Quantity and Quality manuscript No.
(will be inserted by the editor)

Tackling selection bias in sentencing data analysis: A new approach based on a scale of severity

the date of receipt and acceptance should be inserted later

Abstract For reasons of methodological convenience statistical models analysing judicial decisions tend to focus on the duration of custodial sentences. These types of sentences are however quite rare (7% of the total in England and Wales), which generates a serious problem of selection bias. Typical adjustments employed in the literature, such as Tobit models, are based on questionable assumptions and are incapable to discriminate between different types of non-custodial sentences (such as discharges, fines, community orders, or suspended sentences). Here we implement an original approach to model custodial and non-custodial sentence outcomes simultaneously avoiding problems of selection bias while making the most of the information recorded for each of them. This is achieved by employing Pina-Sánchez et al. (2019) scale of sentence severity as the outcome variable of a Bayesian regression model. A sample of 7,242 theft offences sentenced in the Crown Court is used to further illustrate: (a) the pervasiveness of selection bias in studies restricted to custodial sentences, which leads us to question the external validity of previous studies in the literature limited to custodial sentence length; and (b) the questionable adequacy of Tobit models and similar methods used in the literature to adjust for such bias.

Keywords sentencing · selection bias · severity · paired comparison · Bayesian statistics · Tobit models

This work was supported by the National Centre for Research Methods [RIS14241/06]

Address(es) of author(s) should be given

1 Introduction

Core questions to the discipline of criminology, such as the effectiveness of prison as a deterrent or the presence of racial discrimination in the criminal justice system, rely heavily on statistical models of sentence data. Such models are commonly based on the specification of one of two variables: the duration of custodial sentences or whether the offender was sentenced to prison (Bushway and Piehl, 2001; Ostrom et al., 2008). These variables can be taken as proxies for sentence severity, yet they present important limitations. Reducing the complexity of sentencing to a binary choice (custody or other) represents a vast loss of information, whereas focusing on the sentence length involves dropping non-custodial outcomes, which in turn creates a problem of selection bias; for example, only 7.2% of cases in England and Wales are sentenced to custody (Ministry of Justice, 2018).

This is not a new problem; the presence of selection bias in sentence data analysis was widely acknowledged (Berk and Ray, 1982; Greenberg, 1977; Hopkins, 1977; Klepper et al., 1983) following Heckman's seminal contributions (Heckman, 1976, 1979). Since then, many of the leading criminal justice scholars (Feldmeyer and Ulmer, 2011; Nobiling et al., 1998; Steffensmeier and DeMuth, 2001; Ulmer et al., 2010; Ulmer and Johnson, 2004; Zatz and Hagan, 1985) have implemented Heckman's two-stage estimation to adjust for selection bias in models specifying durations of custodial sentences. Such adjustments are based on a number of assumptions, one of them is that the sentencing process is divided into two steps: first, a decision is made as to whether the offender is sentenced to custody, and then the duration of the sentence is considered. However, this conceptualisation of the sentencing process is disputable, at least for the case of England and Wales, where the disposal type (fine, community order, suspended sentence order, etc.) and its magnitude can be modulated through a series of preliminary steps (see for example Sentencing Council, 2018). Perhaps more importantly, the Heckman process relies on exclusionary data, that is, variables affecting the selection process but not the outcome of interest. Variables meeting these criteria are, however, practically impossible to find (Bushway et al., 2007; Koons-Witt et al., 2014). Aggravating factors that will make the offence worth of custody will be commonly taken into account to define the length of the sentence. As a result, a number of sentencing scholars have explicitly discarded the Heckman model (Hester and Sevigny, 2014; Koons-Witt et al., 2014).

Another group of authors (Albonetti, 1997, 1998; Bushway and Piehl, 2001; Helms and D., 2002; King et al., 2010; Kurlychek and Johnson, 2004, 2010; Muhlhausen, 2004; Nienstedt et al., 1988; Rhodes, 1991) have sought to tackle the problem of selection bias using Tobit models for censored data (Tobin, 1958). Custodial and non-custodial sentences are modelled simultaneously under the assumption that both of them belong to a common underlying normal distribution from which the latter have been left-censored. Unlike the Heckman selection model, this method does not assume a two-stage process to sentencing, however it is affected by other important issues that need to be highlighted. Tobit models will provide biased estimates if the explanatory variables included affect the probability of receiving a custodial sentence and the duration of that sentence differently (Helms and D., 2002; Smith and Brame, 2003). Another problem is the common violation of the assumption of normality made by these types of models, since sentence length, as a form of duration data, tends to be right-skewed.[1]

Finally, both models for censored data (either parametric, or not) and Heckman's two-stage corrections imply a loss of information since they treat all non-custodial outcomes as a homogeneous group, disregarding the differences in severity amongst these types of sentences. This shortfall has become increasingly problematic as the range of non-custodial sentences has grown over time in many Western jurisdictions (Bottoms et al., 2004; Canton and Dominey, 2018).

In the present paper, we suggest taking an alternative route based on the estimation of a scale of sentence severity ranking the different types of sentences available to judges. The use of a scale of severity as a strategy to circumvent the problem of selection bias in sentence data analyses is not an entirely original idea. For a short spell of time, such strategy was repeatedly advocated (Buchner, 1979; Erickson and Gibbs, 1979; Sebba, 1980; Sebba and Nathan, 1984), giving rise to the 'penal metric theory', understood as a specific subdiscipline of 'penal theory' (Tremblay, 1988). However, over the last three decades, this area of research has been nearly abandoned. In this paper we build upon the new scale of severity presented by Pina-Sánchez et al. (2019) to illustrate empirically the magnitude of the problem of selection bias in sentence data analysis, and to provide an original modelling framework capable of eliminating such problem.

Pina-Sánchez et al. (2019) scale is based on Thurstone (1927) method, magistrates' views on the relative punitiveness of different sentences, and the sentencing ladder informally in operation in England and Wales (a working consensus on the ordinal ranking of the main disposal types available to sentencers). By inquiring on the potential exchangeability between certain sentences in terms of their relative severity Pina-Sánchez and colleagues were able to estimate 'severity scores' for the main disposal types used in England and Wales and their duration. In this paper, we improve the scale of severity of Pina-Sánchez et al. (2019) in two important ways. First, we review the questionable assumption of a perfectly proportional relationship between severity and sentence length beyond three months in custody. Secondly, to ensure that the uncertainty associated with the estimation of the scale of severity is adequately propagated to subsequent models where this scale is to be used, we undertake the estimation of the severity scale using Bayesian methods. Specifically, this approach seeks to account for the sampling error stemming from the magistrates' responses and the measurement error associated with the incomplete information available for certain sentence outcomes.

Our approach can then be used to model simultaneously custodial and non-custodial sentences to solve the problem of selection bias, making the most of the data available (that is, discriminating between non-custodial sentences based on their relative severity), while reflecting the uncertainty associated with the estimation of a scale of severity. As such, our approach offers an alternative framework to produce more robust and efficient analyses of sentence data. This is shown through the modelling of a sample of 7,242 theft offences sentenced in the England and Wales Crown Court in 2011. Sentence outcomes from this sample are transformed into severity scores and regressed using a linear model with case and offender char-

[1] This is acknowledged by (Sullivan et al., 2008; Wilson et al., 2018) who suggested semi- and non-parametric approximations of censored models to relax the unrealistic assumption that custodial and non-custodial sentences belong to a same normal distribution, at the expense of not being able to provide accurate measures of uncertainty.

acteristics as the explanatory variables. To illustrate the impact of selection bias, the model summarising the effect of offender and case characteristics on sentence severity is specified using only custodial sentences first; results from this model are then compared against the same model using all sentence outcomes. Finally, to illustrate the effectiveness of adjustments suggested in the literature, we replicate the model for the full sample using a Tobit specification, treating non-custodial sentence outcomes as left-censored. In the next section, we proceed to review the different methods that have been used in the literature to estimate a scale of sentence severity, as this step - estimating sentence severity - lies at the core of the approach to tackle the problem of selection bias suggested in this article.

2 Measuring sentence severity

We have identified four main approaches to scale estimation used in the literature: ordinal scales, magnitude estimation, data-driven methods, and paired comparisons; by reviewing them we aim to convey the complexity of the measurement of sentence severity, to point out the limitations of the most commonly used methods, and to justify the approach adopted here.

The most common scales of sentence severity are based on a ranking of different categories of sentence outcomes on an ordinal scale. For example, Pina-Sánchez et al. (2018) consider each of the five disposal types considered in the sentencing ladder in England and Wales (*discharge* < *fine* < *community order* < *suspended sentence order* < *custodial sentence*). Irwin-Rogers and Perry (2015) and Lightowlers and Pina-Sánchez (2017) use a similar five points scale, only they grouped all non-custodial outcomes in the same category and differentiated between duration of custodial sentences in the remaining four categories. Ordinal scales are, however, problematic for two main reasons, the arbitrariness associated with the choice of the different severity thresholds, and the assumption that jumps from one threshold to the next are equivalent.

