ESRC Research Methods Festival

Tackling Selection Bias in Sentence Data Analysis

Jose Pina-Sánchez (University of Leeds)
Sara Geneletti (London School of Economics)
John Paul Gosling (University of Leeds)
Marco Dorreti (University of Perugia)
Amber Isaac (Sentencing Council)

Sentencing Studies

- The concept of *punishment* is central to Criminology and Criminal Justice
- Sentencing is how this concept is best exemplified
- Analysing sentence data we can explore lots of important research questions

Sentencing Studies

- We can look into the decision-making process of judges
  - e.g. identify the case characteristics considered by judges when passing sentences

Sentencing Studies

- We can look into the decision-making process of judges
  - e.g. identify the case characteristics considered by judges when passing sentences
- Assess whether the sentencing guidelines are being followed
  - e.g. the guidelines indicate that drug/alcohol addictions should be considered as an aggravating factor

Sentencing Studies

- We can look into the decision-making process of judges
  - e.g. identify the case characteristics considered by judges when passing sentences
- Assess whether the sentencing guidelines are being followed
  - e.g. the guidelines indicate that drug/alcohol addictions should be considered as an aggravating factor
- Detect unwarranted disparities in the system
  - the concept of *sentencing postcode lottery*
  - systematic disparities in the form of discrimination

Sentencing Studies

- We can look into the decision-making process of judges
  - e.g. identify the case characteristics considered by judges when passing sentences
- Assess whether the sentencing guidelines are being followed
  - e.g. the guidelines indicate that drug/alcohol addictions should be considered as an aggravating factor
- Detect unwarranted disparities in the system
  - the concept of *sentencing postcode lottery*
  - systematic disparities in the form of discrimination
- Study key concepts such as *penal populism*, *deterrence*, etc.

The Problem

- Five main sentence outcomes (aka disposal types)
  - *discharge* < *fine* < *community order* < *suspended sentence* < *custodial sentence*
- Most of those disposal types use different units of measurement
  - e.g. pounds for fines, days for custodial sentences, conditions for community orders

The Problem

- Five main sentence outcomes (aka disposal types)
  - *discharge* < *fine* < *community order* < *suspended sentence* < *custodial sentence*
- Most of those disposal types use different units of measurement
  - e.g. pounds for fines, days for custodial sentences, conditions for community orders
- For reasons of convenience we tend to focus on custodial sentences
  - However these represent only 7% of the sentences imposed in England and Wales
  - Creating a massive problem of selection bias
- Alternatively some studies focus on the probability of custody
  - This involves reducing the sentence outcome to a (0,1) variable
  - A remarkable loss of information

Introduction

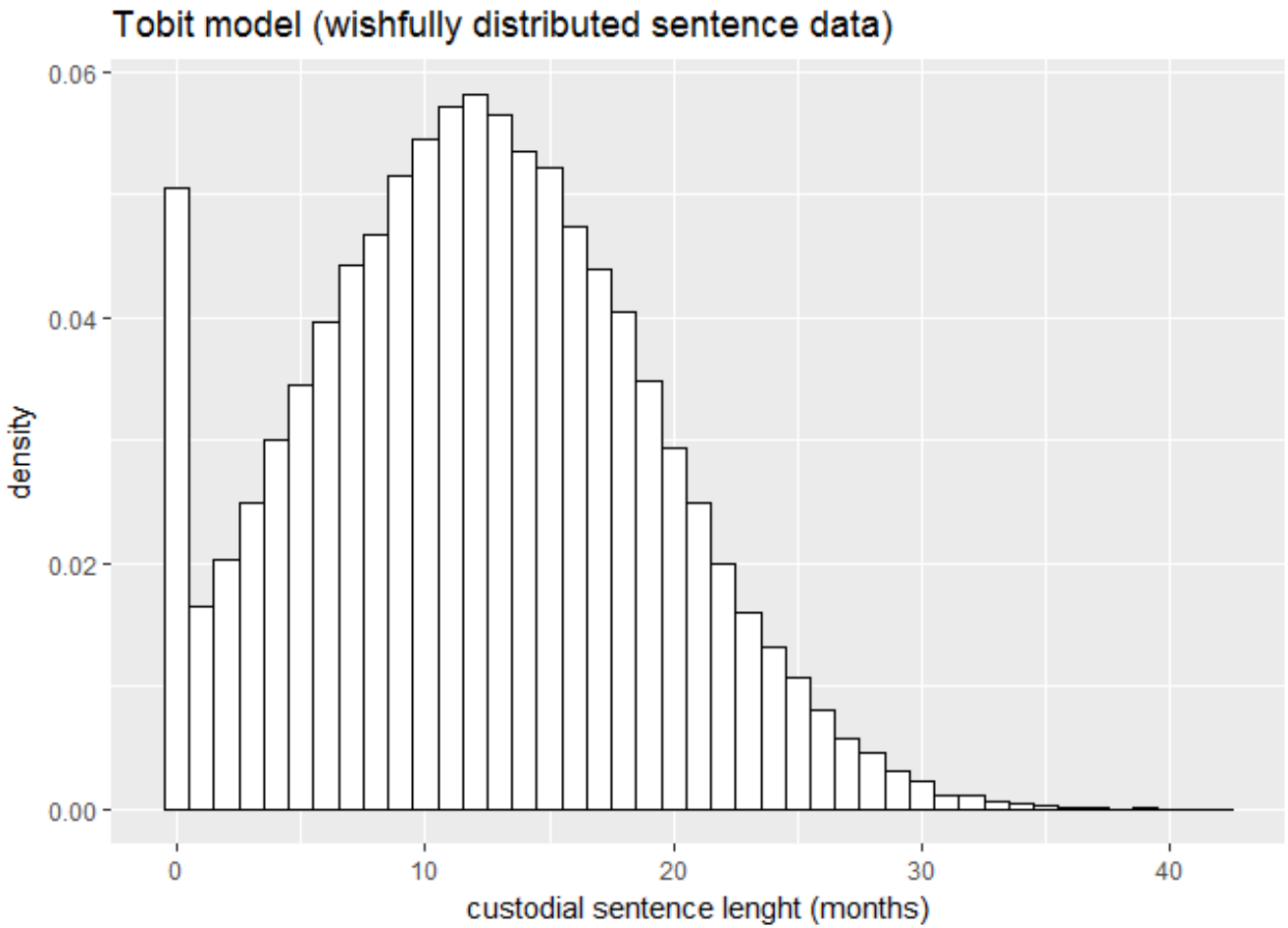
- Various statistical adjustments have been applied to tackle the problem of selection bias
  - But the assumptions upon which they are built are questionable (at least in England & Wales)
  - And keep treating non-custodial cases as a homogeneous group

