckman two-step, a fact which has been demonstrated in numerous simulation studies (Leung and Yu 1996; Maddala 1985; Puhani 2000). However, the FIML relies more heavily on the normality assumption and is therefore less robust than the Heckman two-step to deviance from that assumption. The FIML may have difficulty converging, particularly in the absence of exclusion restrictions, while the Heckman two-step model can almost always be estimated. As we will discuss below, the robustness of the Heckman can be a mixed blessing, since it may provide a false sense of security.

Criminological research has relied exclusively on the Heckman two-step estimator,¹¹ which is based on the recognition that the sample selection problem is really an example of omitted variable bias (Heckman 1979). To illustrate this, the precise form of the Heckman two-step method is presented below, where (Eq. 4.1) is the selection equation and 4.2 the substantive equation of interest.

$$Y_2 = aZ + d \tag{4.1}$$

$$Y_1 = B_0 + B_1X + r\epsilon - aZ + r^0 \tag{4.2}$$

In the selection equation, which is estimated with a probit, Y_2 is the dichotomous dependent variable, Z is the independent variable, a is the coefficient of Z , and d is the normally distributed error term. In the regression equation the value of Y_1 is observed when Y_2 is greater than some threshold T , and it is censored (i.e. missing) if $Y_2 \leq T$. Estimation of 4.2 by simply regressing Y on X will be biased because of the sigma term, which represents the omitted variable. This problem can be solved in

⁹ There is also a large statistical literature that attempts to loosen the assumption of bivariate normality by estimating the first stage semi-parametrically, see especially Ahn and Powell (1993) and Blundell and Powell (2001) as well as Vella (1999) and Newey (1999).

¹⁰ STATA 8.2 provides for both the FIML and Heckman two-step estimators, while LIMDEP 7.0 also provides for a third maximum likelihood estimator of the Heckman two-step, sometimes called the Limited Information Maximum Likelihood (LIML).

¹¹ We are unaware of any case in which the FIML (or the related LIML model) has been used in criminology. This is somewhat ironic because Heckman originally recommended using the two-step approach only to generate starting values for the FIML estimator (Heckman 1976).

Table 1 Use of Heckman's correction for selection bias in criminological research

Author(s)	Year	Heckman calculated/reported	1st stage of analysis	2nd stage of analysis	Used inverse Mills ratio ^b	Reported collinearity problems	Reported exclusion restriction
Bontrager et al.	2005	Calculated, not reported	Unreported	HLM Logit	Hazard	Yes	No
Steen et al.	2005	Calculated, reported	Logit	OLS	Probability	No	No
Ulmer and Johnson	2004	Calculated, reported	HLM Logit	HLM Linear	Hazard	No	No
Stolzenberg et al.	2004	Calculated, not reported	Logit	HLM Logit	Hazard	No	No
Wooldredge and Thistlethwaite	2004	Calculated, reported	Logit	HLM Logit and Linear	Probability	No	No
Demuth	2003	Calculated, reported	Logit	Logit and OLS	Hazard	No	No
Kingsnorth et al.	2002	Calculated, not reported	Probit	OLS	Unreported	No	Yes
Spohn and Holleran	2002	Calculated, reported	Unreported	Logit and Survival Analysis	Probability	No	Yes
Steffensmeier and Demuth	2001	Calculated, reported	Logit	OLS	Hazard	No	No
Engen and Gainey	2000	Calculated, reported	Logit	OLS	Unreported	No	No
Wright and Cullen	2000	Calculated, reported	Logit	OLS	Hazard	No	Yes
MacMillan ^a	2000	Calculated, reported	Probit	OLS	IMR	No	Yes
Steffensmeier et al.	1998	Calculated, reported	Logit	OLS	Hazard	No	No
Felson and Messner	1996	Not calculated, not reported	Unreported	Logit	N/A	Yes	No
Ulmer and Kramer	1996	Calculated, reported	Logit	OLS	Unreported	No	No
Felson	1996	Calculated, not reported	Unreported	Logit	Unreported	Yes	No
Steffensmeier et al.	1993	Calculated, not reported	Unreported	OLS	Probability	Yes	No
Chiricos and Bales	1991	Calculated, not reported	Probit	OLS	Probability	No	No
Erez and Tontodonato	1990	Calculated, reported	Logit	OLS	Probability	Yes	No
Keil and Vito	1989	Calculated, reported	Logit	Logit	Probability	Yes	No
Myers	1988	Calculated, reported	Logit	WLS	Probability	Yes	No
Hagan and Palloni	1986	Calculated, reported	Probit	OLS	IMR	Yes	No
Myers and Talarico	1986	Calculated, reported	Logit	OLS	Probability	Yes	No
Albonetti	1986	Calculated, not reported	Probit	Probit	IMR	No	No
McCarthy and Smith	1986	Calculated, reported	OLS	SEM	Probability	No	No

Note: Our review of *Criminology* studies also included Meithe and Moore (1985), Tibbets and Piquero (1999), Messner et al. (2002), Lundman and Kaufman (2003), and Messner et al. (2005). However, none of these studies dealt explicitly with Heckman's two-step estimator in the selection (non-treatment) equation

^a This study examined selection issues in two separate datasets. Heckman's two-step estimator was examined but not included in analyses using the National Youth Survey because the selection model revealed no significant predictors of sample selection (560). It was included in analyses using the Canadian General Social Survey. The entry in the table refers to these latter analyses

^b Table 1 represents our best effort to identify whether or not the inverse Mills ratio was properly calculated. Many studies are not explicit about this process. Studies coded Unreported do not contain necessary information, studies coded Probability report using predicted probabilities, studies coded Hazard state they calculated a hazard rate but do not say how, and studies coded IMR clearly used the inverse Mills ratio or reported a lambda (k) selection coefficient

two steps. First the selection equation (Eq. 4.1) is estimated using probit and the predicted values are retained as estimates of $T - aZ$. The inverse Mills ratio is then estimated for each case by dividing the normal density function evaluated at $-(T - aZ)$ by one minus the normal cumulative distribution function estimated at $-(T - aZ)$ (Eq. 5).¹² The second step is an ordinary least squares regression with X and the inverse Mills ratio included as regressors. The estimator is consistent when the assumptions are met.

$$k\delta T - aZ\beta \frac{\delta T - aZ\beta}{1 - U\delta T - aZ\beta} \quad \delta\beta$$

Our review of articles in which the technique is applied in *Criminology* suggests potentially serious errors with its application. These problems include the use of a logit rather than probit in the first stage, the application of Heckman to discrete substantive outcomes, the use of predicted probabilities in place of the inverse Mills ratio, and the use of uncorrected standard errors in statistical significance tests. Table 1, which represents our best effort to code the articles in *Criminology* that use the technique, provides a sense of how endemic these problems are in contemporary criminological research.

As Table 1 indicates, most studies that incorporate the Heckman two-step procedure rely on the logit rather than the probit model to estimate the first stage (less than 25% of the studies reviewed used probit). Although the logit and probit are sometimes viewed as interchangeable, they each make different assumptions about functional form—the probit uses normality and the logit uses log normality. All of the features of the Heckman estimator are based on the assumption of bivariate normality and therefore require the use of the probit. Substantively this distinction is not likely to have a substantial impact on one’s findings unless the probability of being selected into the substantive equation is close to 1 or 0; however, the problem highlights the common confusion surrounding the appropriate calculation of Heckman’s two-step estimator.