Other authors have used more refined versions of these scales by either ranking different disposal types and durations within them in sufficient level of detail (Albonetti, 1991; Bernstein et al., 1977; Gibson, 1978, 1980; Gruhl et al., 1981; Holmes et al., 1996; King and Noble, 2005; Newheiser et al., 2012; Tiffany et al., 1975; Uhlman, 1978) or by taking the mean of separate ordinal rankings carried out by multiple experts (Curran, 1983; Kapardis and Farrington, 1981). These scales of severity might take the appearance of an interval level scale since sentence outcomes can now be grouped into more than just a few categories, however they are still originated from ordinal perceptions of severity, which is problematic. For example, experts do not express the relative difference in severity between sentence outcomes.

To obtain interval level scales of severity, some researchers have relied on the assignation of numerical values of severity for different sentence outcomes. Hindelang et al. (1975), Schiff (1997), Sorokin (1962) and Zamble and Kalm (1990) have suggested various meaningful ranges (for example, 0 for discharge, 100 for capital punishment) and proceeded to assign values of severity for different sentences within those ranges - or ask samples of experts to do so. To facilitate this task, another group of researchers have used ‘magnitude escalation’, a more systematic approach that involves using a standard stimulus known as a ‘modulus’.

The modulus (for example, a one-month custodial sentence) is assigned a specific value (say, 100), the interviewee is then given a new stimulus (for example, a six-month custodial sentence suspended for a year) and asked to estimate its intensity relative to the modulus. This has been the most common approach across studies set out to measure sentence severity (see for example, Apospori and Alpert, 1993; Erickson and Gibbs, 1979; Harlow et al., 1995; Leclerc and Tremblay, 2016; McClelland and Alpert, 1985; Spelman, 1995; Tremblay, 1988; Warr et al., 1983), but it is still affected by important limitations. Arguably, the biggest of them all is the not always tenable expectation of interviewees' numeracy (when piloting this method through focus groups we observed that participants did not always understand 'four times bigger than 100' equals 400, while 'four times smaller than 100' equals 25). The vast variances in individual assignations of severity scores reported in the literature are a reflection of this problem (see Tremblay, 1988).

A third group encompasses different approaches based on sentence data and sentencing guidelines, as opposed to subjective assessments of severity. For example, Croyle (1983) estimated the equivalence of probation and imprisonment sentences using the average prison time experienced by offenders sentenced to probation who failed to meet the conditions in their sentence and ended up spending time in jail. This is a straightforward approach, but its coverage is limited to probation and prison sentences.

Until recently the Sentencing Council for England and Wales employed a scale of severity derived from their sentencing guidelines. In essence this scale exploits the relationship between different sentence outcomes attached to increasing levels of offence seriousness coded in the guidelines. The overall functional form linking levels of seriousness to custodial sentence outcomes (measured in days in prison) across different offence-specific guidelines was first estimated and then extrapolated to non-custodial outcomes (see Pina-Sánchez et al. (2019) for a more detailed explanation of this process). In spite of its original design, this scale relies on arbitrary choices (such as the choice of a value of severity for a one-month custodial sentence to ensure that severity scores for non-custodial outcomes are not expressed as negative days in prison) and unrealistic assumptions (such as taking changes in seriousness across different offences to be equivalent).

Francis et al. (2005) and McDavid and Stipack (1981) managed to create scales of severity covering all disposal types using latent variable estimation methods such as canonical correlation and correspondence analysis. Both of these methods are similar in that they search for latent scales summarising the relationship between different types of sentences and crimes recorded in two-way frequency tables. These methods possess a clear advantage in that they do not rely on subjective views, yet, they are still subject to important limitations. Canonical correlation analysis assumes a linear perfect correlation between the latent variables for crime seriousness and sentence severity. This is equivalent to assume that sentences are entirely determined by the seriousness of the crime, disregarding the effect that personal mitigating factors or other non-retributive considerations have on the final sentence. Perhaps more importantly, these type of methods seem to give rise to nonsensical values, questioning their validity. For example, Francis et al. (2005) scale identified 'tagging plus a community punishment of less than one year' as a more severe sentence than 'tagging and community punishment of one year or more'.

A last group of studies has relied on different forms of paired comparisons. As conceived by Thurstone (1927), paired comparisons is the simplest approach to elucidate subjective views on sentence severity (Spelman, 1995). Subjects are presented with a series of choices; for each choice, they are asked to identify the option they perceive to be more intense (severe, in our case). As such, the rationale of the application of Thurstone's paired comparisons to the estimation of sentence severity resides on the concept of 'penal exchangeability', which points at the existence of overlaps in severity between different disposal types (Erickson and Gibbs, 1979; Harlow et al., 1995; McClelland and Alpert, 1985; Sebba, 1978; Sebba and Nathan, 1984; Leclerc and Tremblay, 2016; Lovegrove, 2001; Petersilia and Deschesnes, 1994a,b; Spelman, 1995; Tremblay, 1988). For example, high community orders with onerous conditions attached might in some instances be more severe than short suspended sentences. Under the assumption that severity scores for every sentence considered are normally distributed, the number of times an offence is judged more serious than another is then used to locate each of the severity distributions for each of the sentences considered, and to obtain their respective mean severity. In spite of its simplicity and capacity to offer interval and well-founded severity scores, we are only aware of a single study that has applied paired comparison in its original format for the estimation of sentence severity (Buchner, 1979), which is now four decades old. Pina-Sánchez et al. (2019) - the scale of severity adopted in this study - borrows several elements of paired comparison methods, but it also relies on additional processes which we proceed to cover next.

2.1 The method used in Pina-Sánchez et al. (2019)

The scale developed by Pina-Sánchez et al. (2019) relies on self-completed questionnaire responses from 21 magistrates recruited using snowball sampling. Two magistrates operating in two different courts - in the north and south of England - were initially contacted. These magistrates then circulated the questionnaire amongst contacts from their own and nearby magistrates' courts. The specific courts where the 21 responses were gathered cannot be disclosed to protect respondents' anonymity. The content of the questionnaire was determined by the sentence outcomes recorded in the England and Wales official sentence datasets, namely the Ministry of Justice sentencing statistics (Ministry of Justice, 2018), and the Crown Court Sentencing Survey, published by the Sentencing Council for England and Wales (Sentencing Council, 2015). In their original format, these datasets distinguish the different disposal types used in England and Wales, but not the specific details of those sentences beyond the duration of custodial and suspended sentences (for example, the type of conditions attached to community orders or suspended sentence orders is not known).

These restrictions in the official data limits the number of specific sentence outcomes that can be studied. Still, there are several pairwise combinations of sentences that could potentially be considered based on the available data. To minimise the duration of the questionnaire, only those sentence comparisons deemed essential were included. Two focus groups were convened with two and four magistrates to explore the types of questions where overlaps in severity could be present. Paired comparisons were discarded when it was agreed that 'almost always' one of the sentences will be more severe (for example, an immediate custody sentence was considered to be always more severe than a community order). At this point, the final set of sentences to be compared was extended to include enough cases of suspended and immediate custody, so the severity score for any duration of those types of sentences could be extrapolated in a second stage. Under such criteria the sentence outcomes included in the questionnaire are: conditional discharge, fine, community order, one-month custodial sentence suspended for six months, one-month custodial sentence suspended for twelve months, six-month custodial sentence suspended for six months, twelve-month custodial sentence suspended for 24 months, and one-, two-, and three-month immediate custodial sentences.

Rather than wording paired comparisons in absolute terms, as originally designed by Thurstone (1927), Pina-Sánchez et al. (2019) phrased their questions in relative terms. That is, instead of asking *'which of the two is more severe'*, the authors asked *'how often can sentence A be more punitive than sentence B'*.[2] This change was introduced to prevent participants from systematically choosing the sentence considered more severe according to the sentencing ladder. Clearly, this modification makes the questionnaire more cognitively demanding, although still simpler to answer than any of the other alternative methods used in the literature, while providing the necessary measurements of the severity overlap between crucial sentence outcomes.

The information retrieved from the 21 magistrates is summarised in Table 1, where each cell indicates the sample mean describing how often the sentence at the top is deemed more severe than the sentence in the left margin. Only those cells showing values different from 1s, 0s, and 0.5s report sample means retrieved from the questionnaire; cells featuring any of the three figures listed above represent sentence combinations not included in the questionnaire as they were deemed unnecessary (one of the sentences considered always more severe, 1, always less severe, 0, or comparisons of a sentence with itself, 0.5). Absolute discharge was not included in the questionnaire either, but it was included in this 'matrix of severity' to set up a meaningful zero in the scale of severity.

To estimate severity scores for each of the sentences types considered, the proportions presented in Table 1 were employed in a Thurstone model for paired comparisons. The statistical model that underpins the method associates a normal distribution with same variance and changing mean for each of the sentence types. Each of the means in these normal distributions can be thought of as severity scores. Intuitively, the amount of overlap between the distributions - informed by the proportions in Table 1 - dictates their closeness on the severity scale. To estimate the means of the underlying normal distributions and, hence, the severity scores, a least squares approach is utilised where the reported proportions are compared with the probabilities of one sentence being greater than another conditional on different sets of means. The specific model used is Thurstone model Case V (Mosteller, 1951), as configured under the 'psych' package in R (Revelle, 2018). This is the simplest form of the Thurstone model, it assumes that the underlying normal distributions for the severity of each sentence type have equal variance of 0.5 so that the differences between sentences have a variance of 1.

[2] The full questionnaire can be accessed here:
`https://leeds.onlinesurveys.ac.uk/scale-of-sentence-severity-open`.