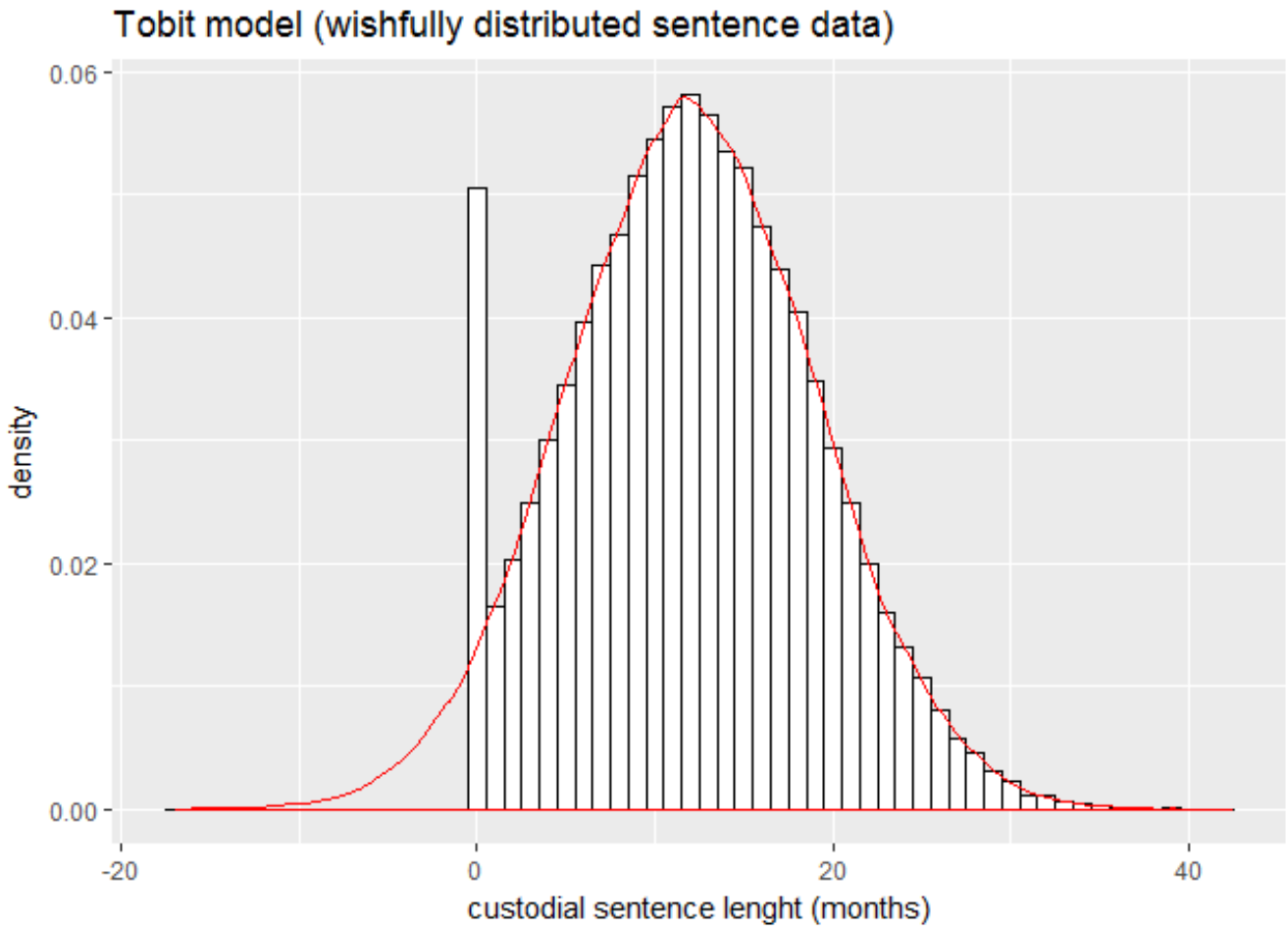
Introduction

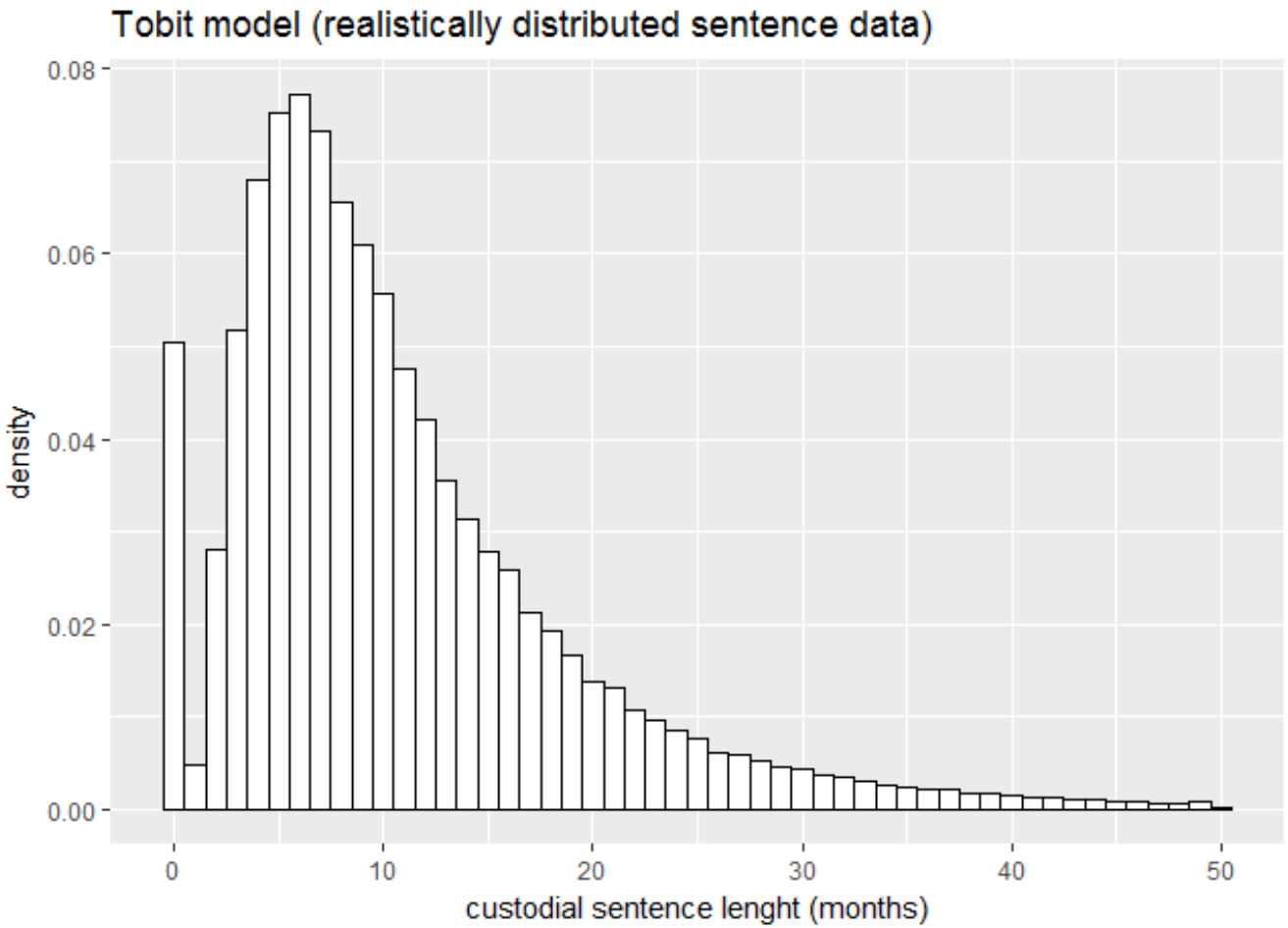
- Various statistical adjustments have been applied to tackle the problem of selection bias
  - But the assumptions upon which they are built are questionable (at least in England & Wales)
  - And keep treating non-custodial cases as a homogeneous group
- Two stage processes (Heckman selection model)
  - Assumes that sentencing is undertaken in two steps
  - Requires variables that meet the exclusion criteria

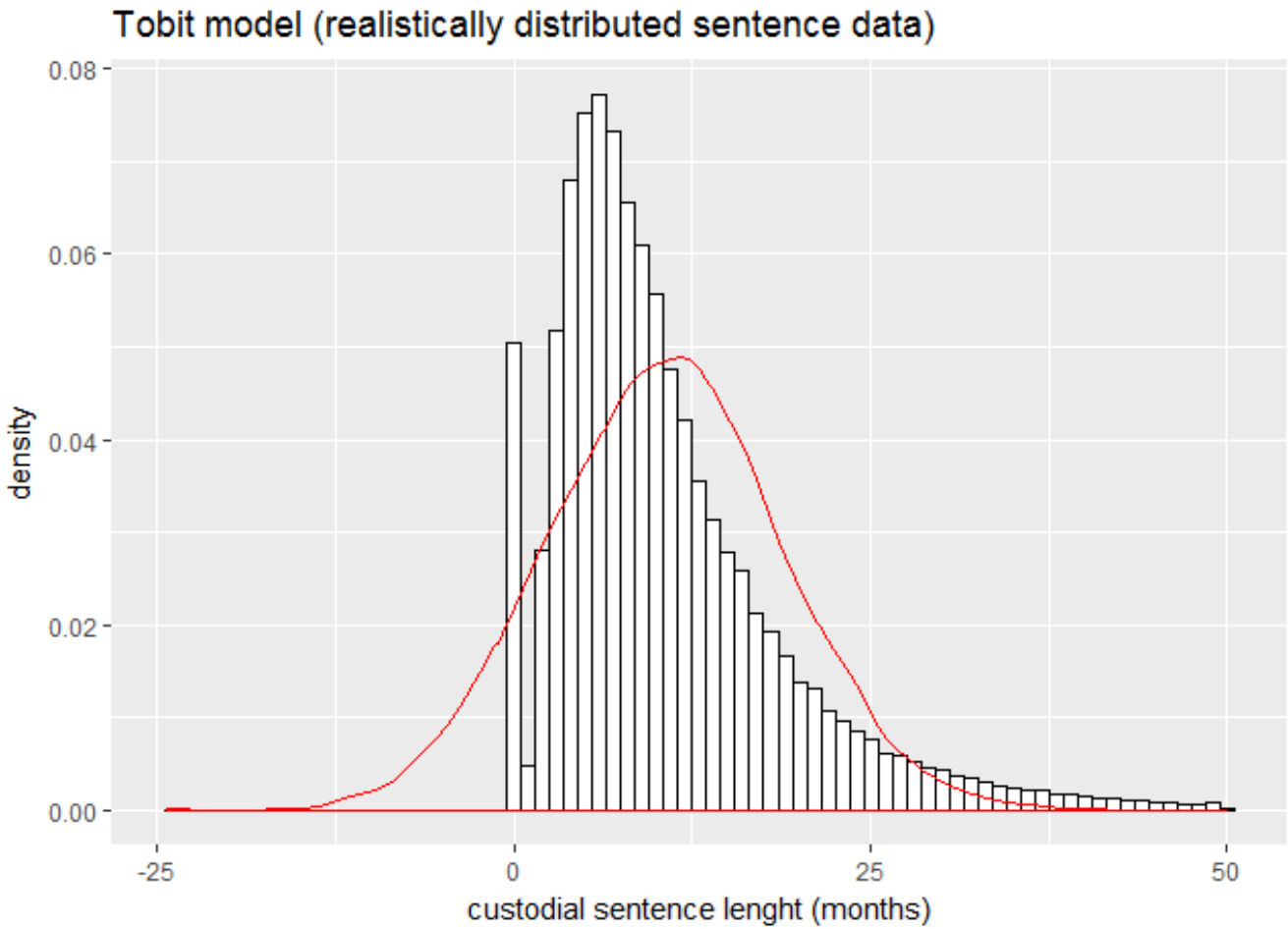
Introduction

- Various statistical adjustments have been applied to tackle the problem of selection bias
  - But the assumptions upon which they are built are questionable (at least in England & Wales)
  - And keep treating non-custodial cases as a homogeneous group
- Two stage processes (Heckman selection model)
  - Assumes that sentencing is undertaken in two steps
  - Requires variables that meet the exclusion criteria
- Models for censored data (Tobit model)
  - Assumes that sentencing is a one-step decision process
  - Assumes that non-custodial sentences are part of the same distribution (normal) as custodial durations









Introduction

- We suggest alternative approaches based on the estimation of a scale of severity
  - Advocated in the 80s (Buchner, 1979; Erickson and Gibbs, 1979; Sebba, 1980; Sebba and Nathan, 1984)
  - Strangely abandoned since then (a few exceptions; Tremblay, 2016)
  - Recently picked up by the Sentencing Council for England and Wales

Introduction

- We suggest alternative approaches based on the estimation of a scale of severity
  - Advocated in the 80s (Buchner, 1979; Erickson and Gibbs, 1979; Sebba, 1980; Sebba and Nathan, 1984)
  - Strangely abandoned since then (a few exceptions; Tremblay, 2016)
  - Recently picked up by the Sentencing Council for England and Wales
- key benefit: the analysis of 100% of the offences, while making the most of the information available
  - MoJ data captures disposal types, and durations of suspended and custodial sentences

Introduction

- We suggest alternative approaches based on the estimation of a scale of severity
  - Advocated in the 80s (Buchner, 1979; Erickson and Gibbs, 1979; Sebba, 1980; Sebba and Nathan, 1984)
  - Strangely abandoned since then (a few exceptions; Tremblay, 2016)
  - Recently picked up by the Sentencing Council for England and Wales
- key benefit: the analysis of 100% of the offences, while making the most of the information available
  - MoJ data captures disposal types, and durations of suspended and custodial sentences
- key challenge: to estimate the relative severity of different sentence outcomes

Our approaches

- We used a two-pronged approach to measuring severity

1 We used the scale the council developed (Application 1) which is based on the custodial sentences only

- and extrapolated it to the non-custodial sentences

Our approaches

- We used a two-pronged approach to measuring severity

1 We used the scale the council developed (Application 1) which is based on the custodial sentences only

- and extrapolated it to the non-custodial sentences

2 We developed a Thurstone paired comparison scale (Applications 2 and 3) which is based on comparing the non-custodial and low valued custodial sentences

- and extrapolated it to the higher valued custodial senctences

Our approaches

- We used a two-pronged approach to measuring severity

1 We used the scale the council developed (Application 1) which is based on the custodial sentences only

- and extrapolated it to the non-custodial sentences

2 We developed a Thurstone paired comparison scale (Applications 2 and 3) which is based on comparing the non-custodial and low valued custodial sentences

- and extrapolated it to the higher valued custodial senctences
- The approaches are complementary in the sense that they give us full information about one type of sentence outcome
- In the future we'd like to combine them

The Council's Scale

- The Sentencing Council (2015) has created their own scale to evaluate the impact of their sentencing guidelines
- Based on the starting points for different levels of seriousness encoded in the guidelines

The Council's Scale: Rationale

Grievous Bodily Harm with Intent

Offence Category	**Starting Point** *(Applicable to all offenders)*
Category 1	12 years' custody
Category 2	6 years' custody
Category 3	4 years' custody

The Council's Scale: Rationale

Grievous Bodily Harm

Offence Category	**Starting Point** *(Applicable to all offenders)*
Category 1	3 years' custody
Category 2	1 year 6 months' custody
Category 3	High level community order

The Council's Scale: Formula

$$y = (28.8 + (\log((\text{csl}) + 8.9) - \log(8.9)) \times (60.71))$$

- Where csl is the sentence length in years
- This score takes into account the law of diminshing returns (or in this case the law of diminishing severity)
- The difference between one and two years in custody is larger than the difference between 10 and 11 years.