In addition, the Heckman commonly has been used for cases of incidental selection even when the dependent variable of interest is not continuous. As one example, Wooldredge and Thistlewaite (2004) recently estimated logit models of dichotomous case processing outcomes for a sample of intimate assaults. Specifically, they studied whether or not a charged individual was prosecuted, whether or not a prosecuted person was convicted and whether or not a convicted person was jailed. In each model, they attempt to use the Heckman two-step method by incorporating the predicted probability of making it to the prior stage of justice processing. Other criminological scholars have taken the same approach for different dichotomous outcomes (e.g. Albonetti 1986; Demuth 2003; Keil and Vito 1989; Stolzenberg et al. 2004). However, the Heckman two-step estimator is specifically a probit model followed by a linear regression, and there is no simple analog of the Heckman

¹² Considerable confusion surrounds the predicted probability, inverse Mills ratio, and what is often referred to as the ‘hazard rate’. The probability is simply the predicted value from the probit equation. To calculate the inverse Mills ratio, this value is multiplied by negative one and inserted into Eq. 5. This inverse Mills ratio, then, represents the hazard rate, or the instantaneous probability of exclusion for each observation conditional on being at risk (see Berk 1983).

method for discrete choice models despite the logical appeal of the process (Dubin and Rivers 1990, p. 411).¹³

Alternative estimators, such as the bivariate probit, have been suggested by Heckman, among others. Lundman and Kaufman (2003) correctly estimate the bivariate probit, but the model fails to converge because of the absence of exclusion restrictions. They then attempt to apply the ‘logic of the Heckman’ to their model, by including a variable measuring other police contact to control for selection into traffic stops.

An additional problem involves the calculation of the inverse Mills ratio. The convention in much research is to substitute the predicted probability from the first stage (or the negative or inverse of the probability) rather than to calculate the actual hazard rate based on the inverse Mills ratio. As Table 1 suggests, only three papers offer clear evidence that the inverse Mills ratio was actually applied (Hagan and Palloni 1986; Albonetti 1986; MacMillan 2000). Because most studies are not explicit about these calculations, it is often difficult to assess whether the inverse Mills ratio was properly calculated, but it appears as though the predicted probability is often substituted, even in studies that purport to calculate a hazard rate.¹⁴ While the two are clearly related, they are not identical. In fact, when the same variables are used to model the selection and substantive equations (i.e. when exclusion restrictions are not utilized), the model is only identified by the non-linearity inherent in the inverse Mills ratio. A model that includes the predicted probability from an OLS regression when $Z = X$, then, is not formally identified and should not be estimated.¹⁵

Another common error in the Heckman approach is a failure to properly correct for misestimated standard errors. As Heckman himself stated, ‘the standard least squares estimator of the population variance ... is downward biased’ and therefore ‘the usual formulas for standard errors for least squares coefficients are *not* appropriate except in the important case of the null hypothesis of no selection bias’ (Heckman 1976, pp. 157 – 158, emphasis in original). Because the data are censored, the variance estimates obtained will be smaller than the true population variance. This, in turn, produces underestimated standard errors in the second stage of the Heckman two-step model. As a result, researchers need to correct these standard errors using a consistent errors estimator (often referred to as robust standard

¹³ Dubin and Rivers (1990) discuss alternative formulations of both a two-step and maximum likelihood adaptation of Heckman’s model to the case of dichotomous outcomes, but we are unaware of any criminological application of these procedures.

¹⁴ As one example, Nobiling et al. (1998, p. 470) describe this convention as follows: ‘We used logistic regression to estimate the likelihood that the offender would be sentenced to prison. For each case, the logistic regression produced its predicted probability of exclusion from the sentence length model—the hazard rate.’ An anonymous referee suggested this approach represents an altogether different estimator based on propensity scoring, and should therefore not be confused with the Heckman approach. Given that each study reviewed cited Heckman (1976, 1979) or Berk (1983) and no study made explicit reference to the literature on propensity scoring, we feel justified in associating this approach with the Heckman correction. It is possible that there is statistical merit in such an approach, but we leave it to other researchers to verify it in this context.

¹⁵ In the treatment effects version of the Heckman estimator, there are actually two different formulas for the inverse Mills ratio depending on whether an individual case is selected. Failure to correctly specify the inverse Mills ratio can bias one’s findings in the face of selection. Spohn and Holleran (2002) is an example of an application of the treatment model where the probability was substituted for the inverse Mills ratio. Correspondence with the authors indicated that the coefficient on the inverse Mills ratio was not significant when the model was re-estimated correctly, suggesting an absence of selection bias in their study. We thank Cassia Spohn and David Holleran for their willingness to engage in a constructive discussion regarding their paper.

errors). Although this adjustment is automatically performed in readily available statistical packages (e.g. Limdep 7.0 and STATA 8.2), in our experience, calculations are usually done by hand and standard errors are not adjusted.¹⁶ Because the correction of standard errors is rarely mentioned, it was not possible to code its correct application. In fact, we found no articles in our *Criminology* sample that discussed corrected standard errors, and only one in the larger literature by Kingsnorth et al. (1999). This suggests the possibility that statistical significance is likely to be overstated in many of the prior applications of Heckman's two-step correction for selection bias. Moreover, the magnitude of overstated significance can be substantial. As discussed below, standard errors were underestimated by 10–30% depending on the model specification.

Exclusion Restrictions and Model Identification

In all of the cases discussed above, the errors in application, while troubling, are easily fixed. Researchers today have access to standard statistical packages, such as Limdep 7.0 and STATA 8.2, which automatically calculate Heckman's estimator correctly. The virtual explosion in computer power in recent years has made the once taxing calculations commonplace, and these programs now provide fast and easy calculations of Heckman's two-step and FIML estimators that preclude common user mistakes. However, one additional problem remains that cannot be solved so easily—the inclusion of the inverse Mills ratio often results in multicollinearity that can have profound consequences for model estimates.

Because the inverse Mills ratio is estimated by the non-linear probit model, the correction term k will not be perfectly correlated with X , even in the absence of exclusion restrictions (i.e. when all the variables used to estimate Step 1 (Z in Eq. 4.1) are the same as covariates in Step 2 (X in Eq. 4.2)). This is the essential feature of the two-step estimator that allows the model to be identified without exclusion restrictions. If the first stage was linear, the model would not be identified and could not be estimated. It is the non-linearity of the probit in the first stage that allows us to generate an answer for the Heckman two-step estimator. However, the probit model will be linear for the mid-range values of X , and is truly non-linear only when X takes on extreme values. As evidence of this problem, scholars often report very high correlations between k and regressors in the substantive equation, which lead to large standard errors. For example, Myers (1988) and Myers and Talarico (1986) reported correlations of .9 or higher between k and crime severity and were only able to estimate the model by excluding crime severity from the sentence length equation.

A large Monte Carlo literature, summarized in Puhani (2000), demonstrates that both the FIML and the Heckman suffer from inflated standard errors when the covariates in the selection and regression equations are identical. For these reasons, Stolzenberg and Relles (1990) conclude that even though estimates from the uncorrected TPM will always be biased in the presence of selection, they may be preferred to estimates obtained using the Heckman correction without exclusion restrictions given that estimates from the later model are inefficient. In other words, the TPM may represent a better approach in some cases because any one estimate

¹⁶ The majority of studies in Table 1 use logit rather than probit, a feature not available in the standard software programs. Also, it is unlikely that any paper that describes using the probability rather than the IMR used standard software.

will be closer to the ‘right’ answer than a given estimate from the Heckman, even though the Heckman estimates remain unbiased and consistent.¹⁷ Researchers in criminology have frequently acknowledged the problem of inflated standard errors (i.e. inefficiency) in Heckman selection models, often using it as justification for not using the Heckman estimator (e.g. Bontrager et al. 2005; Felson 1996; Felson and Messner 1996; Leiber and Jamieson 1995; Lundman and Kaufman 2003; Nobiling et al. 1998; Sorenson and Wallace 1999; Steffensmeier et al. 1993).¹⁸