Table 1: Matrix of proportions stemming from paired comparisons for each of the sentence outcomes considered

	absolute discharge	conditional discharge	fine	community order	1-month custody 6-month suspended	1-month custody 12-month suspended	6-month custody 6-month suspended	12-month custody 24-month suspended	1-month immediate custody	2-month immediate custody	3-month immediate custody
absolute discharge	0.5	1	1	1	1	1	1	1	1	1	1
conditional discharge	0	0.5	0.69	1	1	1	1	1	1	1	1
fine	0	0.31	0.5	0.77	1	1	1	1	1	1	1
community order	0	0	0.23	0.5	0.39	1	1	1	1	1	1
1-month custody 6-month suspended	0	0	0	0.61	0.5	1	1	1	1	1	1
1-month custody 12-month suspended	0	0	0	0	0	0.5	0.74	1	1	1	1
6-month custody 6-month suspended	0	0	0	0	0	0.26	0.5	1	1	1	1
12-month custody 24-month suspended	0	0	0	0	0	0	0	0.5	0.41	0.51	0.61
1-month immediate custody	0	0	0	0	0	0	0.59	0.5	1	1	1
2-month immediate custody	0	0	0	0	0	0	0	0.48	0	0.5	1
3-month immediate custody	0	0	0	0	0	0	0	0.38	0	0	0.5

The sentence types included in the questionnaire reached up to a three-month immediate custodial sentence. After obtaining severity scores for the sentence types included Pina-Sánchez et al. (2019) proceeded to estimate longer custodial sentences. To do so, they used a linear extrapolation using the severity scores for one-, two- and three-month immediate custody. This approach is questionable though. Different studies on the penal metric literature (Leclerc and Tremblay, 2016; Spelman, 1995) find non-linear relationships between days in prison and severity. Most commonly, they find diminishing returns of severity for every additional day of punishment or fine amount.

This modelling challenge is undertaken in the next section, where we also cover how to estimate the uncertainty associated with the development of a scale of severity based on paired comparisons. The former upgrade is key to obtain a realistic and valid scale of severity, however it will be applied, the latter is driven by our interest to use this scale of severity as a tool to eliminate the problem of selection bias in sentence data. To do so we need to find a way to estimate and propagate the uncertainty associated with the estimation of our scale of severity into any sentencing model of interest where the scale of severity will be adopted.

2.2 Accounting for uncertainty in the estimation of severity

Pina-Sánchez et al. (2019) only accessed a small subset of magistrates from the entire population of sentencers in England and Wales. As such, we are uncertain as to whether the proportions reported in Table 1 are truly representative. If we assume the individual magistrates form a random sample, we can use a Bayesian model for the individual proportions retrieved from the questionnaire to capture our uncertainty. In practice, we assume that each of those proportions, ϕ, is drawn from a beta distribution whose parameters, α and β, are selected to reflect the variability in the magistrates' judgements. Giving relatively non-informative prior distributions for both α and β

$$\alpha \sim \mathrm{IG}(0.1, 0.1) \text{ and } \beta \sim \mathrm{IG}(0.1, 0.1),$$

we can estimate the uncertainty about ϕ. To undertake these estimations we used the Bayesian package Stan (Carpenter et al., 2017); the code used is shown in the supplementary files. In subsequent analyses, this uncertainty about all of the unknown proportions in Table 1 is propagated to the final results. Choices that we have made in this modelling step and the sample of magistrates we have obtained have a clear effect on the uncertainty distributions; we accommodate this fact by conducting robustness analyses on the model outputs accounting for reasonable deviations in the modelling. In Figure 1, an example of the Bayesian update for the proportion of times a fine is judged to be more punitive than a community order is given. It can be seen from Figure 1 that, despite the relatively few observed proportions, there is an appreciable level of learning about the parameters of the underlying beta distribution.

Posterior severity scores are obtained by sampling from the posterior distributions for all of the proportions and applying the Thurstone model. The posterior median severity scores obtained for each of the sentence outcomes considered are shown in Table 2.

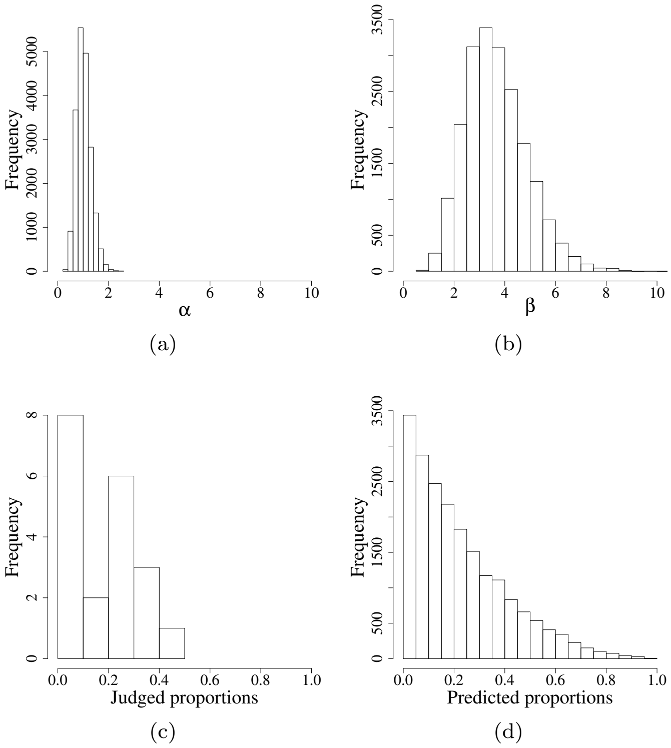

Fig. 1: Posterior samples for α and β from the model for the proportion of times a fine is judged to be more punitive than a community order (top histograms); data collected from the magistrates (bottom left); and sampled proportions from the predictive distribution based on the model and data (bottom right).

Table 2: Posterior median scale of severity

Sentence outcome	Severity score
absolute discharge	0
conditional discharge	0.96
fine	1.32
community order	2.14
1 month custody 6 months suspended	2.34
1 month custody 12 months suspended	3.6
6 months custody 6 months suspended	3.84
12 months custody 24 months suspended	5.76
1 month custody	5.03
2 months custody	5.75
3 months custody	6.45
12 months custody	12.77
5 years custody	45.55
20 years custody	156.01

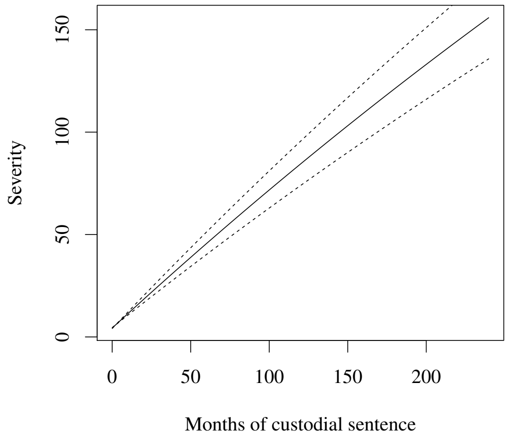

Fig. 2: Posterior median (solid line), 5th and 95th percentiles (dashed lines) for the function covering extrapolation from a three-month custodial sentence up to 20 years.

Immediate custodial sentences longer than three months were estimated using a new extrapolation model that accounted for a law of diminishing severity. The necessary information was elicited from our sample of magistrates. For example, on average, the change in severity going from a three-month custodial sentence to a four-month custodial sentence was reported to be greater than the change in severity going from a twelve-month custodial sentence to a thirteen-month custodial sentence. The functional form chosen for the extrapolation was

$$\text{Severity} = f(\text{Custodial sentence in months} = c) = p_1 \left(1 - \exp\left(-\frac{c}{p_2}\right)\right) + p_3$$

where the parameters p_1, p_2 and p_3 are determined by optimisation given the three severity scores for one-, two- and three-month custodial sentences and information about three different comparators that capture the diminishing severity effect. Figure 2 shows the uncertainty in the extrapolation to longer sentencing lengths given the uncertainty in the proportions that were used in the Thurstone model.

A similar approach could be adopted to estimate severity scores for suspended sentences of different durations from those included in the original survey. This process was not undertaken here since it was not possible to obtain the original datasets from the Ministry of Justice or the Sentencing Council where the information on the suspended sentences is available. In our subsequent analysis, in the absence of further information, we allocated one of the four severity values for suspended sentences in Table 2 at random with each having equal probability of allocation.

3 Modelling sentence severity

In this section, we proceed to use our new scale of severity to illustrate the impact of selection bias in sentence data analysis and to assess the effectiveness of adjustments based on Tobit models. To do so, we use a sample of 7,242 theft offences sentenced at the Crown Court in 2011 recorded by the Crown Court Sentencing Survey. This sample was chosen as it is one of the few sentencing datasets in England and Wales for which the specific duration of custodial sentences is publicly available. The focus on theft offences is explained by the relative lower seriousness of these offence types compared to other crimes sentenced at the Crown Court, which offers a large enough group of offences receiving non-custodial sentences. Specifically, within the 7,242 records there are 151 conditional discharges, 74 fines, 989 community orders, 1,806 suspended sentences, and 4,220 immediate custodial sentences. Figure 3 shows the distribution of those sentences once transformed into severity scores.

