Essays in Cluster Sampling and Causal Inference

Susanna Maria Mäkelä

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ABSTRACT

Essays in Cluster Sampling and Causal Inference

Susanna Maria Mäkelä

This thesis consists of three papers in applied statistics, specifically in cluster sampling, causal inference, and measurement error. The first paper studies the problem of estimating the finite population mean from a two-stage sample with unequal selection probabilies in a Bayesian framework. Cluster sampling is common in survey practice, and the corresponding inference has been predominantly design-based. We develop a Bayesian framework for cluster sampling and account for the design effect in the outcome modeling. In a two-stage cluster sampling design, clusters are first selected with probability proportional to cluster size, and units are then randomly sampled within selected clusters. Methodological challenges arise when the sizes of nonsampled cluster are unknown. We propose both nonparametric and parametric Bayesian approaches for predicting the cluster size, and we implement inference for the unknown cluster sizes simultaneously with inference for survey outcome. We implement this method in Stan and use simulation studies to compare the performance of an integrated Bayesian approach to classical methods on their frequentist properties. We then apply our propsed method to the Fragile Families and Child Wellbeing study as an illustration of complex survey inference.

The second paper focuses on the problem of weak instrumental variables, motivated by estimating the causal effect of incarceration on recidivism. An instrument is weak when it is only weakly predictive of the treatment of interest. Given the well-known pitfalls of weak instrumental variables, we propose a method for strengthening

a weak instrument. We use a matching strategy that pairs observations to be close on observed covariates but far on the instrument. This strategy strengthens the instrument, but with the tradeoff of reduced sample size. To help guide the applied researcher in selecting a match, we propose simulating the power of a sensitivity analysis and design sensitivity and using graphical methods to examine the results. We also demonstrate the use of recently developed methods for identifying effect modification, which is an interaction between a pretreatment covariate and the treatment. Larger and less variable treatment effects are less sensitive to unobserved bias, so identifying when effect modification is present and which covariates may be the source is important. We undertake our study in the context of studying the causal effect of incarceration on recividism via a natural experiment in the state of Pennsylvania, a motivating example that illustrates each component of our analysis.

The third paper considers the issue of measurement error in the context of survey sampling and hierarchical models. Researchers are often interested in studying the relationship between community-levels variables and individual outcomes. This approach often requires estimating the neighborhood-level variable of interest from the sampled households, which induces measurement error in the neighborhood-level covariate since not all households are sampled. Other times, neighborhood-level variables are not observed directly, and only a noisy proxy is available. In both cases, the observed variables may contain measurement error. Measurement error is known to attenuate the coefficient of the mismeasured variable, but it can also affect other coefficients in the model, and ignoring measurement error can lead to misleading inference. We propose a Bayesian hierarchical model that integrates an explicit model for the measurement error process along with a model for the outcome of interest for both sampling-induced measurement error and classical measurement error. Advances in Bayesian computation, specifically the development of the Stan probabilistic programming language, make the implementation of such models easy and straightforward.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisor, Professor Andrew Gelman, for all of his guidance and support during my time in graduate school. I have admired his work since before I was a graduate student. I was inspired to pursue statistics many years ago from his blog, which I would read when I was bored at work. Now having the opportunity to study under and learn from him at Columbia has been a true privilege. His dedication to the intersections of statistics, the social sciences, teaching, and statistical communication is an inspiration, and I am deeply grateful to him for his encouragement, insights, and mentorship.

I would also like to thank Professor José Zubizarreta for introducing me to the world of causal inference. Studying causality has been one of the most intellectually rewarding parts of my graduate study, and it was his talk in one of our student seminars that first got me hooked. I am sincerely thankful for and appreciative of his guidance, advice, and kindness over the last five years.

Thank you to Professor Daniel Nagin for giving me the opportunity to study causal inference via the question of incarceration and recidivism. Working on a problem as important and impactful as this has motivated me every step of the way, and I am grateful to have learned from his expertise in criminology.

I would also like to thank Professors David Madigan and Michael Sobel for their willingess to serve on my thesis committee and provide their valuable insights and feedback. Their comments have helped improve the contents of this thesis.

Professor Yajuan Si has provided unwavering support and insight over the last several years, and I am so thankful for all that she has done for me. Her mentorship, patience, and guidance have been invaluable, and her thoughtful advice and comments have been an integral part of this dissertation. Its contents are greatly improved

thanks to her input.

To my New York women PhD friends – Charlotte, Shawn, Shira, Krista – I would never have made it through the last five years without you. You are a constant source of joy, strength, and inspiration, and I am so lucky to have you in my life. Alice, Becca, Lena, Sarah – thank you for supporting me from afar, reminding me of who I am outside of graduate school, and living on the West Coast where the time difference means I can call you at midnight just to say hi.

Finally, there are not enough words to express my love and appreciation for my family. None of this would have been possible without you. None of this means anything without you.

To my beloved niece Eeva. I hope that when you get older, you like math as much as your favorite aunt does.

Chapter 1

Bayesian Inference under Cluster Sampling

1.1 Introduction

We develop a Bayesian paradigm for survey inference under cluster sampling, particularly in the absence of design information for nonsampled clusters. Cluster sampling increases cost efficiency when partial clusters are included in the probability sampling framework. Bayesian inference in this context is essentially outcome prediction for nonsampled units in sampled clusters and for all units in nonsampled clusters. It is important to account for design information in the model, but it is often unknown or inaccessible for nonsampled clusters. We introduce estimation strategies for design information and connect multilevel regression models to sampling design as a unified Bayesian framework for survey inference.

We consider two-stage cluster sampling, which involves first sampling primary sampling units (PSUs) and then sampling secondary sampling units (SSUs) within selected PSUs. This sampling design requires a complete listing of PSUs and a complete listing of units only within selected PSUs and is thus widely used when generating a sample frame of every unit in the population is infeasible or impractical. For example, in designing a nationally representative household survey, generating a complete listing of every household in the country requires essentially as much effort as a complete census of all households. Instead, the sampling proceeds in stages, first sampling PSUs such as counties, cities, or census tracts. The PSUs are sampled with probability proportional to a measure of size, which is commonly the number of secondary units in the PSU but can be a more general measure of size, such as annual revenue or agricultural yield. SSUs are then randomly selected within selected PSUs, often with a fixed number or proportion. This design assumes invariance and independence of the second-stage sampling design (Särndal *et al.*, 1992). Invariance means that the sampling of SSUs is independent of which PSUs are sampled, and independence means sampling of SSUs in a given PSU is independent of sampling in other PSUs. In contrast, a two-*phase* design is one in which one or both of these assumptions do not hold.

Our motivating application survey, the Fragile Families (FF) study, was collected via a multi-stage sampling design, where cluster sampling was as a key step. The Fragile Families and Child Wellbeing Study (Reichmann *et al.*, 2001) aims to examine the conditions and capabilities of new unwed parents and the wellbeing of their children. To obtain a nationally representative sample of non-marital births in large U.S. cities, the study sequentially sampled cities, hospitals, and births. The sampling of cities used a stratified random sample of all U.S. cities with 200,000 or more people, where the stratification was based on policy environments and labor market conditions in the different cities. Inside each stratum, cities were selected with probability proportional to the city population size. In the selected cities, all hospitals in the small cities were included, while a random sample of hospitals or the hospital with the largest number of non-marital births was selected in large cities. Lastly, a predetermined number of births were selected inside each hospital. Classical weighting adjustment for the complex study design results in highly variable weights (Carlson, 2008), which lead to unstable inferences.

Our goal is to develop hierarchical models that account for design effects to yield robust survey inference. Bayesian hierarchical models are well-equipped to handle the multi-stage design and stabilize estimation via smoothing. As an intermediate step, two-stage cluster sampling is crucial in the FF study to select cities and hospitals. However, cluster sampling presents unique methodology challenges in the Bayesian context, as little information is available on the nonsampled clusters. In this work, we use the FF study as an illustration and focus on Bayesian cluster sampling inference to build a unified survey inference framework. The unified framework can be extended under a complex sampling design, as discussed in Section 1.5.

We illustrate finite population inference with the estimation of a population mean in a two-stage cluster sample. Specifically, we consider a population of J clusters, with each cluster j containing N_j units and a total population size of $N = \sum_{j=1}^{J} N_j$. Let I_j denote the inclusion indicator for cluster j and $I_{i|j}$ denote the inclusion indicator

for unit i in cluster j, $i = 1, \ldots, N_{j[i]}$, where $j[i]$ denotes the cluster to which unit i belongs. Clusters are sampled with probability proportional to the measure of size M_j, which is known to the analyst only for the sampled clusters. Our goal is to estimate the finite population mean of the survey variable y, which, for a continuous variable is defined as

$$\overline{y} = \sum_{j=1}^{J} \frac{N_j}{N} \overline{y}_j, \tag{1.1}$$

where $\overline{y}_j$ represents the mean (proportion) of y in cluster j. For a binary outcome, we seek to estimate the population proportion, given by

$$\overline{y} = \sum_{j=1}^{J} \frac{y_{(j)}}{N}, \tag{1.2}$$

where $y_{(j)}$ is the population total in cluster j.

Classically, inference in survey sampling has been design-based. The design-based approach treats the survey outcome y as fixed, with randomness arising solely from the randomization distribution of the inclusion indicator I. Design-based estimators have the advantage of being design-consistent, where design-consistency means that the estimator will converge to the true value as the population and sample sizes increase under the given sampling design. However, they are often unstable with large standard errors. For estimating the finite population mean of an outcome y_i, the classical design-based estimator for a single-stage sample s of size n is the Hájek estimator (Särndal *et al.*, 1992) $\widehat{\theta}^H = \frac{\sum_{i=1}^{n} y_i/\pi_i}{\sum_{i=1}^{n} 1/\pi_i}$, where π_i is the inclusion probability of unit i. In the two-stage sample s, when J_s out of J clusters are selected with n_j sampled SSUs, the estimator becomes

$$\widehat{\theta}^H = \frac{\sum_{j=1}^{J_s} \left(\sum_{i=1}^{n_j} y_i/\pi_{i|j} \right) /\pi_j}{\sum_{j=1}^{J_s} N_j/\pi_j}, \tag{1.3}$$

where π_j is the selection probability of cluster j, and $\pi_{i|j}$ is the selection probability of unit i in cluster j given that cluster j was sampled (Särndal *et al.*, 1992).

One major challenge with design-based estimators is variance estimation. Expressions for the variance of design-based estimators generally require knowledge of not

only the inclusion probability π_i for a given unit i, but also the joint inclusion probability $\pi_{ii'}$ for any two units i and i'. This information is often unknown in practice, as in the case of unknown measures of size for nonsampled clusters under the PPS setting. Joint inclusion probabilities can be challenging to compute even for straightforward sampling designs, and variance estimators for design-based estimators are often based on simplifications and approximations. In addition, the inverse-probability weighting often leads to highly variable estimators.

Bayesian inference, in contrast, directly models both the inclusion indicators I and the survey outcomes y. The Bayesian approach to survey inference has many advantages over the design-based approach, including the ability to handle complex design features like clustering, better inference for small-sample problems, incorporation of prior information, and large-sample efficiency (Little, 2004). In addition, if we are able to include the design variables in our model, the selection mechanism becomes ignorable and we can model the outcomes y alone, instead of jointly modeling y and the inclusion vector I (Rubin, 1983; Gelman *et al.*, 2013). The importance of including design variables in the model has also been emphasized for missing data imputation (Schafer, 1997; Reiter *et al.*, 2006).

Unfortunately, in many (arguably most) practical situations, the set of design variables is not known for the entire population and is instead known only for sampled clusters or units. In the case of PPS sampling, in which the design variables consist of the cluster measures of size $\{M_j\}_{j=1}^J$, we as the survey analyst may only have access to M_j (or, equivalently, the probability of selection π_j) for the sampled clusters. This missing data is a problem in the Bayesian setting because we cannot predict the values of y for the nonsampled clusters without it. We need to model the values of M_j for nonsampled clusters before we can make inferences about $\overline{y}$ conditional on the design information.

Recent Bayesian approaches to this problem (Zangeneh *et al.*, 2011a; Zangeneh and Little, 2015) consider the case of a single-stage PPS sample. In addition, they

separate estimation of the missing measure sizes and inference for the finite population quantities into two steps. In contrast, we propose an approach that integrates these steps into one model for a two-stage cluster sample. Our model allows for both cluster- and unit-level information to be used in certain cases. For the remainder of this paper, we assume the measure of size is equal to the cluster size N_j and use N_j in place of M_j.

The rest of this paper proceeds as follows. Section 1.2 first gives an overview of current approaches to estimating finite population totals under PPS and then describes our approach and its advantages. In Section 1.3, we describe a simulation study to investigate the performance of our method and other proposed methods. We apply our method to data from the Fragile Families study in Section 1.4 and discuss the results and extensions in Section 1.5.

1.2 Methods

In two-stage cluster sampling, a fixed number J_s of clusters are sampled with PPS, so that the probability of cluster j being included in the sample is proportional to N_j: $\Pr(I_j = 1 \mid N_j) \propto N_j$. We only observe N_j's for the clusters in the sample, that is, the empirical distribution of $(N_j|I_j = 1)$. Our proposed procedure simultaneously models the population cluster sizes and the outcome. Let x_i denote the auxiliary variables that are predictive for the outcome.

The observed data are $(y_{obs}, x_{obs}, N_{obs}, \overline{x}_{1:J}, N, J, J_s)$, where $\overline{x}_{1:J}$ is the cluster-level mean of the covariate x for all clusters $j = 1, \ldots, J$, and N, J, and J_s are the total population size, total number of clusters, and number of sampled clusters, respectively. The subscript *obs* denotes the observed portions of the variables: $y_{obs} = \{y_i \mid i = 1, \ldots, n_{j[i]}, j = 1, \ldots, J_s\}$, $x_{obs} = \{x_i \mid i = 1, \ldots, n_{j[i]}, j = 1, \ldots, J_s\}$, $N_{obs} = \{N_j \mid j = 1, \ldots, J_s\}$, where for convenience we number the sampled clusters $j = 1, \ldots, J_s$ and the nonsampled clusters as $j = J_s + 1, \ldots, J$.

The goal is to estimate the finite population mean $\overline{y}$, defined for a continuous outcome as

$$\overline{y} = \sum_{j=1}^{J} \frac{N_j}{N} \overline{y}_j = \frac{1}{N} \left(\sum_{j=1}^{J_s} \frac{n_j \overline{y}_{obs,j} + (N_j - n_j)\overline{y}_{exc,j}}{N_j} + \sum_{j=J_s+1}^{J} N_{exc,j} \overline{y}_{exc,j} \right),$$

where $\overline{y}_{obs,j}$ is the mean of the sampled units in sampled cluster j, $\overline{y}_{exc,j}$ is the mean of the nonsampled units in cluster j, and $N_{exc,j}$ is the size of nonsampled cluster j. For a binary outcome, the population proportion is

$$\overline{y} = \sum_{j=1}^{J} \frac{y_{(j)}}{N} = \frac{1}{N} \left(\sum_{j=1}^{J_s} \left(y_{obs,(j)} + y_{exc,(j)} \right) + \sum_{j=J_s+1}^{J} y_{exc,(j)} \right),$$

where $y_{(j)}$ is the total of all units in cluster j, $y_{obs,(j)}$ is the total of sampled units in sampled cluster j and $y_{exc,(j)}$ is the total of the binary outcome in nonsampled units in cluster j.

We assume the continuous survey outcome y is related to the covariate x and cluster sizes N_j in the following way:

$$y_i \sim \mathrm{N}(\beta_{0j[i]} + \beta_{1j[i]} x_i, \sigma_y^2) \tag{1.4}$$

$$\beta_{0j} \sim \mathrm{N}(\alpha_0 + \gamma_0 \log^c(N_j), \sigma_{\beta_0}^2) \tag{1.5}$$

$$\beta_{1j} \sim \mathrm{N}(\alpha_1 + \gamma_1 \log^c(N_j), \sigma_{\beta_1}^2) \tag{1.6}$$

$$N_j \sim p(N_j \mid \phi), \tag{1.7}$$

where ϕ are the parameters governing the distribution of the cluster sizes N_j. The model assumes the regression coefficients are cluster-varying and depend on the cluster sizes. We use random-effects model to borrow information across clusters. While fixed-effects model with cluster membership indicators can also be used to quantify the cluster effect, fixed cluster effects models may increase the variance, as shown by Reiter *et al.* (2006) and Andridge (2011) in the context of missing data imputation. In addition, predictions cannot be made for nonsampled clusters using fixed-effects models.

Our model for a binary outcome is identical, except that we modify (1.4) to be

$$\Pr(y_i = 1) = \text{logit}^{-1}(\beta_{0j[i]}) \tag{1.8}$$

and omit (1.6). We do not include a unit-level covariate in the binary case because the nonlinear nature of the inverse logit makes it challenging to make use of data at the unit level. Specifically, predicting $\overline{y}_{exc,j}$ requires knowing x_i for all nonsampled units in cluster j, and if we knew this, clearly we would also know N_j for nonsampled clusters j as well.

We use the centered logarithms of the cluster sizes $\log^c(N_j)$ as predictors; we work on the log scale to better accommodate large cluster sizes and center for interpretation convenience. The sampling is assumed to be ignorable after including the design variables in the outcome model. We assign an estimation model $p(N_j \mid \phi)$ to the cluster sizes, which we observe only for the sampled clusters. We develop both nonparametric and parametric modeling strategies to predict the cluster sizes of nonsampled clusters.

We use ψ to denote the regression parameters $\psi = (\alpha_0, \gamma_0, \sigma_0, \alpha_1, \gamma_1, \sigma_1, \sigma_y)$, ψ to denote the parameters of the cluster size distribution, and θ for all parameters of interest: $\theta = (\psi, \phi)$. The likelihood for the observed data is

$$p(y_{obs} \mid x_{obs}, N_{obs}, \theta) \propto p(y_{obs} \mid x_{obs}, N_{obs}, \psi)p(N_{obs} \mid \phi),$$

and the posterior distribution is

$$p(\theta \mid y_{obs}, x_{obs}, N_{obs}) \propto p(y_{obs} \mid x_{obs}, N_{obs}, \psi)p(N_{obs} \mid \phi)p(\psi)p(\phi),$$

where we assume that ψ and ϕ are independent, allowing us to write $p(\theta) = p(\psi)p(\phi)$.

Because of the independence and invariance assumptions in the two-stage cluster sampling, the distribution of the outcome y, given the design variables, is the same in the sample and the population; that is, the observed data likelihood is the same as the complete data likelihood,

$$p(y_{obs} \mid x_{obs}, N_{obs}, \psi) = p(y \mid x, N, I = 1, \psi) = p(y \mid x, N, \psi),$$

where $p(y \mid x, N, \psi)$ is specified by (1.4)–(1.6) for continuous y and by (1.5), (1.6), and (1.8) for binary y.

The challenge lies in estimating the distribution of the N_j's when the sampling is informative. Under PPS sampling, the probability of sampling a cluster of size N_j is $\Pr(I_j = 1|N_j) \propto N_j$, with the population distribution $p(N_j)$ of N_j as specified in (1.7). The probability of observing a cluster of size N_j in the PPS sample is then

$$\begin{aligned} p(N_j \mid I_j = 1) &\propto \Pr(I_j = 1 \mid N_j) p(N_j) \\ &\propto N_j p(N_j). \end{aligned} \tag{1.9}$$

We consider both nonparametric and parametric modeling strategies for the population distribution $p(N_j)$. First, we introduce the Bayesian bootstrap algorithm as a nonparametric approach to predicting the unobserved N_j's. Second, we investigate two parametric distributional assumptions for $p(N_j)$, the negative binomial and lognormal distributions. Here our goal is to directly model the distribution of the cluster sizes accounting for the fact that the observed distribution is biased from the complete population distribution. Following Patil and Rao (1978), we refer to these parametric choices as size-biased distributions.

Bayesian bootstrap

For a nonparametric model of the sampled cluster sizes, we take the Bayesian bootstrap algorithm in Little and Zheng (2007) that was modified by Zangeneh and Little (2015) for one-stage PPS sampling and apply it two-stage PPS sampling. Without making parametric assumptions about $p(N_j)$, this approach connects $p(N_j \mid I_j = 0)$ with $p(N_j \mid I_j = 1)$ through the empirical distributions under PPS sampling. Assume the N_j's observed for the sampled clusters have B unique values $N_1^*, \ldots, N_B^*$, and let $k_1, \ldots, k_B$ be the corresponding counts of these unique sizes, such that $\sum_b k_b = J_s$. Let ψ_b denote the probability of observing a cluster of size N_b^* in the sample: $\psi_b = \Pr(N_j = N_b^* \mid I_j = 1)$. We can then model the counts $k = (k_1, \ldots, k_B)$ as multinomi-

ally distributed with parameters $\psi = (\psi_1, \ldots, \psi_B)$. The observed likelihood is

$$\begin{aligned} & L_{obs}(\psi) \\ & \quad = \Pr\left(k_1 = \sum_{j=1}^{J_s} I(N_j = N_1^*), \ldots, k_B = \sum_{j=1}^{J_s} I(N_j = N_B^*) \mid I_j = 1, j = 1, \ldots, J_s\right) \\ & \quad \propto \prod_{b=1}^{B} \psi_b^{k_b}, \end{aligned}$$

where $I(\cdot)$ is an indicator function, $I(\cdot) = 1$ if the inside expression is true and 0 otherwise. The ψ's are given a noninformative Haldane prior: $p(\psi_1, \ldots, \psi_B) = \text{Dirichlet}(0, \ldots, 0)$. The posterior distribution of ψ is then

$$p(\psi_1, \ldots, \psi_B | k_1, \ldots, k_B) = \text{Dirichlet}(k_1, \ldots, k_B).$$

Assume the unique values of N_j's cover all possible values in the population. We let $k_b^\star$ denote the number of nonsampled clusters with size N_b^*, for $b = 1, \ldots, B$ and let $\psi_b^\star$ denote the probability of an unobserved cluster having size N_b^*: $\psi_b^\star = \Pr(N_j = N_b^* \mid I_j = 0)$. Then the counts of the B unique sizes among the nonsampled clusters, $(k_1^\star, \ldots, k_B^\star)$, follow a multinomial distribution with total $J - J_s$ and probabilities $(\psi_1^\star, \ldots, \psi_B^\star)$:

$$p(k_1^\star, \ldots, k_B^\star \mid J - J_s, \psi_1^\star, \ldots, \psi_B^\star) \propto \prod_{b=1}^{B} \psi^{\star k_b^\star}$$

Using Bayes' rule, we can write $\psi_b^\star$ as

$$\begin{aligned} \psi_b^\star &= \Pr(N_j = N_b^* \mid I_j = 0) \\ &\propto \Pr(N_j = N_b^* \mid I_j = 1) \frac{\Pr(I_j = 0 | N_j = N_b^*)}{\Pr(I_j = 1 | N_j = N_b^*)} \\ &= \psi_b \frac{1 - \pi_b}{\pi_b}, \end{aligned} \tag{1.10}$$

where $\pi_b = \Pr(I_j = 1 | N_j = N_b^*) = J_s N_b^* / N$ is the conditional cluster selection probability known in the PPS sample, J_s is the number of sampled clusters, and N is the population size. This approach adjusts the probability of resampling an observed

size N_b^* by the odds of a cluster of that size not being sampled, so that smaller sizes are upweighted relative to larger ones.

