Arielle Herman
Experimental Studies for Policy and Economics
Professors Alessandra Casella and Donald Green
Fall 2022

Replication Assignment

Abstract

In this paper, I replicate elements of Green's and Winik's 2010 study, "Using random judge assignment to estimate the effects of incarceration and probation on recidivism among drug offenders". Green and Winik analyze a natural experiment in which 1003 defendants charged with drug-related offenses are assigned to nine judicial calendars, each correlated with a set of judges of the DC Superior Court, between June 1, 2002 and May 9, 2003. They argue that judges' sentencing behaviors varied sufficiently such that random assignment to their court calendars serves as an instrumental variable to measure the causal impact of one month of incarceration on recidivism rates. In a replication study, "The impacts of incarceration on crime," David Roodman examines the instrument for weakness with the Anderson-Rubin test and explores the sensitivity of the findings to the definition of the follow-up interval.

In this paper, I explore the assumption that calendars can be ranked, replicate the instrumental variable regressions of *laterarr* on *toserve* (TSLS and and LIML), replicate part of the Roodman's Anderson-Rubin test, and replicate Roodman's sensitivity analysis for the measurement of follow-up interval. Overall, my findings support Green and Winik analysis, and Roodman's replication.

Data and Measures

Data. Green and Winik compile data from public lockup lists and case file records from the DC Superior court, supplemented by the Court's public electronic case management database. They restrict observations to defendants charged with at least one felony drug offense or at least one non-drug-related misdemeanor (e.g. panhandling or public intoxication) between June 1, 2002 and May 9, 2003. In order to avoid exposure to multiple treatments (i.e. different sets of judges), they additionally exclude a small number of instances when a defendant was sentenced or disposed for multiple cases simultaneously. Green and Winik employ robust cluster standard errors to account for dependency in the observations, as 172 codefendants were assigned the same judge. Finally, the majority of defendants are assigned to the Felony II docket, and Green and Winik therefore exclude the cases when defendants were assigned to the Accelerated Felony Calendar (AFTC) pre-randomization.

Sentencing (endogenous variables, treatment). Green and Winik measure sentences using continuous and binary variables. Possible sentences within the data include incarceration (*incarc* in months), probation (*probat* in months), or both. While probation may be sentenced independently of incarceration, frequently, a pre-defined portion of the defendant's incarceration is indefinitely suspended (*suspend* in months) to be imposed only if the defendant fails to comply with the conditions of their probation. Green and Winik estimate a defendant's time imprisoned (*toserve* in months) as the difference between *incarc* and *probat*. These variables measure the intended treatment of the study.

Variable	Name	Description
Incarceration	incarc incarcerate	Months of incarceration initially sentenced (potentially including some portion of suspended time) Binary indicating the defendant was sentenced to incarceration at their disposition
Probation	probat probatnonzero	Months of probation, which may be revoked and replaced with incarceration if the defendant violates the terms of their parole. Binary indicating the defendant was sentenced probation
Non-suspended sentence	toserve	Portion of total incarceration that is not suspended or dependent on fulfill the conditions of probation.

Covariates (exogenous variables). Green and Winik verify the random assignment of judges by examining age, binary and dummy demographic factors (e.g. *female*, *nonblack*), and dummies describing previous criminal history. Notably, the prevalence of criminal history among the defendants indicates that previous judicial treatments have not been successful in deterring crime, implying that the population may be biased. They find no systematic relationships that would undermine their causal inference.

Recidivism (outcome variables). Green and Winik track recidivism as a binary indicator of rearrest (*laterarr*) within four years from the defendant's initial disposition in the data. For robustness, they additionally code binaries that describe types of arrests and convictions. Green and Winik note that starting the follow-up period on the date of disposition confounds the effects on recidivism of deterrence and incapacitation. However, they argue that this effect should be small given that 97.8% of the defendants had at least one year to recidivate upon release within the timeframe of the study. Roodman examines the result's sensitivity to the follow-up interval by repeating the analysis multiple times while varying the duration of the follow-up interval between 2 days and 4 years from the initial date of disposition. This is further discussed below.

Judge Assignments (instrumental variable). A mechanical wheel randomly assigns defendants to calendars, associated with a set of judges for a given year. Thus, random assignment of the calendars exposes defendants to discrete sets of judges. Notably, the judges assigned to each calendar may rotate annually at the beginning of the year. Random assignment therefore is of the calendar, and not the set of judges. Online research has failed to clarify the assignment mechanism, as consistent online documentation of local rule changes do not precede 2015.[1] I was able to find documentation regarding the random assignment of the DC District Court (see appendix 1 for further documentation on judicial assignment).

[1] "Prior Rule Changes | District of Columbia | United States District Court," accessed December 25, 2022, https://www.dcd.uscourts.gov/prior-rule-changes.