In addition to the sentence outcome, a set of explanatory variables was retrieved from the Crown Court Sentencing Survey. This covers two offender characteristics (age and gender), the main offence committed (divided in 22 offence types) and two additional case characteristics (number of previous convictions and whether a guilty plea was entered). Descriptive statistics for the variables used are shown in Table 3.

Our analytical strategy involves the specification of severity scores under three different models. To adjust for the right skewness in the distribution of the severity scores shown in Figure 3 this variable is log-transformed. The log transformation of severity scores is then regressed using the rest of the variables listed in Table 3 as the set of explanatory variables (with burglary in a dwelling taken as the reference category). To illustrate the effect of selection bias we compare results from Models 1 and 2. Model 1 is restricted to cases sentenced to immediate custody, while Model 2 is based on the whole sample.

Model 2 eliminates the problem of selection bias by using all sentence outcomes available while making the most of the information within them: that is, distinguishing severity scores for different non-custodial outcomes. The uncertainty associated with the estimation of severity scores to be used in both models is propagated as explained in Section 2.2. Yet, another modelling challenge needs to be tackled in the specification of Model 2. Since we do not know the exact figure in pounds sterling for fines imposed or the conditions attached to conditional discharges or community orders, our scale of severity takes each of those three sentence outcomes as homogeneous groups. This is equivalent to the introduction of Berkson measurement errors (Berkson, 1950) in the severity scores that we are using as a dependent variable, which is known to bias the measures of uncertainty from models' estimates downward (Armstrong, 1998; Heid et al., 2004). To account for this unobserved heterogeneity in conditional discharges, fines and community orders, we use the latent distributions of severity used in the Thurstone method for each of those sentence outcomes. Specifically, for each of the scales of severity obtained through the iterative process described above, severity scores for cases sentenced to conditional discharge, fine or community order were taken from three different normal distributions with mean equal to the severity score obtained for each of those disposal types and a variance of 0.5.

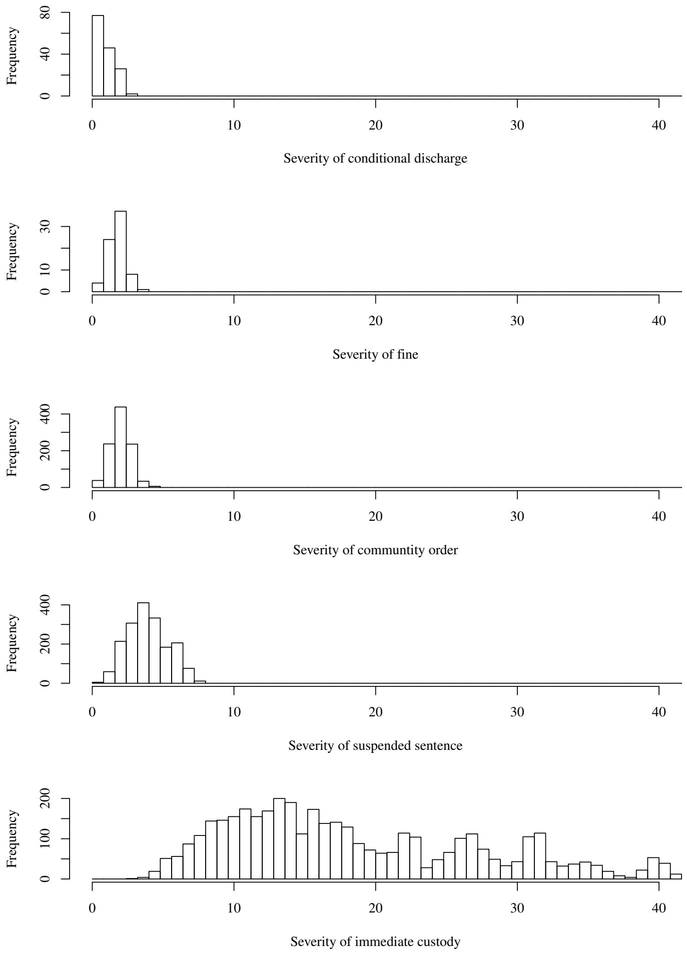

Fig. 3: Distribution of values of severity in our dataset given a single random scale of severity taken from our model. Note that to facilitate comparability the x-axis as been capped at 40.

The overall workflow highlighting the different stages and inputs to the modelling is given in Figure 4. The final step of the process shown in Figure 4 is a linear regression model fitted through conjugate Bayesian methods. In particular, we regress each set of severity scores (s_i for $i = 1, \dots, 7242$ when considering all sentencing outcomes) on the various case characteristics that we have available

Table 3: Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Max
Offender characteristics					
age of the defendant	7,242	32.423	11.024	18	83
male defendant	7,242	0.852	0.355	0	1
Case characteristics					
severity	7,242	13.116	12.363	0.95	105.84
guilty plea	7,242	0.847	0.360	0	1
previous convictions 1 to 3	7,242	0.252	0.434	0	1
previous convictions 4 to 9	7,242	0.164	0.370	0	1
previous convictions 10 or more	7,242	0.170	0.375	0	1
Main offence					
aggravated burglary dwelling	7,242	0.004	0.063	0	1
aggravated burglary not in a dwelling	7,242	0.006	0.076	0	1
attempted theft	7,242	0.005	0.072	0	1
burglary in a dwelling	7,242	0.316	0.465	0	1
commercial burglary	7,242	0.079	0.269	0	1
conspiracy to commit burglary	7,242	0.003	0.057	0	1
conspiracy to commit fraud	7,242	0.007	0.084	0	1
conspiracy other	7,242	0.002	0.048	0	1
conspiracy to steal	7,242	0.008	0.088	0	1
dishonest representation	7,242	0.066	0.248	0	1
going equipped	7,242	0.007	0.085	0	1
false passport	7,242	0.035	0.184	0	1
handling stolen goods	7,242	0.011	0.106	0	1
immigration offence	7,242	0.004	0.066	0	1
money laundering	7,242	0.016	0.124	0	1
other fraud offence	7,242	0.140	0.347	0	1
other theft offence	7,242	0.040	0.196	0	1
receiving stolen goods	7,242	0.066	0.248	0	1
theft from a person	7,242	0.048	0.215	0	1
theft from a shop	7,242	0.061	0.239	0	1
theft in breach of trust	7,242	0.062	0.242	0	1
theft of a vehicle	7,242	0.005	0.071	0	1

($\mathbf{x}_i$ covering gender, offence type etc.):

$$\begin{aligned} s_i &= \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i \text{ for } i = 1, \ldots, 7242, \\ \epsilon_i &\sim \mathrm{N}(0, \sigma^2), \\ \boldsymbol{\beta} | \sigma^2 &\sim \mathrm{MVN}(\mathbf{0}, (10{,}000\sigma^2) I), \\ \sigma^2 &\sim \mathrm{InvGamma}(0.001, 0.001). \end{aligned}$$

Following this process, we have posterior distributions for the parameters of the regression model for each sampled severity scale. Using the law of total variance, we are able to combine the uncertainty from the scale fitting with the uncertainty from the regression.

To assess the effectiveness of selection bias adjustments commonly undertaken in the literature, we estimate a third model using a Tobit specification. Model 3 replicates Model 2 with the only difference that severity scores for non-custodial sentences are taken as left-censored - reflecting the main rational of adjustments based on Tobit models - while using severity rather than sentence length as the scale to facilitate comparisons with Models 1 and 2.

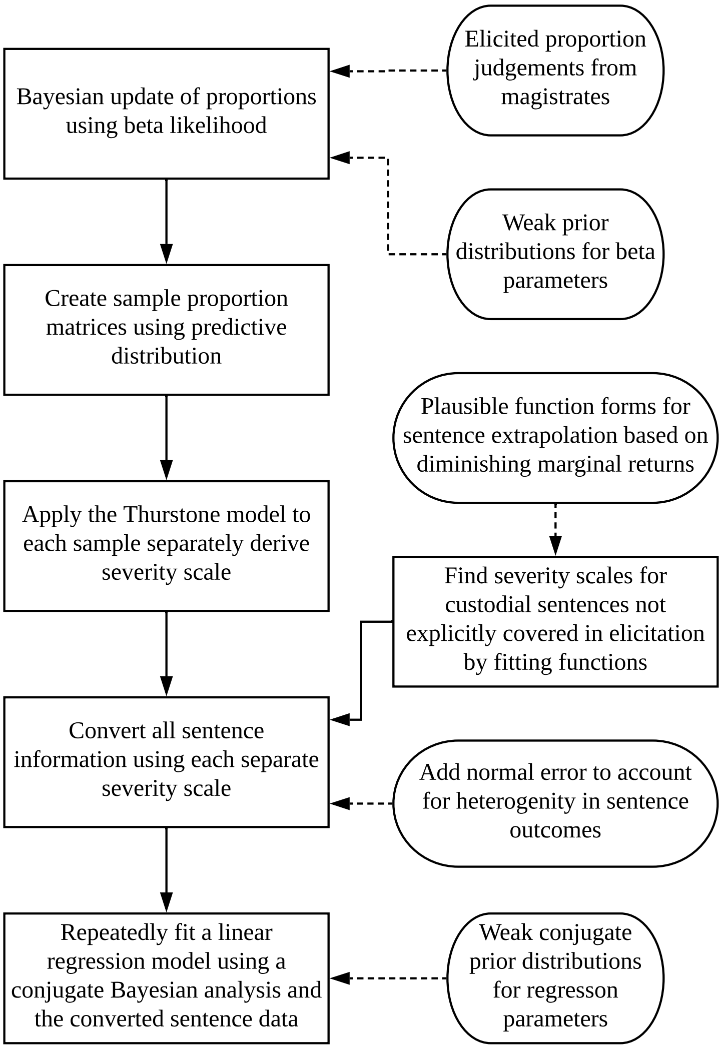
Elicited proportion judgements from magistrates
Bayesian update of proportions using beta likelihood
Weak prior distributions for beta parameters
Create sample proportion matrices using predictive distribution
Plausible function forms for sentence extrapolation based on diminishing marginal returns
Apply the Thurstone model to each sample separately derive severity scale
Find severity scales for custodial sentences not explicitly covered in elicitation by fitting functions
Convert all sentence information using each separate severity scale
Add normal error to account for heterogenity in sentence outcomes
Repeatedly fit a linear regression model using a conjugate Bayesian analysis and the converted sentence data
Weak conjugate prior distributions for regresson parameters

Fig. 4: The different stages of the analysis with rounded boxes showing inputs to the process and sharp boxes giving computational steps.