The best solution to this problem is to incorporate exclusion restrictions. With a valid exclusion restriction, the inverse Mills ratio and the X vector in the substantive equation will be less correlated, reducing multicollinearity among predictors as well as the correlation between error terms.¹⁹ This also facilitates model identification. Admittedly it can be difficult to identify appropriate exclusion restrictions; however, the potential benefits justify the effort, and we believe criminologists will be able to identify good exclusion restrictions if they begin to look for them. Excluding a variable from the X vector which clearly has a strong impact on the substantive equation does not solve the problem, and neither does including a variable in the selection equation which has little power to predict selection (Heckman et al. 1999), so careful thought must be devoted to the choice of exclusions.²⁰

Ultimately, the argument about whether or not a variable can properly be excluded from the outcome equation is one that must be made on substantive rather than technical grounds. That is, researchers must make convincing theoretical justifications for their choice of exclusion restrictions. This is rare in criminology.²¹ For example, Wright and Cullen (2000) correct for selection bias in occupational delinquency by first modeling the probability that an adolescent will be employed, and then including it as a hazard term in subsequent models of delinquency. They use measures of adolescent depression and dating behavior to predict employment

¹⁷ Desirable estimators have three properties—they are unbiased, efficient and consistent. An unbiased estimate means the difference between the true parameter and the expected value of the estimator of the parameter is 0. This has to do with how well our point estimate represents the true population value of interest. Efficiency has to do with the standard error of the estimate. More efficient estimates have smaller standard errors, meaning any one estimate is more likely to approximate the true population value. Finally, consistent estimates are estimates that become unbiased as N approaches infinity. While the TPM estimates are biased, they will be more efficient, whereas the Heckman estimates will be unbiased and consistent, but less efficient. Which estimate better captures the true parameter of interest depends on the degree of bias and inefficiency that exists in each.

¹⁸ This list includes articles outside of our *Criminology* sample. This list was compiled using a search of articles that cite Stolzenberg and Relles.

¹⁹ Exclusion restrictions are very similar to instrumental variables (Angrist 2001). In each case, we try to identify factors that affect one variable but do not affect a second variable. The ultimate exclusion restriction would be experimental or random assignment. Clearly, if people are randomly assigned to prison, then there will be no correlation between the error terms of the prison and sentence length equations, and there will be no selection bias. In the absence of experiments, we need to find quasi-experiments or elements of the selection process that are uncorrelated with the substantive decision of interest.

²⁰ One possible exclusion restriction for sentencing research is strength of evidence. Early studies suggest that this factor affects the likelihood of conviction, but not the severity of the sentence post-conviction (Albonetti 1991).

²¹ To be more comprehensive, we attempted to identify additional papers outside the journal *Criminology* that used the Heckman with exclusion restrictions. We did find four other examples of models using the Heckman correction and incorporating exclusion restrictions (Albonetti 1991; Maxwell et al. 2002; Robertson et al. 2002; Worrall 2002), but none of the articles we reviewed defended or discussed their exclusion restrictions in any detail.

but not to predict delinquency (except through their impact on employment). These two factors therefore represent potentially important exclusion restrictions, but no justification is provided regarding their relevance for employment but not delinquency. This is problematic because the substantive and theoretical justification is the essential component that enables one to claim that the selection process has been accurately modeled.

Far more common is the case where no exclusion restrictions are incorporated (21 out of the 25 articles we reviewed in *Criminology*, see Table 1).²² In these cases, it is not uncommon for the authors to report little difference in the results with and without the Heckman correction. These types of results have been offered as evidence against substantial selection bias in some work, but an alternative and equally compelling interpretation is that the estimator has failed to capture the selection effect. This may be particularly true if the estimator has been implemented incorrectly. Criminologists have adopted the first explanation, while economists subscribe to the latter. When economists cannot find a valid exclusion restriction, they simply estimate both the admittedly biased TPM and the unconditional model, and acknowledge they have a problem with selection (see Moffitt 1999). A substantial number of papers in criminology, cited above in ‘Exclusion restrictions and model identification’, follow a similar approach, and it is likely others would have followed suit if not pushed by reviewers to include a selection correction. We believe it is better to estimate the TPM and admit the problem than to pretend to solve it by mechanically applying the Heckman correction. Ideally, researchers should strive to model selection with natural experiments when possible, but in the absence of natural experiments, the goal should be to develop more detailed intuition about the threat of selection bias in one’s data. In the following section, we detail two methods for doing so—one that examines the overall potential for Heckman’s correction to provide improved model estimates, and one that offers specific insight into the amount of bias present in specific coefficients of interest.

An Empirical Example: Selection Bias in Criminal Sentencing

A number of recent attempts have been made to develop intuition about the size of selection bias faced in any given problem, and a number of studies have evaluated the relative performance of competing estimators (e.g. Klepper et al. 1983; Maddala 1985; Puhani 2000; Stolzenberg and Relles 1997; Winship and Mare 1992). Knowing something about the size and underlying reasons for the selection problem helps one to better understand how important the problem is, and how beneficial or detrimental different approaches will be in attempting to address it. Because Heckman’s correction has been applied frequently in research on criminal sentencing, we utilize that example as a vehicle for our discussion. In what follows, we continue to maintain the assumption of normality adopted by Heckman.²³

²² In Table 1, we counted those articles which arbitrarily excluded variables that were highly correlated with the correction factors in order to estimate the model as not having exclusion restrictions.

²³ There is a substantial literature on alternative estimation frameworks, including non-parametric frameworks, which avoid some of the problems associated with the restrictive parametric assumptions required by the bivariate normal models (Ahn and Powell 1993, Chay and Lee 2001). While we do not deny the potential usefulness of these alternative frameworks, we have chosen to remain within a framework known to most criminologists.

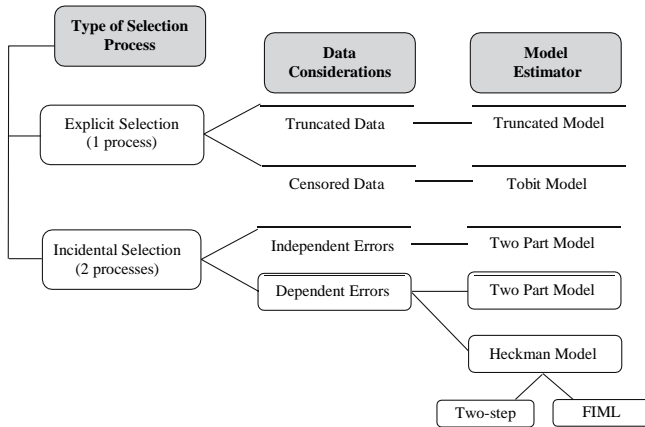


Fig. 3 Flow chart of common selection decisions in the choice of model estimators

As Fig. 1 suggests, the first step in selecting an appropriate modeling strategy is to identify the population of interest. Assuming we are interested in modeling the length of incarceration for convicted offenders, we need to select a population to whom we will generalize our findings. In typical models of sentence length it is presumed, though seldom explicitly stated, that we are interested in unconditional estimates—that is in estimates for some larger population of interest beyond those offenders sentenced to incarceration. However, it is often unclear whether the true population of interest is all criminals, all arrested criminals, all convicted criminals, or some other undefined group. Although some sentencing scholars have suggested that the totality of criminal events is the appropriate population of interest (Klepper et al. 1983), it is difficult to imagine a dataset that would allow reliable inferences to this elusive populace. We find the argument for the population of all arrested offenders compelling (Klepper et al. 1983; Reitz 1998; Smith 1986), but for the present illustration, we follow standard practice and focus on the population of convicted offenders. In this context, the main concern is that the coefficients predicting sentence length will be biased because we only observe sentence length for those individuals receiving incarceration.