Statistical Model

In addition to the usual necessary assumptions for experiments (i.e. random assignment of the treatment, absence of spillover effect or the stable unit treatment value assumption), Green and Winik assume the following to motivate their instrumental variable analysis:

1. Judges influence recidivism only through sentencing of the defendant.
2. The sets of judges vary consistently in their sentencing behavior, meaning that calendar assignment has a nonzero effect on sentence length.
3. Sets of judges may be monotonically ranked by severity of sentencing patterns.

Green and Winik support assumptions 2 and 3 by describing the variation in the relevant endogenous variables, as summarized in their table 3, where higher values indicate greater harshness.

Table 1. Replicated Table 3

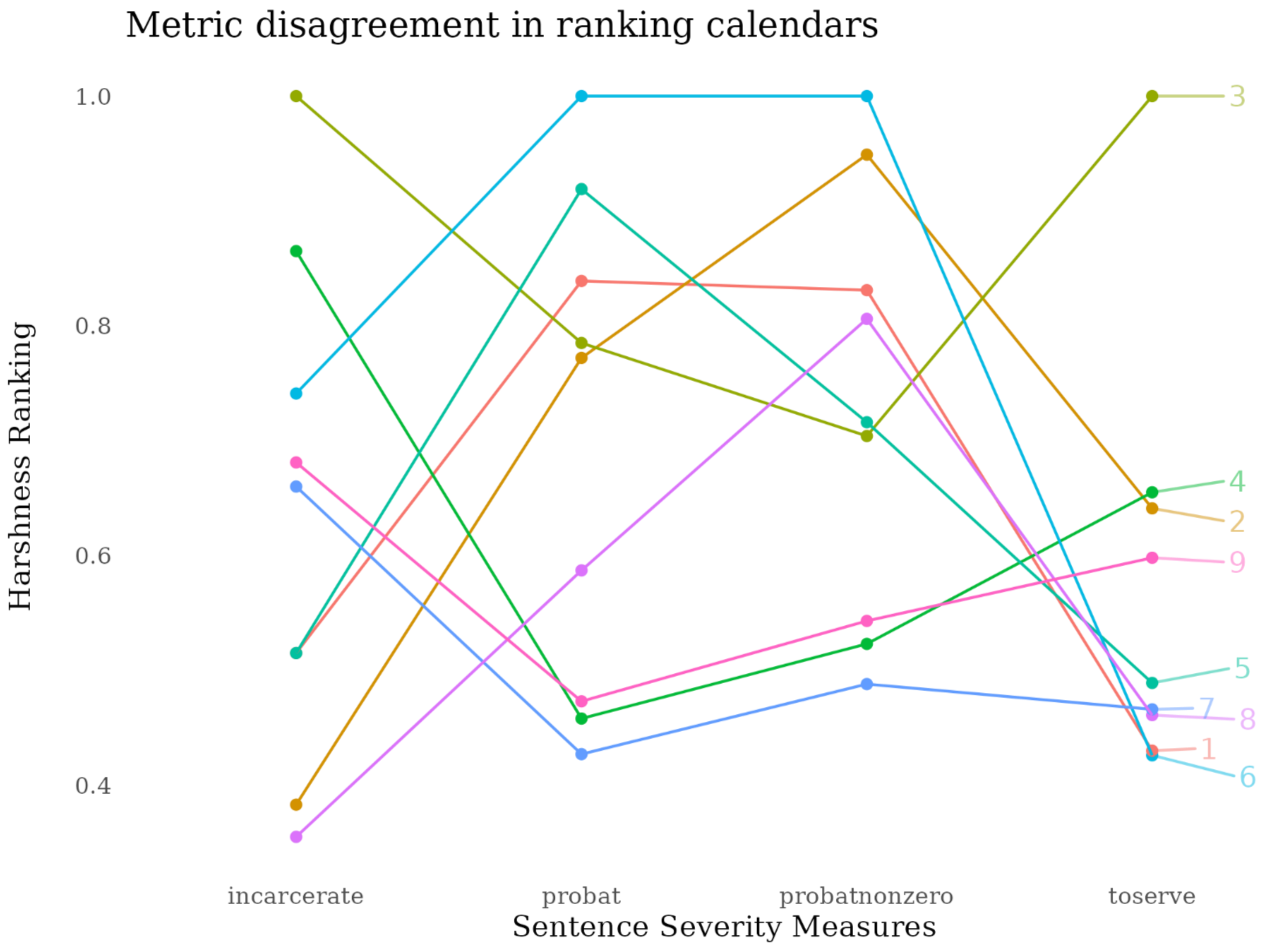
Endogenous	1	2	3	4	5	6	7	8	9
Toserve	5.120	7.630	11.900	7.790	5.820	5.070	5.550	5.490	7.120
probatnonzero	0.500	0.571	0.424	0.315	0.431	0.602	0.294	0.485	0.327
Probat	12.500	11.500	11.700	6.830	13.700	14.900	6.360	8.740	7.050
incarcerate	0.336	0.250	0.653	0.565	0.336	0.484	0.431	0.232	0.445
Incarc	24.600	82.600	28.200	11.200	18.800	21.000	11.600	24.500	14.200