Lastly, to assess the robustness of our findings regarding the presence of selection bias in models restricted to custodial sentences and the effectiveness of Tobit models to adjust for it, the same three-model comparison is replicated using a different scale of severity. The scale of severity developed by the Sentencing Council is used for such sensitivity analysis since it covers most of the sentence outcomes captured in the sentence data used in our analysis. Other scales of severity developed for the jurisdiction of England and Wales are too outdated (Kapardis and Farrington, 1981), or considers different sentence outcomes (Francis et al., 2005).

The severity scores that conform the Council's scale and results from Models 1, 2, and 3 when the log of severity scores derived from this scale is used as the response variable are reported in the Appendix.

4 Results

Table 4 shows the results for Models 1 and 2. Looking first at Model 1, where only custodial sentences are used, we can see that all case characteristics point in the expected direction (for example, previous convictions increase sentence severity while pleading guilty reduces it) and are found to be statistically significant. Age and gender are also found to be positive and significant, indicating that older, male offenders seem to receive harsher sentences. This could however be due to confounding effects from having missed important aggravating factors related to the offence seriousness or other personal mitigating factors. For example, the models presented are not controlling for whether the offender has caring responsibilities, which tend to be disproportionally associated with women. Hence, it would be unwise to claim that the positive effect observed for male offenders is due to sentencing discrimination against men.

The impact of selection bias in sentence data analyses limited to custodial sentences can be illustrated through the comparison of the regression coefficients obtained in Models 1 and 2. We can see how, in spite of a larger sample size, standard errors in Model 2 are bigger than in Model 1. More importantly, there are substantial differences in the effects of most of the variables considered. In particular, the effect size for previous convictions and for gender is at least four and three times bigger when considering all sentence outcomes than in the model focused on custodial sentences. This result indicates that both of these variables play a more important role at determining the severity of the disposal type to be imposed than the specific duration of custodial sentences. On the other hand, the effect of guilty plea is smaller in Model 2, pointing at this factor being more decisive when it comes to determine sentence duration than the specific disposal type to be used. This difference in the effect of a guilty plea resonates well with how plea reductions are structured in England and Wales. Specifically, both the Sentencing Guidelines Council (2007) and the more recent Sentencing Council (2017) guilty plea guidelines establish reductions as a percentage of the final sentence, which suggests that the reduction is to be applied once the specific disposal type has been decided, and only to sentences expressed numerically, such as days in prison. Importantly, these differential effects observed for guilty pleas and previous convictions across Model 1 and 2 operate in direct violation of the assumptions of the commonly used Tobit models for censored data.

It is also worth noting that each of these effects attributed to selection bias are corroborated in our sensitivity analysis when severity scores are derived from the Council's scale of severity (see Table A2 in the appendix).

Besides differences in the direction that selection bias can operate, we should also note the magnitude of that bias. To facilitate that, we can compare the effect of some of the factors considered under different hypothetical scenarios. The simplest of those would be the reference case of burglary in a dwelling committed by a female offender with no previous records pleading not guilty. The predicted severity for such case in Model 1 would be 16.22 (equivalent to 16.95 months in custody). The

Table 4: Regression coefficients (posterior median) from each model with standard deviation in parentheses

	Dependent variable: log(severity)		
	Model 1 - custody	Model 2 - all sentences	Model 3 - Tobit
constant*	2.786 (0.066)	1.838 (0.068)	1.615 (0.070)
Offender characteristics			
age of the defendant	0.006 (0.001)	0.005 (0.001)	0.006 (0.001)
male defendant	0.054 (0.026)	0.185 (0.036)	0.275 (0.041)
Case characteristics			
guilty plea	-0.132 (0.020)	-0.104 (0.032)	-0.152 (0.035)
previous convictions 1 to 3	0.095 (0.021)	0.477 (0.034)	0.600 (0.035)
previous convictions 4 to 9	0.187 (0.024)	0.732 (0.041)	0.880 (0.041)
previous convictions 10 or more	0.199 (0.024)	0.835 (0.042)	0.990 (0.041)
Main offence			
aggravated burglary dwelling	0.855 (0.087)	1.235 (0.173)	1.300 (0.185)
aggravated burglary not in a dwelling	0.737 (0.073)	1.185 (0.143)	1.234 (0.152)
attempted theft	-0.719 (0.125)	-1.242 (0.167)	-1.446 (0.194)
commercial burglary	-0.337 (0.028)	-0.434 (0.046)	-0.466 (0.048)
conspiracy to commit burglary	0.221 (0.099)	0.430 (0.189)	0.470 (0.205)
conspiracy to commit fraud	0.290 (0.072)	0.651 (0.132)	0.706 (0.144)
conspiracy other	0.126 (0.114)	0.675 (0.224)	0.753 (0.243)
conspiracy to steal	0.212 (0.068)	0.430 (0.127)	0.501 (0.137)
dishonest representation	-0.587 (0.052)	-0.714 (0.063)	-1.071 (0.070)
going equipped	-0.728 (0.099)	-1.009 (0.137)	-1.198 (0.157)
handling stolen goods	-0.426 (0.081)	-0.692 (0.112)	-0.881 (0.128)
immigration offence	-0.339 (0.092)	0.289 (0.167)	0.370 (0.183)
money laundering	-0.015 (0.063)	-0.200 (0.095)	-0.218 (0.104)
other fraud offence	-0.262 (0.027)	-0.370 (0.041)	-0.386 (0.044)
other theft offence	-0.626 (0.049)	-0.935 (0.067)	-1.037 (0.071)
receiving stolen goods	-0.520 (0.041)	-1.030 (0.058)	-1.209 (0.058)
theft from a person	-0.558 (0.042)	-0.883 (0.062)	-0.921 (0.063)
theft from a shop	-0.827 (0.049)	-1.247 (0.064)	-1.332 (0.060)
theft in breach of trust	-0.294 (0.039)	-0.417 (0.056)	-0.534 (0.062)
theft of a vehicle	-0.564 (0.113)	-0.923 (0.162)	-1.117 (0.182)
false passport	-0.580 (0.043)	-0.043 (0.066)	0.027 (0.071)
Observations	4,220	7,242	7,242

*Reference categories: female offender, no guilty plea entered, no previous convictions, main offence burglary in a dwelling

predicted severity from the same model for a similar case where the offender is a man would be 17.12 (18.24 months in custody), which equals a 5.5% increase in severity. Hence, a researcher set out to assess any potential discriminatory effect against men in the Crown Court based on a model for sentence length might interpret the gender effect as relatively negligible. However, a similar comparison based on Model 2 shows that male offenders receive 20.3% more severe sentences than women. Whether that is due to genuine discriminatory practices or to a different composition of aggravating and mitigating factors across genders remains to be explored. What is undeniable is how fundamentally misleading would have been to rely on the effect size from a model based on custodial sentences.

The impact of selection bias on the gender estimate is dwarfed in comparison by the much bigger effect observed for previous convictions. Using the reference case described above and results from Model 1, we can estimate an average increase in severity of 10% when the defendant has one to three previous convictions compared to the same case when no previous convictions were noted. The estimated increase for cases with four to nine, and for ten or more previous convictions, compared to the same case with no previous convictions, reaches 20.6% and 22%, respectively. The same comparisons based on results from Model 2 show increases of 61.1%, 107.9% and 130.5%. It is therefore clear how analyses based on sentence length would have severely underestimated the effect of previous convictions on sentence severity. Substantively, the fact that previous convictions can more than double the severity with which the same offence is dealt with, questions the assumed principle of offence-based proportionality in England and Wales, and demonstrates that this issue is not confined to US jurisdictions (Hester et al., 2018). It should also be noted that the marginal effect is strongest when transitioning from no previous convictions to one to three previous convictions, with that effect decaying as the number of previous convictions increases. This corroborates the 'progressive loss of mitigation' model described by von Hirsch (2010) (see also Roberts, 2008; Roberts and Pina-Sánchez, 2014), and refutes the 'cumulative sentencing' model, under which the marginal increase in severity for every additional relevant previous conviction is constant.