The Council's Scale

- The minimum severity y for custodial outcome is 29.4 for a sentence length of 28 days as this is the minimum recommended for custodial sentences
- This is OK for custodial outcomes (although somewhat arbitrary) but cannot be calculated for non-custodial outcomes
- The council assumed fixed values for each category of non-custodial outcomes

The Council's Scale: Pros and Cons

- It tries to take into account the law of diminishing returns

The Council's Scale: Pros and Cons

- It tries to take into account the law of diminishing returns
- It assumes that the jumps in seriousness have an equivalent increase in severity across different offences
- To anchor the function of severity an arbitrary choice is made for 1-year in custody
- Doesn't consider suspended sentences (although these can be included with some additional considerations)

Application 1: Constraints and Anchors in a Bayesian Framework

- We take the severity score developed by the council for the custodial sentences and extrapolate to the non-custodial sentences
- The problem with doing that is that we have NO severity data for the non-custodial sentences
- We do three things to overcome this:
  - We assume that the relationship between covariates and severity is the same for custodial and non-custodial sentences

Application 1: Constraints and Anchors in a Bayesian Framework

- We take the severity score developed by the council for the custodial sentences and extrapolate to the non-custodial sentences
- The problem with doing that is that we have NO severity data for the non-custodial sentences
- We do three things to overcome this:
  - We assume that the relationship between covariates and severity is the same for custodial and non-custodial sentences
  - We use constraints built into the Bayesian MCMC machinery to "disallow" incorrectly imputed values for the non-custoial sentences

Application 1: Constraints and Anchors in a Bayesian Framework

- We take the severity score developed by the council for the custodial sentences and extrapolate to the non-custodial sentences
- The problem with doing that is that we have NO severity data for the non-custodial sentences
- We do three things to overcome this:
  - We assume that the relationship between covariates and severity is the same for custodial and non-custodial sentences
  - We use constraints built into the Bayesian MCMC machinery to "disallow" incorrectly imputed values for the non-custoial sentences
  - We use "anchors" a small number of fake data points (cheating)

Our approach

- How can we use the rich covariate information data to say something about the non-custodial data
- What assumptions do we need to make?
- Can we build a regression model that learns in a simple way how to assign a severity to non-custodial outcomes?
- The severity for non-custodial outcomes is entirely unidentified
- Can we still use methods from the missing data literature?
- Can we use the fact that we know what the sentence outcome is?

Ad-hoc frequentist model

- To start us off:
- We ran an initial regression of severity on the covariates (except sentence outcome) for the custodial outcomes.
- The coefficients of the covariates made sense and the fit was very good. We called this model M1.

Ad-hoc frequentist model

- To start us off:
- We ran an initial regression of severity on the covariates (except sentence outcome) for the custodial outcomes.
- The coefficients of the covariates made sense and the fit was very good. We called this model M1.
- Using M1 we predicted the severities for the non-custodial outcomes.
- As expected these predictions were not good
- i.e. they were too high and overlapped with custodial sentence severity too much

Ad-hoc frequentist model

- To start us off:
- We ran an initial regression of severity on the covariates (except sentence outcome) for the custodial outcomes.
- The coefficients of the covariates made sense and the fit was very good. We called this model M1.
- Using M1 we predicted the severities for the non-custodial outcomes.
- As expected these predictions were not good
- i.e. they were too high and overlapped with custodial sentence severity too much
- We ran a second regression where severity for the non-custodial outcomes was the severity predicted using M1
- We added sentence outcome to the predictors. We named this M2

M2

- M2 fit the data (+ predicted from M1) well
- The coefficients for the sentence outcome were based entirely on the data predicted in M1

M2

- M2 fit the data (+ predicted from M1) well
- The coefficients for the sentence outcome were based entirely on the data predicted in M1
- They were negative which makes sense as non-custodial sentences are less severe than custodial sentences
- This means that the covariates have quite a lot of information on severity

Algorithm

Starting from M2 we ran the follwing algorithm

1. While the error rate was above 6%
2. change the value of the non-custodial outcomes which were outside of their pre-defined range to a value within the range using a set of rules (see below)
3. Re-run the regression

Algorithm

Starting from M2 we ran the follwing algorithm

1. While the error rate was above 6%
2. change the value of the non-custodial outcomes which were outside of their pre-defined range to a value within the range using a set of rules (see below)
3. Re-run the regression
   a) If the root mean square error for the regression was too far from the RMSE of M2 then we repeated step 2).
   – The motivation for this was to ensure that the regression would remain good at predicting the custodial outcomes.

Algorithm

Starting from M2 we ran the follwing algorithm

1. While the error rate was above 6%
2. change the value of the non-custodial outcomes which were outside of their pre-defined range to a value within the range using a set of rules (see below)
3. Re-run the regression
   a) If the root mean square error for the regression was too far from the RMSE of M2 then we repeated step 2).
   - The motivation for this was to ensure that the regression would remain good at predicting the custodial outcomes.
   - Once an acceptable RMSE was obtained we generated the error rate and started the process from 1) again.
4. Repeat until all criteria are fulfilled. Call the final regression M3

The rules

- Choose hard limits for the values of the severity of each non-custodial sentence type
- Discharge: -100-5, Fine: 5-15, Community Order: 15-29.4
- Aside: Discharge should always be 0 but this makes the model struggle to fit
  1. If the predicted score is within the corresponding range do nothing

The rules

- Choose hard limits for the values of the severity of each non-custodial sentence type
- Discharge: -100-5, Fine: 5-15, Community Order: 15-29.4
- Aside: Discharge should always be 0 but this makes the model struggle to fit
  1. If the predicted score is within the corresponding range do nothing
  2. If the score is below the range then assign to it a score that is close to the lower limit (the further away the closer to the limit)

The rules

- Choose hard limits for the values of the severity of each non-custodial sentence type
- Discharge: -100-5, Fine: 5-15, Community Order: 15-29.4
- Aside: Discharge should always be 0 but this makes the model struggle to fit
  1. If the predicted score is within the corresponding range do nothing
  2. If the score is below the range then assign to it a score that is close to the lower limit (the further away the closer to the limit)
  3. If the score is above the range then assign to it a score that is close to the upper limit (the further away the closer to the limit)

M3

- M3 fits the data **ridiculously** well ($R^2 = 1$!!)
- But of course it's not an acceptable solution.

M3

- M3 fits the data **ridiculously** well ($R^2 = 1$!!)
- But of course it's not an acceptable solution.
item However it does help to highlight some valuable points to identify the severity for non-custodial outcomes:
  1. We need to assume that the relationship between the covariates and the severity stays the same (or is modified in a deterministic way)
  2. A good model should be able to predict the custodial outcomes well too
  3. The constraints are useful for training/learning
- We tried one type of constraint in the Ad-hoc approach but there are other ways

Bayesian modelling

- We adopt a Bayesian approach.
- The idea is similar except the training happens within each MCMC run

Bayesian modelling

- We adopt a Bayesian approach.
- The idea is similar except the training happens within each MCMC run
- Because in an MCMC run each parameter is sampled conditional on the values of the other parameters
- If we don't allow certain values this affects the remaining parameters

Bayesian modelling

- We adopt a Bayesian approach.
- The idea is similar except the training happens within each MCMC run
- Because in an MCMC run each parameter is sampled conditional on the values of the other parameters
- If we don't allow certain values this affects the remaining parameters
- This is not like the Frequentist approach where data are re-used
- We use JAGS, Burnin of 5,000, run for 10,000 and update 1000
- Convergence is OK but not great for the parameters associated with the Sentence Outcome and for the imputed values

Basic model

Model

$$y_i \sim \mathrm{N}(\mu_i, \tau)$$
$$\mu_i = \alpha + \mathrm{X}_i\boldsymbol{\beta}$$

Priors

$$\alpha \sim \mathrm{N}(30, 0.001)$$
$$\beta_k \sim \mathrm{N}(30, 0.001)$$
$$\tau \sim \Gamma(1 \times 10^{-3}, 1 \times 10^{-3})$$

- y the severity for the custodial sentences
- $\mathbf{X}$ the design matrix
- The model then automatically imputes the severity for the non-custodial outcomes as these are passed as missing to JAGS

Posterior means for Basic model

	$\hat{\beta}$	se($\hat{\beta}$)
Intercept	32.55	1.37
start.CO	9.76	1.32
start.<1	14.51	1.41
Start.<3	18.83	1.41
Start.<3	34.92	1.37
SO.Dis	-29.67 (-29.5)	31.55
SO.Fine	-19.76 (-16.6)	31.47
SO.CO	-2.55 (-13.3)	31.65
TAMF	0.50	0.07
RGP.ND	0.49	1.04
RGP.<20%	-0.39	0.48
RGP.>21%	-2.80	0.42
LOS.2	-10.08	0.55
LOS.3	-12.05	0.64
PC.1-3	-0.27	0.59
PC.None	-1.51	0.40

Posterior means for Basic model

- The parameters values for Discharge and Fine are similar to those in M2
- The parameter for Community Order is very different
- The variances are very large – this makes sense because the model is taking into account that there is no data to estimate these parameters

Non custodial correction model

Model

$$\begin{aligned} \mathrm{y}_i &\sim \mathrm{N}(\mu_i, \tau) \\ \mu_i &= \alpha + \mathrm{X}_i\boldsymbol{\beta} && \text{for } i \in \{1, ..., n\} \\ \mathrm{y}_i &\sim \mathrm{N}(\mu_i^*, \tau) \\ \mu_i^* &= f(\mathrm{lp}_j, \mathrm{ll}_j, \mathrm{ul}_j) && \text{for } j \in \{n+1, ..., N\} \end{aligned}$$