Notably, these various measures suggest conflicting rankings for calendar severity. Thus, I make the below plot. For each variable, I aggregate the means by calendar and normalize to 1 by dividing a given variable by its maximum mean value.[2]

If the measures all agreed, the plot would exhibit horizontal lines, associating each calendar (labeled on the right) with a single ranking (y-axis). Instead, the plot captures the extent of the disagreement. For example, the measures *incarc* (total sentence in months), *probat* (months of probation), and *probatnonzero* (binary indicating sentence included probation), rank calendar 4 as the least or second-least severe calendar; whereas, *incarcerate* (binary indicating incarceration) and *toserve* (months of non-suspended sentence) rank calendar 4 as the second-most severe calendar. While each of the proposed measures for severity may differentiate the sentencing behavior of one calendar from another, the disagreement suggests either that these measures do not capture the full extent of calendar harshness or that the calendars cannot be monotonically ranked.

[2] Notably, the ranking plot suggests the possibility of outliers in calendar 2 for *incarc*. The presence of outliers is confirmed by a density distribution (see Appendix 2), which indicates 23 outlying sentences of 324 months each.

Figure 1. Metric disagreement in ranking calendars



We can further explore the sentencing behaviors of the calendars by analyzing the monthly trends of each sentencing measure. I measure the monthly average because the continuous endogenous variables are measured in months. It is necessary to take an average over some base unit of time because judges for a given calendar sometimes depose multiple cases in one day. Even though the measures disagree on ranking, if the sentencing patterns are stationary over time and distinct from each other, we may still conclude that the calendars may be monotonically ranked, without knowing all the variables that contribute to calendar harshness. Further, this analysis will confirm whether the sentencing patterns of calendars vary with the new year, at the moment judges may be reassigned.

First, I run the Augmented Dickey-Fuller (ADF) test on each endogenous variable of interest (*incarc*, *incarcerate*, *probat*, *probatnonzero*, *toserve*) averaged per month. We can reject the null hypothesis that any variable is a unit root (a characteristic of a nonstationary process) at $p < 0.05$. This suggests that the sentencing trend observations are independent and that the calendars do exhibit discernable trends.

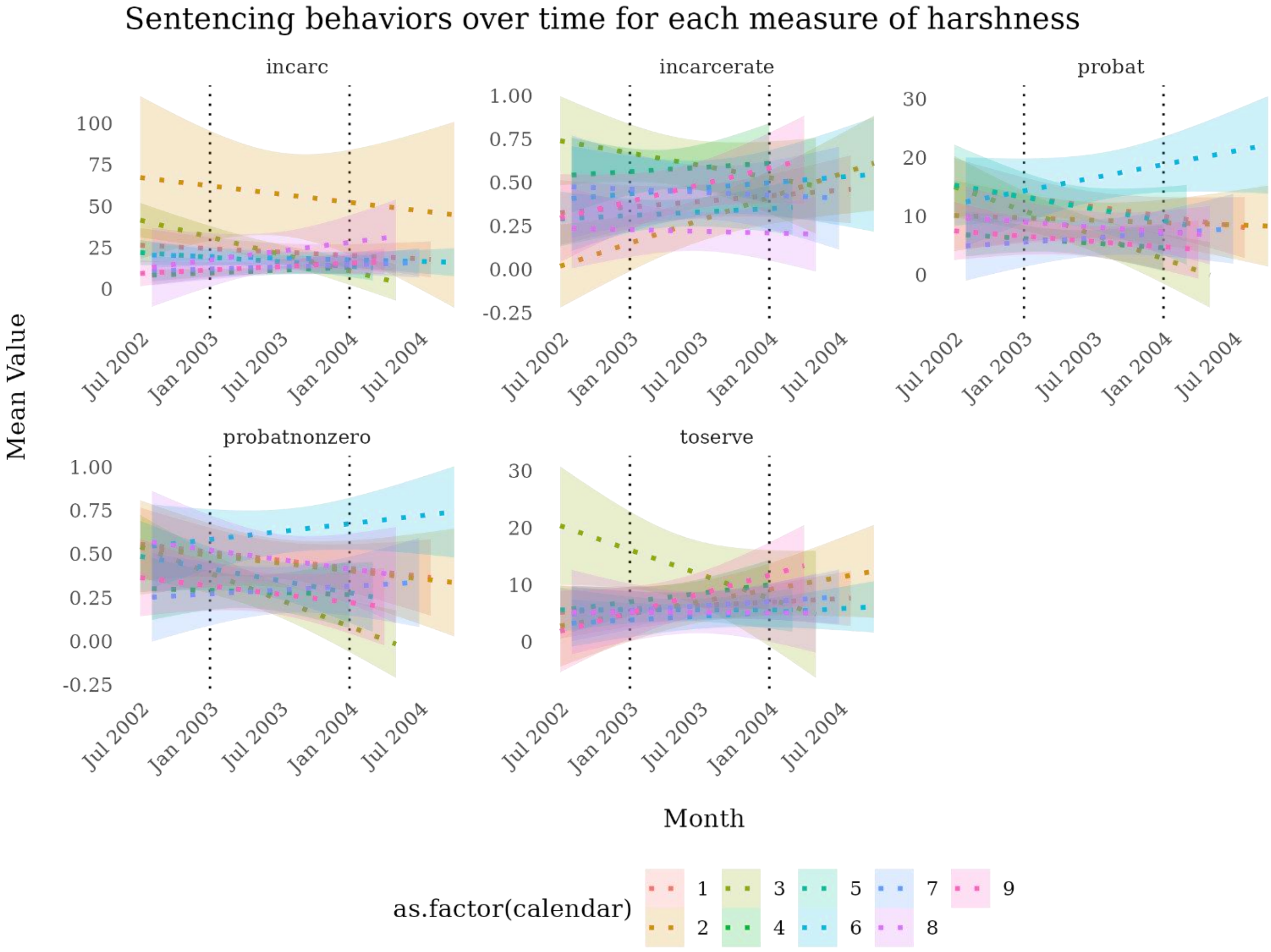
Table 2. Augmented Dickey-Fuller Test results