Finally, results from Model 3 can be used to illustrate the effectiveness of methods used to adjust for selection bias in sentence data analysis. This model replicates Model 2, but sets non-custodial sentences as left-censored, following the logic of adjustments based on Tobit models. We can see how this approach manages to rescale estimates for previous convictions, which are now roughly in line with the estimates observed for Model 2 when the whole sample size is used. However, we can also note how the 'adjustments' obtained for male defendant and guilty plea end up making things worse. The effect of male in Model 3 is much stronger than what was observed in Model 2. So much so that the difference between the estimate from Model 3 and Model 2 is wider than the difference between Model 2 and 1, attributed to the original problem of selection bias. Yet, limitations of adjustments are even more noticeable when we look at the effect of guilty plea. In this case the adjustment contributes to augment the original bias. This illustrates how ill-equipped adjustments based on Tobit models are when it comes to estimate the effect of factors that affect custodial and non-custodial outcomes differently.

It needs to be noted, however, that after replicating Models 2 and 3 using the Council's scale of severity we observe a better performance of the Tobit model, which manages to adjust a larger share of the selection bias in most of the regression coefficients (see Table A2 in the appendix). One exception being guilty plea, for which the Tobit model remains ineffective. This general better performance of the Tobit model for severity scores derived from the Council's scale can be attributed to two specific features of that scale, which conform with two key assumptions invoked by the Tobit model. Namely, that there is no overlap in terms of severity between custodial and non-custodial sentences (that is, the former will always be more severe than the latter), and that both custodial and non-custodial sentences can be considered to be stemming from the same normal distribution (see Table A1 in the appendix). Our scale on the other hand allows for severity overlaps between long suspended and short immediate custodial sentences and considers different distributions of severity for each of the main disposal types employed (see Figure 3).

5 Discussion

In this study, we have used a scale of severity and a sample of theft offences to illustrate empirically the magnitude of the problem of selection bias affecting studies limited to custodial sentences. Even though the sample used comprises mainly custodial sentences (71.6% of the total), the magnitude of the bias is substantial, which corroborates previous warnings raised in the literature pointing at the problem of selection bias as one of the most pervasive methodological challenges affecting the analysis of sentence data (Baumer, 2013; Bushway et al., 2007; Ulmer, 2012; Zatz and Hagan, 1985). The more immediate conclusion that should be taken from this study is that findings from studies based on samples of custodial sentences cannot be generalised to the whole of the sentencing practice. Given the magnitude of the bias and the fact that it can operate in different directions, it is difficult to anticipate the true effect of different legal or extralegal factors on the final sentence using just custodial sentences.

From our specific application using a sample of theft offences processed in the Crown Court in 2011, we can also provide various other substantial findings regarding the effect of guilty pleas, gender and previous convictions, three old questions in the field of sentencing, for which analyses have been hampered by selection bias. Using our new scale of severity we can now obtain more accurate estimates of the effect of these case characteristics, and shed some light on how previous literature might have been misled by the problem of selection bias. Specifically, we have noted how studies based on custodial sentence length will overestimate the effect of guilty plea reductions on sentence severity and underestimate the effect of gender and previous convictions. Gender, as an extralegal factor, should not really affect sentence severity all other things equal, and that is what would have been roughly inferred had we based our analysis on sentence length. Yet, when considering all types of sentences we detected a 20.3% increase in severity for men. Still, the impact of selection bias was best exemplified for the effect of previous convictions, for which the increase in severity appears five to ten times smaller when analyses are limited to custodial sentences.

Besides illustrating empirically the potential magnitude of selection bias in sentence data analyses, the goal of this study is to offer an alternative modelling approach to tackle this methodological challenge. We have shown how the severity scale can be estimated and used as the response variable in a sentencing data model of interest while propagating the uncertainty associated with the estimation of the severity scale adequately. Using this approach the problem of selection bias in sentence data analyses can be eliminated, while providing important advantages compared to the methods currently used in the literature for this purpose.

We argued in the introduction that the assumptions invoked by the methods for the adjustment of selection bias used in the literature are unrealistic. Our approach is also based on a number of assumptions, many of them clearly quite restrictive. Equal variance in the distributions of severity scores for the different sentence outcomes, or the assumed normality of those distributions are two important assumptions that we invoke. As a sensitivity analysis we replaced the Thurstone model for the estimation of the scale of severity with the Bradley-Terry model (Bradley and Terry, 1952) using the 'BradleyTerryScalable' (Kaye and Firth, 2017) R package, which involves assuming logistic rather than normal distributions of severity for each sentence outcome. Results are remarkably similar, which offer some reassurance towards the robustness of our approach. We have used these two distributions as they are the most commonly applied in the paired comparison literature. Yet, these are only two of the potentially infinite distributional forms that severity scores could be following. As a further sensitivity analysis we replicated our analytical strategy using an alternative scale of severity developed by the Sentencing Council for England and Wales. We found similar manifestations of the impact of selection bias - underestimated measures of uncertainty and biased regression coefficients - only stronger in magnitude.

The main advantages of our approach over the Heckman and Tobit models resides in the possibility of discriminating between non-custodial sentence outcomes and in recognising overlaps in severity across different disposal types. Something that these two approaches so commonly used in the literature for the adjustment of selection bias (and any other approach that we are aware of; for example, two-part models (Koons-Witt et al., 2014) or hurdle models (Hester and Hartman, 2017)) fail to achieve. Treating all non-custodial sentences as a homogeneous group represents an important loss of information, equivalent to the introduction of Berkson measurement errors, which besides the obvious loss of statistical power, it is also capable of generating biases of its own. By left-truncating severity scores for non-custodial sentence outcomes, we reproduced a typical adjustment based on a Tobit model. Comparing results for this model with those obtained from a model using all the available information as estimated in our scale of severity, we found that solutions based on Tobit models, far from adjust for the problem of selection bias, they can in some instances exacerbate it by reinforcing patterns seen in the non-censored data.

5.1 Future avenues of research

The extent of the selection bias problem is proportional to the share of cases receiving non-custodial sentences. The use of a scale of severity overcomes this problem, and it offers an innovative approach with which to refocus the attention of sentencing research towards areas that have been relatively neglected, such as sentencing in lower courts, or less serious offences. One particularly relevant case would be the study of sentencing in the magistrate's court in England and Wales, where only 3.7% of offenders were sentenced to custody in 2017, compared to a 58.1% rate in the Crown Court according to Ministry of Justice (2018) statistics. After decommissioning the Crown Court Sentencing Survey, the Sentencing Council has focused its data-collection strategy on the magistrate's court. Analyses of these new Sentencing Council datasets will benefit from the approach suggested here.

Beyond the possibility of shedding new light on those courts where most of the sentencing practice takes place, the release of these new Sentencing Council datasets will offer a new opportunity to further refine the scale of severity presented here. The Council is now recording community orders by their type (low, medium and high) and fines by their band (A, B, C, D, E and F). These more dissaggregated sentence outcomes will provide a fantastic opportunity to discriminate even more within non-custodial sentence and make the approach to tackle selection bias suggested here even more informative, and statistically efficient.

The potential to expand the severity scale to encompass this new information recorded by the Sentencing Council also illustrates the main limitation of our approach, namely scales of severity are place- time- and topic-dependent. Different jurisdictions use different types of sentences, furthermore, the sentences available to judges in a particular jurisdiction, and their relative severity, change across time. Our scale of severity will need to be replicated in the medium term to maintain the accuracy of our approach for studies using England and Wales data in the future. Studies focused on a different jurisdiction interested in adopting our approach to deal with the problem of selection bias will need to estimate their own scale of severity. We believe that the methodology for the estimation of a scale of severity presented here offers important advantages over other methods used in the literature and we would like to encourage researchers to adopt it. Our approach avoids problems of extreme unreliability plaguing methods entirely based on subjective views, such as magnitude estimation, while simultaneously avoiding the generation of nonsensical results shown in entirely data-driven applications such as canonical correlation. Furthermore, we have also shown how given the adequate constraints in the pairs to be compared, a small sample of participants and questions would be enough to estimate a highly reliable scale of severity. Finally, in this study, we have relied on the views of sentencers to the estimation of our scale of severity since the most immediate application is the study of sentencing practice. Researchers pursuing different topics, such as penal populism, or deterrence, could use samples of the general public or offenders.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Albonetti CA (1991) An integration of theories to explain judicial discretion. Social Problems 38(2):247–266

Albonetti CA (1997) Sentencing under the federal sentencing guidelines: Effects of defendant characteristics, guilty pleas, and departures on sentence outcomes for drug offenses. Law and Society Review 31:789–824

Albonetti CA (1998) The role of gender and departures in the sentencing of defendants convicted of a white-collar offense under the Federal Sentencing Guidelines. Sociology of Crime, Law and Deviance 1:3–48

Apospori E, Alpert G (1993) Research note: The role of differential experience with the criminal justice system in changes in perceptions of severity of legal sanctions over time. Crime and Delinquency 39(2):184–194

Armstrong BG (1998) Effect of measurement error on epidemiological studies of environmental and occupational exposures. Occupational and Environmental Medicine 55:651–656

Bangs M (2016) Research outputs: developing a crime severity score for England and Wales using data on crimes recorded by the police. Tech. rep., Office for National Statistics, URL https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/

researchoutputsdevelopingacrimeseverityscoreforenglandand-walesusingdataoncrimesrecorded bythepolice/2016-11-29