Once we have decided to focus on the convicted population, we need to make a decision about the appropriate selection model and its estimator. Figure 3 provides a heuristic device for conceptualizing this process. The first distinction to be made is between explicit and incidental selection processes. If judges are sentencing on a continuum that includes probation and incarceration, then an explicit selection process is occurring. Offenders sentenced to probation or any other alternative to incarceration have a sentence which is censored to the researcher. This censoring is an empirical problem. In the case of criminal sentencing data, information is typically available for the covariates in both incarceration and non-incarceration cases. The distribution is therefore censored (rather than truncated) and the Tobit model provides an appropriate modeling strategy. Although the use of the Tobit model has become increasingly popular in sentencing research (e.g. Albonetti 1997; Bushway and Piehl 2001; Kurlychek and Johnson 2004), we will follow the more traditional assumption that judicial decision making occurs in two distinct stages: judges decide whether or not to incarcerate and *then* they decide the length of the sentence for

those receiving incarceration (Wheeler et al. 1982). This type of theoretical argument supports an incidental selection process.²⁴ Ultimately, it is up to the researcher to provide clear theoretical justification for the type of selection occurring in their particular research context. There is no empirical test that can ‘prove’ one approach is better than the other.

If the incarceration and sentence length decisions are independent of one another, then their error terms will not be correlated and selection bias is not a problem. In this case, estimates from the simple TPM will be unbiased and preferred. Such a scenario, however, relies on the unlikely assumption that the selection equation accurately and wholly reflects the selection process. Recall that to the extent that the same unobservable characteristics influence both processes, the error terms from the two equations will remain correlated. Because sentencing data (and criminological data in general) seldom contain the full gamut of operative influences in the decision making process, it is prudent to assume selection effects exist when modeling these outcomes. Given that judicial sentencing decisions involve incidental selection with correlated error terms in the incarceration and sentence length decisions, the researcher is left with two different modeling options. One is to implement a correction model like the Heckman estimator and the other is to simply rely on the uncorrected estimates from the TPM. This is a common dilemma faced not only by sentencing scholars, but by researchers concerned with a broad array of criminological research questions, and as we demonstrate, it can have important consequences for the conclusions one draws about different relationships of interest.

Two Approaches to Assessing Selection Bias

In the face of incidental selection, criminological research seldom provides adequate justification for incorporating or failing to incorporate Heckman’s estimator. At least in part, this is the product of a lack of clear criteria for assessing its appropriateness. We focus here on two complementary approaches that are likely to provide useful insights for researchers concerned with selection bias. The first approach was explicated by Stolzenberg and Relles (1997). Earlier work by Stolzenberg and Relles (1990) is sometimes cited by criminologists for its finding that Heckman’s correction can lead to highly unstable estimates if used without exclusion restrictions. It is typically employed as justification for not using Heckman’s estimator even when selection bias is explicitly acknowledged as a potential problem (Felson and Messner 1996; Kautt 2002; MacMillan 2000; Sorenson and Wallace 1999). Unfortunately, more recent follow-up work by Stolzenberg and Relles (1997), which provides useful techniques for assessing the amount of selection bias in one’s estimates, has garnered less attention from the field. Our systematic review of 64 articles that cite their work failed to identify a single study that has employed their method for assessing selection bias. Our goal here is to reacquaint the field with this method in the hope that criminologists concerned with selection bias will begin to carefully assess the

²⁴ The assumption of a two-stage decision making process in criminal sentencing largely grew out of early research on white collar offenders in federal districts that argued ‘the first and hardest decision the judge makes is whether the person will go to prison or not,’ which is ‘experienced as qualitatively different from the decision as to how long an inmate should stay in prison’ (Wheeler et al. 1982, pp. 642, 652). This assumption has been challenged recently and may be less valid for cases sentenced in states with presumptive sentencing guidelines (Bushway and Piehl 2001).

size of the problem, and then carefully apply methods which might correct the problem when appropriate.

Stolzenberg and Relles outline a methodology that can be used to calculate the amount of bias we would expect in our estimation if we did not correct for selection. We provide only a brief explanation of these calculations because a more detailed treatment is available in Stolzenberg and Relles (1997). In the following equations $\hat{\theta}_{B_1}$ represents this predicted bias.

$$\hat{\theta}_{B_1} = \frac{r q_{ed} q_{x_1 \hat{Z}_{x_1 \dots x_p}} k' T - a_1 Z_1 - \dots - a_q Z_q}{1 - R_{x_1 \dots x_p}^2} \frac{\delta SZ = Sx_1}{1 - R_{x_1 \dots x_p}^2} \delta \theta_{B_1}$$

In (Eq. 6), x_1 is the independent variable for which we want to estimate the effects of selection bias, and x_2 through x_p are other independent variables in the regression equation. Z represents the covariates in the selection equation and \hat{Z} the predicted values from the selection equation. The r represents the standard deviation of the regression equation error, and q_{ed} remains the unknown correlation between the selection and substantive equation error terms. $k' T - a_1 Z_1 - \dots - a_q Z_q$ is the first derivative of the function k , evaluated at the mean of the predicted values from the probit equation Z , multiplied by -1 . $\delta SZ = Sx_1$ is calculated by dividing the standard deviation of Z by the standard deviation of the covariate of primary interest, x_1 .

Finally, $\delta R_{\hat{Z}_{x_1 \dots x_p}}^2$ is the squared multiple correlation between Z and $x_2 \dots x_p$ (i.e. the R^2 for the regression of \hat{Z} on $x_2 \dots x_p$) and $\delta R_{x_1 \dots x_p}^2$ is the squared multiple correlation between x_1 and $x_2 \dots x_p$ (i.e. the R^2 from the regression of x_1 on $x_2 \dots x_p$). Stolzenberg and Relles (1997) recommend evaluating this bias under a variety of assumptions about the size of q_{ed} , the unknown correlation between the selection and substantive equation error terms. All other terms are known.²⁵

To determine if selection bias is relatively large or small, Stolzenberg and Relles advocate calculating the ratio of the estimated selection bias for the variable of interest to this estimate's sampling error. If evidence is found that selection bias may be substantial, the first choice is to implement a selection model (either the Heckman two-step or FIML) with exclusion restrictions. However, in the absence of exclusion restrictions, Leung and Yu (1996, 2000) suggest that it still may be possible to obtain reasonable answers using the Heckman two-step or FIML, under certain conditions. Their approach hinges on the condition number, which is a measure of how sensitive the solution is to data errors/changes. A regression equation with a matrix of independent variables with a high degree of multicollinearity will be very sensitive to small changes in the data, and is therefore unstable, or 'ill-conditioned.'²⁶ The general issue is whether the X vector has a broad enough range to make the inverse Mills ratio effectively non-linear, since the standard normal model is only truly non-linear in the tails of the distribution (when Z is less than -3 or greater than 3). Leung and Yu suggest that the Heckman/FIML can be productively

²⁵ The STATA do file used to estimate these coefficients is available from the authors by request.

²⁶ Statistically, the condition number is the ratio of the maximum Eigenvalue to the minimum Eigenvalue for the matrix of independent variables in an analysis (see Belsley et al. 1980, pp. 100 - 104, for a technical treatment of the condition number). It can be calculated in STATA 8.2 with the *collin()* function and is also provided in standard output for collinearity diagnostics in SPSS and other statistical software packages.

Table 2 Estimates of selection bias in the sentence length model for race, ethnicity and gender coefficients

q_{ed}	Black			Hispanic			Male		
	@ _b	$j@_{b_1}=S_{b_1}j$	Months	@ _b	$j@_{b_1}=S_{b_1}j$	Months	@ _b	$j@_{b_1}=S_{b_1}j$	Months
0.25	-0.24	4.16	0.36	-0.42	3.96	0.58	-0.46	6.73	0.75
0.50	-0.47	8.32	0.73	-0.84	7.91	1.16	-0.93	13.47	1.49
0.75	-0.71	12.47	1.09	-1.25	11.87	1.73	-1.39	20.20	2.24
1.00	-0.94	16.63	1.45	-1.67	15.83	2.31	-1.85	26.93	2.98

Note: @_b = estimate of selection bias

$j@_{b_1}=S_{b_1}j$ = ratio of selection bias to the sampling error of the independent variable of interest

estimated without exclusion restrictions if the condition number is less than 20, though this represents a rough guideline rather than an absolute threshold (Belsley et al. 1980). It is up to the researcher to balance the relative merits of Heckman’s correction against the added collinearity it introduces, but very large condition numbers clearly suggest instability in regression coefficients.