Variable	Test Statistic	P-value	Alternative
probat	-3.693952	0.02607521	stationary
toserve	-5.403712	0.01000000	stationary
incarcerate	-4.592618	0.01000000	stationary
probatnonzero	-3.767729	0.02192651	stationary
incarc	-4.132075	0.01000000	stationary

Next, we can examine the trend line over time, as fitted with a glm model. The below plot visualizes these trends and the associated standard errors. The extent of the overlap of standard errors imply that only a few of the calendars can be distinguished by these metrics over time.[3] Going forward, I will refer to *toserve* as this measure most directly concerns the research question of the paper.

[3] 23 outliers may be responsible for the apparent differentiation of calendar2 in *incarc* (see Appendix 2)

Figure 2. Sentencing Behaviors over time for each measure of harshness



Analysis

Of Green and Winik's analysis, I replicate the instrumental variable regressions in 2SLS and LIML, looking at only one endogenous variable, *toserve*. I prefer this variable because out of all the measures for sentence explored above, this variable most directly targets the research question regarding the effect of an additional month on future recidivism rates. There are also limitations in the implementation of LIML in R, which permits only endogenous variable. I additionally focus on just one binary outcome variable for recidivism, *laterarr*, which measures if the defendant is rearrested within four years of the disposition date. This is the preferred outcome variable in Green and Winik, and I feel that Roodman sufficiently explores the other most compelling binary outcome variable, which indicates if defendants are convicted within the follow-up interval, to corroborate Green and Winik's analysis.

The equation I use in the analysis uses the instrumental variable, *calendar*, to estimate the causal impact of *toserve* on *laterarr*, while controlling for the relevant covariates.

To implement the analysis in R, I utilize the packages "plm" and "ivmodel." The former offers only 2SLS but permits multiple endogenous variables and robust clustered standard errors, while the second offers both 2SLS and LIML, but only permits one endogenous variable. I replicate Green and Roodman's table 7 with package "plm." See full results in Appendix 3.

Table 3a. Replicated Regression Coefficients for 2SLS from packages "plm"

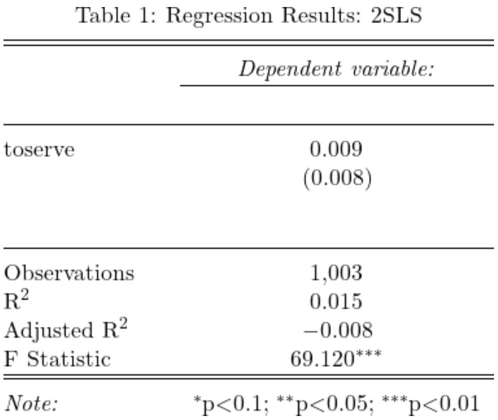


Table 1: Regression Results: 2SLS

	Dependent variable:
toserve	0.009
	(0.008)
Observations	1,003
R^2	0.015
Adjusted R^2	−0.008
F Statistic	69.120***
Note:	*p<0.1; **p<0.05; ***p<0.01