Given the posterior draws of $\psi_b^\star$'s and $k_b^\star$'s, we create k_b^* replicates of the size N_b^*, yielding a sample of the nonsampled cluster sizes from their posterior predictive distribution. The Bayesian bootstrap for cluster sampling is similar to the "two-stage Polya posterior" approach proposed by Meeden (1999), which simulates draws that form an entire population of clusters and then an entire population of elements within each cluster. Zhou *et al.* (2016) incorporated weights in Bayesian bootstrap for multiple imputation in two-stage cluster samples. Si *et al.* (2015) uses a similar approach to estimating the poststrafication cell sizes constructed by the survey weights.

The Bayesian bootstrap avoids parametric assumption on the population distribution $p(N_j)$ and use the empirical distribution in the observed clusters. However, this approach restricts the draws for the nonsampled cluster sizes to come from the set of observed cluster sizes, where small clusters may be omitted under PPS sampling. This implicitly introduces a noninformative prior distribution on N_j's. While the Bayesian bootstrap is a robust algorithm for predicting the unknown N_j's, we can achieve efficiency gains with a parametric distribution on $p(N_j)$, especially in combination with prior information.

Size-biased distributions

Inducing parametric sized-biased distributions follows the superpopulation concept in the model-based survey inference literature. Sized-biased distributions were considered by Patil and Rao (1978) for population size estimation. In practice, we may have some knowledge about the cluster sizes, such as the distribution in a similar population. We can incorporate this additional information through the prior distribution specification. We consider both a discrete and a continuous distribution as candidates for modeling the size distributions. Using (1.9), we can derive the observed likelihood based on the considered population distribution.

For the discrete case, we assume the population cluster sizes N_j follow a negative binomial distribution: $N_j \sim \text{NegBin}(k, p)$, with $k > 0$ and $p \in (0, 1)$. By normalizing the distribution in (1.9) and completing the algebra shown as below, we see that the sizes in the PPS sample can be written as $N_j = 1 + W_j$, where $W_j \sim \text{NegBin}(k+1, p)$ (Patil and Rao, 1978).

Let N_j denote the size variables in the population, $N_j \sim \text{NegBin}(k, p)$, with $k > 0$, $p \in (0, 1)$. For $m = 0, 1, 2, \ldots$, the probability of observing $N_j = m$ in the PPS sample is therefore

$$\begin{aligned}
\Pr(N_j = m \mid I_j = 1) &= \frac{\Pr(I_j = 1 \mid N_j = m)\Pr(N_j = m)}{\Pr(I_j = 1)} \\
&= \frac{m\binom{m+k-1}{m}p^k(1-p)^m}{\sum_{m=0}^{\infty} m\binom{m+k-1}{m}p^k(1-p)^m} \\
&= \frac{m\binom{m+k-1}{m}p^k(1-p)^m}{\mathbb{E}[N_j]} \\
&= \frac{m\binom{m+k-1}{m}p^k(1-p)^m}{(1-p)k/p} \\
&= \frac{((m-1)+(k+1)-1)!}{(m-1)!\,k!}p^{k+1}(1-p)^{m-1} \\
&= \binom{(m-1)+(k+1)-1}{m-1}p^{k+1}(1-p)^{m-1} \\
&= \Pr(W = m-1),
\end{aligned}$$

where $W \sim \text{NegBin}(k+1, p)$.

For the continuous case, we use the lognormal distribution. If the population distribution is $N_j \sim \text{LogNormal}(\mu, \tau^2)$, then $(N_j \mid I_j = 1) \sim \text{LogNormal}(\mu + \tau^2, \tau^2)$ (Patil and Rao, 1978). To see this, let $p(N_j)$ denote the pdf of the size variables N_j in the population and let $w > 0$ denote a particular realization of N_j. Then the pdf

of N_j in the PPS sample is

$$\begin{aligned} p(N_j \mid I_j = 1) &= \frac{\Pr(I_j = 1 \mid N_j) p(N_j)}{\Pr(I_j = 1)} \\ &= \frac{(1/\sqrt{2\pi\tau^2}) \exp\left(-\frac{(\ln w - \mu)^2}{2\tau^2}\right)}{\int_0^\infty (1/\sqrt{2\pi\tau^2}) \exp\left(-\frac{(\ln w - \mu)^2}{2\tau^2}\right) dw} \\ &= \frac{\exp\left(-\frac{(\ln w - \mu)^2}{2\tau^2}\right)}{\int_0^\infty \exp\left(-\frac{(\ln w - \mu)^2}{2\tau^2}\right) dw}. \end{aligned} \tag{1.11}$$

We can now simplify the denominator:

$$\begin{aligned} &\int_0^\infty \exp\left(-\frac{(\ln w - \mu)^2}{2\tau^2}\right) dw \\ &\quad = \exp\left(\mu + \frac{\tau^2}{2}\right) \int_0^\infty \exp\left(-\frac{(\ln w - (\mu + \tau^2))^2}{2\tau^2}\right) \frac{1}{w} dw \\ &\quad = \exp\left(\mu + \frac{\tau^2}{2}\right) \int_{-\infty}^\infty \exp\left(-\frac{(z - (\mu + \tau^2))^2}{2\tau^2}\right) dz \qquad \text{(substitute } z = \ln w\text{)} \\ &\quad = \sqrt{2\pi\tau^2} \exp\left(\mu + \frac{\tau^2}{2}\right) \end{aligned} \tag{1.12}$$

Now, substitute (1.12) for the denominator in (1.11):

$$\begin{aligned} p(N_j \mid I_j = 1) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln w - \mu)^2}{2\tau^2} - (\mu + \frac{\tau^2}{2})\right) \\ &= \frac{1}{w\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\ln w - (\mu + \tau^2))^2}{2\tau^2}\right). \end{aligned}$$

Thus, the distribution of sampled sizes in the PPS sample is $(N_j | I_j = 1) \sim \text{LogNormal}(\mu + \tau^2, \tau^2)$ (Patil and Rao, 1978).

Regardless of the parametric model we choose, in order to generate predictions of the nonsampled cluster sizes, we need to draw from $p(N_j \mid I_j = 0)$. Zangeneh *et al.* (2011b) give the following derivation for $p(N_j \mid I_j = 0)$ in the context of a PPS sample. Again denoting a realized value of N_j by w, let $p(w \mid \psi)$ denote the marginal density of the cluster sizes N_j indexed by parameters ψ and let $p(w, \iota \mid \psi)$

denote the joint density of cluster sizes N_j and the sampling indicator I_j. Clearly $p(w, 0 \mid \psi) + p(w, 1 \mid \psi) = p(w \mid \psi)$.

Under PPS sampling, $p(w, 1 \mid \psi) = cwp(w \mid \psi)$ for some constant c. The unconditional probability of selection is $\Pr(I_j = 1) = J_s/J$, where J_s is the number of sampled clusters and J is the total number of clusters in the population, so

$$\frac{J_s}{J} = \int p(w, 1 \mid \psi) d\nu(w) = c \int wp(w \mid \psi) d\nu(w) = c\ \mathbb{E}[N_j].$$

Then $c = J_s/(J\mathbb{E}[N_j])$ and

$$p(w, 1 \mid \psi) = \frac{J_s}{J\mathbb{E}[N_j]} wp(w \mid \psi).$$

Since $p(w, 0 \mid \psi) + p(w, 1 \mid \psi) = p(w \mid \psi)$, we can write

$$p(w, 0 \mid \psi) = p(w \mid \psi) - p(w, 1 \mid \psi) = \left(1 - \frac{Kw}{J\mathbb{E}[N_j]}\right) p(w \mid \psi). \tag{1.13}$$

As shown previously, the conditional density $p(w \mid 1, \psi)$ is

$$p(w \mid 1, \psi) = \frac{p(w, 1 \mid \psi)}{\Pr(I_j = 1 \mid \psi)} = \frac{wp(w \mid \psi)}{J_s/J}. \tag{1.14}$$

Combining (1.13) and (1.14), we get

$$p(w \mid 0, \psi) = \frac{p(w, 0 \mid \psi)}{1 - \Pr(I_j = 1 \mid \psi)} = \frac{J\mathbb{E}[N_j] - J_s w}{J\mathbb{E}[N_j]} \frac{1}{1 - J_s/J}\ p(w \mid \psi). \tag{1.15}$$

If we follow Zangeneh *et al.* (2011b) and make the assumption that $\mathbb{E}[N_j]$ is equal to the finite population mean cluster size N/J, where $N = \sum_{j=1}^{J} N_j$, then (1.15) simplifies to

$$p(w \mid 0, \psi) = \frac{N - J_s w}{N} \frac{1}{1 - J_s/J}\ p(w \mid \psi). \tag{1.16}$$

Given the posterior distribution of $p(w \mid \psi)$, we use rejection sampling to get posterior samples from $p(w \mid 0, \psi)$.

Bayesian Estimation

For the models specified in (1.4)-(1.6), we use the following weakly informative prior distributions as recommended by Gelman (2006),

$$\alpha_0, \gamma_0, \alpha_1, \gamma_1 \overset{ind}{\sim} \mathrm{N}(0, 10)$$
$$\sigma_{\beta_0}, \sigma_{\beta_1}, \sigma_y \overset{ind}{\sim} \mathrm{Cauchy}^{+}(0, 2.5).$$

Here $\mathrm{Cauchy}^{+}(0, 2.5)$ denotes a Cauchy distribution with location 0 and scale 2.5 restricted to positive values.

For the parameters governing the distribution of N_j, here (k, p) or (μ, τ), we can use noninformative priors when the number of clusters sampled is large. However, when only a few clusters are sampled, we need stronger priors to counteract the sparsity of the data. This is particularly true when using a model for the cluster sizes that includes implicit assumptions about the data, such as the negative binomial. As an overdispersed extension of the Poisson distribution, the negative binomial assumes that the data come from a distribution whose mean is larger than the variance. However, in a sample of only, say, five clusters, it may well be that the sample mean is less than the sample variance, making it difficult for Stan the negative binomial distribution to the data without strong prior information. We therefore reparameterize the negative binomial as a Gamma mixture of Poissons and place a prior on the coefficient of variation (CV), the standard deviation divided by the mean. In this case, the CV works out to the the reciprocal of the square root of the scale parameter of the Gamma distribution (**?**). With a small number of clusters, we expect the CV to be close to one and therefore use an exponential prior with rate 1. For the lognormal distribution, we place a $\mathrm{Cauchy}^{+}(0, 2.5)$ prior on the scale parameter τ.

In nonsampled clusters j, the posterior predictive distribution for $\overline{y}_{mis,j}$ is

$$(\overline{y}_{mis,j} \mid \cdot) \sim \mathrm{N}\left(\beta_{0j} + \beta_{1j}\overline{x}_j, \sigma_y^2/N_j\right),$$

where we assume $\overline{x}_j$ is known. Specifically, we draw new values of β_{0j}, β_{1j}, σ_y, and N_j from their posterior distributions and then draw $\overline{y}_{mis,j}$ from the above distribution.

In sampled clusters, the posterior predictive distribution for the nonsampled units is

$$\overline{y}_{exc,j} \sim N\left(\beta_{0j} + \beta_{1j}\overline{x}_j, \sigma_y^2/(N_j - n_j)\right).$$

When N_j is large compared to n_j, as is the case in many large-scale surveys and specifically in the Fragile Families survey, $\overline{y}_{exc,j}$ is close to the cluster mean $\overline{y}_j$ and is well approximated by $\beta_{0j} + \beta_{1j}\overline{x}_j$, which we calculate using the posterior means of β_{0j} and β_{1j}.

The posterior computation is implemented in Stan (Stan Development Team, 2016b), which conducts full Bayesian inference and generates the posterior samples. The models for the outcome and the cluster sizes are integrated into the posterior computation, which allows for uncertainty in both the outcome and cluster size models to be propagated throughout the parameter estimates, in contrast to other approaches (e.g., Little and Zheng, 2007; Zangeneh and Little, 2015).

To understand the importance of explicitly controlling for all design variables in this context, we also fit a model similar to (1.4)–(1.7) but with γ_0 and γ_1 set to 0. Such a model accounts for the hierarchical cluster nature of the data by allowing β_0 and β_1 to vary by cluster, but does not account for the sampling design since the cluster sizes N_j are excluded from the model:

$$\begin{aligned} y_i &\sim \mathrm{N}(\beta_{0j[i]} + \beta_{1j[i]}x_i, \sigma_y^2) \qquad \text{(continuous)} \\ \Pr(y_i = 1) &= \mathrm{logit}^{-1}(\beta_{0j[i]}) \qquad \text{(binary)} \\ \beta_{0j} &\sim \mathrm{N}(\alpha_0, \sigma_{\beta_0}^2) \\ \beta_{1j} &\sim \mathrm{N}(\alpha_1, \sigma_{\beta_1}^2) \end{aligned} \tag{1.17}$$

1.3 Simulation study

Design

We perform a simulation study to compare the performance of our integrated model and classical design-based estimators. We generate a fixed population from which

we take repeated two-stage cluster samples under PPS and use each of the methods to estimate $\overline{y}$. In this way, we compare the approaches on the statistical validity of the finite population inference. We generate a population consisting of $J = 100$ clusters, with cluster sizes N_j drawn from either a Poisson distribution with rate 500 or a multinomial distribution over scaled Gamma-distributed sizes. Specifically, we draw $J = 100$ candidate cluster sizes N_j as $N_j = 100W_j$, where $W_j \sim \text{Gamma}(10, 1)$. We then take a multinomial draw from these 100 unique sizes, with the J-vector of probabilities drawn from a Dirichlet distribution with concentration parameter 10, which disperses probability mass roughly equally across the $J = 100$ components. In both cases, to avoid clusters that would be selected with probability 1, we resample the J cluster sizes until none are so large as to be selected with certainty.

For the continuous outcome variable, we draw a value y_i for each unit in the population from the following model:

$$\begin{aligned}
y_i &\sim \text{N}(\beta_{0j[i]} + \beta_{1j[i]} x_i, \sigma_y^2) \\
\beta_{0j} &\sim \text{N}(\alpha_0 + \gamma_0 \log^c(N_j), \sigma_{\beta_0}^2) \\
\beta_{1j} &\sim \text{N}(\alpha_1 + \gamma_1 \log^c(N_j), \sigma_{\beta_1}^2) \\
\alpha_0, \alpha_1, \gamma_0, \gamma_1 &\sim \text{N}(0, 1) \\
\sigma_{\beta_0} &\sim \text{N}^+(0, 0.5) \\
\sigma_{\beta_1} &\sim \text{N}^+(0, 0.5) \\
\sigma_y &\sim \text{N}^+(0, 0.75),
\end{aligned} \tag{1.18}$$

where $\text{N}^+(\mu, \sigma)$ denotes the positive part of the normal distribution with mean μ and standard deviation σ. The model for binary y is identical, except that the first line of (3.1) is replaced with $y_i \sim \text{Bern}(\text{logit}^{-1}(\beta_{0j[i]}))$ (and we omit β_{1j}). We generate x_i by sampling from the discrete uniform distribution between 20 and 45 (as might be for a survey of reproductive-age women, for example) and center it by subtracting the mean.

To understand how the performance of various models is affected by dependence

between the outcome y and the cluster size N_j, we generate two populations using the above model. In the first, we draw γ_0 and γ_1 from normal distributions as shown above, and in the second we set $\gamma_0 = 0 = \gamma_1$. In this latter case, there is no population-level dependence between the outcome y and the cluster size N_j. Comparing the results between these two population structures allows us to evaluate the importance of including design variables and cluster sizes in each of the candidate models. Specifically, comparing the results from fitting the model in (1.17) to a population where $\gamma_0 \neq 0 \neq \gamma_1$ to one where $\gamma_0 = 0 = \gamma_1$ will show what happens when we incorrectly omit the cluster sizes from the model. Similarly, comparing the results from fitting the model in (1.4)-(1.5) and assuming all population cluster sizes are known will show what happens when we include cluster sizes in the model, even when the outcome is independent of the sizes.

We assume that x_i is known for all sampled units, and that $\overline{x}_j$ is known for all clusters. If x is a demographic covariate, in practice it's often the case that we know demographic characteristics of clusters even if we don't know the cluster size. We also assume that the total population size $N = \sum_{j=1}^{J} N_j$ and N_j's in the sampled clusters are known, but N_j's for the nonsampled clusters are not known.

We sample $J_s < J$ clusters using random systematic PPS sampling with probability proportional to the cluster size N_j and n_j units in each selected cluster j via SRS. We consider values of $J_s \in \{10, 50\}$ and $n_j \in \{0.1N_j, 0.5N_j, 10, 50\}$. Note that when $n_j \in \{10, 50\}$, the sample is self-weighting, meaning each unit has an equal probability of selection. To see this, recall that the probability of sampling cluster j is $\pi_j \propto N_j$. Since within-cluster sampling is done with SRS, the probability of sampling unit i in cluster j given that cluster j is selected is $\pi_{i|j} = n_j/N_j = n/N_j$ since n_j is the same for all clusters. The marginal probability of sampling unit i is therefore $\pi_i = \pi_j \pi_{i|j} \propto N_j \cdot (n/N_j) = n$, which is constant across units and clusters.

For each combination of J_s and n_j, we draw $L = 100$ two-stage samples from the finite population. For each two-stage sample, we then estimate the finite population

mean using the methods described below. We emphasize that we draw L samples from one fixed finite population instead of one sample from each of L finite populations, because our goal is to evaluate the performance of each method in terms of finite population inference.

We use the following methods to estimate the finite population mean $\overline{y}$.

- `negbin`: The negative binomial size-biased distribution as described in Section 1.2;
- `lognormal`: The lognormal size-biased distribution as described in Section 1.2;
- `bb`: The Bayesian bootstrap as described in Section 1.2;
- `hajek`: The Hájek estimator in (1.3);
- `greg`: The generalized regression estimator (Deville and Särndal, 1992), which leverages a unit-level covariate to improve prediction. We only use this estimator for continuous y. To estimate the variances of the Hájek and generalized regression estimators, we use the formulas given in Chapter 8 of Särndal *et al.* (1992);[1]
- `cluster_inds`: The model in (1.17), which accounts for the hierarchical nature of the data via random cluster effects but does not use the cluster sizes as a cluster-level predictor in modeling β_{0j} and β_{1j}, and therefore does not fully account for the sampling design. We expect this model to perform well in the cases where $\gamma_0 = \gamma_1$ in the data generation model and when the sample is self-weighting;

[1]In some cases, the sample size is so large as to make calculating the design-based variance under a non-self-weighting design difficult. This is due to the $\check{\Delta}_{k\ell}$ term in equations 8.6.3 and 8.9.27 in Särndal *et al.* (1992), which requires generating an $n \times n$ matrix, where $n = \sum_{j=1}^{J_s} n_j$. When $J_s = 50$ and $n_j = 0.5N_j$, n can easily be 20000 or larger, making the matrix prohibitively large to compute. In these cases, we estimate the variance by randomly selecting 100 units via SRS in each sampled cluster and using those units to compute the required matrix.

- `knowsizes`: The model in (1.4)—(1.6), where we additionally assume the cluster sizes are known for the entire population. This is a best-case scenario and will serve as a benchmark for the other Bayesian methods;

There are three main comparisons that we make in evaluating the results of the simulation study. First, we measure the performance of our proposed integrated Bayesian approach against that of the classical design-based estimators; we do this by comparing the performance of `negbin`, `lognormal`, and `bb` to that of `hajek` and `greg`. Second, among the Bayesian methods, we want to understand when the parametric models `negbin` and `lognormal` outperform the nonparametric Bayesian bootstrap `bb`. Third, we compare the performances of both `cluster_inds` and `knowsizes` when $\gamma_0 \neq 0 \neq \gamma_1$ to when $\gamma_0 = 0 = \gamma_1$ in order to understand the importance of explicitly including cluster sizes as cluster-level predictors in (1.5) and (1.6). In this case, we assume that cluster sizes are known for all clusters in the population and focus on the effects of incorrectly excluding or including the cluster sizes as cluster-level predictors in the model.

We carefully monitor the sampling diagnostics for each simulation. Stan is unique in providing detailed warnings and diagnostics to inform the user when posterior inferences may be unreliable due to difficulties in sampling from the posterior. Divergent transitions indicate that the sampler is unable to explore a portion of the parameter space, which can lead to significant bias in the resulting posterior distribution and ultimately unreliable inferences (Stan Development Team, 2016c). Stan reports the number of divergent transitions for each chain, and even one divergent transition indicates that the results may be suspect (Stan Development Team, 2016a). If divergent transitions occur, we follow the recommendation of Stan developers and iteratively increase the target acceptance rate `adapt_delta` (Stan Development Team, 2016a). If divergent transitions occur even with `adapt_delta` $= 0.99999$, we switch to the noncentral parameterization and follow the same procedure for increasing adapt delta as necessary. The noncentral parameterization is a mathematically equivalent

formulation for the model that can avoid posterior geometries that are difficult for HMC to explore; see Betancourt and Girolami (2013) and Stan Development Team (2016c). If divergent transitions remain, we discard the simulation.

We also monitor the estimated potential scale reduction factor $\widehat{R}$ for each parameter. This diagnostic assesses the mixing of the chains; at convergence, $\widehat{R} = 1$. If $\widehat{R} \geq 1.1$ for any parameter, we increase the number of iterations by 1000 until all values of $\widehat{R}$ are less than 1.1, up to 4000 iterations. If values of $\widehat{R} \geq 1.1$ remain with 4000 iterations, we discard the simulation. The results presented here are based on a minimum of 85 simulations for each combination of number of clusters sampled, number of units sampled, and whether $\gamma_0 = 0 = \gamma_1$ or not.

Results

The results of the simulation study are in Figures 1.1 to 1.4, with each figure displaying a different combination of outcome type (continuous or binary) and population cluster size model (Poisson or multinomial). In each figure, there are six panels displaying the six metrics with which we evaluate the methods: relative bias, relative root mean squared error (RRMSE), coverage of 50% and 95% uncertainty intervals, and the average relative widths of the 50% and 95% uncertainty intervals. The first four of these describe the performance of the point estimator, while the coverage rates and relative widths of the uncertainty intervals help evaluate the efficiency of each method; ideally, a method will have high (or close to nominal) coverage rates and narrow average uncertainty intervals. The relative bias is calculated as $\frac{1}{L}\sum_{\ell=1}^{L}\frac{\overline{y}-\widehat{\overline{y}}_\ell}{\overline{y}}$, where $\overline{y}$ is the true population mean, $\widehat{\overline{y}}_\ell$ is the estimated value from the ℓ-th simulation, and L is the number of simulations. (We aim for $L = 100$, but this is not achieved in every instance as explained above. However, $L \geq 85$ in all cases.) RRMSE is calculated as $\sqrt{\frac{1}{L}\sum_{\ell=1}^{L}\left(\frac{\overline{y}-\widehat{\overline{y}}_\ell}{\overline{y}}\right)^2}$. For the Bayesian methods `negbin`, `lognormal`, `bb`, `cluster_inds` and `knowsizes`, the 50% (95%) intervals are calculated from the 25th and 75th (2.5th and 97.5th) percentiles of the posterior predictive distribution for $\overline{y}$. For the classical

design-based methods `hajek` and `greg`, we rely on asymptotic normal theory and the variance estimators given in Results 8.6.1 and 8.9.2 of Särndal *et al.* (1992). The relative widths of the uncertainty intervals are calculated by dividing the width of the uncertainty interval by the true $\overline{y}$ and averaging across the L simulations.