Function f

$$\begin{aligned} f(\mathrm{lp}_i, \mathrm{ll}_i, \mathrm{ul}_i) = {} & \mathrm{I}(\mathrm{ll}_i < \mathrm{lp}_i < \mathrm{ul}_i)\mathrm{lp}_i && \text{if within range leave estimate} \\ & + \{1 - \mathrm{I}(\mathrm{ll}_i < \mathrm{lp}_i < \mathrm{ul}_i)\} && \text{if not in range then..} \\ & \times [\{\mathrm{I}(\mathrm{ul}_i < \mathrm{lp}_i)U_i\} && \text{if above upper limit assign } U_i \\ & + \{\mathrm{I}(\mathrm{ll}_i > \mathrm{lp}_i)L_i]\}] && \text{if below lower limit assign } L_i \end{aligned}$$

- $lp_j = \alpha_z + X_j\beta_{\mathbf{z}}$
- ul_j is the vector or upper limits of the range
- ll_j is the vector of lower limits of the range

L_j and U_j

- If the imputed severity y is not in the correct range then it is assigned a value as follows ($\delta = 10$):
  - If it's above the limit then:

$$U_j = \text{ul}_j - \frac{\text{lp}_j - \text{ul}_j}{\delta}. \quad (1)$$

  - If it's below its lower limit then:

$$L_j = \frac{\text{ll}_j - \text{lp}_j}{\delta} + \text{ll}_j \quad (2)$$

- The idea then is that the further away from the limit an estimated value is, the closer to the edge of the range it should be set to
- This takes into account to some extent the contribution of the covariates
- The limits for discharge aren't 0 and 0 because this hampered the model too much

Posterior parameter means for NC correction

	basic		NC correction	
	$\hat{\beta}$	se($\hat{\beta}$)	$\hat{\beta}$	se($\hat{\beta}$)
Intercept	32.55	1.37	32.65	1.4
start.CO	9.76	1.32	9.72	1.34
start.<1	14.51	1.41	14.47	1.5
start.<3	18.83	1.41	18.84	1.48
start.>3	34.92	1.37	34.88	1.44
SO.Dis	-29.67	31.55	-30.42	31.8
SO.Fine	-19.76	31.47	-20.49	30.36
SO.CO	-2.55	31.65	-7.21	32.54
TAMF	0.5	0.07	0.5	0.08
RGP.ND	0.49	1.04	0.44	1.09
RGP.<20%	-0.39	0.48	-0.41	0.48
RGP.>21%	-2.8	0.42	-2.85	0.39
LOS.2	-10.08	0.55	-10.08	0.56
LOS.3	-12.05	0.64	-12.07	0.69
PC.1-3	-0.27	0.59	-0.38	0.58
PC.None	-1.51	0.4	-1.55	0.42

Selection model with rules

Model for outcome

$$
\begin{aligned}
y_i &\sim \mathrm{N}(\mu_i, \tau) \\
\mu_i &= \alpha + \mathrm{X}_i\boldsymbol{\beta} \text{ for } i \in \{1, ..., n\} \\
y_i &\sim \mathrm{N}(\mu_i^*, \tau) \\
\mu_i^* &= f(\mathrm{lp}_j, \mathrm{ll}_j, \mathrm{ul}_j) \text{ for } j \in \{n+1, ..., N\}
\end{aligned}
$$

Model of missingness

$$
\begin{aligned}
\kappa_i &= \gamma_0 + \mathrm{X}_i\boldsymbol{\gamma} + \gamma_1 y_i \\
\mathrm{logit}(p_i) &= \kappa_i \\
R_i &\sim Bern(p_i)
\end{aligned}
$$

- This is a standard expression for a selection model where missingness is not at random
- We currently have an issue with γ_1. JAGS will only run if the distribution of γ_1 has positive support

Tables of correctly predicted

- Severity is imputed correctly for the mean and median in all models
- The table below shows what happens in the 2.5, 25, 75 and 97.5 percentiles in one set of runs. The picture is similar over other runs except for the selection model
- In the 2.5 % all of them correctly impute Discharge but none of the others
- The selection models do better in the 97.5% than the others but worse in the 25%
- In other runs the selection model does better in the 2.5% and worse in the 75%

%ile	basic	NC correction	sel rules
2.5	8.2	8.2	8.2
25	100	100	18.6
75	100	100	99.7
97.5	0	0	55.7

Some comments

- Overall the results aren't very good in terms of variability
- And where they are better they are very sensitive to prior information
- However the posterior means for the parameters are pretty consistent across runs
- If we use the mean imputed severities as "predicted" severity we might have problems related to ordering
- This means that one sentence of custodial outcome might be more severe in one run than in another
- This is not good if the aim is comparison
- However if we take the mean parameter values and "plug in" the values for the covariates we obtain far more stable severities

Anchors

- The models we have developed so far have very variable results
- Essentially we cannot get something for (almost) nothing: i.e. the constraints are not sufficiently informative to get severities that are stable
- In a final bid to get something more out of this approach we decided to introduce anchors
- The idea is that if we believe that the value for discharge should be 0 then pretend e.g. that the severity for one of the discharges (one that sits in the middle if predicted using the frequentist model) is 0 – we have data
- Similarly, if we think that the median value for community order is 20 then we invent yet another data point
- We created 3 - 6 data points like this
- Then we ran the Selection model again

Further work

- Using the the Thurstone scale to determine the value of the hard limits – currently these are somewhat arbitrary
- Combining the two approaches we have outlined would seem like the next step
- The Thurstone score tells us all about the non-custodial outcomes
- The sentence lenght (or a function thereof) tells us all about the custodial sentences
- Extrapolating in either direction depends on many assumptions
- Some clever re-scaling should make them compatible

Methods Used in the Literature

- Four main methods have been used:
  - Direct ratings (Hindelang, et al., 1975)
  - Magnitude escalation (Leclerc and Tremblay, 2016)
  - Thurstone pair-comparisons (Buchner, 1979)
  - Canonical correlation / correspondence analysis (Francis et al., 2005)

Methods Used in the Literature

- Four main methods have been used:
  - Direct ratings (Hindelang, et al., 1975)
    arbitrary; unreplicable
  - Magnitude escalation (Leclerc and Tremblay, 2016)
  - Thurstone pair-comparisons (Buchner, 1979)
  - Canonical correlation / correspondence analysis (Francis et al., 2005)
    assumes perfect linear correlation between crime seriousness and sentence severity; generates nonsensical values

Methods Used in the Literature

- Four main methods have been used:
  - Direct ratings (Hindelang, et al., 1975)
    arbitrary; unreplicable
  - Magnitude escalation (Leclerc and Tremblay, 2016)
    assumes numeracy of subjects; vast variability in responses
  - Thurstone pair-comparisons (Buchner, 1979)
    no variability across certain comparisons
  - Canonical correlation / correspondence analysis (Francis et al., 2005)
    assumes perfect linear correlation between crime seriousness and sentence severity; generates nonsensical values

Exercise: Piloting Severity Questions

How much more severe is a **2-years custodial sentence** than a **1-year custodial sentence**?

- ○ 1 (equally severe)
- ○ 2 (twice as severe)
- ○ 5 (five times more severe)
- ○ 10 (ten times more severe)
- ○ I cannot answer this question
- ○ Other

a. If you selected Other, please specify:

Exercise: Piloting Severity Questions

Which is more severe a **fine band-C (150% weekly income)** or a **low community order (60 hours of unpaid work)**?

- ○ fine band-C (150% weekly income)
- ○ low community order (60 hours of unpaid work)

Thurstone Approach

- Thurstone model and a sample of 21 magistrates
  - Rather than asking to compare pairs of sentences
  - We ask how often a particular disposal type can be more punitive than other
  - This gives us a matrix of severity similar to that obtained through pair comparisons
- The questionnaire includes eleven sentence outcomes
  - Not all combinations of pairs were included
  - Only those where an overlap in the level of severity is expected
  - e.g. high community orders attaching multiple and long requirements can be harsher than suspended sentences with no onerous conditions attached

Thurstone Approach: Question Format

How often can a **fine** be more punitive than a **community order**? (Please consider all possible combinations of amounts that can be raised by a fine, conditions that can be attached to a community order, and circumstances of the offender)

- ○ Never (0% of times)
- ○ As often (50% of times)
- ○ Always (100% of times)
- ○ I cannot answer this question
- ○ Other

a. If you selected Other, please specify:

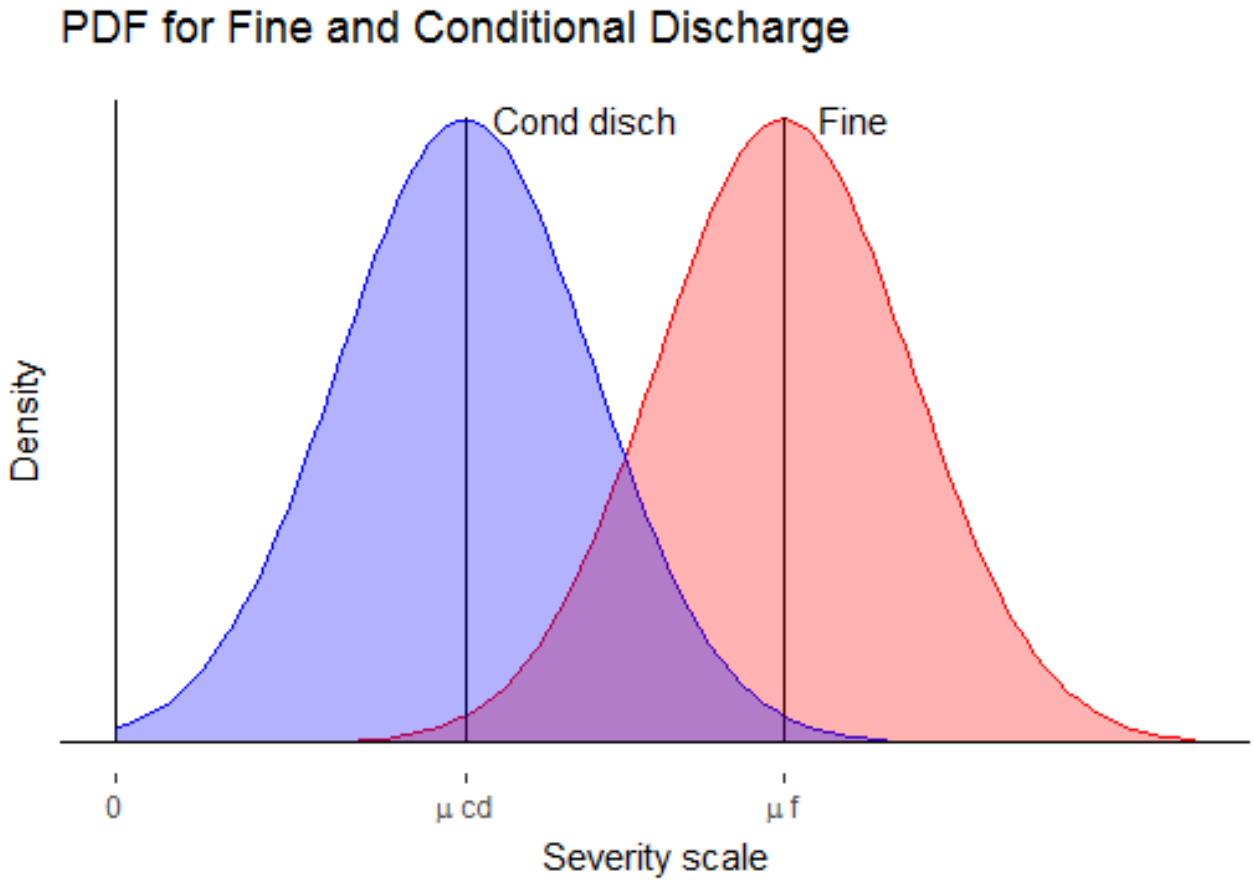
Matrix of Severity

	absolute discharge	conditional discharge	fine	community order	6month suspended 1month custody	12month suspended 1month custody	6month suspended 6month custody	24month suspended 12month custody	1month inmediate custody	2month inmediate custody	3month inmediate 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	0.78	1	1	1	1	1	1	1
community order	0	0	0.22	0.5	0.37	1	1	1	1	1	1
6month susp 1month cust	0	0	0	0.63	0.5	1	1	1	1	1	1
12month susp 1month cust	0	0	0	0	0	0.5	0.73	1	1	1	1
6month susp 6month cust	0	0	0	0	0	0.27	0.5	1	1	1	1
24month susp 12month cust	0	0	0	0	0	0	0	0.5	0.41	0.52	0.62
1month inmediate custody	0	0	0	0	0	0	0	0.59	0.5	1	1
2month inmediate custody	0	0	0	0	0	0	0	0.48	0	0.5	1
3month inmediate custody	0	0	0	0	0	0	0	0.38	0	0	0.5

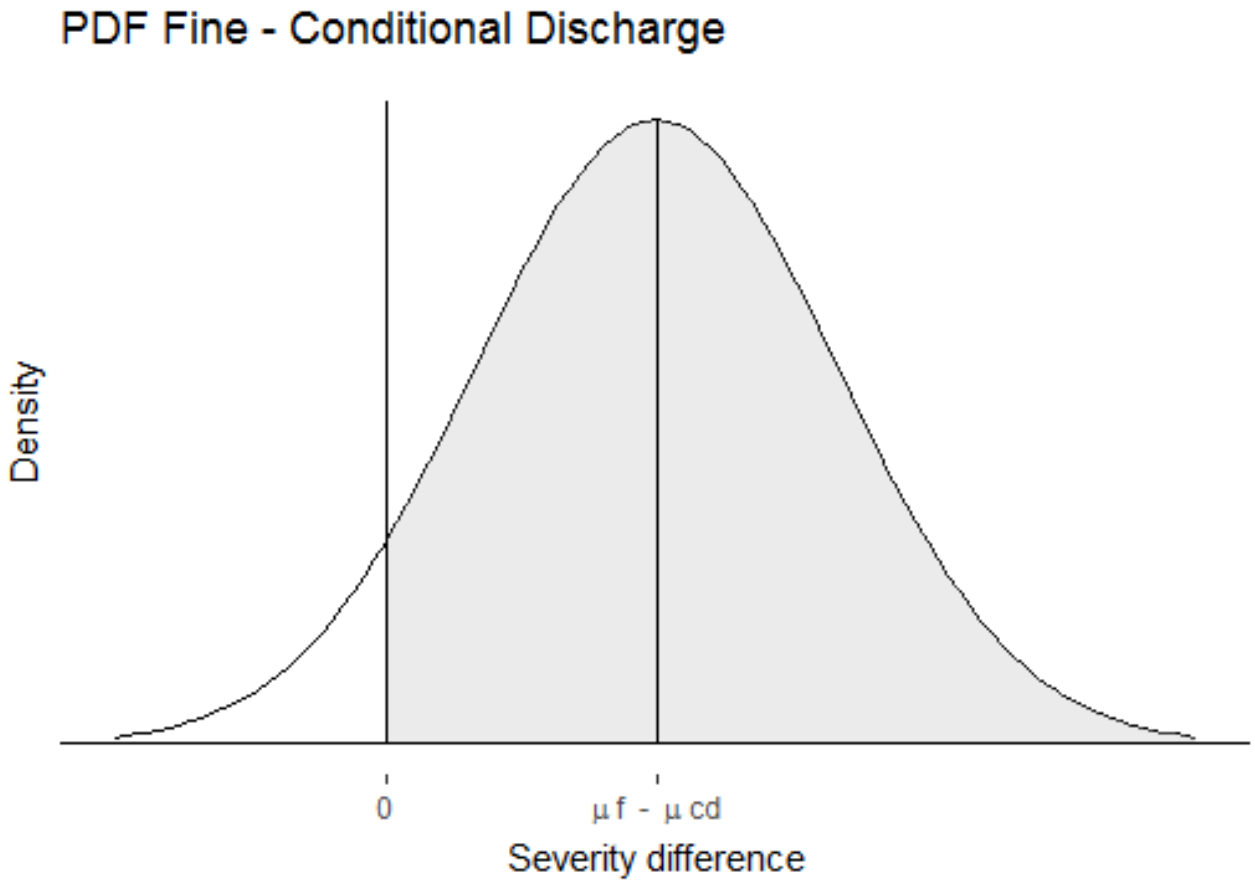
Thurstone Model: Intuitively

- We use the Thurstone-Mosteller model (Type V) to convert the proportions from pairwise comparisons into a severity scale
- Based on latent normal distributions for each sentence outcome included
- Each of those normal distributions will have its own mean, μ_s, and identical variance
- The amount of overlap between the distributions determines their closeness on the severity scale, i.e. their severity score (μ_s)

Thurstone Model: Visually



Thurstone Model: Visually



Severity Scores

Sentence outcome	Severity score
absolute discharge	0
conditional discharge	0.97
fine	1.33
community order	2.13
1-month custody 6-months suspended	2.34
1-month custody 12-months suspended	3.66
6-months custody 6-months suspended	3.78
12-months custody 24-months suspended	5.74
1-month custody	5.05
2-months custody	5.75
3-months custody	6.45
12-months custody	
5-years custody	
20-years custody	

Severity Scores

Sentence outcome	Severity score
absolute discharge	0
conditional discharge	0.97
fine	1.33
community order	2.13
1-month custody 6-months suspended	2.34
1-month custody 12-months suspended	3.66
6-months custody 6-months suspended	3.78
12-months custody 24-months suspended	5.74
1-month custody	5.05
2-months custody	5.75
3-months custody	6.45
12-months custody	13.45
5-years custody	47.05
20-years custody	173.05

Assumptions Invoked

- Our methodology relies on a number of assumptions
  - We assume that the severity scores for each sentence outcome are normally distributed and of equal variance
  - We assume independence across the participants and their various judgements
  - There are no diminishing returns in severity for every additional month in custody
- We are currently undertaking sensitivity analyses to assess the robustness of these assumptions
  - Exploring a Bradley−Terry model, based on logistic distributions
  - Including diminishing returns of severity based on responses to our questionnaire

Face Validity

Sentence outcome	Council's scale	Our scale
absolute discharge	0	0
conditional discharge	0	0.97
fine	10.25	1.32
community order	21.65	2.12
1-month custody 6-months suspended	-	2.34
1-month custody 12-months suspended	-	3.66
6-months custody 6-months suspended	-	3.78
12-months custody 24-months suspended	-	5.74
1-month custody	29.37	5.05
2-months custody	29.93	5.75
3-months custody	30.48	6.45
12-months custody	35.26	13.45
5-years custody	55.89	47.05
20-years custody	100	173.05

Face Validity

Sentence outcome	Council's scale	Our scale
absolute discharge	0	0
conditional discharge	0	0.97
fine	10.25	1.32
community order	21.65	2.12
1-month custody 6-months suspended	-	2.34
1-month custody 12-months suspended	-	3.66
6-months custody 6-months suspended	-	3.78
12-months custody 24-months suspended	-	5.74
1-month custody	29.37	5.05
2-months custody	29.93	5.75
3-months custody	30.48	6.45
12-months custody	35.26	13.45
5-years custody	55.89	47.05
20-years custody	100	173.05

Face Validity

Sentence outcome	Council's scale	Our scale
absolute discharge	0	0
conditional discharge	0	0.97
fine	10.25	1.32
community order	21.65	2.12
1-month custody 6-months suspended	-	2.34
1-month custody 12-months suspended	-	3.66
6-months custody 6-months suspended	-	3.78
12-months custody 24-months suspended	-	5.74
1-month custody	29.37	5.05
2-months custody	29.93	5.75
3-months custody	30.48	6.45
12-months custody	35.26	13.45
5-years custody	55.89	47.05
20-years custody	100	173.05

Face Validity

Sentence outcome	Council's scale	Our scale
absolute discharge	0	0
conditional discharge	0	0.97
fine	10.25	1.32
community order	21.65	2.12
1-month custody 6-months suspended	-	2.34
1-month custody 12-months suspended	-	3.66
6-months custody 6-months suspended	-	3.78
12-months custody 24-months suspended	-	5.74
1-month custody	29.37	5.05
2-months custody	29.93	5.75
3-months custody	30.48	6.45
12-months custody	35.26	13.45
5-years custody	55.89	47.05
20-years custody	100	173.05

Application 2: Have the Guidelines Increased Severity?