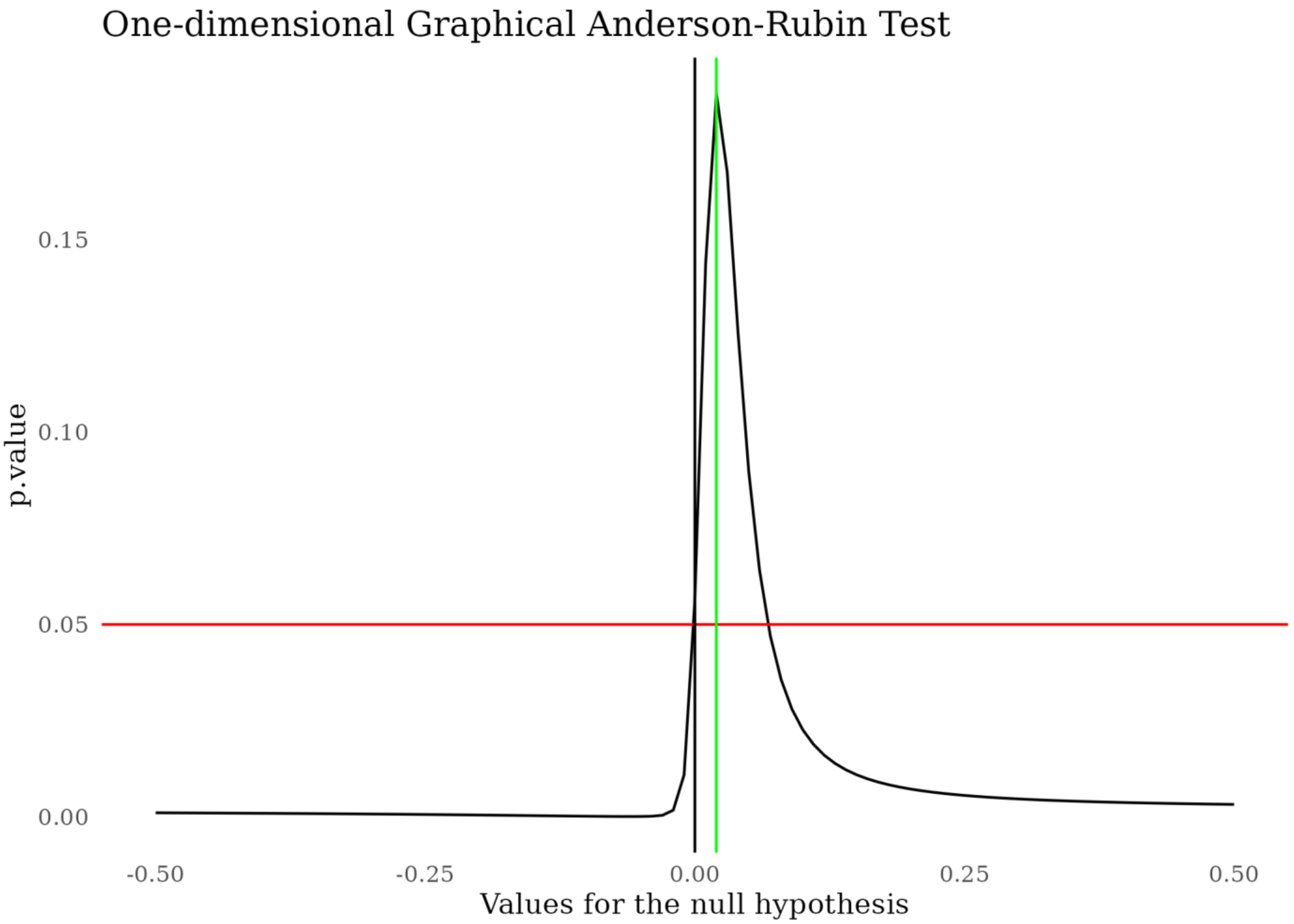
Below, I provide the results of the LIML regression using the packages "ivmodel". Uniquely, this "ivmodel" function calculates both LIML and 2SLS. Therefore, the table below permits direct comparison of the two estimated impacts.

Table 3b. Replicated Regression Coefficients for LIML and 2SLS from package "ivmodel"

Regression	Endogenous	k	Point Estimate	Standard Error	P-value
LIML	toserve	1.01	0.199	0.0183	0.277
kClass (2SLS)	Toserve	1	0.00818	0.00814	0.315

Next, I perform a one-dimensional Anderson-Rubin test. While Roodman performs a two-dimensional graphical Anderson-Rubin test, this was not yet possible in R as an iv regression with two endogenous variables has not yet been implemented. I use the estimate from the LIML regression and compare it to multiple null hypotheses, for every 0.01 unit value between -0.5 and 0.5. Roodman takes an additional step and bootstraps the AR test. The coefficient is indicated on the plot with the green line, and the results of the test suggest that we can strongly reject the null hypotheses that the estimated coefficient is negative.