In each panel, the top row of plots is for the case where $\text{Corr}(y, N_j) \neq 0$ (i.e. $\gamma_0 \neq 0 \neq \gamma_1$), labeled "Dependent", and the bottom row is for $\text{Corr}(y, N_j) = 0$ (i.e. $\gamma_0 = 0 = \gamma_1$), labeled "Independent". The columns are for different within-cluster sampling schemes. The left two columns represent fixed-percentage schemes, where $n_j = \rho N_j$ for $\rho = 0.1$ and $\rho = 0.5$, $j = 1, \ldots, J_s$. The right two columns represent the self-weighting samples, with $n_j = 10$ and $n_j = 50$, $j = 1, \ldots, J_s$. The colors of the circles represent different first-stage sample sizes J_s, $J_s \in \{10, 50\}$.

As described earlier, there are three main comparisons we make in evaluating the results: Bayesian vs. design-based methods, parametric vs. nonparametric models, and including vs. excluding the cluster sizes as cluster-level predictors. We now describe the results for these three comparisons for each combination of outcome type (continuous and binary), population cluster size model (Poisson and multinomial distributions), and whether .

For continuous y, the Bayesian models outperform the design-based estimators, both for the Poisson and the multinomially distributed population cluster sizes in Figures 1.1 and 1.2, respectively. The Hajek estimator has surprisingly high bias, particularly when the sample is self-weighting (right two columns in each panel), but including auxiliary information via the GREG estimator helps reduce it. The classical estimators yield unstable results, evident in their high RRMSE, particularly when y_i and N_j are independent (bottom row of plots in each panel) in Figure 1.1 and when they are dependent (top row) in Figure 1.2. In Figure 1.2, the RRMSEs and uncertainty interval lengths are much larger for the design-based estimators compared to the Bayesian methods when y_i and N_j are dependent (top row), but when they are independent this difference largely disappears. Overall, the Bayesian methods are

clearly preferable to the design-based estimators: they are lower in bias and RRMSE, and yield short uncertainty intervals whose coverage rates are very close to or above the nominal level.

Estimating the finite population proportion for binary y is somewhat simpler because $\Pr(y_i = 1)$ is constant within cluster in (1.8). In this case, there is generally little difference between the best Bayesian method and the Hajek estimator when the number of sampled clusters is large, $J_s = 50$; this holds for both the Poisson-distributed cluster sizes in Figure 1.3 and the multinomially distributed cluster sizes in 1.4. When the number of sampled clusters is small, the Hajek estimator generally outperforms the Bayesian methods in terms of bias and has comparable RRMSE. However, the coverage rates for the Hajek estimator are often below the nominal level, particularly when the sample is not self-weighting (left two plots in each panel).

The parametric models `negbin` and `lognormal` perform comparably to the non-parametric `bb` for continuous y. While both are about equally unbiased in Figures 1.1 and 1.2, particularly when the number of sampled clusters J_s is large, coverage is generally highest for `lognormal`. RRMSE and uncertainty interval lengths are the same for the parametric and nonparametric models. For binary y, there is again little difference between the parametric and nonparametric models when J_s is large. For small J_s, `bb` has high bias when the sample is self-weighting in Figure 1.1 and when y_i and N_j are dependent (top row in each panel) in Figure 1.2. In coverage rates and especially in RRMSE and uncertainty interval lengths, the parametric and nonparametric models again perform equally well.

Interestingly, incorrectly omitting cluster sizes as cluster-level predictors – that is, using `cluster_inds` instead of `knowsizes` – has little impact when y is continuous in either Figure 1.1 or 1.2. The bias, RMSE, and coverage rates are for the two methods are very similar for both Poisson- and multinomially distributed cluster sizes. The differences between `cluster_inds` and `knowsizes` are quite minor for binary y as well; `cluster_inds` does not perform appreciably worse than `knowsizes` in either

Figure 1.3 or 1.4.

1.4 Application

To evaluate the performance of our method in a more realistic context, we use a modified version of the Fragile Families study design in conjunction with a simulated outcome to estimate the finite population mean/proportion. The Fragile Families study design (Reichmann *et al.*, 2001) categorized the 77 U.S. cities with 1994 populations of 200,000 or greater into nine strata based on their policy environments and labor markets. Eight of the strata were for cities with extreme values in at least one of the three policy dimensions under consideration (labor markets, child support enforcement, and welfare generosity), and the ninth stratum was for cities that had no extreme values. One city was selected via PPS in each of the eight extreme strata, with a target sample size of 325 births in each city. In the last stratum, eight cities were selected via PPS, with a target sample size of 100 births in each. (There was an intermediate stage of selecting hospitals, which we ignore for our purposes; see Reichmann *et al.* (2001) for exact details of the Fragile Families study design.)

We generate a population of cities with sizes equal to the observed 1994 populations of the 77 cities in the Fragile Families sampling frame. For the purposes of this simulation, we use the city populations (divided by 100 for computational convenience) as both the measure of size M_j and the number of units in the cluster, N_j, though the ultimate unit of sampling in the study was births. We exclude the three cities that would be selected with probability one for a total of $J = 74$ cities. For each unit in the population, we create an outcome y according to our model in (3.1). While the original Fragile Families sampling design involved nine strata, we combine them into a single stratum. As in the actual study design, we sample 16 cities via PPS. In each sampled city, we sample either 100 or 325 births, depending on whether the city is a large- or small-sample city, as designated in Reichmann *et al.* (2001).

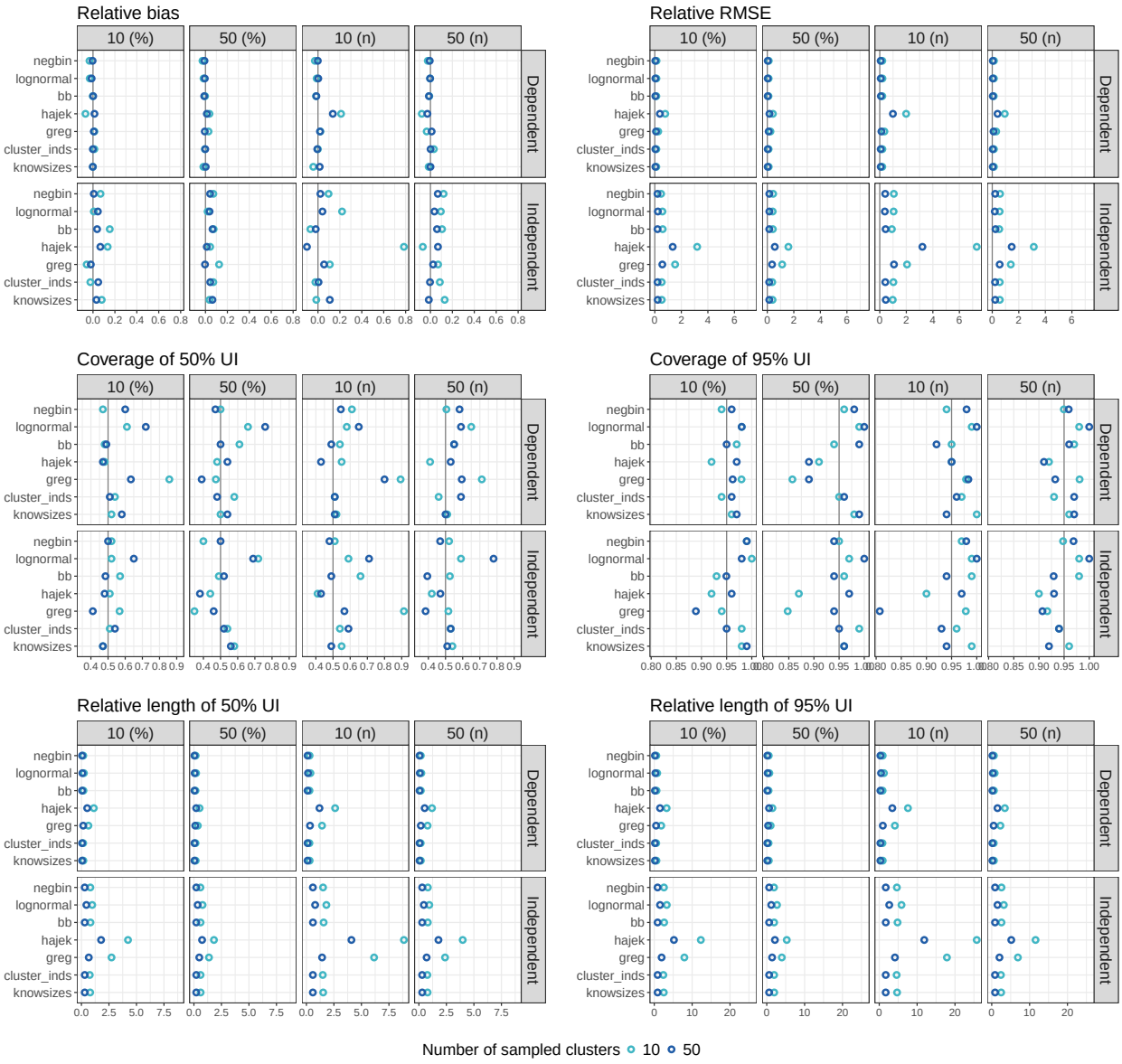


Figure 1.1: Results for continuous y with cluster sizes N_j drawn from a Poisson distribution.

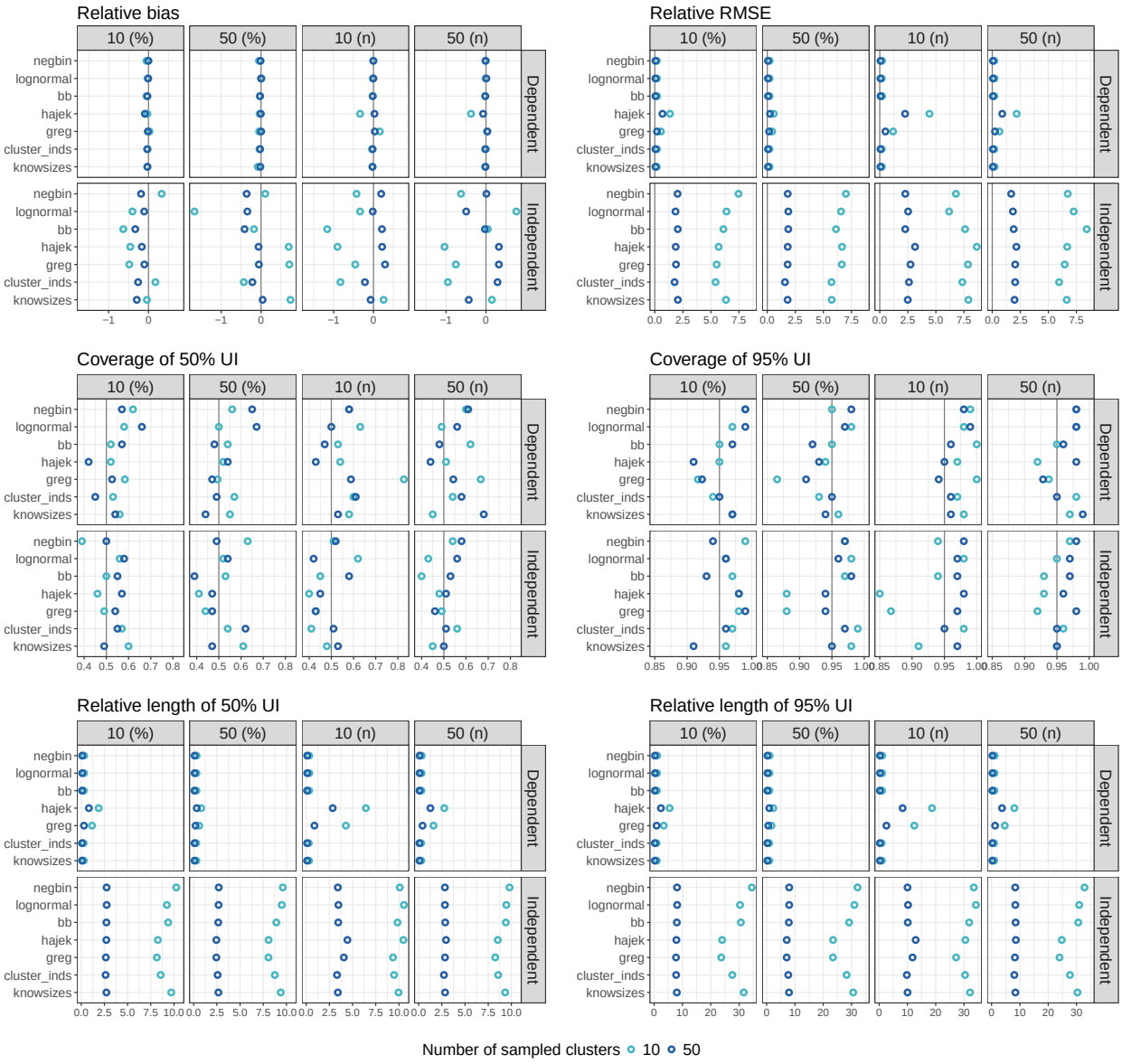


Figure 1.2: Results for continuous y with cluster sizes N_j drawn from a multinomial distribution.

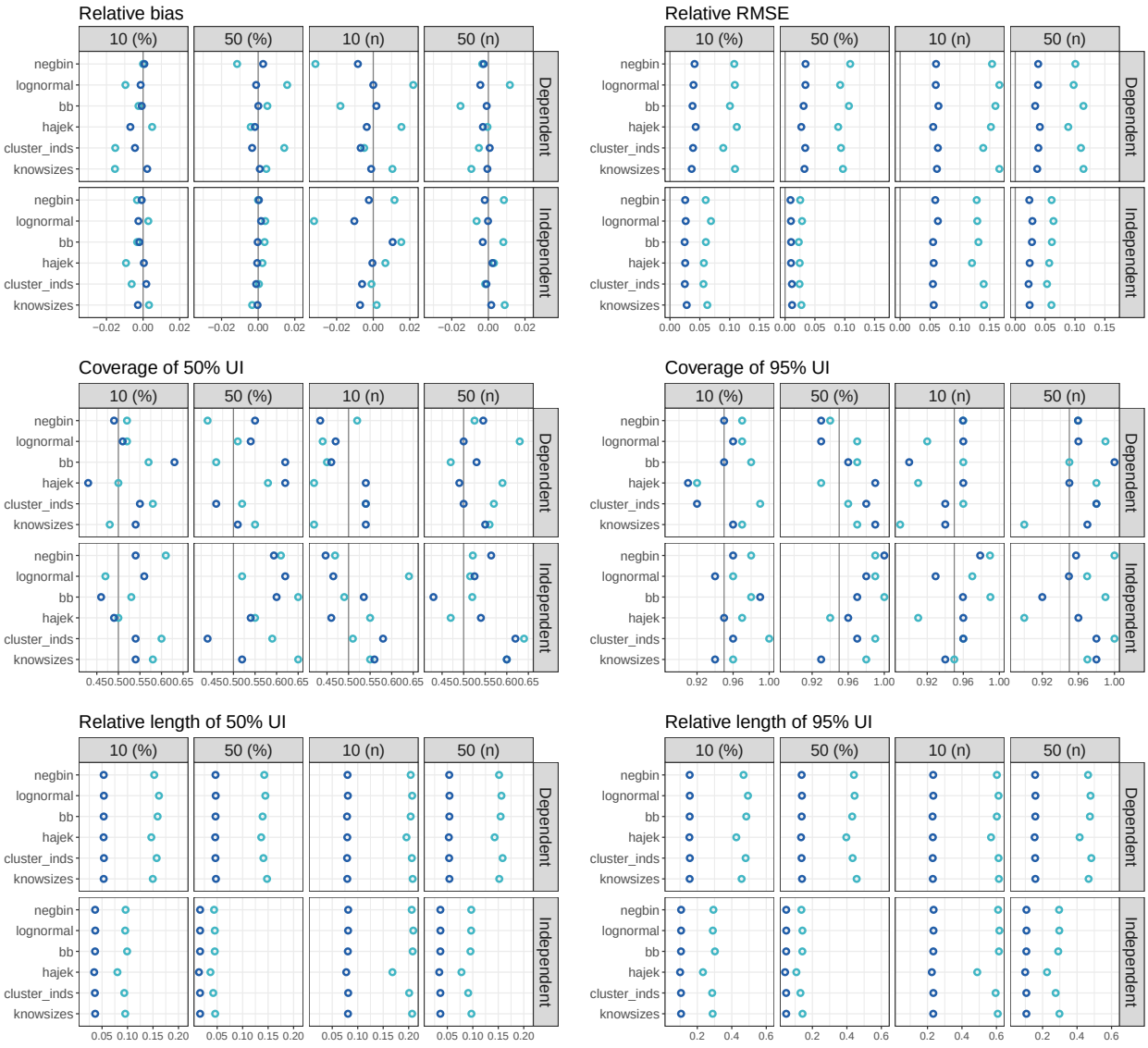


Figure 1.3: Results for binary y with cluster sizes N_j drawn from a Poisson distribution.

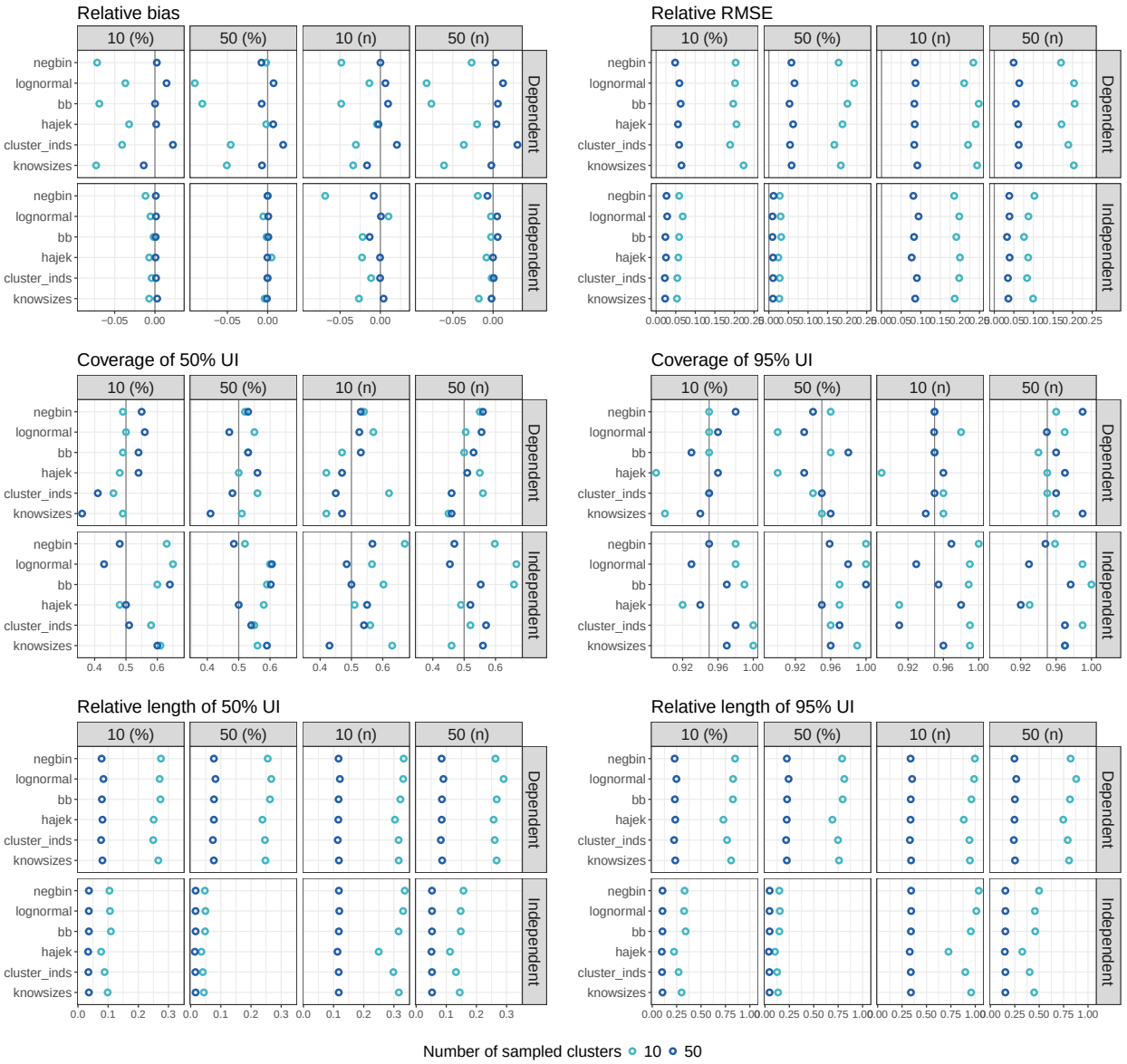


Figure 1.4: Results for binary y with cluster sizes N_j drawn from a multinomial distribution.

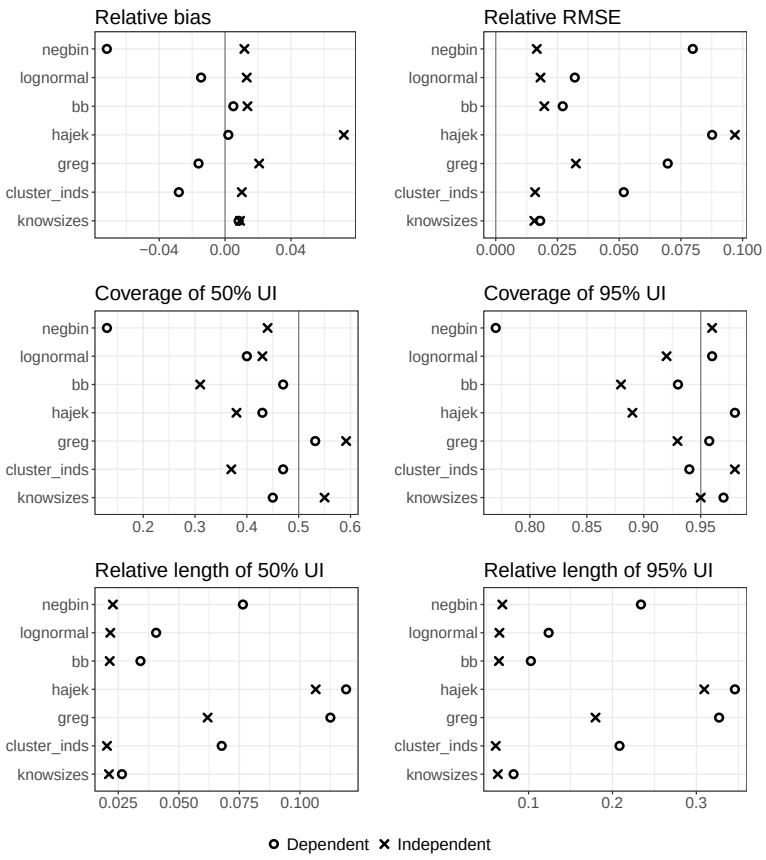


Figure 1.5: Results for continuous y with cluster sizes N_j taken from the Fragile Families study design.