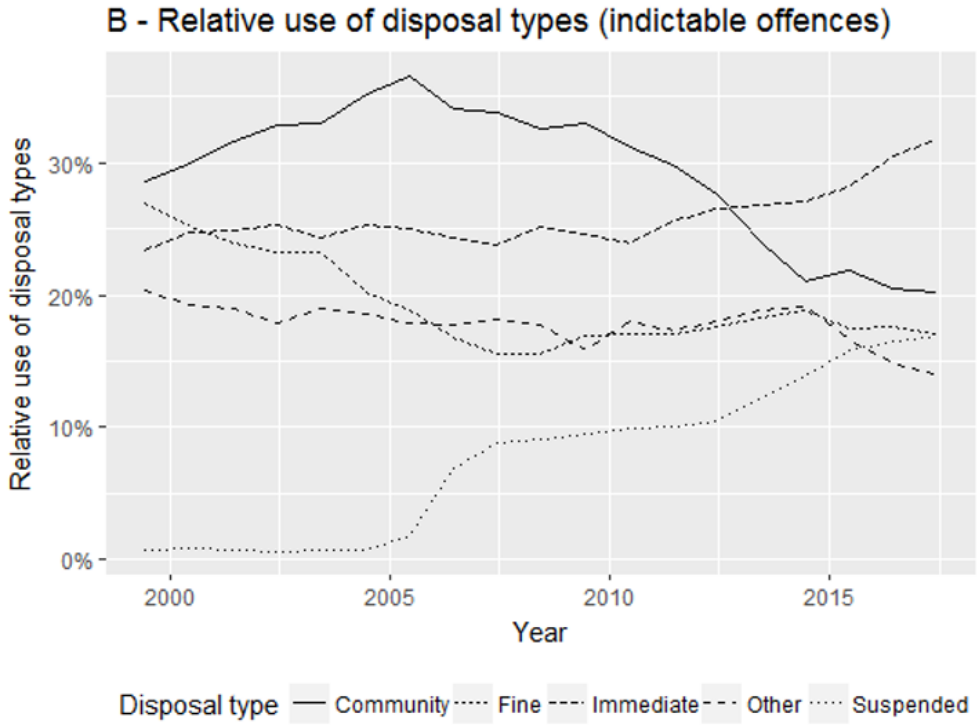
- We explore the increase in sentence severity in E&W (Roberts and Ashworth, 2016)

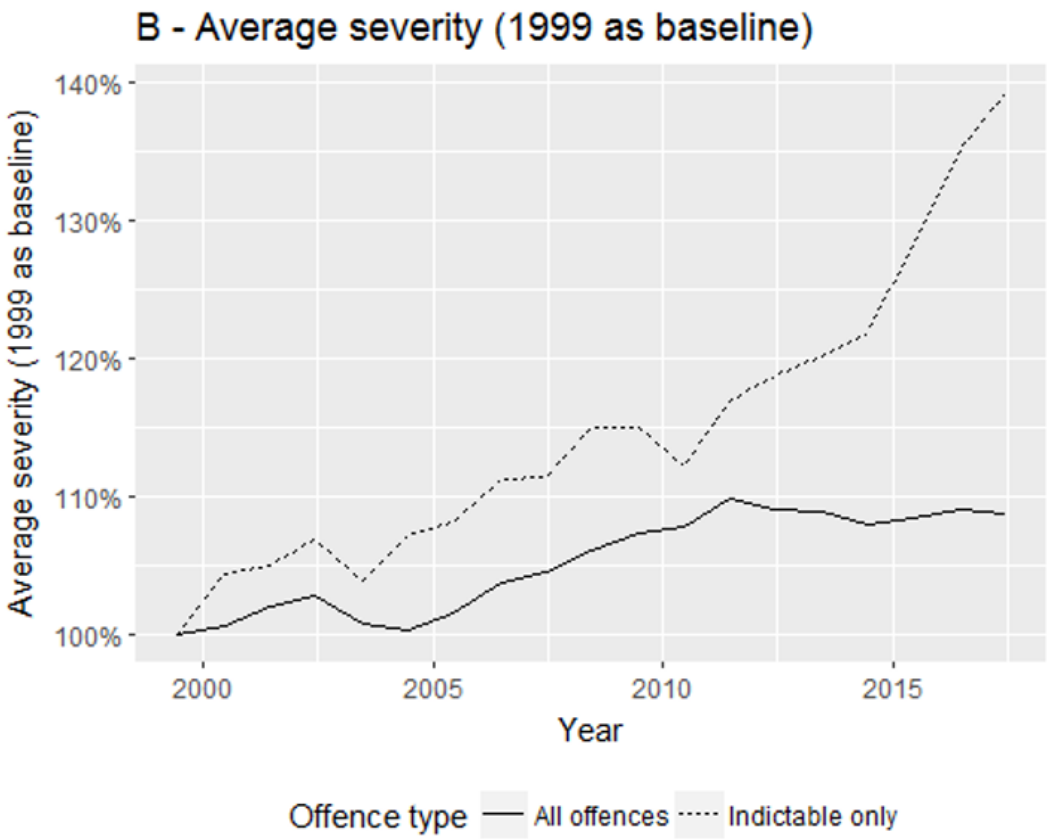
Application 2: Have the Guidelines Increased Severity?

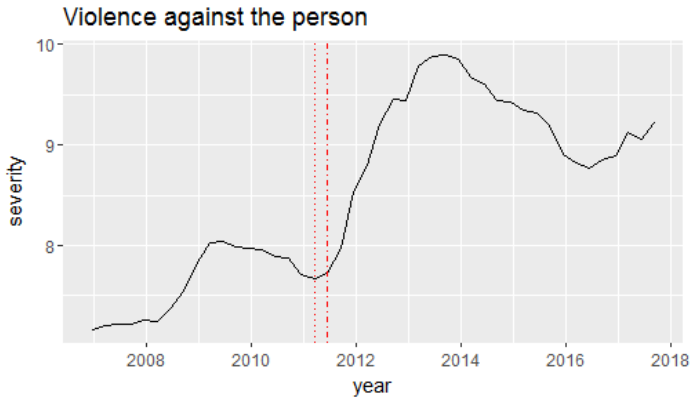
- We explore the increase in sentence severity in E&W (Roberts and Ashworth, 2016)
- Test whether the new sentencing guidelines are to be blamed (Allen, 2016)
  - Two phenomena that coincide in time
  - But it is possible that the increase in severity reflects an on-going trend

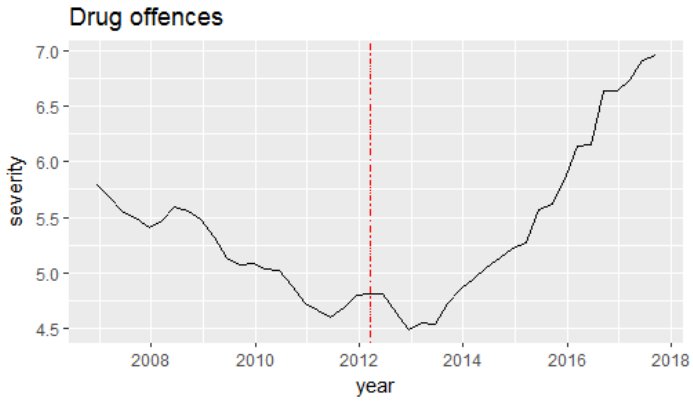
Application 2: Have the Guidelines Increased Severity?

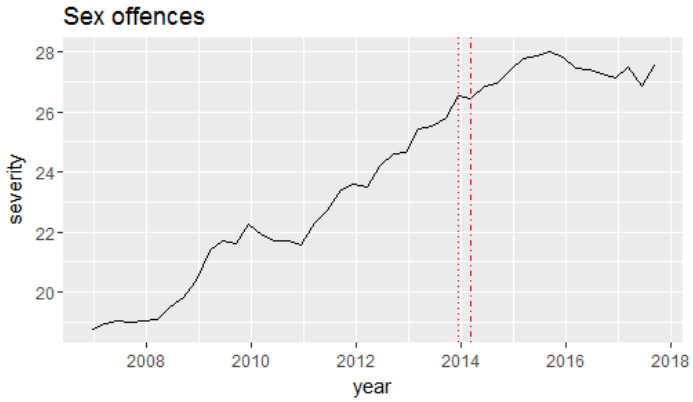
- We explore the increase in sentence severity in E&W (Roberts and Ashworth, 2016)
- Test whether the new sentencing guidelines are to be blamed (Allen, 2016)
  - Two phenomena that coincide in time
  - But it is possible that the increase in severity reflects an on-going trend
- We use descriptive statistics to represent the overall trend in severity
- And simple time series (ARIMA) models to explore whether the level of severity following the application of the guidelines falls within the expected region