Although the condition number can provide useful intuition into the potential dangers of collinearity for Heckman’s estimator, it is important to note that econometricians have questioned its use in the absence of exclusion restrictions (see Vella 1999 for a cogent statement of these concerns). The primary concern is that the distribution in the tails, which identifies the Heckman estimator, is heavily dependent on the normality assumption of the probit.²⁷ Because the Leung and Yu papers are based on simulation data where the normality assumption is met by construction, they do not assess the sensitivity of their results to violations of normality. Given this controversy, we believe that any reliance on the condition number in criminology research should be exercised with caution, and only in conjunction with other approaches that provide a sense of the magnitude of the problem. Ideally, application of these diagnostic techniques should only ensue after a comprehensive search for valid exclusion restrictions. We demonstrate our recommended approach in the next section.

Empirical Example Using Pennsylvania Data

To demonstrate the utility of the above procedures, we examine two recent years (1999 – 2000) of sentencing data from the Pennsylvania Commission on Sentencing (PCS). These data have a high degree of censoring, combined with relatively good model fit in a large dataset. Just over half of the 148,890 criminal cases sentenced during these years received incarceration while the rest were sentenced to some form of alternative sanction, such as probation or intermediate punishment. The amount of explained variation in our sentence length regression is almost .7, which is large by standard assessments. From these indicators alone, it is difficult to know whether the amount of selection bias in our estimates warrants the use of Heckman’s

²⁷ This is commonly understood by criminologists in another context. The logit and probit models will provide very similar effect size estimates when the probability of the dependent variable being a success is not close to 1 or 0. However, as the proportion of successes approaches 1 or 0, which requires extreme values of Z, the two models can provide very different results because of the different parametric assumptions.

Table 3 Comparison of regression estimates for the TPM and Heckman two-step estimator (with and without standard error correction)

	Two-part model (TPM) <i>B</i> (S.E.)	Heckman (no error correction) <i>B</i> (S.E.)	Heckman (error correction) <i>B</i> (S.E.)
Black	0.63* (.087)	1.47* (.088)	1.47* (.096)
Hispanic	2.56* (.146)	3.78* (.147)	3.78* (.162)
Male	0.37* (.111)	2.33* (.117)	2.33* (.126)

Note: Coefficients reported from the full multivariate model controlling for presumptive sentence, mandatory, age, offense type, mode of conviction, and sentencing year ($N = 148,890$ for incarceration; $N = 79,444$ for sentence length)

* Coefficients significant at $p \leq 0.001$

correction. While poor model fit in a well-specified selection equation suggests small selectivity bias (or a poor understanding of the selection process), good model fit in the substantive equation of interest is desirable for less biased coefficients (Stolzenberg and Relles 1997, p. 499).

Table 2 provides the results from our application of (Eq. 6) to the PCS data for race, ethnicity, and gender, net of the presumptive sentence, mandatory sentences, sentencing year, offender age, crime type (violent, property, drug) and mode of conviction (see Johnson 2006 for a detailed description of these data). We focus on these three effects for demonstrative purposes because they remain at the heart of an extensive literature on discrimination in the criminal justice system (Zatz 2000). Because it is impossible to know the correlation between the error terms in the selection and regression equations (i.e. $\rho_{\epsilon d}$), we repeat the analysis allowing this correlation to vary between .25 and 1. As the correlation between the error terms increases, so does the amount of selection bias. In order to gauge the relative size of the selection bias, we follow Stolzenberg and Relles (1997) and divide the estimates of bias by the sampling error in their estimates. This provides a more intuitive measure of the magnitude of bias in the coefficients.

Several conclusions can be drawn from these results. First, the substantive effect of selection bias for being black is relatively modest, even in the worst case scenario. The estimated bias for black ranges from $-.24$ when the correlation between the regression and the selection errors is .25 to a maximum of $-.94$ when the error terms are perfectly correlated. When these estimates are divided by their standard errors they range from 4.16 to 16.63. This suggests that the predicted bias at its greatest is about 16 times the standard error of the black coefficient in the uncorrected TPM. We estimated the uncorrected TPM in the first column of Table 3 and found a standard error on the black coefficient of .087. In substantive terms, then, the black coefficient would be biased at most by 1.45 months ($16.63 \times .087 = 1.45$) given perfect correlation between the error terms in the two equations.

Second, the amount of bias is somewhat larger for Hispanic ethnicity and gender. For ethnicity, the upper bound of the selection bias effect is 2.31 months whereas for gender it is almost three full months. Substantively speaking, then, selection bias is about twice as strong for gender as for race. These results suggest that the necessity of correcting for selection bias, at least in these data, is greater for research questions focusing on gender than race. The extent to which these findings are specific to the PCS data examined here is an empirical question that needs to be addressed in future applications.

For the sake of this exercise, suppose our primary interest is in the effect of gender on sentence length. We could simply use the Stolzenberg and Relles bias estimates under the most conservative assumption ($q_{ed} = 1$) to bound our estimates from the TPM. In this case, we would conclude that males are sentenced to somewhere between .37 and 3.35 months of additional incarceration. While a useful exercise, this still provides a fairly broad range of possible effect sizes. Ideally, therefore, we would prefer to estimate a Heckman correction with exclusion restrictions. As one example, if our population of interest was all arrested offenders and we had data on conviction decisions, we might use strength of evidence to model selection into conviction but not to predict sentence length. Evidentiary factors are strong predictors of the probability of conviction, but once convicted should be theoretically unrelated to sentence severity (see Albonetti 1991 for an application of this example). When exclusion restrictions are not available, though, the researcher must choose between not correcting for bias (i.e. using the TPM), or correcting for it with the Heckman or FIML estimator. While the FIML will always be more efficient, it is more sensitive to the normality assumption than the Heckman two-step. A reasonable strategy therefore is to compare the estimates from these two estimators, and if the estimates vary dramatically it suggests there may be a problem with the model assumptions.

As discussed, a first step in assessing the potential validity of either the Heckman two-step or FIML when exclusion restrictions are not available is to examine the condition number as suggested by Leung and Yu (1996). These scholars suggest that a condition number below 20 indicates multicollinearity is not a concern. Under these conditions, Leung and Yu suggest that the corrected estimates will be superior to those from an uncorrected TPM.

Our estimated condition number of 19.66 is very close to the critical threshold of 20, which suggests that the collinearity introduced by the inverse Mills ratio might cause problems for the Heckman estimator without exclusion restrictions. Next, we attempted to estimate the FIML model, which failed to converge to a solution for our chosen specification. Although the FIML is a more efficient estimator than the Heckman two-step, it is also more sensitive to violations of underlying model assumptions, particularly the bivariate normality assumption. To investigate this, we re-estimated the FIML using a logged measure of sentence length which should better satisfy the normality assumption. This alternative specification did converge to a solution, offering some support for this explanation.

In this case, we have two facts which allow us to make an a priori decision about the use of Heckman. We found a marginally high condition number, suggesting the X variables may not create enough variation in the underlying Z parameter to provide stable estimates, and the FIML estimator failed to converge, suggesting that the bivariate normal assumptions may not be valid. Together, these pieces of information suggest the Heckman estimator is unlikely to provide more reliable estimates compared to the simple Two Part Model.