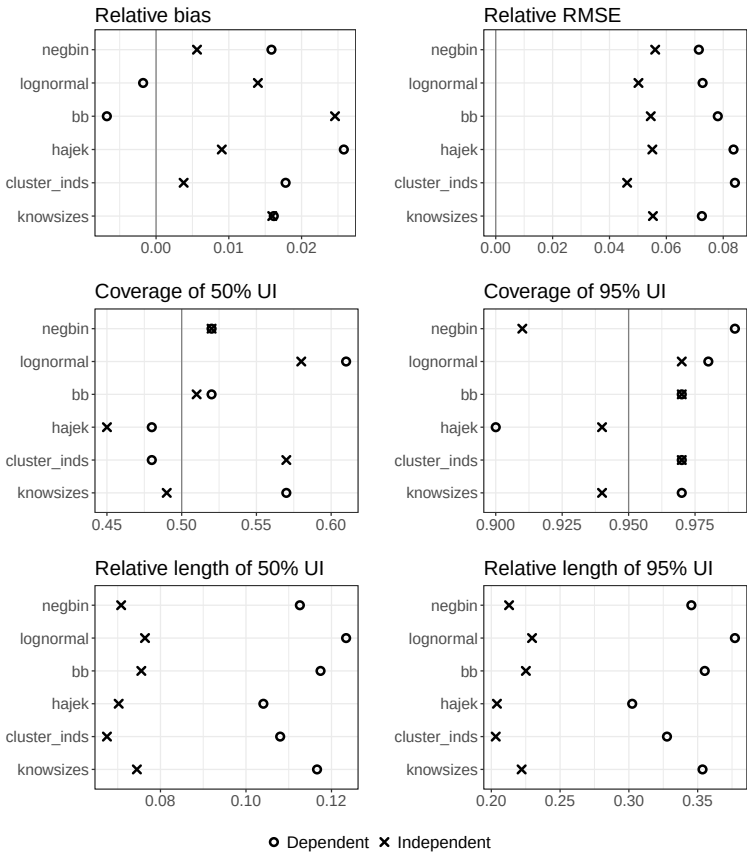


Figure 1.6: Results for binary y with cluster sizes N_j taken from the Fragile Families study design.

Figures 1.5 and 1.6 show the results for estimating the finite population mean and proportion, respectively. As before, they display relative bias, RRMSE, and coverage rates and relative widths of 50% and 95% uncertainty intervals. The circles are for the case when y_i and N_j are dependent and the crosses are for when y_i and N_j are independent.

For continuous y in Figure 1.5, the performance of several of the methods varies greatly depending on whether y_i and N_j are dependent or independent. When they are dependent, `negbin` performs poorly on every metric. The other Bayesian and classical methods are comparable in terms of bias, but the RRMSE of the Bayesian methods is much smaller. The coverage rates are close to the nominal levels, though `lognormal` and `hajek` have slightly low rates for 50% intervals. The Bayesian methods are more efficient, however, because they yield much shorter uncertainty intervals than the design-based methods. When y_i and N_j are independent, the Bayesian methods are clearly preferable to the design-based ones. Bias and RRMSE are lower for all three Bayesian methods than for `hajek` and `greg`, and while the coverage rates of the Bayesian methods are slightly below the nominal level, they again yield much shorter uncertainty intervals.

While the performance of `negbin` is very poor when y_i and N_j are dependent, it is the best of the Bayesian methods when they are independent with the lowest bias and RRMSE, highest coverage, and uncertainty intervals that are barely longer than those of `lognormal` and `bb`. On the other hand, when y_i and N_j are dependent, the nonparametric `bb` is the best Bayesian method.

Similarly, the effect of incorrectly omitting cluster sizes from the model with `cluster_inds` compared to `knowsizes` depends on whether y_i and N_j are independent. When they are, there is no difference between the two, except for the low 50% interval coverage rate of `cluster_inds`. When y_i and N_j are dependent, omitting the cluster sizes leads to clearly poorer performance: `cluster_inds` has higher bias and RRMSE and longer uncertainty intervals than `knowsizes`.

For binary y in Figure 1.6, the Bayesian methods again outperform the design-based `hajek`. The bias of `hajek` is particularly high when y_i and N_j are dependent, and its coverage rates are consistently below the nominal levels (though, it should be noted, the undercoverage is quite small for all methods in Figure 1.6). The difference between the Bayesian methods and `hajek` is smaller when y_i and N_j are independent. Among the Baysian methods, `lognormal` has consistently the best performance across the six metrics. For independent y_i and N_j, `negbin` has the lower bias and shortest uncertainty intervals, though its 95% uncertainty intervals have slightly lower than nominal coverage rates. Omitting cluster sizes from the model when y_i and N_j are dependent does not have drastically negative effects, but including the cluster sizes via `knowsizes` when y_i and N_j are independent leads to higher bias, RRMSE, poorer coverage, and longer uncertainty intervals.

1.5 Discussion

We propose an integrated Bayesian model for estimating the finite population mean from a two-stage PPS sample. Our method combines predicting measures of size for nonsampled clusters with inference for the population mean into a single approach that propagates uncertainty from both steps. We propose both parametric and nonparametric models for cluster sizes. The parametric models directly account for the unequal selection probabilities by using the closed-form size-biased version of the underlying population distribution, while the nonparametric Bayesian bootstrap draws from the observed cluster sizes with probabilities that are weighted by the odds of that cluster not being selected.

While design-based approaches are common in survey inference, estimating the variance of design-based estimators is often challenging. Current approaches include various jackknife methods (Wolter, 2007; Zheng and Little, 2005; Chen *et al.*, 2010), the Brewer method in the R `survey` package (Lumley, 2004), and the analytical

expressions for variances of design-based estimators under two-stage PPS sampling derived in Särndal *et al.* (1992) and used in our simulation study. In contrast, our integrated approach yields the full posterior distribution for the finite population mean, from which uncertainty intervals, variances, and any other quantities of interest can easily be computed.

In our simulation study, the Bayesian methods outperform the design-based estimators, particularly for continuous y and when the number of sampled clusters is small. The design-based estimators often have very high RRMSE and low rates of coverage for uncertainty intervals. They also have high bias for continuous y when y_i and N_j are independent, which is surprising given that one of the main advantages of design-based estimators is their design-consistency and approximate unbiasedness (Särndal *et al.*, 1992). For binary y, the Bayesian methods were less clearly superior to classical methods in estimating the finite population proportion. However, when the cluster sizes are highly skewed, as in the Fragile Families case, the Bayesian methods were decidedly better, particularly in terms of bias and coverage.

The performance of the parametric methods `negbin` and `lognormal` is largely comparable to that of the nonparameteric Bayesian bootstrap `bb`. One important factor in favor of the parametric methods is that they are simpler to implement in Stan, which makes them more accessible to researchers whose expertise is in areas outside of statistics or programming. Because the results for Bayesian vs. design-based and parametric vs. nonparametric methods are much more similar when $J_s = 50$ than when $J_s = 10$ in many of the scenarios our simulation study considered, we recommend using the parametric methods, at least as an initial step.

For continuous y when $J_s = 10$, RRMSE and uncertainty interval lengths are much larger in Figure 1.2, across all methods, when y_i and N_j are independent than when they are dependent. In the dependent case,

$$\mathbb{E}[y_i \mid \overline{x}_j, N_j] = (\alpha_0 + \alpha_1 \overline{x}_j) + (\gamma_0 + \gamma_1 \overline{x}_j) \log(N_j), \tag{1.19}$$

and in the independent case where $\gamma_0 = 0 = \gamma_1$, (1.19) simplifies to

$$\mathbb{E}[y_i \mid \overline{x}_j, N_j] = \alpha_0 + \alpha_1 \overline{x}_j. \tag{1.20}$$

In populations with clustered data, larger clusters naturally contribute more toward the population mean; this fact is part of the rationale for using PPS sampling in the first place. When y_i and N_j are dependent as in (1.19), larger clusters dominate the finite population mean not just in the sense of contributing more units, but also in the sense of having larger values of y_i in the first place. In addition, we can see in Figure 1.7 that the spread of population cluster sizes is much larger for multinomially distributed cluster sizes than for Poisson-distributed cluster sizes; in the latter, the variance by definition equals the mean, whereas in the former, the cluster sizes are selected from 100 unique sizes drawn from a scaled Gamma(10, 1) distribution whose variance is much larger than their mean. Repeated PPS sampling from the multinomial population is therefore more likely to sample the largest clusters more often than repeated PPS sampling from the Poisson cluster sizes. This combination of PPS sampling and the dependence of y_i on N_j may explain why, in Figure 1.2, bias, RRMSE, and uncertainty interval length are smaller for the case where y_i and N_j are dependent than when they are independent.

We do not see this pattern for the Fragile Families simulation; the magnitudes of RRMSE and uncertainty interval length are much smaller in Figure 1.5 than in Figures 1.1 and 1.2. However, here the specific structure of the Fragile Families population is important. The population sizes for the Fragile Families clusters are even more skewed than for the multinomial case (see Figure 1.7a), leading to more skewed cluster selection probabilities. However, we sample a larger fraction of the Fragile Families cities (16 out of 74, about 22%) than in the case where $J_s = 10$ for the multinomial cases (10 out of 100), and under repeated sampling from the fixed population, we end up sampling more of the smaller clusters in the Fragile Families scenario than for the multinomial ones. Thus, for the Fragile Families case, we are highly likely to sample the largest clusters that contribute most to the mean in terms

of number of units and, when y_i and N_j are dependent, in the magnitude of y_i, and we are also more likely than in the multinomial case to sample the smallest clusters under repeated sampling, both of which contribute to better estimates of the finite population mean.

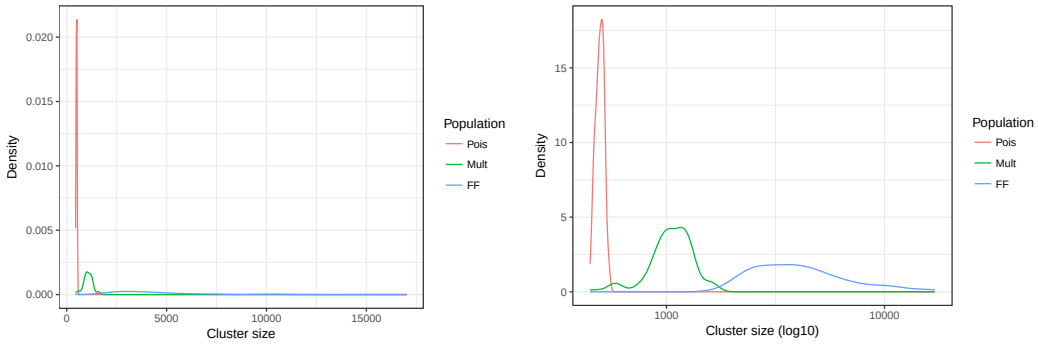
In contrast to the continuous case in Figure 1.2, when y is binary, RRMSE and uncertainty interval lengths are larger in both Figures 1.3 and 1.4 when y_i and N_j are dependent. When y_i and N_j are dependent,

$$\Pr(y_i = 1) = \text{logit}^{-1}\left(\alpha_0 + \gamma_0 \log(N_j) + \eta_j\right), \tag{1.21}$$

where $\eta_j \sim N(0, \sigma^2_{\beta_0})$. It may be that the nonlinear relationship between N_j and $\Pr(y_i = 1)$ amplifies any error in predicting the cluster sizes in addition to error in the estimated regression coefficients, making the estimates of the finite population proportion much more variable than when y_i and N_j are independent and $\gamma_0 = 0$. On the other hand, this nonlinearity may also account in part for why the magnitudes of bias, RRMSE, and uncertainty interval lengths are so much larger for continuous y than binary y. Even if our predictions for the smallest N_js are poor, the decreasing slope of the inverse logit as a function of $\log(N_j)$ means that errors in small values of N_j lead to smaller errors in $\Pr(y_i = 1)$ than for large N_j; for continuous y, we have no such cushion.

In addition to comparing the performance of Bayesian vs. design-based and parametric vs. nonparametric methods, our simulation also explored the importance of explicitly including or excluding design information in the model when we know the design variables for the entire population. Specifically, we considered the importance of including cluster sizes N_j as predictors for the cluster-level parameters β_{0j} and β_{1j} when the population is generated such that there is no relationship between β_{0j} and β_{1j} and N_j (and hence none between y and N_j) but the sampling is PPS in both cases.

Conventional model-based wisdom says to include all relevant design variables in the model, but the results of our simulation study suggest that allowing β_{0j} and β_{1j} to



(a) Original scale. (b) Log10 scale.

Figure 1.7: Density plot of population cluster sizes drawn from a Poisson distribution with rate 500 (Pois), a Gamma/multinomial distribution (Multi) as detailed in Section 1.3, and the Fragile Families study design (FF). The x-axis is on the original scale in the left panel and the log10 scale in the right panel.

vary by cluster without explicitly including the cluster sizes (`cluster_inds`) does not lead to drastically worse results than when the sizes are included in the model, even when y_i and N_j are dependent, though following conventional wisdom and including the cluster sizes anyway (`knowsizes`) does not hurt. This result naturally leads to the question of why, if including cluster sizes in the model makes no difference in terms of predicting the finite population mean, we would wish to bother with predicting cluster sizes for nonsampled clusters. It may be that for the simple models we consider here, excluding cluster sizes from the model is fine, but for more complicated models this may not be the case. In addition, the models we use are correctly specified, and it may be that under severe model misspecification, excluding cluster sizes can lead to worse estimates of the population mean.

An interesting exception to this result is in Figure 1.6, where including the cluster sizes via `knowsizes` when y_i and N_j are independent leads to higher bias, RRMSE, undercoverage, and uncertainty interval length. Understanding the population cluster size distributions and outcome models that lead to this situation is an important area of future research.

There are a number of interesting directions in which the current research could be extended. First, our simulation has not considered the case where the measure of size M_j is not the same as the number of units in cluster j, N_j. The natural next step would be to extend the Fragile Families simulation to include the case where the measure of size M_j is the city population, but the cluster size N_j itself is the total number of births in the city. In doing so, we must make some additional assumptions about what we as the analyst do and don't know. So far, we have assumed that we know M_j only for the sampled clusters, but we would have to make assumptions about our knowledge of N_j as well. If we assume we only know N_j for the sampled clusters, we would need a way of predicting both M_j and N_j for the entire population. One simple idea is to assume that N_j is a linear function of M_j and use regression to predict N_j given M_j, perhaps the on the log scale to avoid predicting negative cluster sizes and difficulties with cluster sizes ranging over several orders of magnitude. In the the Fragile Families study, the correlation between the log of city population M_j and log of total births N_j is 0.78, so this seems like a promising strategy. Additional information on the determinants of N_j, such as historical fertility rates in the Fragile Families context, would further improve predictions of N_j from M_j.

Another direction would be to consider a stratified PPS design as in the original Fragile Families study design. This extension introduces a new challenge in that we would need to adjust for the strata in our model. In doing so, however, we change the interpretation of the other coefficients to be conditional on stratum membership; in this way, the coefficients estimated from the model would not strictly be comparable to those used to generate the data in this simulation. For the parametric cluster size models, we would need to partially pool the size parameters (μ, ϕ in the negative binomial model, μ, σ in the lognormal) across strata, adding another layer of complexity to the model.

Finally, we did not consider the case of estimating a finite population proportion for binary y with a unit-level predictor x. This is a challenging problem because

$\Pr(y_i = 1)$ is a nonlinear function of x_i, so even if the cluster-level means $\overline{x}_j$ are assumed known for all clusters, we cannot directly use them to estimate the cluster-level proportion $\overline{y}_j$ as we could for the continuous case where we assume y is normally distributed. Extending the current work to handle this scenario would greatly increase its practical utility.

Chapter 2

Weak Instrumental Variables in the Context of Recidivism

2.1 Background

The United States has the highest incarceration rate in the world, with 2.3 million people held in correctional facilities in 2017, over 1.4 million of whom are in state or federal prisons (Wagner and Walsh, 2016; Wagner and Rabuy, 2017). Incarceration is thought to affect recidivism through three main mechanisms: 1) incapacitation, meaning offenders are unable to offend because they are locked up; 2) general deterrence, or the threat of incarceration that causes would-be offenders to reconsider their actions so as to avoid incarceration; and 3) specific deterrence, by which the experience of incarceration itself is such that it discourages future criminal behavior. However, existing literature does not give a clear answer to the question of whether incarceration increases or decreases subsequent recidivism.

The gold standard for studying causality is with a randomized experiment, but since random assignment of incarceration is ethically and legally impossible, research on recidivism must use observational data. However, in some jurisdictions, offenders are randomly assigned to judges, and several studies have taken advantage of this natural experiment to study the causal effect of incarceration on recidivism. Green and Winik (2010) use data on felony drug offenders in the District of Columbia and find no effect of incarceration or probation on subsequent rearrest dates. Loeffler (2013) also finds no detectable effect of imprisonment on either recidivism or labor market participation among felony offenders in Chicago. Similarly, Nagin and Snodgrass (2013) conclude that there is little evidence of incarceration impacting rearrest rates in five counties in Pennsylvania. Using data from Harris County, Texas, Mueller-Smith (2015) finds increases in the frequency and severity of recidivism due to incarceration, as well as additional negative effects on both labor market outcomes and dependence on public assistance. Aizer and Doyle (2015) conclude that juvenile incarceration leads to higher adult incarceration rates among juvenile offenders in Chicago.

With the exception of Nagin and Snodgrass (2013), all of these studies use two-

stage least squares (2SLS) as their main estimation strategy. However, 2SLS is known to produce unreliable estimates when the instrumental variable is weak, meaning its correlation with the treatment is low. This study uses alternative methods that are robust to many of the weaknesses of 2SLS. We also directly address the issue of weak instrumental variables and demonstrate a method for strengthening an instrument.

Two recent studies (Nieuwbeerta *et al.*, 2009; Snodgrass *et al.*, 2011) make use of another useful statistical technique, namely matching. Offenders are paired using demographics and other characteristics to construct treatment and control groups that are as similar as possible on observed covariates. Matching creates treatment and control groups that are balanced and have overlap in the distribution of observed covariates, and we can investigate the extent to which our conclusions are sensitive to – that is, can be explained away by – the unobserved confounders (Rosenbaum, 2002b).

We combine several techniques that have often been used separately in the criminology literature. First, we use a natural experiment to study the causal impact of incarceration on recidivism, namely the random assignment of criminal cases to judges in the state of Pennsylvania, and use the harshness of the judge as an instrumental variable. Second, we use matching to generate pairs of offenders that are identical or similar on important demographic and background variables in an effort to create comparable treatment and control groups. Together, these two methods seek to approximate a paired randomized encouragement design, in which one member of a pair is randomly encouraged to receive treatment and the other is not. Third, we avoid the use of two-stage least squares (2SLS) because of its known pitfalls in the presence of a weak instrumental variable.

Further, we directly evaluate the strength of our instrument and demonstrate a new method for improving a weak instrumental variable. We quantify the tradeoff between sample size and instrument strength induced by this method by simulating power and design sensitivity. Finally, in addition to estimating the causal effect of

incarceration on recidivism, we illustrate the use of new methods to determine the extent to which the intention-to-treat (ITT) effect of judge harshness on recidivism is modified by observed covariates.

2.2 Data

2.2.1 Sample

We obtained data on all offenses reported to the Pennsylvania Commission on Sentencing (PCS) between 1998 and 2000. The PCS receives information on all felony and misdemeanor offenses committed in Pennsylvania that are sentenced in Common Pleas Court in a given calendar year and reported to the Commission, with some important exceptions. The first is Philadelphia Municipal Court sentences and offenses sentenced by district magistrates, both of which generally concern driving under the influence (DUI) and other misdemeanor offenses. In addition, Murder 1 and Murder 2 offenses are not required to be reported; these offenses require mandatory life or death sentences. Unfortunately, the PCS does not have an auditing system for determining the extent of non-reporting, so there is no way to know how many cases are missing or whether there is systematic bias in which cases are excluded (Pennsylvania Commission on Sentencing, 1998 2000). Altogether, the missing cases are likely to be the most minor and most serious cases, for which incarceration is rarely and almost always the sentence.

The data record a unique identification code for each offender, along with their date of birth, sex, race, age at offense, age at sentencing, date of offense, date of sentencing, number of prior adjudications and convictions for various felony and misdemeanor offenses, the offense code and label, amount and type of drugs involved (if applicable) and the type and length of sentence imposed. They also include the offender's prior record score (PRS), a measure of the extent and severity of an offender's previous criminal history, as well as an offense gravity score (OGS) for each charge in

a given judicial proceeding. The PRS and OGS are particularly important covariates because they are used as the basis for determining sentence types and durations in the Pennsylvania Sentencing Guidelines sentencing matrix, which gives the standard, aggravated, and mitigated ranges for each PRS and OGS combination. If the offense involved possession of a deadly weapon, youths in drug trafficking, or drug trafficking within 1000 feet of a school, enhanced sentencing ranges apply (Pennsylvania Commission on Sentencing, 1998 2000).

We did extensive cleaning and checking of the data. For example, some offenders appear multiple times in the data under the same identification number with different dates of birth, sexes, and/or races. We corrected this by using the value that appears the most often. We also filtered out offenders using the criteria shown in Figure 2.1. In particular, we removed offenders with missing values for the sentencing judge or county, as well as important covariates like sex, race, date of birth, prior record score, and offense gravity score. We also dropped offenders whose year of birth is before 1918 or after 1997, as this would lead them to be implausibly old or young during the period of observation, as well as offenders whose age at date of sentencing or offense is less than or equal to ten years or greater than or equal to 100 years (these extreme ages can occur because of the date of birth or the date of sentencing). Finally, we kept only the offenders we could locate the the rap sheet data, described in Section 2.2.3 below, and those for whose sentencing judge we could define a binary harshness value as described in Section 2.2.4; it is at these steps that we lose the most offenders, as seen in Figure 2.1. Our final analysis sample consists of 53 318 unique offenders, 51% of the initial 104 532 offenders.