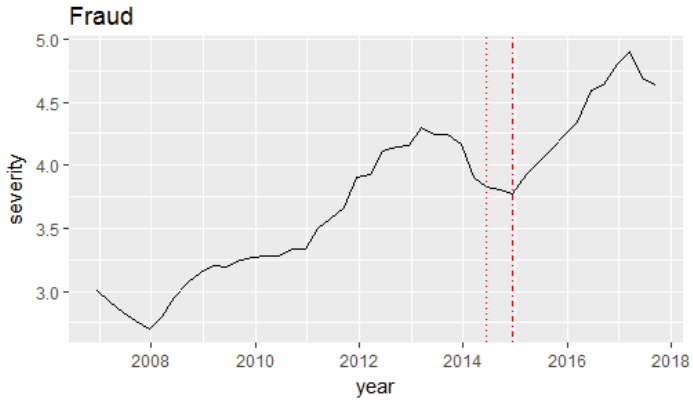


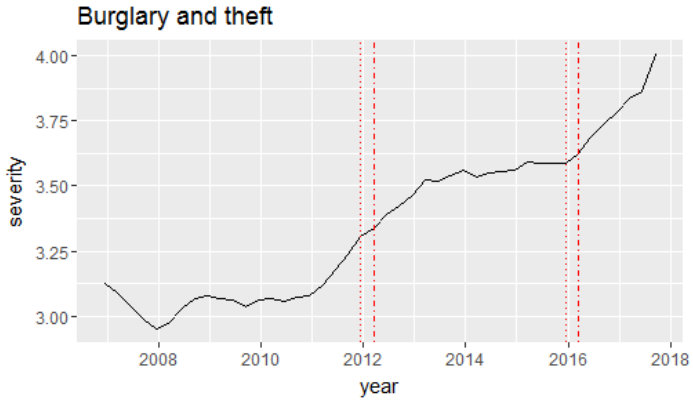


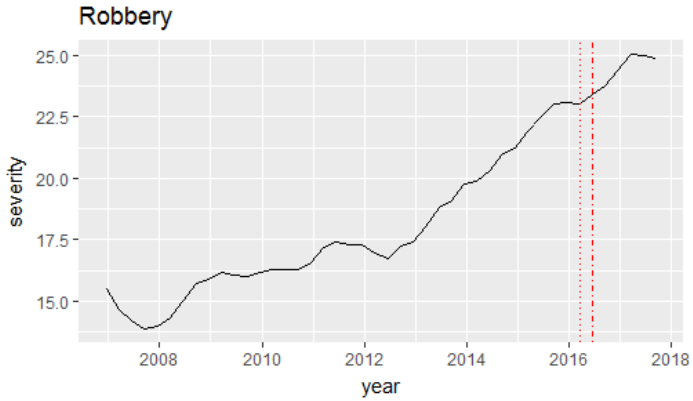


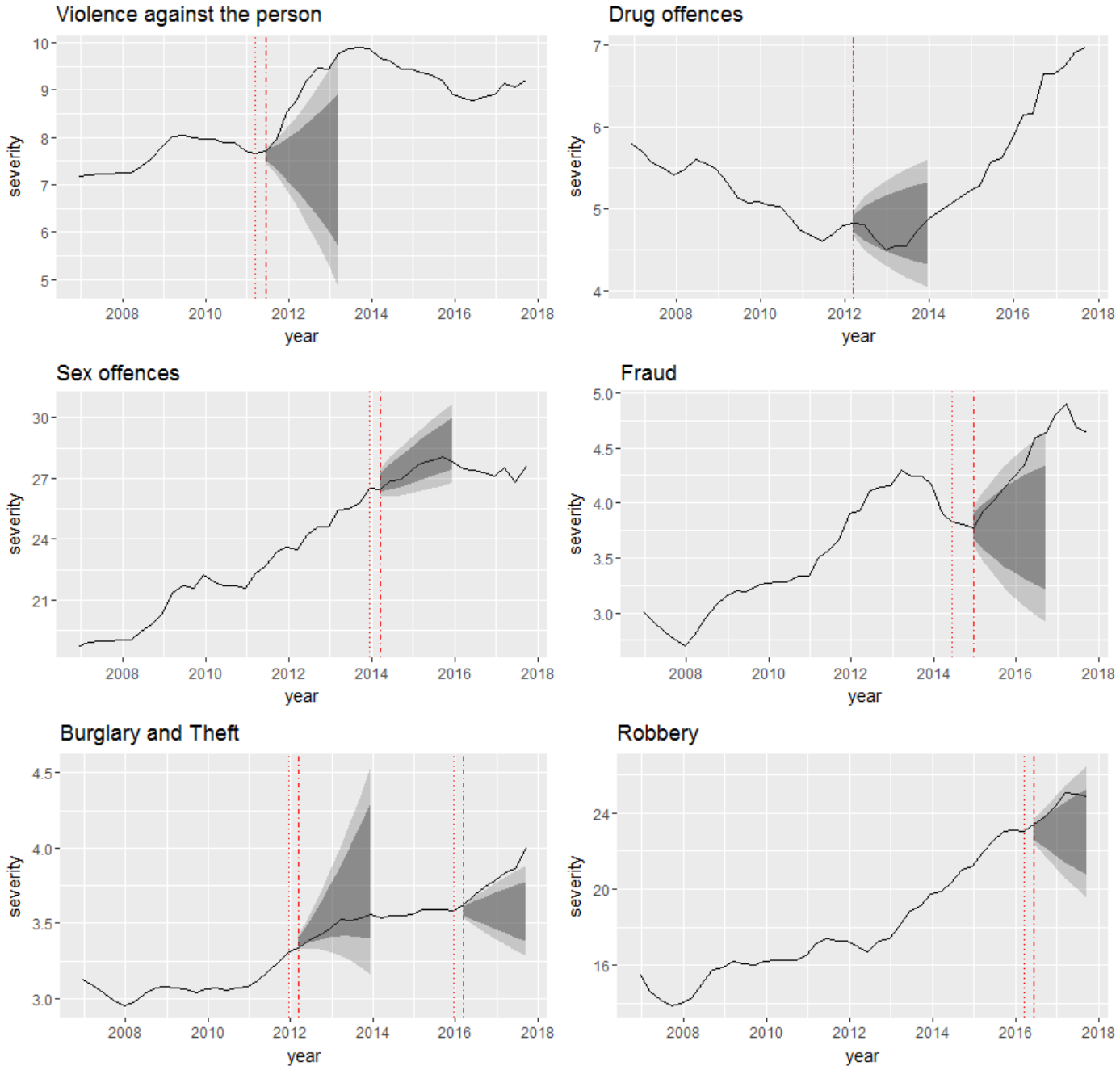












Application 3: Modelling Severity

- Assess the impact of selection bias
  - Model all cases available using their severity scores, compared to a similar model restricted to cases sentenced to custody
- Create a new framework to model sentence data as robustly and efficiently as possible
  - Using multiple dependent models to accommodate different disposal types
  - Measurement error models to account for the unobserved heterogeneity of conditional discharges, fines, and community orders,
  - Propagating the uncertainty associated to the estimation of severity scores properly using Bayesian statistics

A Sample of Theft Offences

- A sample of 7240 offences of theft
- Sentenced at the Crown Court in 2011
- 63.8% received a custodial sentence
  - 151 conditional discharges
  - 74 fines
  - 989 community orders
  - 1806 suspended sentences
  - 4220 custodial sentences

Table: Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Max
severity	7,240	13.116	12.363	0.95	105.84
age	7,240	32.423	11.024	18	83
male	7,240	0.852	0.355	0	1
pc1_3	7,240	0.252	0.434	0	1
pc4_9	7,240	0.164	0.370	0	1
pc10plus	7,240	0.170	0.375	0	1
plea	7,240	0.847	0.360	0	1
PO_aggburgdwell	7,240	0.004	0.063	0	1
PO_aggburgunspec	7,240	0.006	0.076	0	1
PO_atttheft	7,240	0.005	0.072	0	1
PO_commercialburg	7,240	0.079	0.269	0	1
PO_conspburg	7,240	0.003	0.057	0	1
PO_conspfraud	7,240	0.007	0.084	0	1
PO_conspother	7,240	0.002	0.048	0	1
PO_conspsteal	7,240	0.008	0.088	0	1
PO_dishonestrep	7,240	0.066	0.248	0	1
PO_equipped	7,240	0.007	0.085	0	1
PO_handling	7,240	0.011	0.106	0	1
PO_immigration	7,240	0.004	0.066	0	1
PO_laundering	7,240	0.016	0.124	0	1
PO_otherfraud	7,240	0.140	0.347	0	1
PO_othertheft	7,240	0.040	0.196	0	1
PO_receivinggoods	7,240	0.066	0.248	0	1
PO_theftperson	7,240	0.048	0.215	0	1
PO_theftshop	7,240	0.061	0.239	0	1
PO_thefttrust	7,240	0.062	0.242	0	1
PO_theftvehicle	7,240	0.005	0.071	0	1
PO_falsepassport	7,240	0.035	0.184	0	1

	Dependent variable: log(severity)	
	Model 1 - custody	Model 2 - all sentences
age of the defendant	0.006***	
	(0.001)	
guilty plea entered	−0.130***	
	(0.019)	
male defendant	0.052**	
	(0.024)	
1 to 3 prev convictions	0.092***	
	(0.020)	
4 to 9 prev convictions	0.184***	
	(0.022)	
10+ prev convictions	0.191***	
	(0.022)	
constant	2.836***	
	(0.040)	
Observations	4,220	7,240
R^2	0.331	
Adjusted R^2	0.327	
Note:		*p<0.1; **p<0.05; ***p<0.01

	Dependent variable: log(severity)	
	Model 1 - custody	Model 2 - all sentences
age of the defendant	0.006^{***}	0.005^{***}
	(0.001)	(0.001)
guilty plea entered	-0.130^{***}	-0.103^{***}
	(0.019)	(0.028)
male defendant	0.052^{**}	0.182^{***}
	(0.024)	(0.030)
1 to 3 prev convictions	0.092^{***}	0.464^{***}
	(0.020)	(0.027)
4 to 9 prev convictions	0.184^{***}	0.715^{***}
	(0.022)	(0.032)
10+ prev convictions	0.191^{***}	0.812^{***}
	(0.022)	(0.032)
constant	2.836^{***}	1.903^{***}
	(0.040)	(0.053)
Observations	4,220	7,240
R^2	0.331	0.317
Adjusted R^2	0.327	0.314
Note:		*p<0.1; **p<0.05; ***p<0.01

	Dependent variable: log(severity)	
	Model 1 - custody	Model 2 - all sentences
age of the defendant	0.006***	0.005***
	(0.001)	(0.001)
guilty plea entered	−0.130***	−0.103***
	(0.019)	(0.028)
male defendant	0.052**	0.182***
	(0.024)	(0.030)
1 to 3 prev convictions	0.092***	0.464***
	(0.020)	(0.027)
4 to 9 prev convictions	0.184***	0.715***
	(0.022)	(0.032)
10+ prev convictions	0.191***	0.812***
	(0.022)	(0.032)
constant	2.836***	1.903***
	(0.040)	(0.053)
Observations	4,220	7,240
R^2	0.331	0.317
Adjusted R^2	0.327	0.314
Note:		*p<0.1; **p<0.05; ***p<0.01

Transposing Sampling Error

	absolute discharge	conditional discharge	fine	community order	6month suspended 1month custody	12month suspended 1month custody	6month suspended 6month custody	24month suspended 12month custody	1month inmediate custody	2month inmediate custody	3month inmediate custody
absolute discharge	0.5	1	1	1	1	1	1	1	1	1	1
discharge	0	0.5	β_1	1	1	1	1	1	1	1	1
fine	0	$1-\beta_1$	0.5	β_2	1	1	1	1	1	1	1
community order	0	0	$1-\beta_2$	0.5	β_3	1	1	1	1	1	1
6month susp 1month cust	0	0	0	$1-\beta_3$	0.5	1	1	1	1	1	1
12month susp 1month cust	0	0	0	0	0	0.5	β_4	1	1	1	1
6month susp 6month cust	0	0	0	0	0	$1-\beta_4$	0.5	1	1	1	1
24month susp 12month cust	0	0	0	0	0	0	0	0.5	β_5	β_6	β_7
1month inmediate custody	0	0	0	0	0	0	0	$1-\beta_5$	0.5	1	1
2month inmediate custody	0	0	0	0	0	0	0	$1-\beta_6$	0	0.5	1
3month inmediate custody	0	0	0	0	0	0	0	$1-\beta_7$	0	0	0.5