For illustration purposes, Table 3 reports the model coefficients estimated with the uncorrected TPM and Heckman's two-step estimator (with and without standard error corrections). Several important observations can be gleaned from these findings. First, non-trivial differences emerge in the effects of model parameters across estimation strategies. Comparing the TPM to the Heckman model, the effect of black in the latter is more than twice as large, while the effect of gender is over six times as large. Second, the Heckman two-step correction in STATA produces

standard errors that are about 10% larger than the improperly calculated estimates for the Heckman. Examination of additional model specifications, including criminal history and crime severity along with the presumptive sentence (Ulmer 2000), revealed differences as great as 30% (results not shown). Failure to correct standard errors, then, may lead to overstated statistical significance in prior work that employs Heckman. Although this may not be a problem in sentencing research using very large datasets, there are studies in our sample for which such an increase would lead to different conclusions regarding the significance of key coefficients.

Finally, although our a priori tests have led us to conclude that the Heckman without exclusion restrictions is not a valid estimator in this situation, it is interesting to note that the results in Table 3 are in line with what we might expect given the amounts of bias reported in Table 2; the coefficients with the most potential bias are affected the most by selection bias corrections. One reasonable concern about the Stolzenberg and Relles technique is that it relies on the bivariate normality assumption, but rather than providing a single point estimate of the bias it offers upper and lower bounds. It is therefore a useful diagnostic technique that can be used in conjunction with the simple TPM, either when the Heckman cannot be estimated with exclusion restrictions or in the case of high condition numbers. Additional non-parametric or semi-parametric bounding techniques may also prove useful in this regard.

Conclusion

I don't pretend we have all the answers. But the questions are certainly worth thinking about.

-Arthur C. Clarke

We began this paper with a quotation from the noted science fiction author Arthur C. Clarke that suggests advanced technologies can be mistaken for *magical* solutions. Our investigation, however, reveals that the Heckman two-step approach does not offer a magic solution to selection bias—it is not a statistical panacea but rather a particular estimator that requires careful thought and supplemental investigation to adequately assess its appropriateness in any given application. We close the paper with a second quotation from this author, which serves to highlight the fact that we do not have all the answers to this pervasive problem. The only real solution to the problem of selection bias involves striving for a better understanding of its causes and consequences, which can then be utilized to make more informed decisions about the most appropriate way to proceed.

The present work offers some useful guidance in this direction. It set out to investigate the use (and misuse) of Heckman's two-step estimator in criminological research. We began by reviewing recent studies in *Criminology* that use Heckman's approach and found convincing evidence that it is often misapplied. Arguably the most important concern is its routine application to analytical problems for which it simply does not apply, such as discrete dependent variables.

Even when properly applied, though, our investigation suggests that the routine, mechanistic application of Heckman's estimator remains problematic. This is particularly important in the typical case where no exclusion restrictions are employed. When the same predictors are used to model the selection process and substantive

outcome, there will often be substantial correlation between the correction term and the included variables. The presence of serious multicollinearity is a common theme in papers that use the Heckman, but one that is seldom addressed effectively. In some cases, researchers have gone so far as to omit correlated variables, like offense seriousness, from the sentence length equation. This approach gives the false appearance of correcting for selection bias, when in fact the substantive equation has simply been misspecified. It is also common for researchers to report a large but statistically insignificant correction term and to interpret this finding as evidence against substantial bias, even when collinearity substantially inflates the standard error of the correction term, reducing the power of the test for selection bias.²⁸ In the face of severe multicollinearity, simulation research suggests it is better to simply acknowledge the threat of selection and estimate the simple Two-Part Model.

The problem with this approach is that it essentially ignores the selection problem. Our recommendation is that criminologists be more explicit in their attempts to model substantive selection processes. Attempts to collect better data across selection stages are needed. Data on unconditional populations would be useful, for example using indicted samples rather than convicted samples. Better efforts are also needed in identifying valid exclusion restrictions, or theoretically based arguments for variables that affect one part of the selection process and not the other. Natural experiments can provide useful approaches in this respect (see Loeffler (2006) and Owens (2006) for recent examples in the context of sentencing). We are also optimistic about the possibility of experiments or quasi-experiments that can be implemented as part of sentencing reforms. In addition, the Stolzenberg and Relles (1997) technique allows for the investigation of selection bias in particular coefficients of interest, enabling the researcher first to determine if selection bias is likely to affect the variables of interest and second, to bound coefficient estimates. And examination of the condition number provides some general insight into when the Heckman might be used even without exclusion restrictions.

We believe the incorrect application of the Heckman documented in this paper provides a useful warning about the importation of new techniques into criminology. Since the Heckman's introduction into the field, researchers have commonly implemented the technique incorrectly, without referees and editors catching the mistakes. Over time, the errors became embedded in the discipline, and simply became part of how one conducted research. This is a sobering realization, and one that should cause all authors/referees to be especially careful when dealing with new and complicated methods. It should also raise the bar for those who wish to import new methods into the field. It is imperative that such importation be done clearly and carefully so that users are aware of the key issues involved in the application of the method.

²⁸ A similar problem is endemic in comparisons of coefficients across different models affected by selection. In sentencing, key modeling decisions, such as the decision to examine incarceration and sentence length as two distinct outcomes, rest on the finding that coefficients exert different effects on the two outcomes. However, if the error terms in the two models are correlated, the coefficients for sentence length will be biased. This will lead to flawed comparisons if a simple TPM is utilized or if the Heckman is implemented incorrectly. This observation may also apply to the current debate about whether jail and prison can be treated as the same outcome (Holleran and Spohn 2004). To the extent that selection bias exists in the coefficients for jail and prison sentences (i.e. non-random selection processes determine conviction and therefore offender placement), any comparison of these effects across models without good controls for selection must be conducted with caution.

We realize these observations may seem pessimistic, so we conclude the paper with a focus on the steps that can be taken to better address selection issues in criminological work. First, the researcher needs to decide, on theoretical grounds, what kind of selection is present. The Heckman is a potential estimator only if incidental selection is present. The Heckman is not appropriate for other types of selection. Second, if a researcher believes she is dealing with incidental selection, she should think about the amount of potential bias in the coefficients of interest. The Stolzenberg and Relles method is one way this can be accomplished. In the Pennsylvania conviction data, we found little evidence of substantial race bias, whereas the results for gender suggested higher levels of bias. It does not make sense to correct for selection bias in situations where it is relatively small, even in the worst case scenario. Third, researchers should think both substantively and theoretically about the selection process under study in order to find exclusion restrictions to use with the Heckman estimator. The goal is to identify variables that influence the first stage but not the second. In general, criminologists have not devoted enough energy to identifying valid exclusion restrictions when attempting to address issues of selection bias. Research in other disciplines, such as economics, demonstrates that creative, relevant exclusion restrictions can be found for many important research questions. Fourth, when good exclusion restrictions can be identified, implement Heckman's estimator using statistical software (e.g. STATA or LIMDEP) to eliminate user errors. Fifth, if exclusion restrictions cannot be identified, calculate the condition number. If the condition number is less than 20, consider using the Heckman and the FIML, but also estimate the TPM. The failure of the FIML to converge is a useful diagnostic signal that the model may not be appropriate. Take extra care to compare the difference between the Heckman estimators and the TPM with the amount of bias predicted by the Stolzenberg and Relles method. By triangulating these sensitivity analyses, a deeper understanding of the importance and consequence of selection bias in one's analysis is achievable.

Ultimately, continued development of our theoretical understanding of the selection processes underlying criminological phenomenon is needed. In our example of criminal sentencing, this means researchers need to continue to strive to better understand and model the criminal justice process itself, rather than implement an often misunderstood statistical correction. Thinking about the problem in this way makes it clear that the issue of selection is ultimately not an econometric or statistical problem, but rather a substantive and theoretical one. If the ultimate goal of social science research is to understand the processes that lead to the outcomes we observe, then the challenge of selection bias is not a statistical inconvenience, but rather an essential part of our understanding of the basic underlying phenomenon of interest.