2.2.2 Treatment: Imprisonment

Since some offenders have multiple cases reported to the PCS between 1998 and 2000, we retained the earliest recorded case for each offender. For each case, we determined whether the offender was sentenced to (state) prison, (county) jail, time served, or

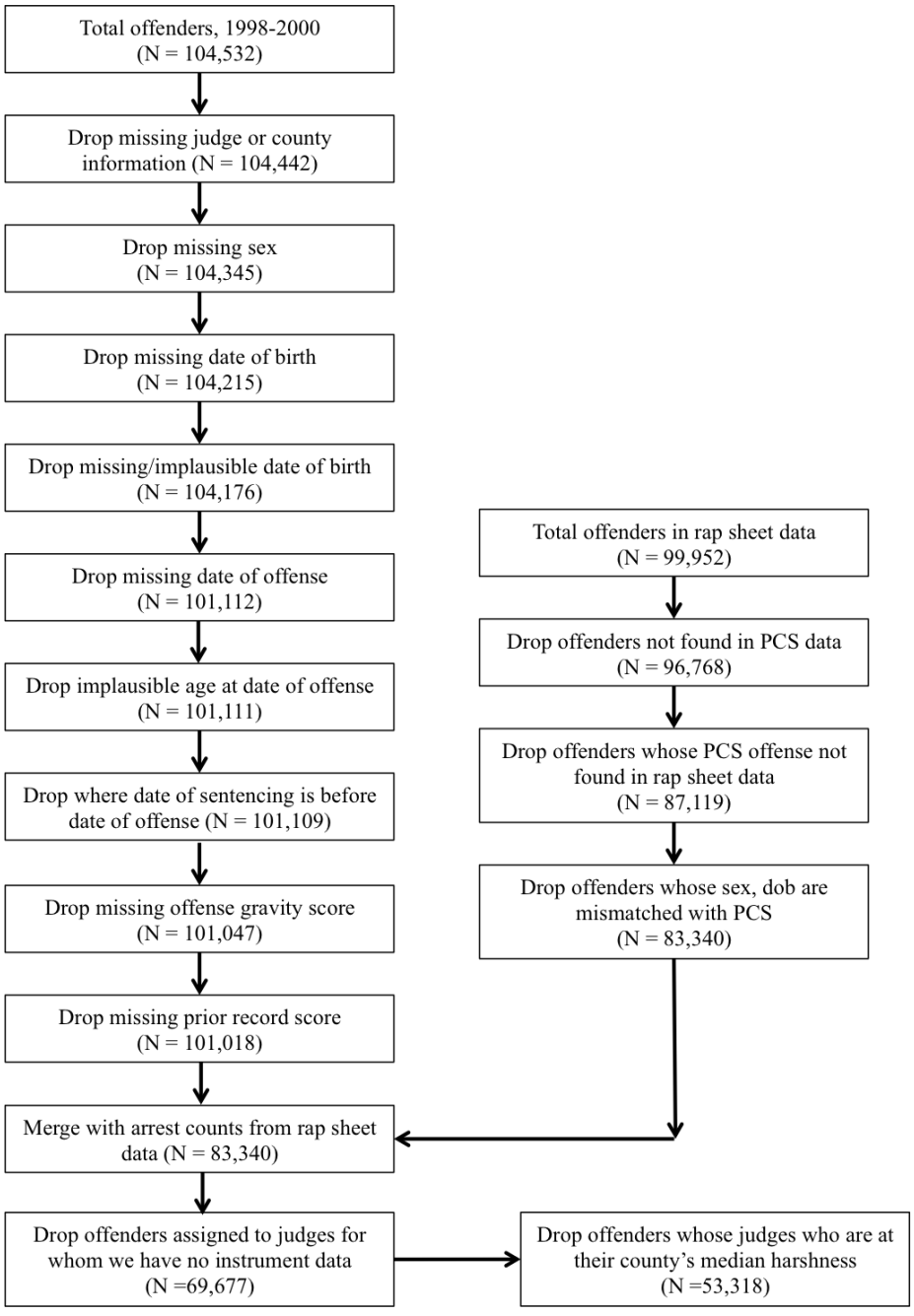


Figure 2.1: Flowchart of inclusion criteria for PCS and rap sheet data.

none of the above. Jail sentences are for periods of two years or less. We define incarceration as a sentence of either prison or jail. When an offender is sentenced to time served, the time they spent in jail before and during their trial is counted towards their total confinement. However, the time spent in jail before and during the trial is, strictly speaking, pre-treatment; by definition, the offender experiences it before sentencing, so whether an offender spends time in jail before or during the trial cannot be affected by the harshness of the judge. For this reason, we treat a sentence of "time served" as equivalent to a non-carceral sentence. Investigating the effect of serving time in prison separately from that of jail, as well as accounting for time spent in confinement for offenders sentenced to time served, is beyond the scope of the current study but an important area of future research.

2.2.3 Outcome: Recidivism

We obtained rap sheet data from the Pennsylvania State Police, from which we can measure the number of arrests in the state of Pennsylvania for each offender from the date of sentencing until November 1, 2013. We define recidivism as the number of arrests in the three years after sentencing. We considered using only felony arrests in defining the outcome, but this information is missing for a substantial portion of the data: 83% of offenders with arrests in the rap sheet data have at least one offense whose felony status cannot be determined. In addition, felony status cannot be determined for 1/3 of the offender-arrest date observations. Counts of felony arrests are thus unreliable.

We merged the PCS offense data with the rap sheet data by matching offenders using their unique identification number. We dropped offenders for whom we could not find any dates in the rap sheet – either an offense, arrest, or disposition date – that fall between the PCS offense and sentencing dates. If we could not locate the PCS offense in the rap sheet in this manner, we were not confident that the rap sheet data contained all of the arrests for the individual. We also ensured that the sex and

year and month of birth match between the PCS and rap sheet data. We did not use race as a criterion for matching offenders between the two datasets, because the race categories are different between the PCS offense and arrest record data. The offense data record race as American Indian, Asian, Black, Hispanic, White, and Other, while the arrest record data use Asian, Black, Indian, White, and Other. Of the 99 952 individuals for whom we have rap sheet data, 96 768 appear in the PCS data, and we were able to find matches in the PCS data for 83 340 (86%) of them.

2.2.4 Instrument: Judge Harshness

In the state of Pennsylvania, Common Pleas judges are elected at the county level, and it is at the county level that offenders are randomly assigned to a judge. Our instrument is the harshness of a judge, which we calculate as the proportion of the cases the judge sentenced to either prison or jail. Specifically, we use data from the PCS for 1997 to calculate, for those judges that saw at least 30 cases in that year, the the proportion of their cases that they sentenced to prison or jail. We then classify a judge as "lenient" if this proportion is below their county's median judge harshness and "harsh" if it is above. In the context of a randomized encouragement design, we refer to an offender assigned to a harsh judge as one who was "encouraged" to receive the treatment (incarceration) and an offender assigned to a lenient judge as one who was "discouraged" from receiving the treatment. However, we also make use of the continuous version of judge harshness (the proportion of 1997 cases they sentenced to prison or jail), and we refer to this measure as the instrument. Thus, "encouragement" and "instrument" refer respectively to the binary and continuous measures of judge harshness. Where the distinction is not clear from context, we specify which one we are referring to.

We considered using only part of the 1997 data to calculate judge harshness and combining the other half with the 1998-2000 data. However, we decided to use all of the 1997 data for several reasons. First, estimates of harshness will be more accurate

with a larger sample size of cases per judge. Second, we cannot track offenders between the 1997 and the 1998-2000 PCS data. Although nearly 20% of offenders appear multiple times in the 1998-2000 data, we only keep the earliest judicial proceeding under which they appear. Because an individual in the 1997 data may reappear in the 1998-2000 data, using part of the 1997 data in the analysis could lead to using the same individual twice, but we would have no way of knowing this. Third, if there are any seasonal trends in judge harshness, for example as a function of seasonality in the types and/or severity of crimes, using an entire year gives the most accurate picture of harshness.

There is one side issue that merits discussion here. Most judges see cases in a single county, but in some cases, a judge will see cases in more than one county, generally in rural counties. Of the 347 judges for whom we have 1997 caseload data, 313 saw cases in one county, 26 in two counties, 6 in three counties, and two in four counties. Regardless of the number of counties in which a judge saw cases, we calculate harshness at the judge level, not the judge-county level, for two reasons. First, in many cases, a judge who saw cases in, say, two counties would see the vast majority of her cases in one county and only a handful of cases in the other. If we were to calculate harshness at the judge-county level, we would have fewer judge-county observations that had the minimum 30 cases, and we would therefore lose from our analysis all offenders in 1998-2000 assigned to those judges. Secondly, while randomization of cases to judges happens at the county level, a judge's harshness, as measured by the proportion of cases she sentences to prison or jail, is a judge-level characteristic. Because the same laws apply across counties and judges who see cases in multiple counties tend to operate in rural counties that have largely the same demographics and crime profiles, there is no reason to believe that a judge would be differentially harsh in one county over the other. However, the binary measure of harshness is a relative measure, and so we classify judges as "lenient" or "harsh" based on the other judges in that county. Thus, it is possible that a judge who saw

cases in county A and county B and sentenced, say, 65% of her total caseload to prison or jail is considered harsh in county A and lenient in county B. This in fact happens for ten of the 347 judges. While it may seem strange that a judge can be considered harsh in one county and lenient in another, there is no logical inconsistency. What matters is whether, at a county level, an offender is more likely to be incarcerated if assigned to one judge over another; by definition, this means that (binary) harshness is relative and depends on the distribution of the (continuous) harshness levels of the judges in that county.

2.3 Instrumental Variables and Judge Harshness

2.3.1 IV Assumptions

Angrist *et al.* (1996) described the five assumptions needed for an instrumental variable to yield valid causal inferences. We describe these five assumptions in the context of incarceration and recidivism.

Assumption 1: Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1974, 1980, 1990)

SUTVA states that 1) there is only one version of the treatment ("no multiple versions of treatment") and 2) the potential outcomes for one unit are unaffected by the treatment assignments of other units. In our context, the first part of SUTVA requires that an offender's potential outcome in terms of recidivism be the same regardless of whether they are sentenced to prison or to jail, or which particular state prison or county jail they are confined to. If we define treatment as the experience of any post-sentencing incarceration, then the distinction between prison and jail becomes less important. However, investigating the specific effects of being sentenced to prison and to jail, as well as the effect of time spent in jail before trial, is an important avenue of further research. If the main mechanisms by which incarceration affects

reoffending rates operate in the same way regardless of facility, then the "no multiple versions of treatment" assumption is plausible as well.

The second part of SUTVA requires that there is no interference between offenders, in the sense that a) whether an offender is incarcerated is not affected by the judge assignment of any other offender; and b) the recidivism of an offender is not affected by the incarceration status of any other offender. It seems unlikely that whether an offender is incarcerated could depend on which judges offenders in other counties are assigned to, so a) is a reasonable assumption in this application. Connections between offenders in the form of friendship, kinship, or criminal association could potentially violate part b). For example, if an offender not sentenced to prison or jail who would otherwise have committed new crimes has a sibling or close friend who is incarcerated, he/she may choose not to reoffend upon observing the effects of incarceration on someone important to them. Alternatively, it is conceivable that an offender not sentenced to prison or jail who would otherwise *not* have committed new crimes may in fact reoffend if someone with whom they were involved in criminal activity was also not incarcerated. We cannot test this part of SUTVA since we do not know the social or familial connections between offenders, and we proceed on the assumption that it holds.

Assumption 2: Exclusion Restriction

The exclusion restriction is an untestable assumption stating that the instrument affects the outcome only through its effect on the treatment. For the exclusion restriction to hold, the causal effect of judge assignment on subsequent recidivism must come only through the causal effect of judge assignment on the sentence received. The exclusion restriction would not hold if, for example, lenient judges were more likely to successfully convince the offenders they saw to utilize job-training resources that then helped offenders maintain employment and avoid reoffending. However, as Loeffler (2013) points out, the large caseload faced by most judges – an average of

nearly 150 cases per judge per year in our data – suggests that there may not be much time to do more than the minimum necessary to move cases through the system.

Assumption 3: Nonzero Effect of Encouragement on Treatment Uptake.

In order for the instrument to be at all useful, its causal effect on treatment uptake must be strictly nonzero on average. This is a quantity we can estimate from our data: 38% of offenders assigned to a harsh judge were incarcerated, compared to 34% of those assigned to a lenient judge, for a difference of four percentage points, or over ten percent.

Assumption 4: Monotonicity (Imbens and Angrist, 1994).

In the context of a randomized encouragement design, this assumption places restrictions on the way subjects can respond to encouragement. The monotonicity assumption cannot be tested, and we must carefully consider whether there are plausible scenarios in which it might not hold. In the context of binary encouragement and treatment, this assumption classifies subjects into always-takers, never-takers, and compliers, and assumes that defiers do not exist. Always-takers (never-takers) are those who would (not) receive treatment regardless of encouragement. Compliers are those who comply with their assigned encouragement: if encouraged, they receive treatment and if not encouraged, they do not receive treatment. Defiers do the opposite of their assigned encouragement: if encouraged, they do not receive treatment and if not encouraged, they receive treatment. In our context, always-takers (never-takers) are those who would (not) be incarcerated regardless of judge harshness. Compliers are those who would only be incarcerated if assigned to a harsh judge, and defiers are those who would only be incarcerated if assigned to a lenient judge. One way of satisfying the monotonicity assumption is to assume that there are no defiers. That is, we assume there are no offenders who would be incarcerated if assigned to a lenient judge and not be incarcerated if assigned to a harsh judge. This

assumption is reasonable in the context of incarceration because all judges follow the same laws and a harsher judge should be strictly more likely to sentence an offender to incarceration than a lenient one.

Assumption 5: As-If Random Assignment.

This assumption states that the assignment of the instrument is as-if random. In the recidivism context, this assumption rests on the assignment of offenders to judges to be truly random, and not affected by subsequent interventions by their attorneys or other factors. We explore the validity the instrument graphically in Section 2.3.2.

2.3.2 Checking Instrument Validity

Our use of judge harshness as an instrumental variable makes several assumptions that are testable, though not provable, with the data at hand. One basic assumption is that there is, in fact, variation in judge harshness within counties in the first place. Figure 2.2 plots judge harshness in 1997 by county. The size of each point is proportional to the number of cases the judge saw, and the color represents whether the judge is classified as lenient (green), harsh (orange), or whether their harshness is undefined (black) because it is exactly the county median. The counties are sorted in descending order of the number of judges in that county. We show only those judges who saw at least 30 cases in 1997 and those counties with at least two such judges. There is considerable variation in judge harshness, with most judges sentencing between 25% and 75% of their caseloads to prison or jail.

As stated by the Pennsylvania Commission on Sentencing, cases are randomly assigned to judges within a county. If this is so, there should be no relationship between the types of cases and offenders a judge sees and how harsh they are. It is important to check for this lack of a relationship in both the data from 1997, which we use to calculate the instrument itself, and in the data from 1998-2000, the data we use for analysis. In the former case, we want to ensure that what we are measuring is the

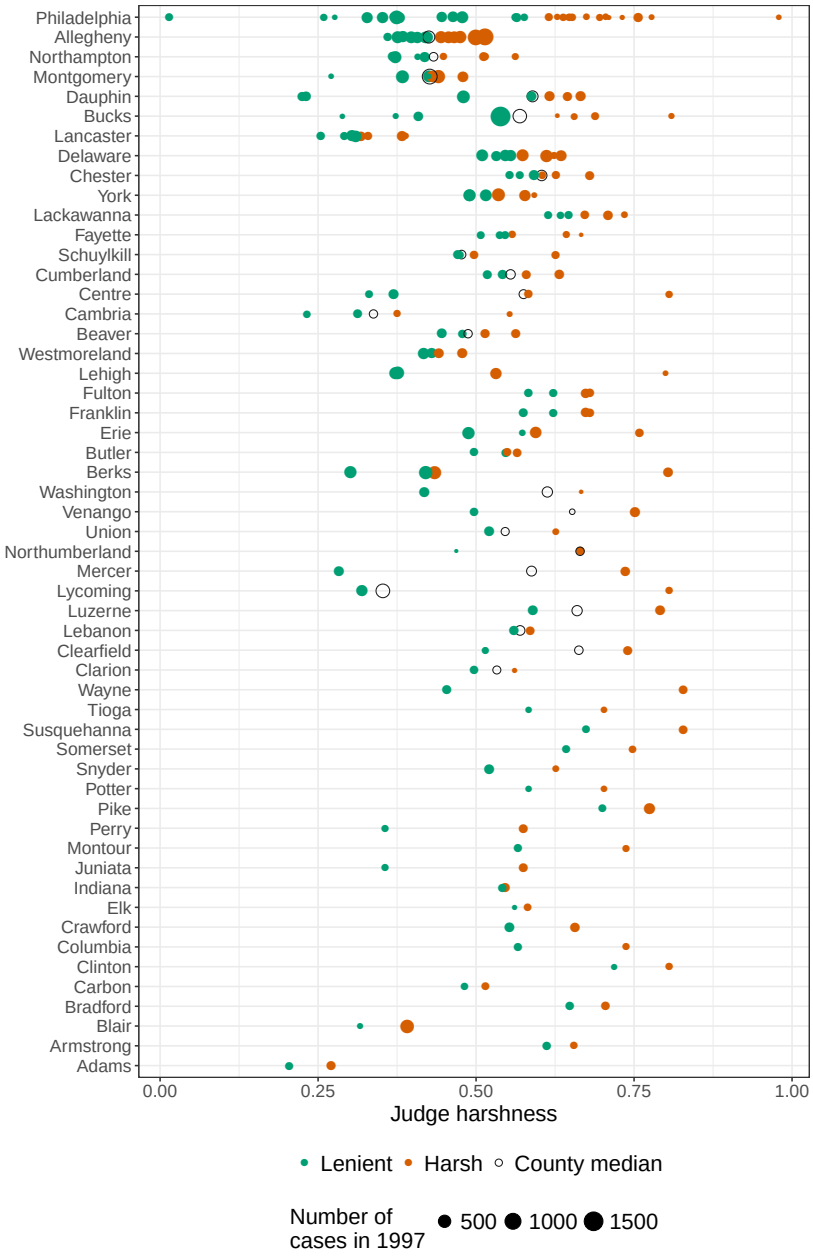


Figure 2.2: Distribution of judge harshness by county as measured in 1997. The size of each circle is proportional to the number of cases seen by that judge in 1997; only judges who saw at least 30 cases and counties with at least two such judges are included. The color of the points represents whether the judge is classified as lenient (green), harsh (orange), or whether the judge's harshness is exactly the county median, in which case their harshness is undefined (hollow black circle). The counties are sorted by the number of judges in that county.

actual harshness of the judge, not the composition of their caseload. In the latter, we want to check the plausibility of Assumption 5 in Section 2.3.1 above, which requires the assignment of the instrument to be as-if random. We investigate the relationship between offense severity and judge harshness in Figures 2.3 and 2.4; corresponding figures for other offender characteristics are in Appendix . In Figures 2.3 and 2.4, the $x-$axis is the (continuous) harshness of the judge in each county and the y-axis is the proportion of each judge's caseload that had a deadly weapon enhancement (DWE), were felonies, and were misdemeanors. The size of each circle is proportional to the number of cases the judge saw in that time period (1997 or 1998-2000); only judges who saw at least 30 cases in a county in each time period and counties with at least two such judges are shown. If cases are indeed randomly assigned, we should not see any strong relationship between harshness and these characteristics and all of the lines would be approximately horizontal. Overall, there is very little relationship between harshness and the various offender and case characteristics.

In addition, we further evaluate the assumption of as-if random assignment by using the binary instrument that classifies a judge as either "harsh" or "lenient". We evaluate the balance of observed covariates between offenders assigned to harsh and lenient judges by plotting the standardized differences in means for observed covariates in Figure 2.5. The standardized difference in means is the difference in means divided by the standard deviation; this puts the covariates on the same scale so that the differences are more easily compared across covariates. The covariates in Figure 2.5 are categorical, and we calculate the standardized difference in proportions for each category. Differences greater than 0 indicate that the proportion of encouraged offenders in a category is larger than the proportion of unencouraged offenders. For example, the proportion of offenders assigned to a harsh judge who are white is about 4% of a standard deviation larger than the proportion of offenders assigned to a lenient judge who are white. The size of each circle represents the proportion of offenders who fall into each category. For example, the circle for men is larger than

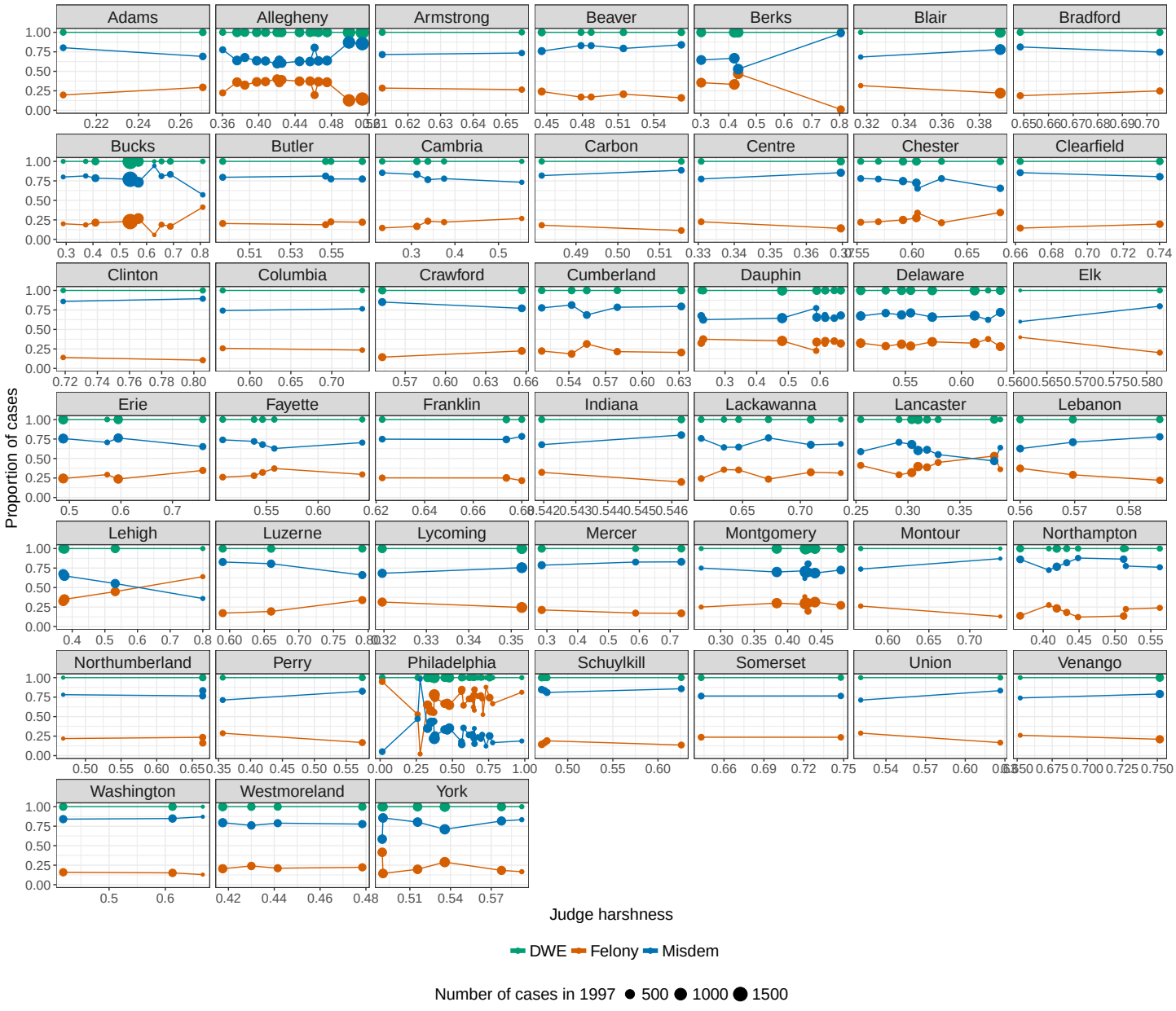


Figure 2.3: Proportion of each judge's 1997 cases that had a deadly weapon enhancement (DWE; green), were felonies (orange), and were misdemeanors (blue), plotted against the harshness of the judge. If cases are, in fact, randomly assigned to judges, we should not see any strong relationship between harshness and the proportion of cases corresponding to each of the crime types. Only counties with at least two judges who saw at least 30 cases in 1997 are shown.