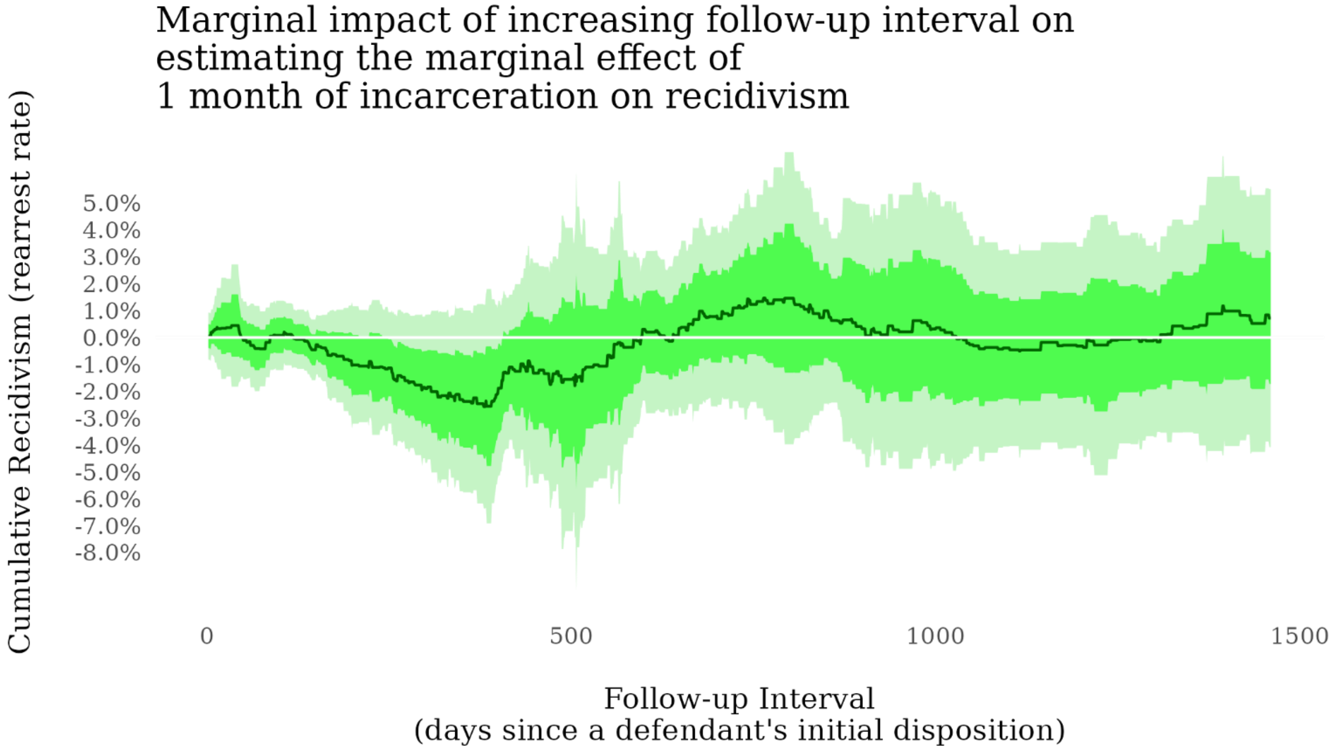
Figure 3. Anderson-Rubin Test



Finally, I replicate Roodman's sensitivity analysis using the same regression formula as earlier in the paper. This means that my results are slightly different from Roodman's, who instead uses two endogenous variables in his LIML equation. But just like Roodman, I vary the interval of the follow-up period and redefine *laterarr* accordingly. Since the number of defendants who have been secondarily arrested does not necessarily change every day, I save computational time by calculating the coefficient for only unique days. I also saved computational time by modifying the ivmodel equation to only estimate the LIML iv model.

The below plot charts the marginal impact of expanding the follow-up interval by one day in estimating the marginal effect of 1 month of incarceration on recidivism. The curve swings below 0 until the beginning of second year, and then it continues to hover around zero, dipping negative again at the beginning of the fourth year. Just as in Roodman's analysis, there is no big positive swing that decays toward the four-year mark, suggesting that the choice of a four-year follow-up interval adequately captures the presence of an effect, if any. And, just like Roodman's analysis, none of the coefficients could be distinguished from zero on the 95% confidence interval, suggesting that 1 month of incarceration has no detectable impact on recidivism no matter the definition of the follow-up period.