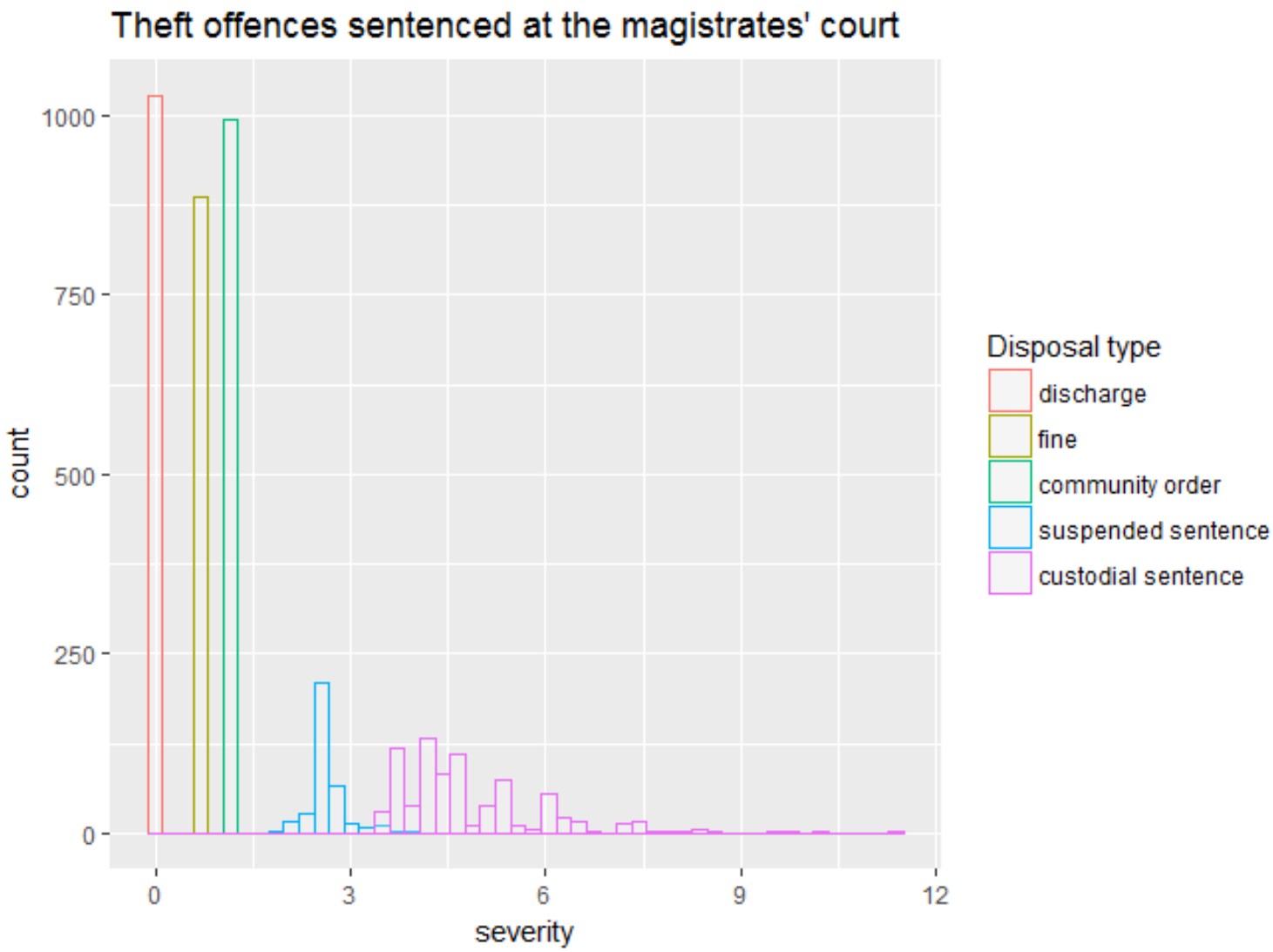
Transposing Sampling Error

- We assume the following model for each paired comparison

$$p_{st,i}|\alpha_{st}, \beta_{st} \sim \text{Beta}(\alpha_{st}, \beta_{st}) \quad \forall i,$$
$$\alpha_{st} \sim \text{IG}(0.1, 0.1),$$
$$\beta_{st} \sim \text{IG}(0.1, 0.1)$$

- Once we estimate all the α and β parameters, we can use their posterior distributions to transpose the uncertainty in the proportions

	Dependent variable: log(severity)		
	Model 1 - custody	Model 2 - all	Model 3 - error
age of the defendant	0.006***	0.005***	0.005***
	(0.001)	(0.001)	(0.001)
guilty plea entered	−0.130***	−0.103***	−0.103***
	(0.019)	(0.028)	(0.031)
male defendant	0.052**	0.182***	0.182***
	(0.024)	(0.030)	(0.035)
1 to 3 prev convictions	0.092***	0.464***	0.466***
	(0.020)	(0.027)	(0.040)
4 to 9 prev convictions	0.184***	0.715***	0.717***
	(0.022)	(0.032)	(0.051)
10+ prev convictions	0.191***	0.812***	0.817***
	(0.022)	(0.032)	(0.054)
constant	2.836***	1.903***	1.911***
	(0.040)	(0.053)	(0.081)
Observations	4,220	7,240	7,240
R^2	0.331	0.317	
Adjusted R^2	0.327	0.314	
Note:			*p<0.1; **p<0.05; ***p<0.01



Application 3: To Do List

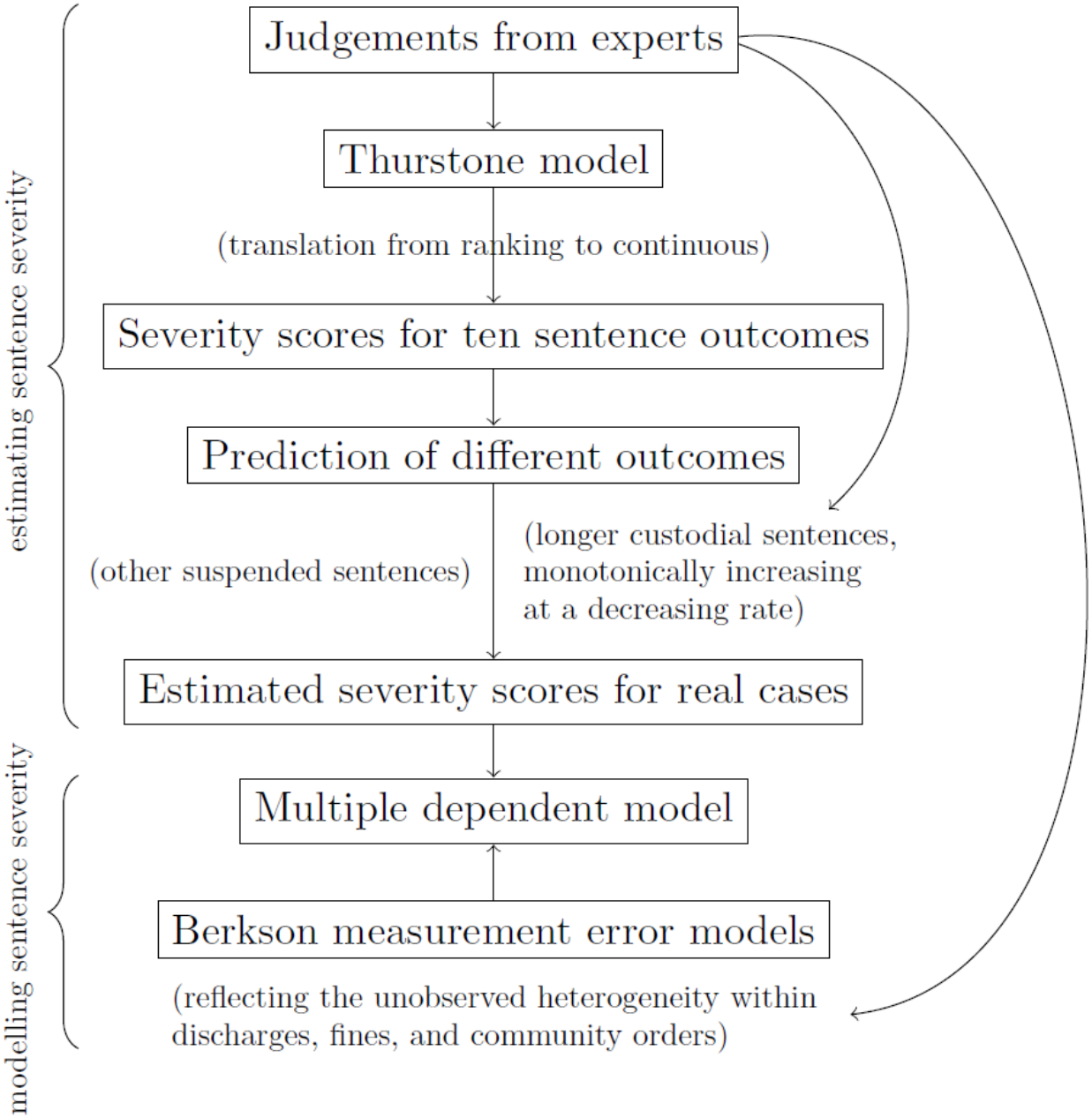
- We are currently exploring multiple dependent models to specify properly the different sentence outcomes
- Figuring out the size of the Berkson measurement error processes for *conditional discharges, fines, and community orders*

Application 3: To Do List

- We are currently exploring multiple dependent models to specify properly the different sentence outcomes
- Figuring out the size of the Berkson measurement error processes for *conditional discharges, fines, and community orders*
- Accounting for the diminishing returns of severity for every additional month in prison
  - Going from 3 to 4 months in custody is more severe than from 120 to 121
  - The magistrates tell us that going from 12 to 13 months, represents 90% of the increase of severity following an extension of 3 to 4 months

Application 3: To Do List

- We are currently exploring multiple dependent models to specify properly the different sentence outcomes
- Figuring out the size of the Berkson measurement error processes for *conditional discharges, fines, and community orders*
- Accounting for the diminishing returns of severity for every additional month in prison
  - Going from 3 to 4 months in custody is more severe than from 120 to 121
  - The magistrates tell us that going from 12 to 13 months, represents 90% of the increase of severity following an extension of 3 to 4 months
- Trying to estimate the whole process simultaneously in Stan
  - This is harder than expected given the difficulty to carry out data manipulation within the model



Discussion

- Selection bias is an extremely pervasive problem in sentence data analyses
- Affecting the validity of studies in key Criminological areas:
  - Effects of sentencing policy (sentencing guidelines)
  - Presence of discrimination (the Lammy review)
  - Deterrence (Von Hirsch et al. 1999)

Discussion

- Selection bias is an extremely pervasive problem in sentence data analyses
- Affecting the validity of studies in key Criminological areas:
  - Effects of sentencing policy (sentencing guidelines)
  - Presence of discrimination (the Lammy review)
  - Deterrence (Von Hirsch et al. 1999)
- Solutions suggested in the literature are based on questionable assumptions and waste information

Discussion

- Selection bias is an extremely pervasive problem in sentence data analyses
- Affecting the validity of studies in key Criminological areas:
  - Effects of sentencing policy (sentencing guidelines)
  - Presence of discrimination (the Lammy review)
  - Deterrence (Von Hirsch et al. 1999)
- Solutions suggested in the literature are based on questionable assumptions and waste information
- The estimation of scale of severity allows us to undertake more realistic/robust sentence data analysis
  - e.g.1 most guidelines cannot be blamed for the increase in severity
  - e.g.2 male defendants are sentenced more harshly than expected

Next Steps: Scaling It Up

- Include subcategories of fines and community orders in the questionnaire
  - *fine A* (50% weekly income), *fine B* (100% weekly income), ... *fine D* (600% weekly income)
  - *low* (40-80 hours), *medium* (80-150 hours), and *high* (150-300) community orders
- Include judges, boost the sample, use simple random sampling
  - We are still waiting for the Judicial Office to reply to our application

Next Steps: Scaling It Up

- Include subcategories of fines and community orders in the questionnaire
  - *fine A* (50% weekly income), *fine B* (100% weekly income), ... *fine D* (600% weekly income)
  - *low* (40-80 hours), *medium* (80-150 hours), and *high* (150-300) community orders
- Include judges, boost the sample, use simple random sampling
  - We are still waiting for the Judicial Office to reply to our application
- Probably we should open up more pairs to be compared
  - e.g. *fine* vs *1 custodial sentence suspended for 6 months*
- Explore applicability in other jurisdictions
  - Scotland ?
  - Maryland ?