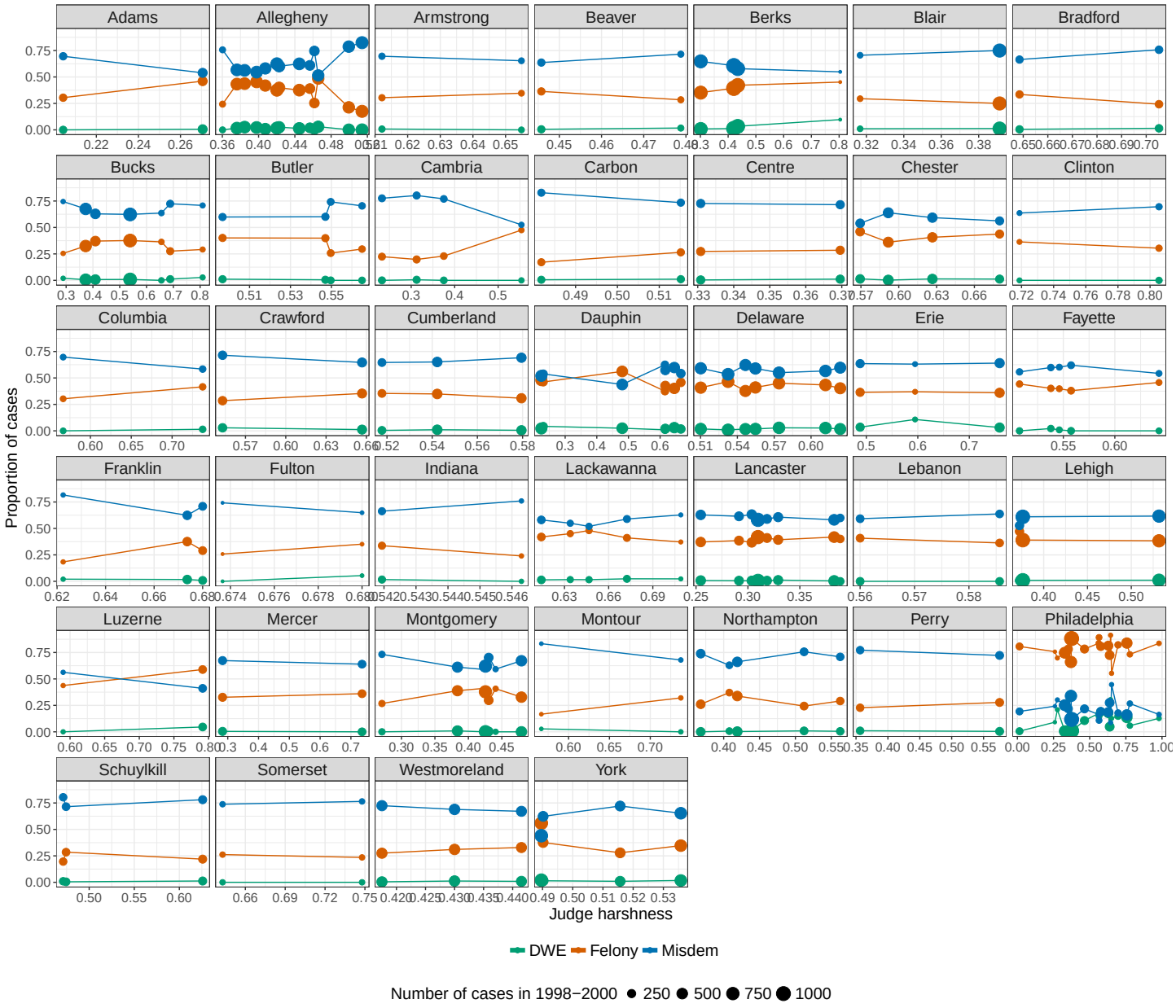


Figure 2.4: Proportion of each judge's 1997 cases that had a deadly weapon enhancement (DWE; green), were felonies (orange), and were misdemeanors (blue), plotted against the harshness of the judge. If cases are, in fact, randomly assigned to judges, we should not see any strong relationship between harshness and the proportion of cases corresponding to each of the crime types. Only counties with at least two judges who saw at least 30 cases in 1998-2000 are shown.

the circle for women because most of the offenders are male. Ideally, standardized differences in means should be below 0.10 (Rosenbaum, 2010), and in this case all of the standardized differences in means are below this threshold.

We also plot the distribution of age at date of offense for offenders assigned to harsh and lenient judges in Figure 2.6. The distributions are quite similar, and together Figures 2.5 and 2.6 indicate that even before matching, there is good balance in observed covariates and reason to believe that Assumption 5, as-if random assignment of encouragement, holds in this case. Still, there are some important variables, like certain OGS categories, for which we may wish to have better balance.

2.3.3 Problems with Weak IVs

A weak instrumental variable is one that does not have a strong relationship with the treatment. In the context of a randomized encouragement trial, a weak instrument is an encouragement that does little to increase uptake of treatment compared to no encouragement. When an instrument is weak, it may not contain much useful information about the causal effect of the treatment on the outcome. One commonly used procedure, two-stage least squares (2SLS), is vulnerable to several serious pitfalls in the presence of a weak instrumental variable. First, the standard errors on the estimate of the causal effect of interest are likely to be large, indicating imprecise measurement of the causal effect of interest; this is essentially a drastic reduction in the effective sample size of the data (Baiocchi *et al.*, 2014). Second, even a small violation of the exclusion restriction can lead to inconsistent estimates when the instrument is weak (Bound *et al.*, 1995). Third, even if the instrument is valid, a 2SLS estimate based on a weak IV can be substantially biased towards the OLS estimate in finite samples (Bound *et al.*, 1995). Fourth, confidence intervals based on 2SLS results using a weak IV have incorrect coverage rates and are too narrow (Imbens and Rosenbaum, 2005).

These problems can be remedied by using different methods. Imbens and Rosen-

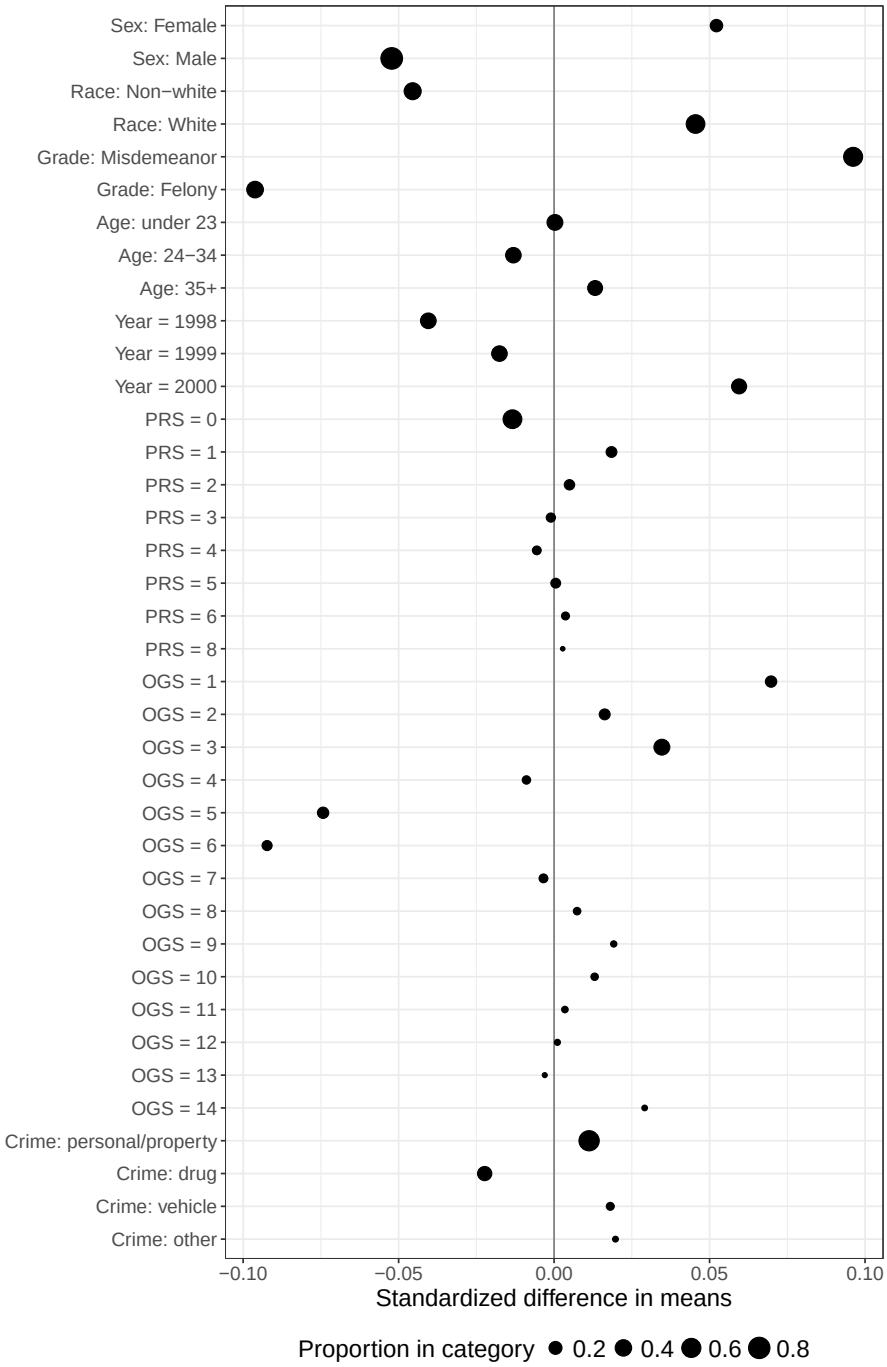


Figure 2.5: Standardized difference in means of observed covariates. The size of each circle represents the proportion of offenders who fall into each category of the covariate.

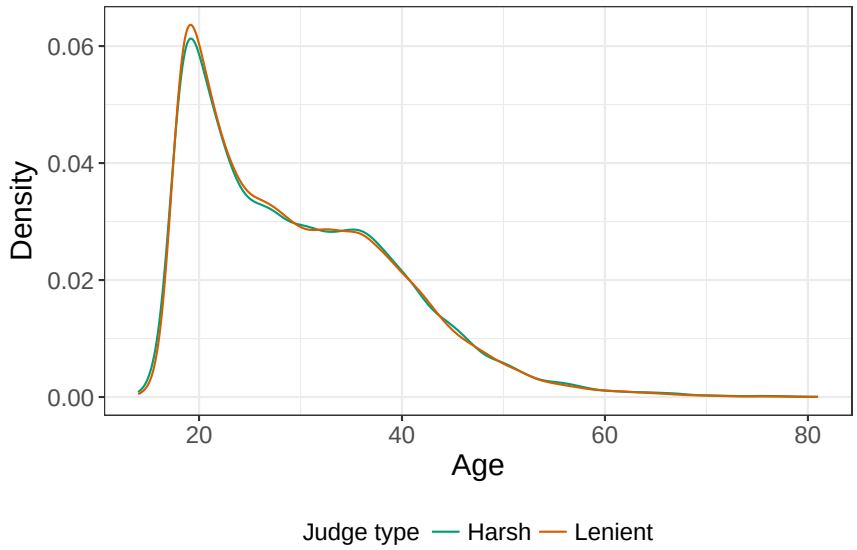


Figure 2.6: Density of age at date of offense for offenders assigned to harsh and lenient judges.

baum (2005) show that permutation-based inferences maintain correct coverage with weak instruments and yield appropriately large confidence intervals that "reflect the limited information in the data". Unfortunately, even with a perfectly valid instrument, the power to detect an effect when it truly exists is often quite low if the instrument is weak (Small and Rosenbaum, 2008). With a large enough sample size, the power does increase (and goes to 1 asymptotically), but ample data do not necessarily exist for every situation.

There is a more serious problem, however, that persists even with an infinitely large sample size. Small and Rosenbaum (2008) show that weak instruments are highly sensitive to small unmeasured biases, even when the true effect size is large. A researcher would have to be either highly confident that her IV is perfectly valid – that is, it satisfies Assumptions 1-5 in Section 2.3.1 above – or be studying huge effects in order to achieve estimates that are not sensitive to unobserved bias. This is perhaps the most serious problem with weak instrumental variables: when the instrument is weak, regardless of how much data we collect, we will not be able to

obtain robust estimates of causal effects that, unbeknownst to the researcher, are in truth substantial. The rarity of truly enormous causal effects in much of the social sciences, combined with the untestability of most of the fundamental assumptions underlying IV estimation, weak instruments pose serious problems.

2.4 The Paired Randomized Encouragement Design

2.4.1 Notation

In a paired randomized encouragement design (Holland, 1986; Rosenbaum, 1996, 2002a), $2I$ subjects are grouped into I pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$. The pairs are matched exactly for observed covariates $\mathbf{x}_{ij}$ so that $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ for every i, but there may be unobserved covariates u_{ij} that are not exactly matched, so $u_{i1} \neq u_{i2}$ for some i. In each pair, one subject is randomly assigned to receive the encouragement, $Z_{ij} = 1$ and the other to receive the control, $Z_{ij} = 0$, such that in each pair, $Z_{i1} + Z_{i2} = 1$. Here the encouragement Z_{ij} is whether offender ij was assigned to a lenient ($Z_{ij} = 0$) or harsh ($Z_{ij} = 1$) judge; we denote by W_{ij} the continuous harshness measure, $W_{ij} \in [0, 1]$. In a randomized experiment, each set of possible treatment assignments that satisfy $Z_{i1} + Z_{i2} = 1$ are equally likely. Specifically, let $\mathbf{Z} = (Z_{11}, \ldots, Z_{I2})^T$ be the vector of length $2I$ containing the treatment assignment for each unit and $\mathcal{Z}$ be the set of all 2^I possible values of $\mathbf{Z}$. We say that $\boldsymbol{z} \in \mathcal{Z}$ if $\boldsymbol{z} = (z_{11}, \ldots, z_{I2})^T$, $z_{ij} \in \{0, 1\}$ for each ij and $z_{i1} + z_{i2} = 1$ for all i. We denote conditioning on the event $\mathbf{Z} \in \mathcal{Z}$ as conditioning on $\mathcal{Z}$. In a randomized experiment, $\Pr(\mathbf{Z} = 1 \mid \mathcal{Z}) = 1/2^I$ for all $\mathbf{Z} \in \mathcal{Z}$, but in an observational study, the assignment of Z_{ij} is not randomized, so the subjects that receive encouragement can be systematically different from the ones that do not.

Under the potential outcomes framework, each unit has two values for the response

and the dosage of treatment received. We observe response r_{Tij} and dose d_{Tij} for an encouraged subject ($Z_{ij} = 1$) and r_{Cij} and d_{Cij} for a non-encouraged subject ($Z_{ij} = 0$). In our context the "dose" – whether the offender is incarcerated – is binary, but in general it can be a continuous quantity. (Note that the T and C in the subscripts refer to the value of the encouragement Z_{ij}, not "treatment" and "control" in terms of the dosage received.) Of course, the fundamental problem of causal inference is that we only observe one of (r_{Tij}, d_{Tij}) or (r_{Cij}, d_{Cij}) depending on whether $Z_{ij} = 1$ or $Z_{ij} = 0$, so we cannot directly estimate the causal effect of encouragement on dose received, $\eta_{ij} = d_{Tij} - d_{Cij}$, or on the response, $\delta_{ij} = r_{Tij} - r_{Cij}$. Instead, we observe response $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})r_{Cij}$ and dose $D_{ij} = Z_{ij}d_{Tij} + (1 - Z_{ij})d_{Cij}$. We write $\mathcal{F} = \{(r_{Tij}, r_{Cij}, d_{Tij}, d_{Cij}, \mathbf{x}_{ij}, u_{ij}), i = 1, \ldots, I, j = 1, 2\}$.

2.4.2 Randomization Inference

One measure of the causal effect of encouragement is the effect ratio λ, defined as the ratio of two average treatment effects: that of the encouragement on the response to that of the encouragement on the dose received (assuming the latter is nonzero):

$$\lambda = \frac{\sum_{i=1}^{I}\sum_{j=1}^{2}(r_{Tij} - r_{Cij})}{\sum_{i=1}^{I}\sum_{j=1}^{2}(d_{Tij} - d_{Cij})}. \tag{2.1}$$

In other words, $\lambda = \overline{\delta}/\overline{\eta}$, where $\overline{\delta} = (1/2I)\sum_{i,j}\delta_{ij}$ and $\overline{\eta} = (1/2I)\sum_{i,j}\eta_{ij}$ (Ertefaie *et al.*). If the exclusion restriction (Assumption 2 in Section 2.3.1) holds, then λ can be interpreted as the causal effect of incarceration on recidivism, among those offenders who are compliers – that is, those offenders who would be incarcerated if assigned to a harsh judge but not if assigned to a lenient judge. In this case, λ is the familiar instrumental variables estimand, variously called the Local Average Treatment Effect (LATE) (Angrist *et al.*, 1996) or the Complier Average Causal Effect (CACE). However, as described by Baiocchi *et al.* (2010), we can interpret and make inferences about λ even if the exclusion restriction does not hold.

We now describe randomization inference for a special case of λ that occurs when

treatment effects are constant and proportional to the dose received. We will use this inference framework for power simulations in Section 2.5.2 to select a match. Our presentation of randomization inference and sensitivity analysis in the rest of Section 2.4 closely follows the discussion in Small and Rosenbaum (2008).

When the treatment effect is constant and proportional to the dose received, we write

$$r_{T_{ij}} - r_{C_{ij}} = \beta(d_{Tij} - d_{Cij}). \tag{2.2}$$

In this case, the exclusion restriction holds and the effect ratio is simply $\lambda = \beta$ (Ertefaie *et al.*), and

$$R_{ij} - \beta D_{ij} = r_{T_{ij}} - \beta d_{Tij} = r_{c_{ij}} - \beta d_{Cij} \equiv a_{ij}. \tag{2.3}$$

Note that while R_{ij} and D_{ij} are observed, $R_{ij} - \beta D_{ij}$ is an unobserved quantity since it depends on the unknown parameter β. However, since a_{ij} is the same regardless of whether $Z_{ij} = 1$ or $= 0$, it is a function of $\mathcal{F}$ and is fixed in that it does not vary with $\mathbf{Z}$. In a randomized encouragement experiment, we can make inferences about β in (2.2) by testing $H_0 : \beta = \beta_0$ using the observed quantity $R_{ij} - \beta_0 D_{ij}$. We can write this quantity as

$$\begin{aligned} R_{ij} - \beta_0 D_{ij} &= Z_{ij}(r_{Tij} - \beta_0 d_{Tij}) + (1 - Z_{ij})(r_{Cij} - \beta_0 d_{Cij}) \\ &= Z_{ij}\left[(r_{Tij} - \beta d_{Tij}) - (\beta - \beta_0 d_{Tij})\right] + (1 - Z_{ij})\left[(r_{Cij} - \beta d_{Cij}) - (\beta - \beta_0 d_{Cij})\right] \\ &= a_{ij} + (\beta - \beta_0)\left[Z_{ij} d_{Tij} + (1 - Z_{ij}) d_{Cij}\right]. \end{aligned}$$

Note that if $H_0 : \beta = \beta_0$ is true, the second term in the last line is zero and $R_{ij} - \beta_0 D_{ij} = a_{ij}$ is fixed with respect to $\mathbf{Z}$ and depends only on $\mathcal{F}$. If H_0 does not hold, then $R_{ij} - \beta_0 D_{ij}$ will vary with Z_{ij}.

We define $V_i^{\beta_0}$ as the encouraged-minus-unencouraged pair difference in the ob-

served quantity $R_{ij} - \beta_0 D_{ij}$:

$$\begin{aligned} V_i^{\beta_0} &= Z_{i1}\left[(R_{i1} - \beta_0 D_{i1}) - (R_{i2} - \beta_0 D_{i2})\right] + (1 - Z_{i1})\left[(R_{i2} - \beta_0 D_{i2}) - (R_{i1} - \beta_0 D_{i1})\right] \\ &= (\beta - \beta_0)\left[Z_{i1}(d_{T_{i1}} - d_{C_{i2}}) + (1 - Z_{i1})(d_{T_{i2}} - d_{C_{i1}})\right] + (2Z_{i1} - 1)(a_{i1} - a_{i2}) \\ &= (\beta - \beta_0)S_i + \epsilon_i, \end{aligned} \tag{2.4}$$

where

$$\begin{aligned} S_i &= Z_{i1}(d_{T_{i1}} - d_{C_{i2}}) + (1 - Z_{i1})(d_{T_{i2}} - d_{C_{i1}}) \qquad \text{and} \\ \epsilon_i &= (2Z_{i1} - 1)(a_{i1} - a_{i2}). \end{aligned} \tag{2.5}$$

If we have a randomized experiment and $H_0 : \beta = \beta_0$ is true, then the first term in (2.4) is 0 and $V_i^{\beta_0}$ is $\pm(a_{i1} - a_{i2})$, each with probability 1/2, for each pair i. In this case, since a_{ij} is fixed with respect to $\mathbf{Z}$, $|V_i^{\beta_0}|$ is also fixed and $V_i^{\beta_0}$ is symmetrically distributed around 0. If H_0 does not hold, then $V_i^{\beta_0}$ is the sum of a quantity that is symmetric about 0, ϵ_i, plus $(\beta - \beta_0)S_i$. If $\beta > \beta_0$, as would happen if $\beta_0 = 0$ but there is in fact a positive treatment effect $\beta > 0$, and if the exclusion restriction holds and there is at least one complier (i.e. at least one ij such that $(d_{Tij}, d_{Cij}) = (1, 0)$), then $(\beta - \beta_0)S_i$ has positive expectation.

One way to test $H_0 : \beta = \beta_0$ is with Wilcoxon's signed-rank statistic, call it T^{β_0}, which we calculate using the ranks of $V_i^{\beta_0}$. If H_0 holds and we have a randomized encouragement experiment, then the distribution of T^{β_0} is (approximately) normal with $\mathbb{E}[T^{\beta_0}] = I(I+1)/2$ and $\text{Var}(T^{\beta_0}) = I(I+1)(2I+1)/6$ for large I.

2.4.3 Sensitivity Analysis

In the previous section, we assumed that we were in the scenario of a randomized encouragement experiment, where Z_{ij} is randomly assigned so that $\Pr(Z_{ij} = 1 \mid \mathcal{Z}, \mathcal{F}) = 1/2$ for all ij. In the context of an observational study, however, we have no guarantee that the encouragement is, in fact, randomly assigned. It is then natural to ask, how large would the departure from random assignment have to be in order to explain away any evidence of a treatment effect that we may find? In other words,

how sensitive are our results to unobserved bias? One approach to sensitivity analysis, described in detail in Rosenbaum (2002b, sec. 4), begins by assuming that we have pairs matched exactly for observed covariates $\mathbf{x}_i$ so that $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ for all pairs i, but it may be that unobserved covariates u_{ij} are not exactly matched, so $u_{i1} \neq u_{i2}$ for some i. Denote by π_{ij} the probability of ij receiving encouragement, $\pi_{ij} = \Pr(Z_{ij} = 1 \mid \mathcal{Z}, \mathcal{F})$, which we assume are independent. We then assume that, because of the unobserved covariate u_{ij}, the odds of a matched pair i may not be exactly equal, but rather differ by a factor of at most $\Gamma \geq 1$:

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}(1-\pi_{ik})}{\pi_{ik}(1-\pi_{ij})} \leq \Gamma, \qquad \forall i, j, k. \tag{2.6}$$

Rosenbaum (2002b, sec. 4) shows that (2.6) is equivalent to assuming a logit model of treatment assignment:

$$\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{Z}, \mathcal{F}) = \frac{\exp(\gamma \mathbf{z}^T \mathbf{u})}{\sum_{\mathbf{b} \in \mathcal{Z}} \exp(\gamma \mathbf{b}^T \mathbf{u})}, \tag{2.7}$$

where $\gamma = \log(\Gamma)$ and $\mathbf{u} = (u_{1,1}, \ldots, u_{I,2}) \in [0,1]^{2I}$. Clearly, when $\Gamma = 1$ so that $\gamma = 0$, $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{Z}, \mathcal{F}) = (|\mathcal{Z}|)^{-1} = (2^I)^{-1}$ and we are in the scenario of a randomized experiment. When we fix $\Gamma > 1$, the encouragement probabilities π_{ij} are unknown but have known bounds. We can therefore use these known bounds to bound test statistics and thus relevant inferential quantities like p-values.