Acknowledgements The authors would like to thank Justin McCrary, Wayne Osgood, Anne Piehl, Jeff Smith and seminar participants at the Economics and Crime summer workshop for helpful comments. We would also like to thank Gary Sweeten for research assistance and useful discussions. The paper is dedicated to the memory of our friend, mentor, and former colleague Douglas A. Smith. Doug helped us to understand the relevant selection issues, and encouraged us to pursue the paper in its formative stages. He is sorely missed.

Appendix

```
Syntax for Regression Estimates with Heckman Two-Step and FIML in STATA 8.2
FIML: HECKMAN depvars indepvars, select (selection indepvars)
TWO-STEP: HECKMAN depvars indepvars, twostep select (selection indepvars)
**depvars is your dependent variable of interest
**indepvars are your independent variables in the regression equation
**selection indepvars are the independent variables in the selection equation
Sample Syntax for Calculating Heckman Two-Step and Condition Number in STATA 8.2
*****HECKMAN TWO-STEP*****
#delimit ;
heckman length presumptive mandatory age agemiss female black hispanic
otherace violent property drugs negplea bench jury dispmiss yr1999,
twostep select (presumptive mandatory age agemiss female black hispanic
otherace violent property drugs negplea bench jury dispmiss yr1999);
predict lambda, mills;
*****CONDITION NUMBER*****
#delimit ;
collin guidemin manapply newage agemiss female black
hispanic otherace violent property drugs negplea bench jury
dispmiss yr1999 lambda if incar = = 1;
```

References

- Ahn H, Powell J (1993) Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *J Econom* 58:3 – 29
- Albonetti C (1986) Criminality, prosecutorial screening, and uncertainty: toward a theory of discretionary decision making in felony case processing. *Criminology* 24:623 – 644
- Albonetti C (1991) An integration of theories to explain judicial discretion. *Soc Probl* 38:247 – 266
- Albonetti C (1997) Sentencing under the federal sentencing guidelines: an analysis of the effects of defendant characteristics, guilty pleas, and departures on sentencing outcomes for drug offenses. *Law Soc Rev* 31:601 – 634
- Amemiya T (1985) *Advanced econometrics*. Harvard University Press, Cambridge, MA
- Angrist J (2001) Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice. *J Bus Econ Stat* 19:2 – 16
- Belsley David, Kuh E, Welsch R (1980) *Regression diagnostics: identifying influential data and sources of collinearity*. John Wiley & Sons, New York
- Berk R (1983) An introduction to sample selection bias in sociological data. *Am Soc Rev* 48:386 – 398
- Berk R, Ray S (1982) Selection bias in sociological data. *Soc Sci Res* 11:352 – 398
- Blundell R, Powell J (2001) Endogeneity in nonparametric and semiparametric regression models. CEMMAP Working Paper No CWP09/01
- Bontrager S, Bales W, Chiricos T (2005) Race, ethnicity, threat and the labeling of convicted felons. *Criminology* 43:589 – 622
- Bushway S, Piehl A (2001) Judging judicial discretion: legal factors and racial discrimination in sentencing. *Law Soc Rev* 35:733 – 767
- Chay KY, Powell JL (2001) Semiparametric censored regression models. *J Econ Persp* 15:29 – 42
- Chesher A, Irish M (1987) Residual analysis in the grouped and censored normal linear model. *J Econom* 34:33 – 61
- Chiricos T, Bales W (1991) Unemployment and punishment: an empirical assessment. *Criminology* 29:701 – 724
- D' Alessio S, Stolzenberg L (2003) Race and the probability of arrest. *Soc Forces* 84:1381 – 1398
- Demuth S (2003) Racial and ethnic differences in pretrial release decisions and outcomes: a comparison of Hispanic, black, and white felony arrestees. *Criminology* 41:873 – 908
- Duan N, Manning WG, Morris CN, Newhouse JP (1983) A comparison of alternative models for the demand for medical care. *J Econ Bus Stat* 1:115 – 126
- Duan N, Manning WG, Morris CN, Newhouse JP (1984) Choosing between the sample-selection model and the multi-part model. *J Bus Econ Stat* 2:283 – 289

- Dubin JA, Rivers D (1990) Selection bias in linear regression, logit and probit models. In: Fox J, Scott Long J (eds) *Modern methods of data analysis*. Sage Publications, Thousand Oaks, CA, pp 411 – 443
- Engen RL, Gaaney RR (2000) Modeling the effects of legally relevant and extralegal factors under sentencing guidelines: the rules have changed. *Criminology* 38:1207 – 1228
- Erez E, Tontodonato P (1990) The effect of victim participation in sentencing on sentencing outcome. *Criminology* 28:451 – 474
- Felson RB (1996) Big people hit little people: sex differences in physical power and interpersonal violence. *Criminology* 34:433 – 453
- Felson RB, Messner S (1996) To kill or not to kill: lethal outcomes in injurious attacks. *Criminology* 36:519 – 545
- Gondolf EW (2000) A 30-month follow-up of court-referred batterers in four cities. *Int J Offender Ther Comp Criminol* 44:111 – 128
- Greene WH (1993) *Econometric analysis*, 2nd edn. Macmillan, New York
- Hagan J, Palloni A (1986) 'Club Fed' and the sentencing of white-collar offenders before and after Watergate. *Criminology* 24:603 – 622
- Hagan J, Parker P (1985) White-collar crime and punishment: the class structure and legal sanctioning of security violations. *Am Sociol Rev* 50:302 – 316
- Halaby C (2004) Panel models in sociological research: theory into practice. *Ann Rev Sociol* 30
- Heckman JJ (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables. *Ann Econ Soc Measure* 5:475 – 492
- Heckman JJ (1979) Sample selection bias as a specification error. *Econometrica* 47:153 – 161
- Heckman JJ, LaLonde R, Smith J. (1999) The economics and econometrics of active labor market programs. In: Ashenfelter O, Card D (eds) *Handbook of Labor Economics*, vol 3A. North-Holland, Amsterdam, pp 1865 – 2097
- Holleran D, Spohn C (2004) On the use of the total incarceration variable in sentencing research. *Criminology* 42(1):211 – 240
- Johnson BD (2006) The Multilevel context of criminal sentencing: integrating judge and county level influences in the study of courtroom decision making. *Criminology* 44(2):259 – 298
- Kaatt PM (2002) Location, location, location: interdistrict and intercircuit variation in sentencing outcomes for federal drug-trafficking offenses. *Justice Q* 19:633 – 671
- Keil TJ, Vito GF (1989) Race, homicide severity, and application of the death penalty: a consideration of the Barnett scale. *Criminology* 27:511 – 536
- Kingsnorth R, Lopez J, Wentworth J (1998) Adult sexual assault: the role of racial/ethnic composition in prosecution and sentencing. *J Crim Justice* 26:359 – 371
- Kingsnorth R, MacIntosh R, Sutherland S (2002) Criminal charge or probation violation: Prosecutorial discretion and implications for research. *Criminology* 40:553 – 578
- Kingsnorth R, MacIntosh R, Wentworth J (1999) Sexual assault: the role of prior relationship and victim characteristics in case processing. *Justice Q* 16:275 – 302
- Klepper S, Nagin D, Tierney LJ (1983) Discrimination in the criminal justice system: a critical appraisal of the literature. In: Blumstein A, Cohen J, Martin S, Tonry M (eds) *Research on sentencing: the search for reform*, vol II. National Academy Press, Washington, DC
- Kurlychek M, Johnson BD (2004) The juvenile penalty: a comparison of juvenile and young adult sentencing outcomes. *Criminology* 42:485 – 517
- Leiber M, Mack K (2003) The individual and joint effects of race, gender, and family status on juvenile justice decision making. *J Res Crime Delinq* 40:34 – 71
- Leiber M, Jamieson K (1995) Race and decision making within juvenile justice: the importance of context. *J Quantitative Criminol* 11:363
- Leung SF, Yu S (1996) On the choice between sample selection and two-part models. *J Econometrics* 72:197 – 229
- Leung SF, Yu S (2000) Collinearity and two-step estimation of sample selection models: problems, origins, and remedies. *Comput Econ* 15:173 – 199
- Loeffer, Charles (2006) Using inter-judge sentencing disparity to estimate the effect of imprisonment on criminal recidivism. Harvard University Working Paper
- Lundman R, Kaufman R (2003) Driving while black: effects of race, ethnicity, and gender on citizen self-reports of traffic stops and police actions. *Criminology* 41:195 – 220
- MacMillan R (2000) Adolescent victimization and income deficits in adulthood: rethinking the costs of criminal violence from a life-course perspective. *Criminology* 38:553 – 587
- Maddala GS (1985) A survey of the literature on selectivity bias as it pertains to health care markets. *Adv Health Econ Health Serv Res* 6:3 – 18