Figure 4. Sensitivity Analysis



Conclusion

In this paper, I explored the assumption that calendars can be ranked, replicated the instrumental variable regressions of *laterarr* on *toserve* (TSLS and and LIML), replicated part of the Roodman's Anderson-Rubin test, and replicated Roodman's sensitivity analysis for the measurement of follow-up interval.

My exploration of the monotonicity of the judicial calendars questions the validity of the assumption, but does not discount it. Further research is necessary to evaluate this assumption. The replicated regressions all find weakly positive coefficients on *toserve* that are not statistically significant. This result reinforces the findings of Green and Winik, and Roodman, that an additional month of incarceration as no detectable effect on recidivism. The Anderson-Rubin test further reinforces this finding as it suggests that the estimated impact is statistically different from negative values. The replication of Roodman's sensitivity analysis reinforces his original findings that the definition of the follow-up period doesn't obscure potential impacts visible early in the period.

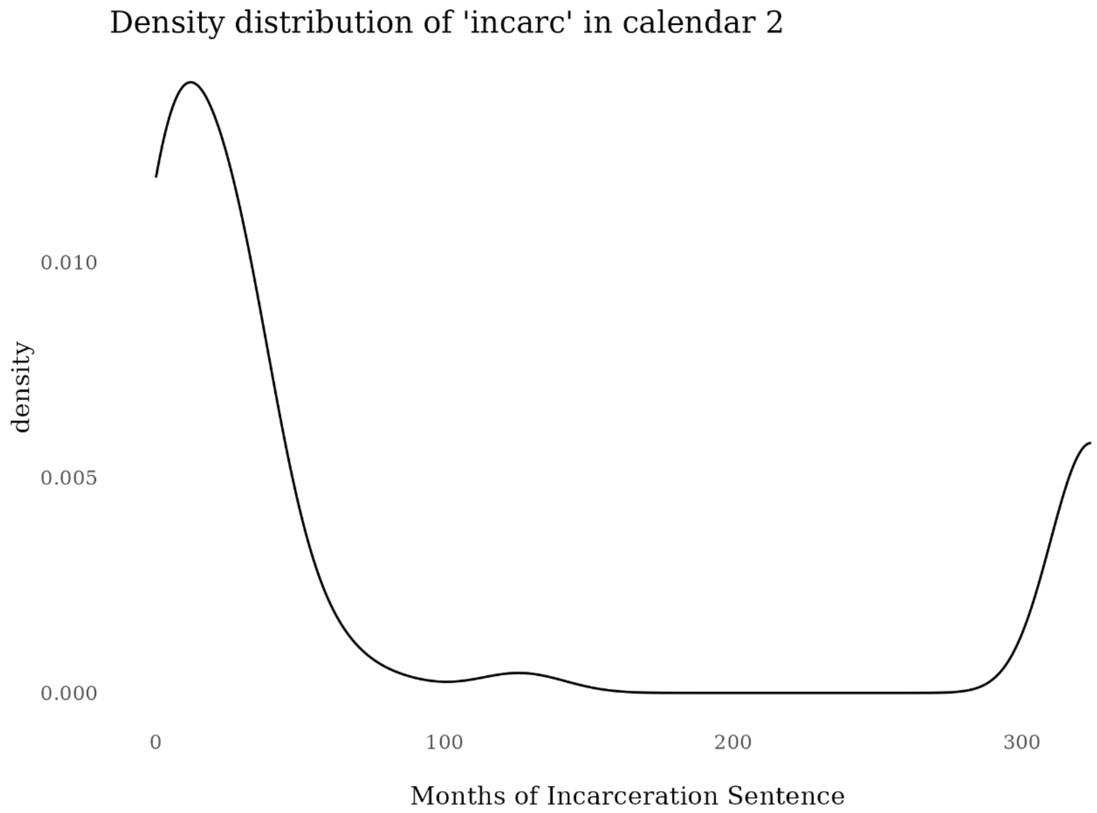
Appendix 1. Rules of the United States District Court for the District of Columbia (2015, updated 2019)