In the case of Wilcoxon's signed-rank statistic, consider two random variables $\overline{T}$ and $\overline{\overline{T}}$ (see Rosenbaum (2002b, sec. 4.3) for a detailed derivation and discussion). Let $\overline{T} = \sum_{i=1}^{I} A_i$ be the sum of I random variables A_i that take on the value 0 with probability $\Gamma/(1+\Gamma)$ and i with probability $1/(1+\Gamma)$, and let $\overline{\overline{T}} = \sum_{i=1}^{I} B_i$ be the sum of I random variables B_i that take on the value 0 with probability $1/(1+\Gamma)$ and i with probability $\Gamma/(1+\Gamma)$. Then when $H_0 : \beta = \beta_0$ holds and for all $\pi = (\pi_{11}, \ldots, \pi_{I2})$ that satisfy (2.6),

$$\Pr(\overline{T} \geq t \mid \mathcal{Z}, \mathcal{F}) \leq \Pr(T^{\beta_0} \geq t \mid \mathcal{Z}, \mathcal{F}) \leq \Pr(\overline{\overline{T}} \geq t \mid \mathcal{Z}, \mathcal{F}). \tag{2.8}$$

From these bounds we can bound both p-values and confidence intervals; when $\Gamma = 1$, the three probabilities above are equal, which gives the standard randomization

distribution of Wilcoxon's signed-rank statistic under H_0. Given a fixed Γ, define $\tilde{t}_\alpha$ as the value such that $\Pr(\overline{\overline{T}} \geq \tilde{t}_\alpha) = \alpha$, where we omit conditioning on $\mathcal{Z}$ and $\mathcal{F}$ for brevity. Then $\Pr(T^{\beta_0} \geq \tilde{t}_\alpha) \leq \Pr(\overline{\overline{T}} \geq \tilde{t}_\alpha) = \alpha$, so we have an upper bound on the p-value $\Pr(T^{\beta_0} \geq \tilde{t}_\alpha)$. If $T^{\beta_0} \geq \tilde{t}_\alpha$, then we can reject H_0 at the α level. We are generally concerned with the upper bound on the p-value since our goal is to understand for which values of Γ the quantity $\Pr(T^{\beta_0} \geq \tilde{t}_\alpha)$ is greater than our chosen significance level α.

To actually calculate the bounds in (2.8), we can appeal to a large-sample approximation and the central limit theorem. In the limit that $I \to \infty$, the distribution of $\overline{\overline{T}}$ is approximately normal, with expectation and variance given by

$$\mathbb{E}[\overline{\overline{T}}] = \frac{\Gamma}{1+\Gamma}\frac{I(I+1)}{2} \qquad \text{and} \qquad \text{Var}(\overline{\overline{T}}) = \frac{\Gamma}{(1+\Gamma)^2}\frac{I(I+1)(2I+1)}{6}. \tag{2.9}$$

We then compare the standardized deviate $(T^{\beta_0} - \mathbb{E}[\overline{\overline{T}}])/\sqrt{\text{Var}(\overline{\overline{T}})}$ to the standard normal distribution to approximate the upper bound in (2.8).

2.4.4 Power of a Sensitivity Analysis and Design Sensitivity

When analyzing observational studies, we can never be certain that our conclusions are unaffected by unobserved bias, and a sensitivity analysis will tell us how large such a bias would have to be in order for our results to be altered. While the data cannot tell us for certain that we are free of unobserved bias, ideally the methods we use would be robust enough to detect a true treatment effect in the absence of unobserved bias and yield a sensitivity analysis stating that it is not sensitive to small biases. Specifically, if $H_0 : \beta = \beta_0$ were far from true – that is, if $\beta - \beta_0$ were large – and if Γ were not too large, then we would like to be able to reject H_0. We reject H_0 when the upper bound on the p-value $\Pr(T^{\beta_0} \geq \tilde{t}_\alpha \mid \mathcal{Z}, \mathcal{F})$ is larger than our chosen significance level α, which happens when $T^{\beta_0} \geq \tilde{t}_\alpha$. The probability that we will be able to correctly reject H_0 in the presence of a treatment effect and a perfect instrument is the power of a sensitivity analysis. More precisely, the power of an

α-level sensitivity analysis for a fixed $\Gamma > 1$ is the probability that $\Pr(\overline{\overline{T}} \geq t \mid \mathcal{Z}, \mathcal{F})$, the upper bound in (2.8), is less than or equal to α, calculated in the scenario of a perfect instrument and a nonzero treatment effect. The power of a sensitivity analysis is analogous to the power of a statistical test; instead of asking about the power of correctly rejecting H_0, the power of a sensitivity analysis asks about the probability of correctly identifying a treatment effect that is robust to small biases (Rosenbaum, 2010).

As the number of pairs $I \to \infty$, the power of the sensitivity analysis approaches 1 for $\Gamma < \tilde{\Gamma}$ and 0 for $\Gamma > \tilde{\Gamma}$; the value $\tilde{\Gamma}$ at which this switch in power occurs is called the design sensitivity, so the design sensitivity is the magnitude of bias to which our study will be sensitive, even in the limit of $I \to \infty$ (Rosenbaum, 2004, 2005). Design sensitivity is useful for comparing specific study designs and methods of analysis in terms of how resistant they are to unmeasured biases; all other things being equal, we would prefer a design and a method that yields higher design sensitivity (Rosenbaum, 2010; Ertefaie *et al.*).

2.5 Near/Far Matching to Strengthen a Weak Instrument

In order to address the problems with weak instrumental variables, we implement a matching method called near/far matching developed by Baiocchi *et al.* (2010) and extended by Zubizarreta *et al.* (2013). The goal is to create matched pairs who are as similar as possible on observed covariates and as dissimilar as possible on the instrument. In our context, we would like pairs of offenders with similar demographics and case characteristics whose judges have very different harshness values. We first review notation and inference for paired randomized encouragement designs, then describe the matching algorithm that yields different matched samples that trade off between sample size and instrument strength, and finally discuss our method for

selecting one of the matches based on simulating the power and design sensitivity for each matched set.

Our goal in near/far matching is to form pairs that are identical or very similar on observed covariates $\mathbf{x}_{ij}$, so that ideally $\mathbf{x}_{i1} = \mathbf{x}_{i2}$, and far on judge harshness so that $W_{i1} - W_{i2}$ is large. This approach strengthens the instrument because it yields pairs in which one unit is strongly encouraged to receive treatment and the other is not. At the same time, the units are comparable on observed covariates. We implement near/far matching via the integer programming algorithm developed by Zubizarreta *et al.* (2013) in the R package `designmatch`. This approach allows us to specify the minimum pair separation in harshness κ that we require for each pair: $W_{i1} - W_{i2} \geq \kappa$. We can also require that the average difference in harshness across the entire matched dataset also meets a certain threshold: $(1/I)\sum_{i=1}^{I}(W_{i1} - W_{i2}) \geq \omega$. In addition, the analyst can specify additional balance requirements such as mean balance, fine balance, or strength-k matching.

2.5.1 Matching Strategy

After data cleaning, the PCS data contain $N = 53\,318$ subjects who committed offenses that were reported to the Sentencing Commission between 1998 and 2000. Of these, 21 739 subjects were assigned to harsh judges and 31 579 to lenient judges. Because the randomization to judges is at the county level, we must match subjects within counties. Without using any other covariates, the maximum number of matched pairs we could possibly achieve is $I_{max} = \sum_{cty} \min(n_{t,cty}, n_{c,cty}) = 18\,874$, where $n_{t,cty}$ ($n_{c,cty}$) denotes the number of offenders assigned to a harsh (lenient) judge in county *cty*.

We form pairs that are matched exactly on binary indicators for sex (male/female), race (white/non-white) and whether the offender was charged with a felony. We also match exactly for three age groups (under 23, 24-34, 35+), three prior record score (PRS) groups (0, 1-2, 3+), and three offense gravity score (OGS) groups (1-2, 3,

4+). For each variable, we chose the groups such that they had roughly the same number of observations in each. We use PRS and OGS because they comprise the Basic Sentencing Matrix, which is contained in the Pennsylvania State Guidelines and gives recommendations on the length and location of confinement. Age and PRS in particular are well-known as important determinants of recidivism (**???**).

To strengthen the instrument, we consider various combinations of the minimum pair separation κ and the average separation ω. Specifically, we use $\kappa \in \{0, 0.02, 0.04, 0.06, 0.08, 0.10\}$ and ω ranging from 0 to 0.3 in units of 0.05 and from 0.3 to 0.5 in units of 0.025. Of course, some of these combinations are illogical, so we exclude them. For example, if we set the average separation ω to 0.05, then clearly the minimum pair separation κ must be $\kappa \leq 0.05$.

We display the results of the matches in Figures 2.7 to 2.9. Figure 2.7 plots the average separation ω on the x-axis and the number of matched pairs on the y-axis. Each circle is a different matched set, and the color and size of the circle is proportional to the difference in incarceration rates between the encouraged and unencouraged offenders, which is both an estimate of the proportion of compliers in the data and a measure of instrument strength. Each panel is for a different value of the minimum pair separation κ. We see that the effect of increasing ω has a consistent pattern across different values of κ. The number of matched pairs drops steeply and instrument strength rises quickly as average separation ω increases from 0 to 0.2, with both quantities changing more slowly as ω goes from 0.2 to 0.4, and we see a small but notable drop in the number of matched pairs going from $\omega = 0.425$ to $\omega = 0.450$.

Figure 2.8 plots the same numbers in a slightly different way. We now have the number of matched pairs I on the x-axis and the difference in incarceration rates on the y-axis. Each point is still a matched set, with the color of the points now representing the average separation ω; the panels are still for different values of the minimum pair separation κ. This figure shows the tradeoff between instrument strength and sample size more explicitly. We see that the instrument strength drops

off quickly as the sample size increases, and for each minimum pair separation there is a point at which instrument strength is at a peak; increasing or decreasing the sample size from this point yields a strictly weaker instrument.

Figure 2.9 shows the absolute standardized difference in means before and after matching for each match, here indexed by the minimum pair separation κ (columns) and the average separation ω (rows). Because we exactly match for sex, race, felony, age group, PRS group, and OGS group, we show here the differences in means for the year of sentencing and the continuous versions of age, PRS, and OGS. We see that in some cases, exact matching can worsen the mean balance quite drastically on the continuous versions of important covariates. This is particularly true for OGS, which has fourteen levels that we categorize into three groups: 1-2, 3, and 4+. While the match is exact on these four categories, the fact that the last category covers OGS scores from 4 to 14 makes the standardized difference in means on OGS poor. However, because high values of OGS are relatively rare in the data – less than 10% of offenders have OGS scores of 8 or higher – this grouping makes the most sense in terms of category size. Because the OGS is used in the Basic Sentencing Matrix, it is an important covariate in the context of recidivism. For this reason, we narrow our focus to the matches resulting from minimum pair separations $\kappa \in \{0.06, 0.08, 0.1\}$ and average separations ω between 0.1 and 0.45.

2.5.2 Power of a Sensitivity Analysis and Selecting a Match

Given these matches with such stark differences in sample size and instrument strengths, how should a researcher choose which one to use in her analyses? If we have a stronger instrument with a smaller sample size, perhaps our results would be more sensitive to unobserved biases. On the other hand, maybe a larger sample size with a weaker instrument gives us more power to detect a treatment effect when it actually exists. How can we understand this tradeoff between sample size and instrument strength?

Our approach is to select a match by calculating for each match the power of a

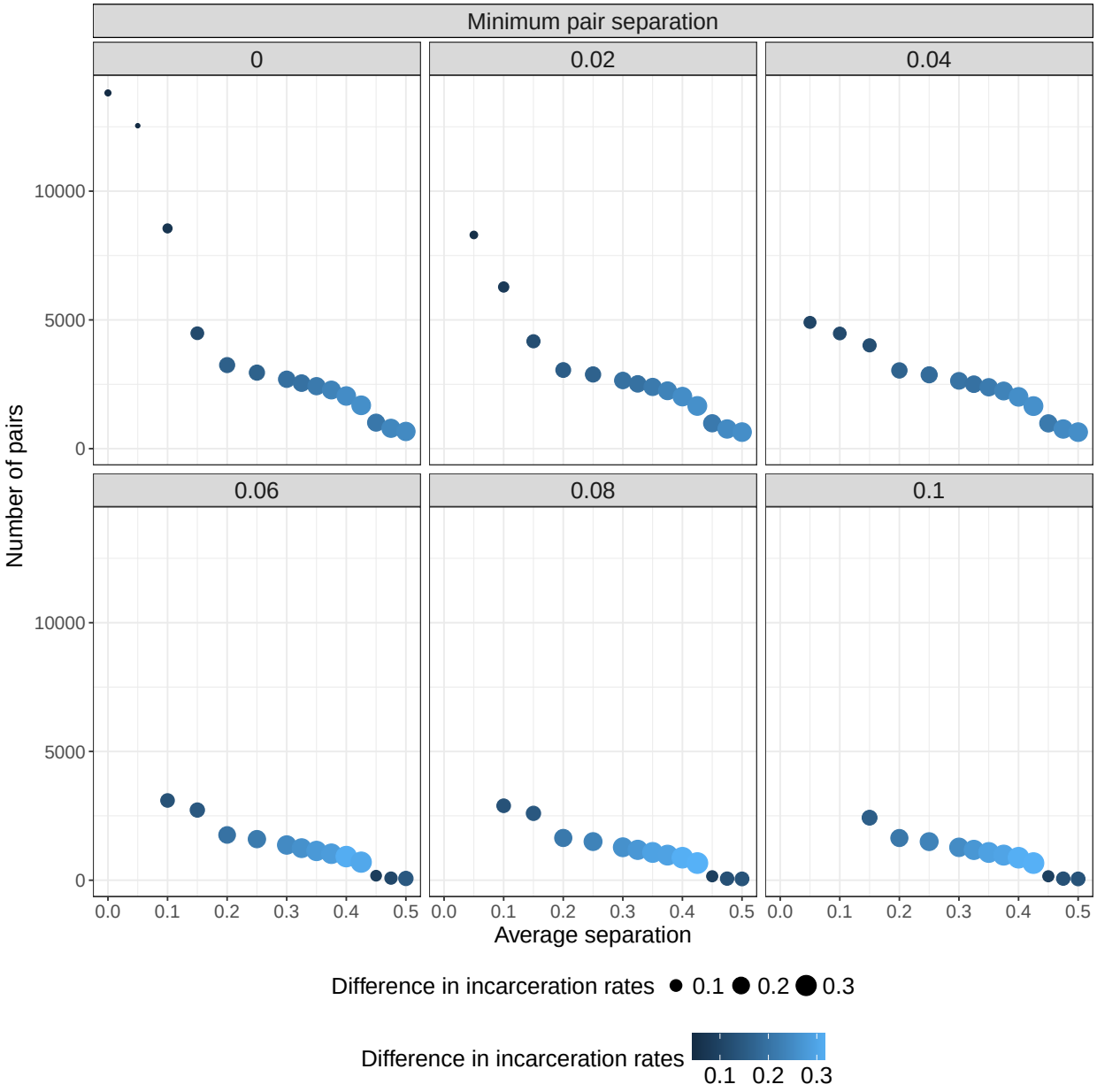


Figure 2.7: Summary of matching results for different combinations of κ (minimum pair separation) and ω (average separation). Each circle represents a matched dataset resulting from one of these combinations. The x-axis is the required average separation in the match and the y-axis is the number of matched pairs. Each panel is a different value of the minimum pair separation. The color and size of each circle represent the instrument strength, measured as the difference in incarceration rates between encouraged and unencouraged offenders.

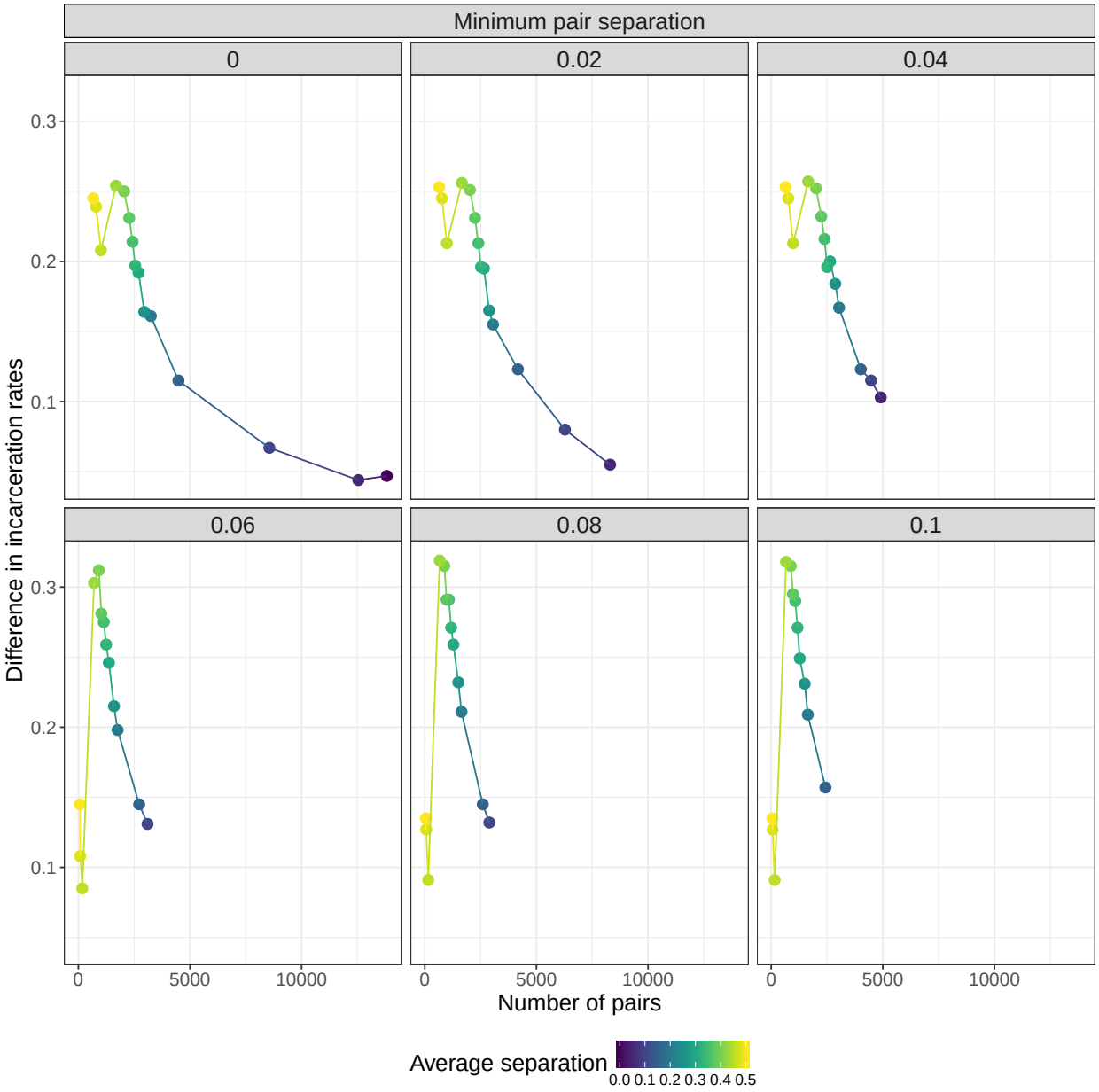


Figure 2.8: Summary of matching results for different combinations of κ (minimum pair separation) and ω (average separation). Each circle represents a matched dataset resulting from one of these combinations. The x-axis is the number of matched pairs and the y-axis is the the instrument strength, measured as the difference in incarceration rates between encouraged and unencouraged offenders. Each panel is a different value of the minimum pair separation. The color each circle represents the required average separation in the match.

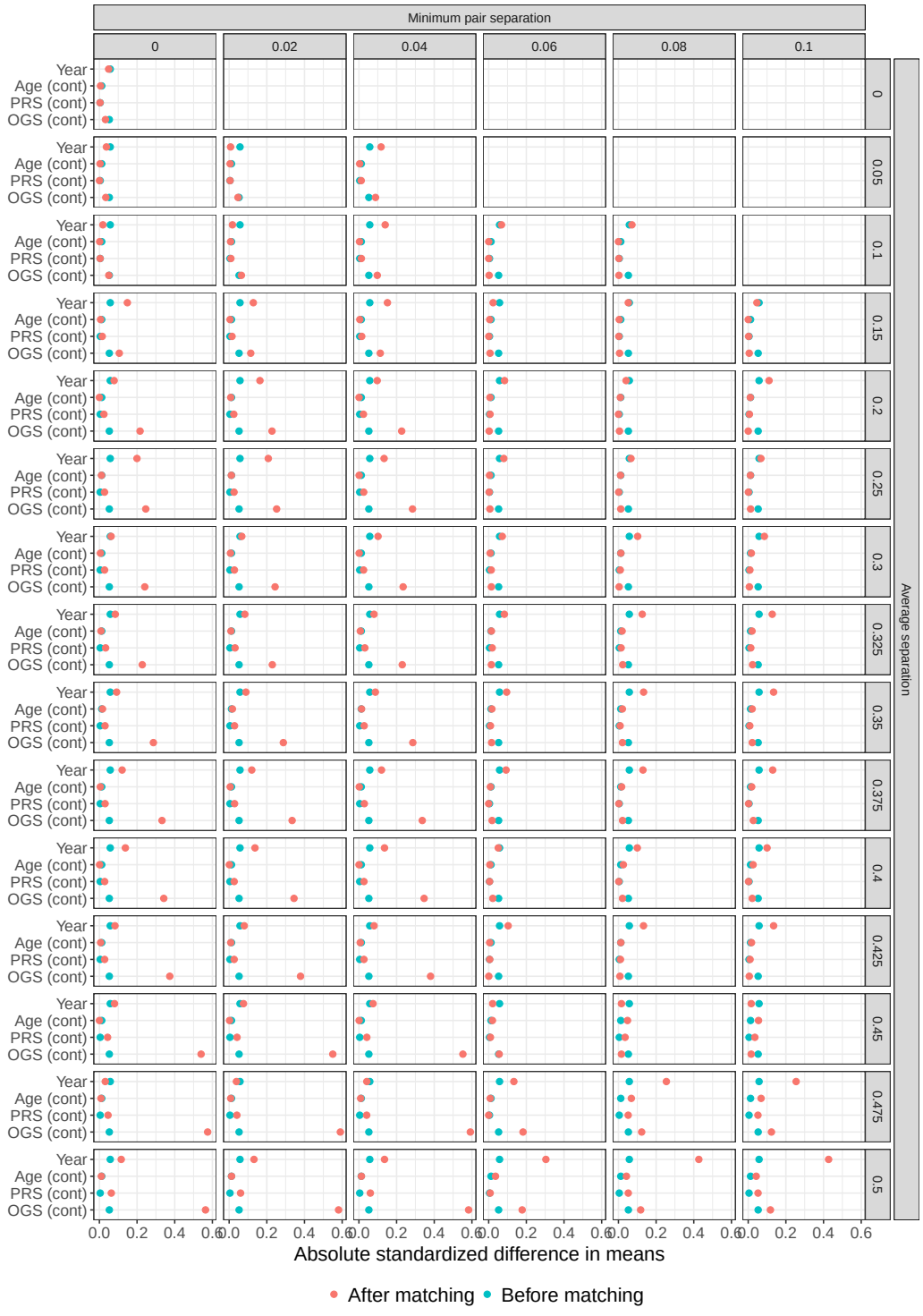


Figure 2.9: Absolute standardized difference in means before (blue) and after (red) matching. Each row is for a different average separation ω and each column for a different minimum pair separation κ.

sensitivity analysis for specific values of Γ. Specifically, we use simulation to calculate the expected power of the Wilcoxon signed-rank statistic as described in general in Rosenbaum (2010) and for the specific case of an observational study with an instrumental variable in Small and Rosenbaum (2008). This procedure uses both the number of matched pairs I and the estimated proportion of compliers (instrument strength) ρ_C to calculate the expected power.