- Maxwell C, Garner J, Fagan J (2002) The preventative effects of arrest on intimate partner violence. *Criminol Publ Policy* 2:51 – 80
- McCarthy B, Smith BL (1986) The conceptualization of discrimination in the juvenile justice process: the impact of administrative factors and screening decisions on juvenile court dispositions. *Criminology* 24:41 – 64
- Messner SF, Deane G, Beaulieu M (2002) A log-multiplicative association model for allocating homicides with unknown victim – offender relationships. *Criminology* 40:457 – 479
- Messner SF, Deane GD, Anselin L, Pearson-Nelson B (2005) Locating the vanguard in rising and falling homicide rates across US cities. *Criminology* 43(3):661 – 696
- Miethe TD, Moore CA (1985) Socioeconomic disparities under determinant sentencing systems: a comparison of preguideline and postguideline practices in Minnesota. *Criminology* 23(2):337 – 363
- Moffitt R (1999) Econometric methods for labor market analysis. In: Ashenfelter O, Card D (eds) *Handbook of labor economics*, vol 3A. North-Holland, Amsterdam
- Myers MA (1988) Social background and the sentencing behavior of judges. *Criminology* 26:649 – 675
- Myers MA, Talarico SM (1986) Urban justice, rural injustice? Urbanization and its effects on sentencing. *Criminology* 24:367 – 391
- Newey WK (1999) Consistency of two step sample selection estimators despite misspecifications of distribution. *Econ Lett* 63:129 – 132
- Nobiling T, Spohn C, Delone M (1998) A tale of two counties: unemployment and sentence severity. *Justice Q* 15:459 – 485
- Osgood DW, Finken L, McMorris B (2002) Analyzing multiple item measures of crime and deviance II: Tobit regression analysis of transformed scores. *J Quantitative Criminol* 18:859 – 866
- Owens, EG (2006) More time, less crime? The incapacitative effect of sentence enhancements. University of Maryland working paper
- Peterson RD, Hagan J (1984) Changing conceptions of race: toward an account of anomalous findings of sentencing research. *Am Sociol Rev* 49:56 – 70
- Puhani PA (2000) The Heckman correction for sample selection and its critique. *J Econ Surveys* 14:53 – 68
- Reitz K (1998) Modeling discretion in American sentencing systems. *Law & Policy* 20:389 – 428
- Rhodes W (1991) Federal criminal sentencing: some measurement issues with application to pre-guideline sentencing disparity. *J Criminal Law Criminol* 81:1002
- Robertson AA, Grimes PW, Rogers KE (2002) A short-run cost-benefit analysis of community based interventions for juvenile offenders. *Crime Delinq* 47:265 – 284
- Smith DA (1986) The plea bargaining controversy. *J Criminal Law Criminol* 77:949 – 967
- Smith DA, Brame R (2003) Tobit models in social science research: some limitations and a more general alternative. *Sociol Methods Res* 31:364 – 388
- Sorenson J, Wallace DH (1999) Prosecutorial discretion in seeking death: an analysis of racial disparity in the pretrial stages of case processing in a Midwestern county. *Justice Q* 16:559 – 578
- Spohn C, Holleran D (2002) The effect of imprisonment on recidivism rates of felony offenders: a focus on drug offenders. *Criminology* 40:329 – 358
- Spohn C, Horney J (1996) The impact of rape law reform on the processing of simple and aggravated rape cases. *J Criminal Law Criminol* 86:861 – 886
- Stata Corp. (2003) *Stata Statistical Software Release 8.0. Reference Manual G-M*. Stata Corporation, College Station, TX, p 71
- Steen S, Engen RL, Gainey RR (2005) Images of danger and culpability: racial stereotyping, case processing, and criminal sentences. *Criminology* 43:435 – 467
- Steffensmeier D, Kramer JH, Streifel C (1993) Gender and imprisonment decisions. *Criminology* 31:411 – 446
- Steffensmeier D, Ulmer JT, Kramer JH (1998) The interaction of race, gender, and age in criminal sentencing: the punishment cost of being young, black, and male. *Criminology* 36:763 – 798
- Steffensmeier D, DeMuth S (2001) Ethnicity and judges' sentencing decisions: Hispanic-Black-White comparisons. *Criminology* 39:145 – 176
- Stolzenberg RM, Relles DA (1990) Theory testing in a world of constrained research design: the significance of Heckman's censored sampling bias correction for nonexperimental research. *Sociol Methods Res* 18:395 – 415
- Stolzenberg RM, Relles DA (1997) Tools for intuition about sample selection bias and its correction. *Am Sociol Rev* 62:494 – 507
- Stolzenberg L, D' Alessio SJ, Eitle D (2004) A multilevel test of racial threat theory. *Criminology* 42:673 – 698

- Tibbets SG, Piquero AR (1999) The influence of gender, low birth weight, and disadvantaged environment in predicting early onset of offending: a test of Moffitt's interactional hypothesis. *Criminology* 37:843 – 878
- Tobin J (1958) Estimation of relationships for limited dependent-variables. *Econometrica* 26:24 – 36
- Ulmer JT (2000) The rules have changed—so proceed with caution: a comment on Engen and Gaaney's methods for modeling sentencing outcomes under guidelines. *Criminology* 38:1231 – 1243
- Ulmer JT, Johnson BD (2004) Sentencing in context: a multilevel analysis. *Criminology* 42:137 – 178
- Ulmer JT, Kramer JH (1996) Court communities under sentencing guidelines: dilemmas of formal rationality and sentencing disparity. *Criminology* 34:383 – 408
- Vella F (1999) Estimating models with sample selection bias: a survey. *J Hum Resour* 33:127 – 169
- Wheeler S, Weisburd D, Bode N (1982) Sentencing the white collar offender: rhetoric and reality. *Am Sociol Rev* 47:641 – 659
- Winship C, Mare RD (1992) Models for selection bias. *Ann Rev Sociol* 18:327 – 350
- Wooldredge JD (1998) Analytic rigor in studies of disparities in criminal case processing. *J Quantitative Criminol* 14:155 – 179
- Wooldredge J (2005) *Introductory economics: a modern approach*. South Western College Publishing
- Wooldredge JD, Thistlewaite A (2004) Bilevel disparities in court dispositions for intimate assault. *Criminology* 42:417 – 456
- Worrall JL (2002) If you build it, they will come: consequences of improved citizen complaint review procedures. *Crime Delinq* 48:355 – 379
- Wright JP, Cullen FT (2000) Juvenile involvement in occupational delinquency. *Criminology* 38:863 – 892
- Zatz MS (2000) The convergence of race, ethnicity, gender, and class on court decision making: looking toward the 21st century. *Policies, processes, and decisions of the criminal justice system*, vol. 3. US Department of Justice, Washington, DC, pp 503 – 552
- Zatz MS, Hagan J (1985) Crime, time, and punishment: an exploration of selection bias in sentencing research. *J Quantitative Criminol* 1:103 – 126