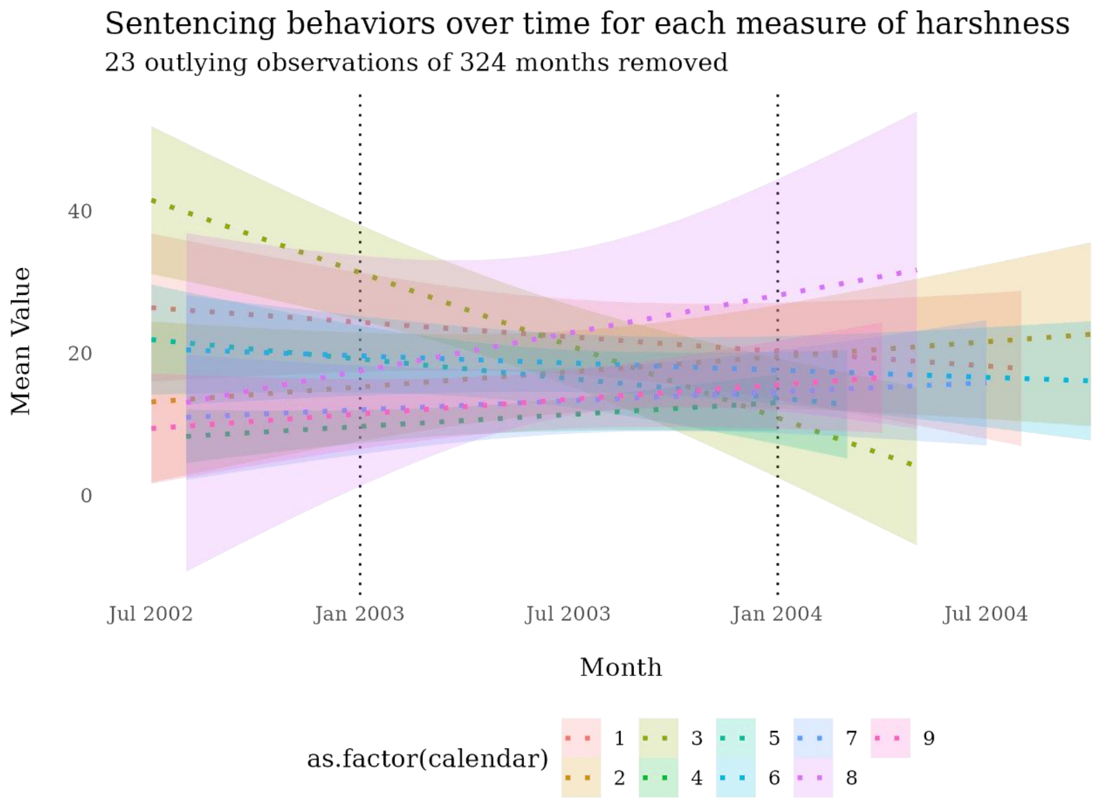
The local rule changes updated as of July 2019 clarify how the Calendar and Case Management Committee handles random assignment of cases currently, and suggests that the set of judges associated with a particular calendar or "deck" are repeatedly sampled randomly without replacement;[4] however, the document does not clearly specify assignment mechanism to the "appropriate" deck or set of judges.

> The Clerk shall create a separate assignment deck in the automated system for each subclassification of civil and criminal cases established by the Court pursuant to LCvR 40.2 of these Rules and a separate deck [or calendar] for miscellaneous cases. The decks will be created by the Liaison to the Calendar and Case Management Committee or the Liaison's backup and access to this function shall be restricted to these individuals to protect the integrity and confidentiality of the random assignment of cases. The Calendar and Case Management Committee will, from time to time determine and indicate by order the frequency with which each judge's name shall appear in each designated deck, to effectuate an even distribution of cases among the active judges.
>
> At the time…information is returned in a criminal case, the case shall be assigned to the judge whose name appears on the screen when the appropriate deck is selected.[5]

[4] E Barrett Prettyman, "EFFECTIVE AS OF SEPTEMBER 2015," n.d.
[5] Prettyman.

Appendix 2. Calendar 2 Outlier in *Incarc*.





Appendix 3. Table 3a Regression Results

Table 1: Regression Results: 2SLS

	Dependent variable:
toserve	0.009
	(0.008)
age	-0.025^{**}
	(0.010)
agesq	0.0002
	(0.0001)
female	-0.001
	(0.064)
nonblack	-0.219^{**}
	(0.109)
priorarr	-0.060
	(0.077)
priordrugarr	0.007
	(0.069)
priorfelarr	0.104
	(0.071)
priorfeldrugarr	-0.100
	(0.073)
priorcon	0.020
	(0.076)
priordrugcon	0.041
	(0.078)
priorfelcon	-0.100
	(0.077)
priorfeldrugcon	0.056
	(0.085)
pwid	0.011
	(0.062)
dist	0.011
	(0.065)
marijuana	0.100^{*}
	(0.058)
cocaine	-0.0002
	(0.058)
crack	0.040
	(0.066)
heroin	0.084
	(0.062)
pcp	0.082
	(0.097)
otherdrug	-0.040
	(0.107)
nondrug	0.001
	(0.050)
Constant	1.012^{***}
	(0.190)
Observations	1,003
R^2	0.015
Adjusted R^2	-0.008
F Statistic	69.120^{***}
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$