Baumer EP (2013) Reassessing and redirecting research on race and sentencing. Justice Quarterly 30(2):231–261

Berk RA, Ray SC (1982) Selection biases in sociological data. Social Science Research 11:352–398

Berkson J (1950) Are there two regressions? Journal of the American Statistical Association 45(250):164–180

Bernstein IN, Kelly WR, Doyle PA (1977) Societal reaction to deviants: The case of criminal defendants. American Sociological Review pp 743–755

Bottoms AE, Rex S, Robinson G (2004) Alternative to Prison: Options for an Insecure Society. Willan

Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika 39:324–345

Buchner D (1979) Scale of sentence severity. The Journal of Criminal Law and Criminology 70(2):182–187

Bushway SD, Piehl AM (2001) Judging judicial discretion: Legal factors and racial discrimination in sentencing. Law and Society Review pp 733–764

Bushway SD, Johnson BD, Slocum LA (2007) Is the magic still there? The use of the Heckman two-step correction for selection bias in criminology. Journal of Quantitative Criminology 23(2):151–178

Canton R, Dominey J (2018) Probation. Routledge, New York

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A (2017) Stan: A probabilistic programming language. Journal of Statistical Software 76(1)

Croyle JL (1983) Measuring and explaining disparities in felony sentences: Courtroom work group factors and race, sex, and socioeconomic influences on sentence severity. Political Behavior 5(1):135–153

Curran DA (1983) Judicial discretion and defendant's sex. Criminology 21(1):41–58

Erickson ML, Gibbs JP (1979) On the perceived severity of legal penalties. The Journal of Criminal Law and Criminology 70(1):102–116

Feldmeyer B, Ulmer JT (2011) Racial/ethnic threat and federal sentencing. Journal of Research in Crime and Delinquency 48(2):238–270

Francis B, Soothill K, Humphreys L, Cutajar Bezzina A (2005) Developing measures of severity and frequency of reconviction. Tech. rep., Lancaster University, URL http://www.research.lancs.ac.uk/portal/en/publications/developing-measures-of-severity-and-frequency-of-reconviction(2e9f9c6b-9f54-4ac9-9145-98b6d7262973).html

Gibson JL (1978) Judges' role orientations, attitudes, and decisions: An interactive model. American Political Science Review 72(3):911–924

Gibson JL (1980) Environmental constraints on the behavior of judges: A representational model of judicial decision making. Law and Society Review pp 343–370

Greenberg DF (1977) Socioeconomic status and criminal sentences: Is there an association? American Sociological Review 42:174–176

Gruhl J, Spohn C, Welch S (1981) Women as policymakers: The case of trial judges. American Journal of Political Science pp 308–322

Harlow RE, Darley JM, Robinson PH (1995) The severity of intermediate penal sanctions: A psychological scaling approach for obtaining community perceptions. Journal of Quantitative Criminology 11(1):71–95

Heckman J (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement 5:475–492

Heckman J (1979) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Econometrica 47:153–161

Heid IM, Küchenhoff H, Miles J, Kreienbrock L, Wichmann HE (2004) Two dimensions of measurement error: classical and Berkson error in residential radon exposure assessment. Journal of Exposure Science and Environmental Epidemiology 14(5):365–377

Helms R, D J (2002) The political context of sentencing: an analysis of community and individual characteristics. Social Forces 81:577–604

Hester R, Hartman T (2017) Conditional race disparities in criminal sentencing: A test of the liberation hypothesis from a non-guidelines state. Journal of Quantitative Criminology 33:77–100

Hester R, Sevigny EL (2014) Court communities in local context: A multi-level analysis of felony sentencing in South Carolina. Journal of Crime and Justice 39:55–74

Hester R, Frase RS, Roberts JV, Mitchell KL (2018) Prior record enhancements at sentencing: Unsettled justifications and unsettling consequences. Crime and Justice 47:209–254

Hindelang M, Dunn C, Sutton P, Aumick AL (1975) Sourcebook of Criminal Justice Statistics–1974. U.S. Department of Justice, Washington, D.C.

von Hirsch A (2010) Proportionality and progressive loss of mitigation: Further reflections. In: Roberts JV, von Hirsch A (eds) Previous convictions at sentencing: Theoretical and applied perspectives, Hart Publishing, Oxford, pp 1–16

Holmes MD, Hosch HM, Daudistel HC, Perez DA, Graves JB (1996) Ethnicity, legal resources, and felony dispositions in two southwestern jurisdictions. Justice Quarterly 13(1):11–30

Hopkins A (1977) Is there a class bias in criminal sentencing? American Sociological Review 42:176–177

Ignatans D, Pease K (2016) Taking crime seriously: Playing the weighting game. Policing: a Journal of Policy and Practice 10(3):184–193

Irwin-Rogers K, Perry TH (2015) Exploring the impact of sentencing factors on sentencing domestic burglary. In: Roberts JV (ed) Sentencing Guidelines: Exploring Sentencing Practice in England and Wales, Palgrave, Basingstoke, pp 213–239

Kapardis A, Farrington DP (1981) An experimental study of sentencing by magistrates. Law and Human Behavior 5(2-3):107–121

Kaye E, Firth D (2017) Bradleyterryscalable: Fits the bradley-terry model to potentially large and sparse networks of comparison data, R package version 0.1.0. Tech. rep., URL https://github.com/EllaKaye/BradleyTerryScalable

King NJ, Noble RL (2005) Jury sentencing in noncapital cases: Comparing severity and variance with judicial sentences in two States. Journal of Empirical Legal Studies 2(2):331–367

King RD, Johnson KR, McGeever K (2010) Demography of the legal profession and racial disparities in sentencing. Law and Society Review 44(1):1–32

Klepper S, Nagin D, Tierney L (1983) Discrimination in the criminal justice system: A critical appraisal of the literature. In: Blumstein A, J C, Martin SE, Tonry MH (eds) Research on Sentencing: The Search for Reform, National Academy Press., Washington D.C.

Koons-Witt BA, Sevigny EL, Burrow JD, Hester R (2014) Gender and sentencing outcomes in South Carolina: Examining the interactions with race, age, and offense type. Criminal Justice Policy Review 25(3):299–324

Kurlychek M, Johnson BD (2004) The juvenile penalty: a comparison of juvenile and young adult sentencing outcomes. Criminology 42:485–517

Kurlychek MC, Johnson BD (2010) Juvenility and punishment: Sentencing juveniles in adult criminal court. Criminology 48(3):725–758

Leclerc C, Tremblay P (2016) Looking at penalty scales: How judicial actors and the general public judge penal severity. Canadian Journal of Criminology and Criminal Justice 58(3):354–384

Lightowlers C, Pina-Sánchez J (2017) Intoxication and assault: an analysis of Crown Court sentencing practices in England and Wales. The British Journal of Criminology 58(1):132–154

Lovegrove A (2001) Sanctions and severity: To the demise of von Hirsch and Wasik's sanction hierarchy. The Howard Journal of Crime and Justice 40(2):126–144

McClelland KM, Alpert GP (1985) Factor analysis applied to magnitude estimates of punishment seriousness: Patterns of individual differences. Journal of Quantitative Criminology 1(3):307–318

McDavid JC, Stipack B (1981) Simultaneous scaling of offense seriousness and sentence severity through canonical correlation analysis. Law and Society Review 16(1):147–162

Ministry of Justice (2018) Criminal Justice System Statistics Quarterly: December 2017. Tech. rep., Ministry of Justice, URL https://www.gov.uk/government/statistics/criminal-justice-system-statistics-quarterly-december-2017

Mosteller F (1951) Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. Psychometrika 16(1):3–9

Muhlhausen DB (2004) The determinants of sentencing in Pennsylvania: do the characteristics of judges really matter? Tech. rep., Center for Data Analysis: Heritage Foundation, URL http://www.heritage.org/Research/Crime/cda04-02.cfm

Newheiser AK, Sawaoka T, Dovidio JF (2012) Why do we punish groups? High entitativity promotes moral suspicion. Journal of Experimental Social Psychology 48:931–936

Nienstedt BC, Zatz M, Epperlein T (1988) Court processing and sentencing of drinking drivers: using new methodologies. Journal of Quantitative Criminology 4:39–59

Nobiling T, C S, M D (1998) A tale of two counties: unemployment and sentence severity. Justice Quarterly 15:459–485

Ostrom B, Ostrom C, Hanson R, Kleiman M (2008) Assessing Consistency and Fairness in Sentencing: a Comparative Study in Three States. National Institute of Justice, Washington

Petersilia J, Deschesnes E (1994a) Perceptions of punishment: Inmates and staff rank the severity of prison versus intermediate sanctions. Prison Journal 74(3):306–328

Petersilia J, Deschesnes E (1994b) What punishes? Inmates rank the severity of prison vs. intermediate sanctions. Federal Probation 58(1):3–8

Pina-Sánchez J, Brunton-Smith I, Guangquan L (2018) Mind the step: A more insightful and robust analysis of the sentencing process in England and Wales under the new sentencing guidelines. Criminology and Criminal Justice

Pina-Sánchez J, Gosling JP, Chung H, Geneletti S, Bourgeois E, Marder I (2019) Have the England and Wales guidelines influenced sentence severity? An empirical analysis using a scale of sentence severity and time-series analyses. British Journal of Criminology

Revelle W (2018) psych: Procedures for personality and psychological research,. Tech. rep., Northwestern University, URL https://www.scholars.northwestern.edu/en/publications/psych-procedures-for-personality-and-psychological-research

Rhodes W (1991) Federal criminal sentencing: some measurement issues with application to pre-guideline sentencing disparity. Journal of Criminal Law Criminology 81:1002–1033

Roberts JV (2008) Punishing persistent offenders. Oxford University Press, Oxford

Roberts JV, Pina-Sánchez J (2014) The role of previous convictions at sentencing in the Crown Court: Some new answers to an old question. Criminal Law Review 8:575–588

Schiff MF (1997) Gauging the intensity of criminal sanctions: Developing the criminal punishment severity scale (CPSS). Criminal Justice Review 22(2):175–206

Sebba L (1978) Some explorations in the scaling of penalties. Journal of Research in Crime and Delinquency 15(2):247–265

Sebba L (1980) Mitigation in sentence in order to deter? Monash Law Review 6:268–293

Sebba L, Nathan G (1984) Further explorations in the scaling of penalties. The British Journal of Criminology 24(3):221–249

Sentencing Council (2015) Crown Court Sentencing Survey. Tech. rep., The Sentencing Council of England and Wales, URL https://www.sentencingcouncil.org.uk/analysis-and-research/crown-court-sentencing-survey/

Sentencing Council (2017) Reduction in sentence for a guilty plea: Definitive guideline. Tech. rep., Sentencing Council, URL https://www.sentencingcouncil.org.uk/publications/item/reduction-in-sentence-for-a-guilty-plea-definitive-guideline-2/

Sentencing Council (2018) Manslaughter Definitive Guideline. Tech. rep., The Sentencing Council of England and Wales, URL https://www.sentencingcouncil.org.uk/publications/item/manslaughter-definitive-guideline/

Sentencing Guidelines Council (2007) Reduction in sentence for a guilty plea: Definitive guideline. Tech. rep., Sentencing Guidelines Council, URL https://www.sentencingcouncil.org.uk/publications/item/reduction-in-sentence-for-a-guilty-plea-definitive-guideline/

Sherman L, Neyroud PW, Neyroud E (2016) The Cambridge crime harm index: Measuring total harm from crime based on sentencing guidelines. Policing: A Journal of Policy and Practice 10(3):171–183

Smith DA, Brame R (2003) Tobit models in social science research: Some limitations and a more general alternative. Sociological Methods and Research 31(3)

Sorokin PA (1962) Social and Cultural Dynamics: Fluctuations of Systems of Truth, Ethics and Law. Bedminster Press, New York

Spelman W (1995) The severity of intermediate sanctions. Journal of Research in Crime and Delinquency 32(2):107–135

Steffensmeier D, DeMuth S (2001) Ethnicity and judges sentencing decisions: Hispanic-Black-White comparisons. Criminology 39:145–176

Sullivan CJ, McGloin JM, Piquero AR (2008) Modeling the deviant Y in criminology: an examination of the assumptions of censored normal regression and potential alternatives. Journal of Quantitative Criminology 24(4):399–421

Thurstone LL (1927) A law of comparative judgement. Psychological Review 34:273–286

Tiffany LP, Avichal Y, Peters GW (1975) A statistical analysis of sentencing in Federal Courts: Defendants convicted after trial, 1967-1968. Journal of Legal Studies 4(2):369–390

Tobin J (1958) Estimation of relationships for limited dependent variables. Econometrica 26:24–36

Tremblay P (1988) On penal metrics. Journal of Quantitative Criminology 4(3):225–245

Uhlman TM (1978) Black elite decision making: The case of trial judges. American Journal of Political Science pp 884–895

Ulmer JT (2012) Recent developments and new directions in sentencing research. Justice Quarterly 29:1–40

Ulmer JT, Johnson BD (2004) Sentencing in context: a multilevel analysis. Criminology 42:137–178

Ulmer JT, Eisenstein J, Johnson BD (2010) Trial penalties in federal sentencing: extraguidelines factors and district variation. Justice Quarterly 27(4):560–592

Warr M, Meier RF, Erickson ML (1983) Norms, theories of punishment, and publicly preferred penalties of crimes. Sociological Quarterly 24:75–91

Wilson T, Loughran T, Brame R (2018) Substantial bias in the Tobit estimator: Making a case for alternatives. Justice Quarterly

Zamble E, Kalm KL (1990) General and specific measures of public attitudes toward sentencing. Canadian Journal of Behavioural Science 22(3):327–337

Zatz MS, Hagan J (1985) Crime, time, and punishment: An exploration of selection bias in sentencing research. Journal of Quantitative Criminology 1(1):103–126

Appendix: sensitivity analysis

Two sensitivity analyses were undertaken to assess the robustness of our findings. These involved the substitution of normal distributions by logistic distributions in each of the pairwise comparisons included in the Thurstone model to estimate our scale of severity, and employing an altogether different scale of severity designed by the Sentencing Council for England and Wales. Here we report results from the latter sensitivity analysis since the former provided almost identical results to those obtained in our main analysis.

Table A1 provides a comparison of the severity scores for our scale and the Council's scale across a range of sentence outcomes. Since the Council's scale did not consider suspended sentences we imputed them by taking the middle point between the severity scores given to a community order and a one-month immediate custodial sentence (25.51). Table A2 reports the results obtained for each of the three models presented in Section 4 when the dependent variable is derived from the Council's scale.

Table A1: Comparison of severity scores for a range of sentence outcomes

Sentence outcome	Our scale	Council's scale
absolute discharge	0	0
conditional discharge	0.96	0
fine	1.32	10.25
community order	2.14	21.65
1 month custody 6 months suspended	2.34	25.51
1 month custody 12 months suspended	3.6	25.51
6 months custody 6 months suspended	3.84	25.51
12 months custody 24 months suspended	5.76	25.51
1 month custody	5.03	29.37
2 months custody	5.75	29.93
3 months custody	6.45	30.48
12 months custody	12.77	35.26
5 years custody	45.55	55.89
20 years custody	156.01	100

Table A2: Sensitivity analysis based on the Council's scale of severity.
Regression coefficients from each model with standard errors in parentheses

	Dependent variable: log(severity)		
	Model 1 - custody	Model 2 - all sentences	Model 3 - Tobit
constant*	3.648 (0.013)	3.258 (0.003)	3.380 (0.017)
Offender characteristics			
age of the defendant	0.002 (0.001)	0.001 (0.001)	0.002 (0.001)
male defendant	0.017 (0.008)	0.101 (0.019)	0.066 (0.010)
Case characteristics			
guilty plea	-0.047 (0.006)	-0.024 (0.017)	-0.047 (0.009)
previous convictions 1 to 3	0.032 (0.007)	0.232 (0.017)	0.146 (0.009)
previous convictions 4 to 9	0.065 (0.007)	0.335 (0.020)	0.220 (0.010)
previous convictions 10 or more	0.067 (0.007)	0.384 (0.020)	0.245 (0.010)
Main offence			
aggravated burglary dwelling	0.335 (0.027)	0.448 (0.097)	0.434 (0.045)
aggravated burglary not in a dwelling	0.287 (0.023)	0.436 (0.080)	0.398 (0.037)
attempted theft	-0.218 (0.038)	-0.516 (0.085)	-0.367 (0.048)
commercial burglary	-0.103 (0.008)	-0.129 (0.024)	-0.128 (0.012)
conspiracy to commit burglary	0.083 (0.031)	0.151 (0.107)	0.138 (0.050)
conspiracy to commit fraud	0.107 (0.023)	0.282 (0.073)	0.198 (0.035)
conspiracy other	0.049 (0.036)	0.256 (0.126)	0.193 (0.060)
conspiracy to steal	0.074 (0.021)	0.110 (0.070)	0.137 (0.034)
dishonest representation	-0.172 (0.015)	-0.166 (0.031)	-0.273 (0.017)
going equipped	-0.220 (0.030)	-0.276 (0.072)	-0.314 (0.039)
handling stolen goods	-0.128 (0.025)	-0.188 (0.059)	-0.226 (0.032)
immigration offence	-0.097 (0.029)	0.189 (0.093)	0.069 (0.045)
money laundering	0.003 (0.020)	0.001 (0.051)	-0.051 (0.026)
other fraud offence	-0.075 (0.008)	-0.112 (0.021)	-0.102 (0.011)
other theft offence	-0.181 (0.013)	-0.358 (0.032)	-0.268 (0.017)
receiving stolen goods	-0.159 (0.011)	-0.427 (0.026)	-0.307 (0.014)
theft from a person	-0.166 (0.012)	-0.434 (0.030)	-0.241 (0.016)
theft from a shop	-0.232 (0.011)	-0.593 (0.028)	-0.340 (0.015)
theft in breach of trust	-0.091 (0.012)	-0.045 (0.029)	-0.141 (0.015)
theft of a vehicle	-0.166 (0.035)	-0.260 (0.086)	-0.287 (0.045)
false passport	-0.169 (0.012)	0.063 (0.036)	-0.025 (0.018)
Observations	4,220	7,242	7,242

*Reference categories: female offender, no guilty plea entered,
no previous convictions, main offence burglary in a dwelling