The formula for an approximation of the expected power of Wilcoxon's signed-rank statistic is given in Rosenbaum (2010, sec. 14) and involves four quantities: the number of pairs I, $p = \Pr(V_i^{\beta_0} > 0)$, $p_1' = \Pr(V_i^{\beta_0} + V_j^{\beta_0} > 0)$, and $p_2' = \Pr(V_i^{\beta_0} + V_j^{\beta_0} > 0 \text{ and } V_i^{\beta_0} + V_k^{\beta_0} > 0)$ for $i < j < k$. We follow Small and Rosenbaum (2008) and estimate these probabilities by simulating one million independent triples $(V_i^{\beta_0}, V_j^{\beta_0}, V_k^{\beta_0})$. For each triple, we calculate three 0/1 indicators corresponding to the events in p, p_1', p_2' to get three binary vectors, each of length one million. We estimate p, p_1', p_2' as the mean of the corresponding binary vector.

To simulate $V_i^{\beta_0}$, we require three components as given in (2.4): the noncentrality parameter $\beta - \beta_0$, which measures the magnitude of the departure from $H_0 : \beta = \beta_0$, the encouraged-minus-unencouraged difference in treatment status S_i, and the error term ϵ_i. We consider three values for $\beta - \beta_0$: $\beta - \beta_0 \in \{0.25, 0.5, 1\}$. For ϵ_i, we consider three distributions, the Normal, Cauchy, and Logistic, all centered at 0 with unit variance (for the Cauchy, we set the scale to one; for the Logistic, we set the scale to $\sqrt{3}/\pi$ so that the variance is one). In this way, $(\beta - \beta_0)/\sigma$ is comparable across the three distributions.

Simulating S_i requires several steps, the first of which is a model for compliance status. We again follow Small and Rosenbaum (2008) and assume that compliance status is multinomially distributed with probabilities ρ_A, ρ_C, and ρ_N for always-takes, compliers, and never-takers, respectively; the monotonicity assumption in Section 2.3.1 means there are no defiers. Note that since treatment is binary, S_i takes on one three values, $S_i \in \{-1, 0, 1\}$, and depends on the compliance statuses of the

encouraged and unencouraged members of pair i. Assuming that both compliance status and encouragement are randomly assigned, we can calculate the probabilities of S_i taking on each of the three values. For example, $S_i = -1$ only when the encouraged member of a pair is a never-taker and the unencouraged member is an always-taker, which occurs with probability $\Pr(S_i = -1) = \rho_N \rho_A$. When the encouraged subject is either a complier or an always-taker and the unencouraged subject is either a complier or a never-taker, we have $S_i = 1$, so $\Pr(S_i = 1) = \rho_C^2 + \rho_A \rho_C + \rho_C \rho_N + \rho_A \rho_N$. Similarly, $\Pr(S_i = 0) = \rho_A^2 + \rho_N^2 + \rho_C \rho_A + \rho_N \rho_C$, and indeed $\Pr(S_i = -1) + \Pr(S_i = 0) + \Pr(S_i = 1) = 1$. We use the estimated proportion of compliers in the matched data as ρ_C: $\rho_C = I^{-1} \sum_{i=1}^{I} U_i$, where $U_i = (Z_{i1} - Z_{i2})(D_{i1} - D_{i2})$ is the encouraged-minus-unencouraged pair difference in treatment outcomes. We assume that the proportions of always- and never-takers are equal, so $\rho_A = \rho_N = (1 - \rho_C)/2$ and $\rho_A + \rho_C + \rho_N = 1$.

We also calculate the design sensitivity $\tilde{\Gamma}$, which for the Wilcoxon signed rank statistic is calculated as $\tilde{\Gamma} = p_1'/(1 - p_1')$; see Rosenbaum (2010, sec. 14) for proof. Ertefaie *et al.* point out that since design sensitivity is an asymptotic quantity, it cannot help us weigh sample size I and bias Γ. However, the proportion of compliers does enter into the calculation for $\tilde{\Gamma}$ through S_i in $V_i^{\beta_0}$, so here the design sensitivity is a function of the instrument strength.

Our approach has several advantages compared to two recent approaches to this problem. Ertefaie *et al.* propose a method that involves splitting the sample in half. The first half is used to determine where the instrument is strong and is then discarded. In the second half, the portion of the data where the instrument was determined to be weak is discarded, and the analysis proceeds with the remainder of the second half. In the context of their application, instrument strength is the frequency of delivery at hospitals with high-level neonatal intensive care units (NICUs) by zip code. They use half of the sample in a zip code to classify the zip code as high, medium, or low frequency, and then use the other half for analysis. This approach has the advantage of not using the data twice, but the disadvantage of losing over half

the sample size. The loss in sample size is justified by their simulations, which use the Bahadur efficiency of a test or sensitivity analysis as the metric for quantifying the sample size-instrument strength tradeoff and show that a "moderate increase in instrument strength is worth more than an enormous increase in sample size", especially for the case of an imperfect instrument that may have small unmeasured biases. However, our strategy of using separate datasets for calculating the instrument and for the analysis sidesteps the issue of using the same data twice without requiring data to be discarded.

Keele and Morgan (2016) combine near-far matching (Baiocchi *et al.*, 2010) with weak instrument tests from the econometrics literature to identify a sample size-instrument strength combination whose performance on these tests indicates that the instrument has been adequately strengthened. They use the F-statistic and the R^2 from a regression of the treatment on the instrument to select a sample size-instrument strength combination. In their application, they note that the match that produces the highest standardized difference in means for the instrument (excess rainfall), the one that many researchers might reasonably select, actually fails the weak instrument test. The advantage of our approach is that, given a willingness to make assumptions about the distribution of ϵ_i, we can directly quantify the sensitivity of each match to unobserved bias, an important consideration in any study of causal effects.

2.5.3 Simulation Results

For each match, we calculate the power of a one-sided level-0.025 sensitivity analysis for several values of $\Gamma \geq 1$ via simulation. Figure 2.10 displays the power plotted as a function of the estimated proportion of compliers (difference in incarceration rates). Each circle represents a match, with the colors distinguishing different departures from the null hypothesis, $\beta - \beta_0$. The size of the circle is proportional to the number of pairs in the match. The columns are for different values of Γ and the rows for the three models for ϵ_i. For clarity, we show the matches resulting from requiring a

minimum pair separation $\kappa = 0.08$.

When the instrument is perfect, $\Gamma = 1$, the probability of detecting a large departure from $H_0 : \beta = \beta_0$ is high, even for matches where the instrument is weakest, and drops off steadily as the departure from H_0 gets smaller. Power decreases as Γ increases, quickly when the errors are Cauchy and when the departure from the null is small, more slowly for Normal and Logistic errors and larger departures from the null. However, power increases sharply when $\Gamma > 1$ as the estimated proportion of compliers increases. The power is not quite monotonic with the proportion of compliers because two matches with very similar instrument strengths may have very different sample sizes; for example, in Figure **??**, one match has 1 138 pairs with an incarceration rate difference of 0.275, while another has double the number of pairs, 2 279, but only a slightly smaller incarceration difference of 0.231. In such cases, the power will be larger for the larger match. Similarly, the sample size for the match with the highest compliance rate is 26% smaller than that for the match with the second-highest rate (667 vs 875), which is why the power is smaller. Figure 2.11 is analogous to Figure 2.10, except that the x-axis shows sample size and the size of each circle is proportional to the estimated proportion of compliers. Power peaks when the compliance rate is large and drops off as sample size increases and the compliance rate decreases.

In Figures 2.12 to 2.14, we plot power as a function of Γ for each of the three error models. Each line represents a single match, with the color of the line denoting the minimum pair separation κ and the linetype denoting the different degrees of departure from $H_0 : \beta = \beta_0$. Each panel is for a separate value of the average separation ω. Each match, indexed by its values of κ and ω, thus appears three times in this figure, once for each of the three values of $\beta - \beta_0$. While Figures 2.10 and 2.11 show power under specific assumptions about the error distribution and the degree of bias Γ, Figures 2.12 to 2.14 follow the power for a single match over many values of Γ. In this way, we can see how quickly the power drops for each match as Γ increases,

under fixed assumptions about the error distribution and the size of $\beta - \beta_0$. Power is higher under the lighter-tailed Normal and Logistic distributions and quite low for the heavy-tailed Cauchy. Across Figures 2.12 to 2.14, the matches formed by requiring the average separation to be at least $\omega = 0.4$ consistently the highest power.

We also calculate design sensitivity $\tilde{\Gamma}$, which is a function of the proportion of compliers in the data. In Figure 2.15, we plot the proportion of compliers against the design sensitivity, with each column representing a different minimum pair separation κ and each row for a different error model. Each point is a match, with the size of the point proportional to the number of pairs in that match and the color denoting the size of the departure from H_0. Design sensitivity $\tilde{\Gamma}$ increases monotonically with the proportion of compliers, so Figure2.15 suggests choosing the match with the largest design sensitivity. This plot is analogous to Table 6 in Small and Rosenbaum (2008), showing design sensitivity under various instrument strength, error distribution, and null departure assumptions. However, producing this figure directly for the matches one is attempting to choose from can help a researcher understand the relationship between $\tilde{\Gamma}$ and instrument strength in her particular data. Another way of looking at this is in Figure 2.16, where we have switched the roles of the number of matched pairs I and the estimated proportion of compliers. As an asymptotic quantity, $\tilde{\Gamma}$ is not a function of I, but because each match can be identified by its number of matched pairs or its estimated proportion of compliers, Figure 2.16 is just a way of projecting the matches along a different dimension. Here we see that the design sensitivity peaks for matches with just under 1000 pairs.

Having seen in Figures 2.12 to 2.14 that power is consistently highest among the matches with average separation $\omega = 0.4$, we plot design sensitivity for each of the three values of $\beta - \beta_0$ and three error models in Figure 2.17. The value of $\tilde{\Gamma}$ is consistently highest for the match with minimum pair separation $\kappa = 0.08$.

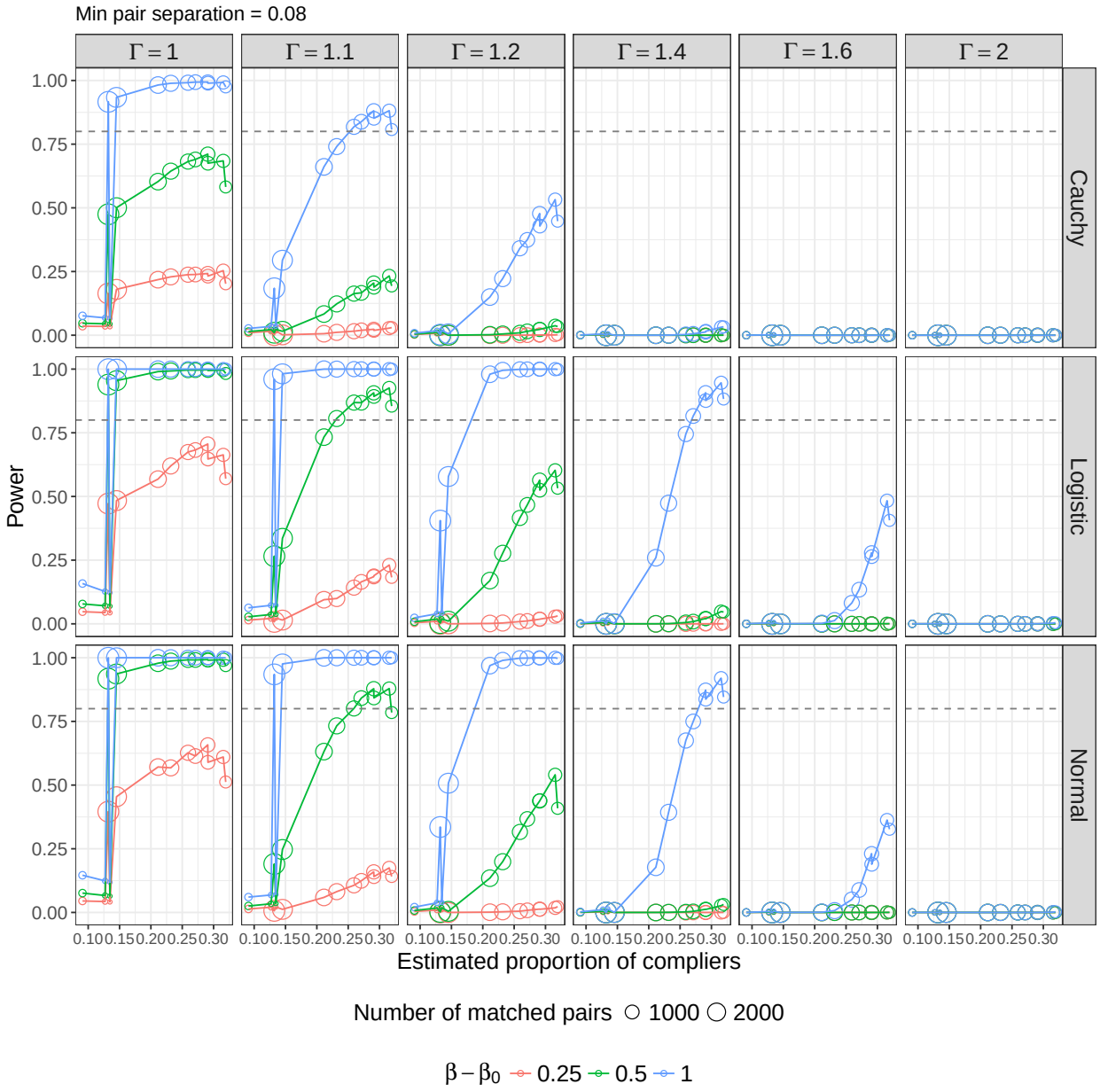


Figure 2.10: Power of a one-sided level-0.025 sensitivity analysis plotted against difference in incarceration rates (instrument strength). Each point represents a matched set resulting from setting the minimum pair separation to 0.08 and varying the average separation. The size of the circle is proportional to the number of pairs in that match. The columns are for different value of Γ and the rows are the three error models. The colors represent different degrees of departure from $H_0 : \beta = \beta_0$. The dashed horizontal line shows power = 0.8.

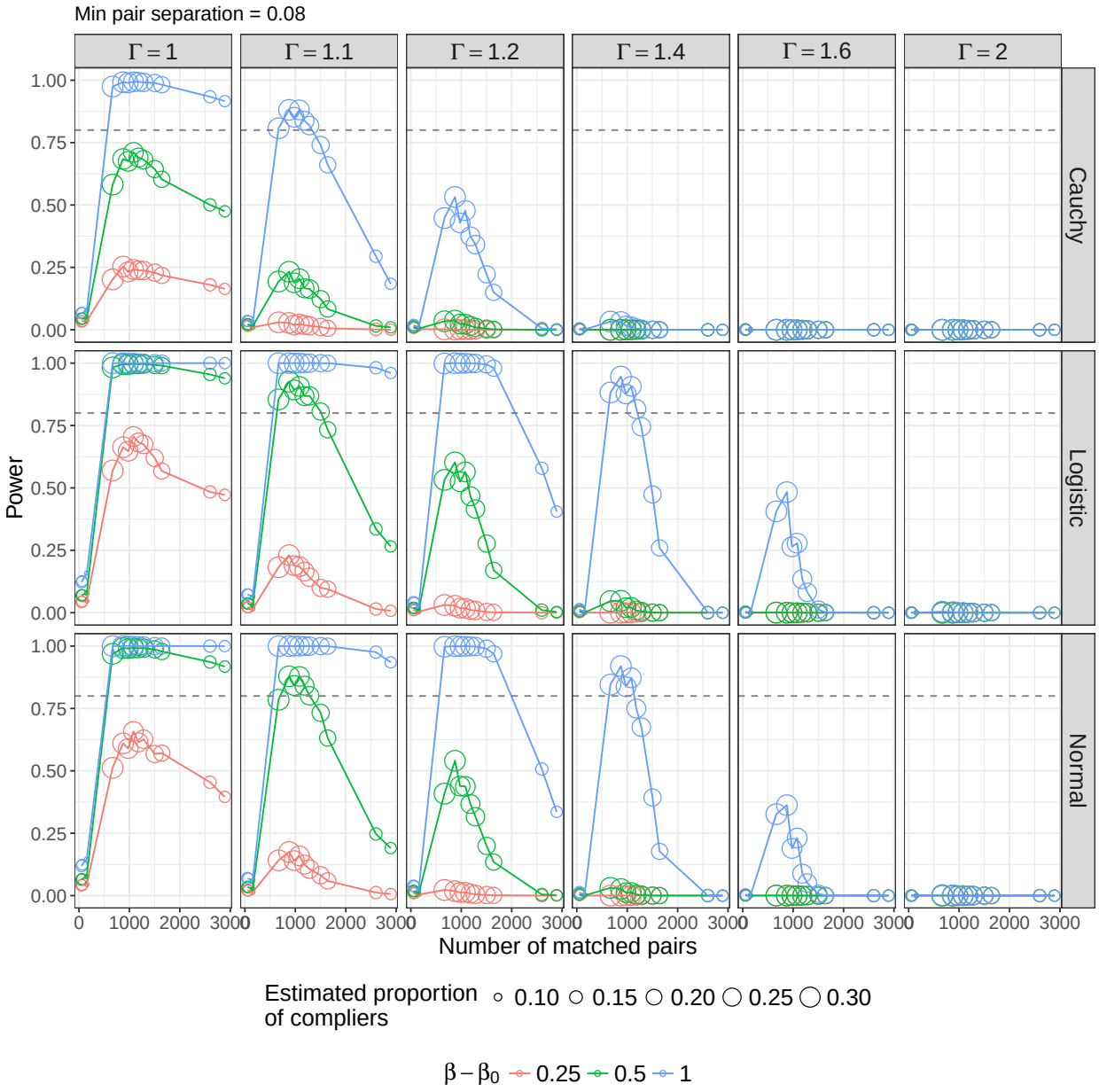


Figure 2.11: Power of a one-sided level-0.025 sensitivity analysis plotted against difference in incarceration rates (instrument strength). Each point represents a matched set resulting from setting the minimum pair separation to 0.08 and varying the average separation. The size of the circle is proportional to the difference in incarceration rates (instrument strength) in that match. The columns are for different value of Γ and the rows are the three error models. The colors represent different degrees of departure from $H_0 : \beta = \beta_0$. The dashed horizontal line shows power = 0.8.

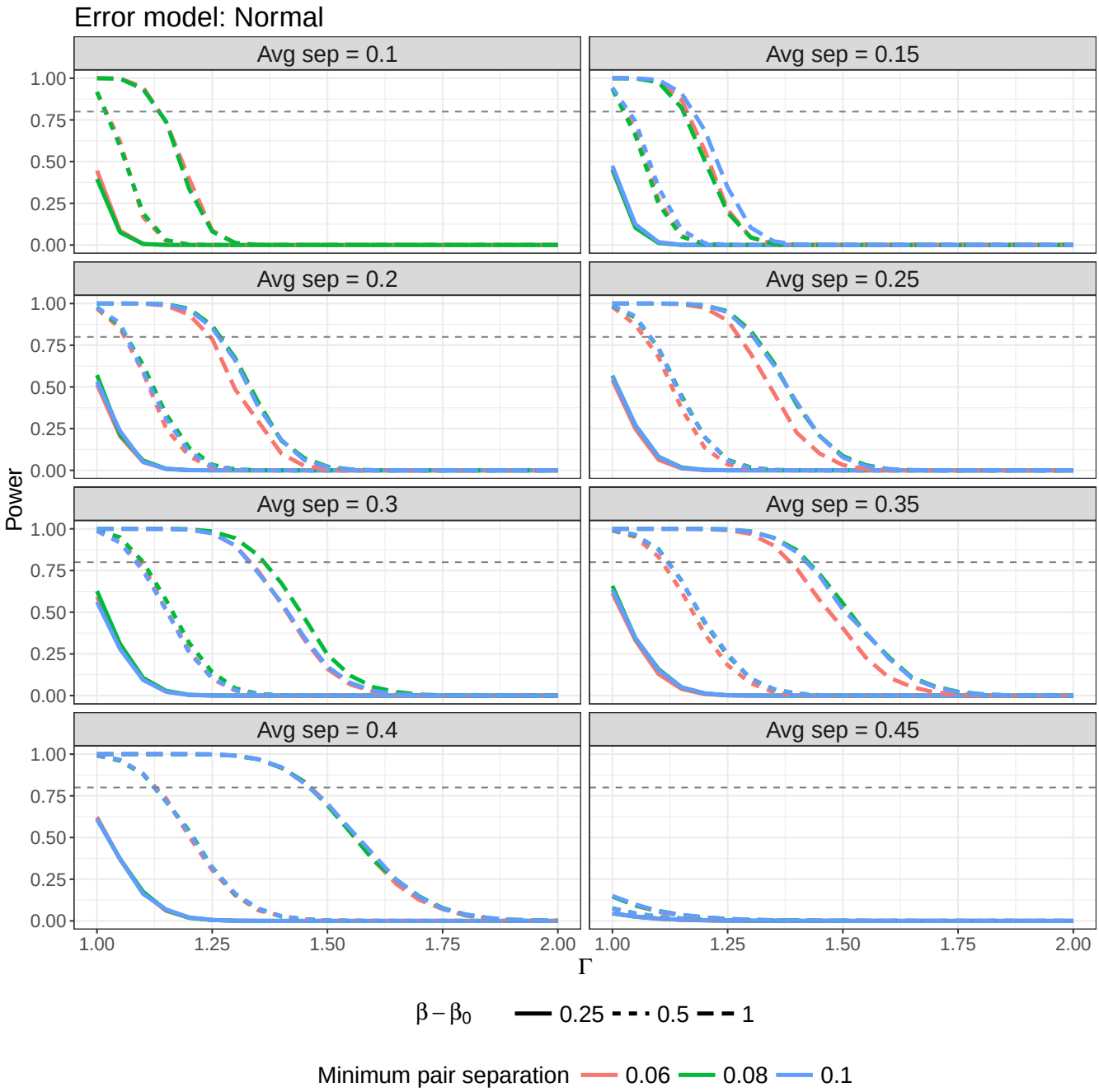


Figure 2.12: Power of a one-sided level-0.025 sensitivity analysis plotted against Γ for the Normal error model. Each line represents a matched set, with the color representing the estimated proportion of compliers. The rows are for different values of the minimum pair separation and the columns are for the degrees of departure from $H_0 : \beta = \beta_0$. The dashed horizontal line denotes where power = 0.8.

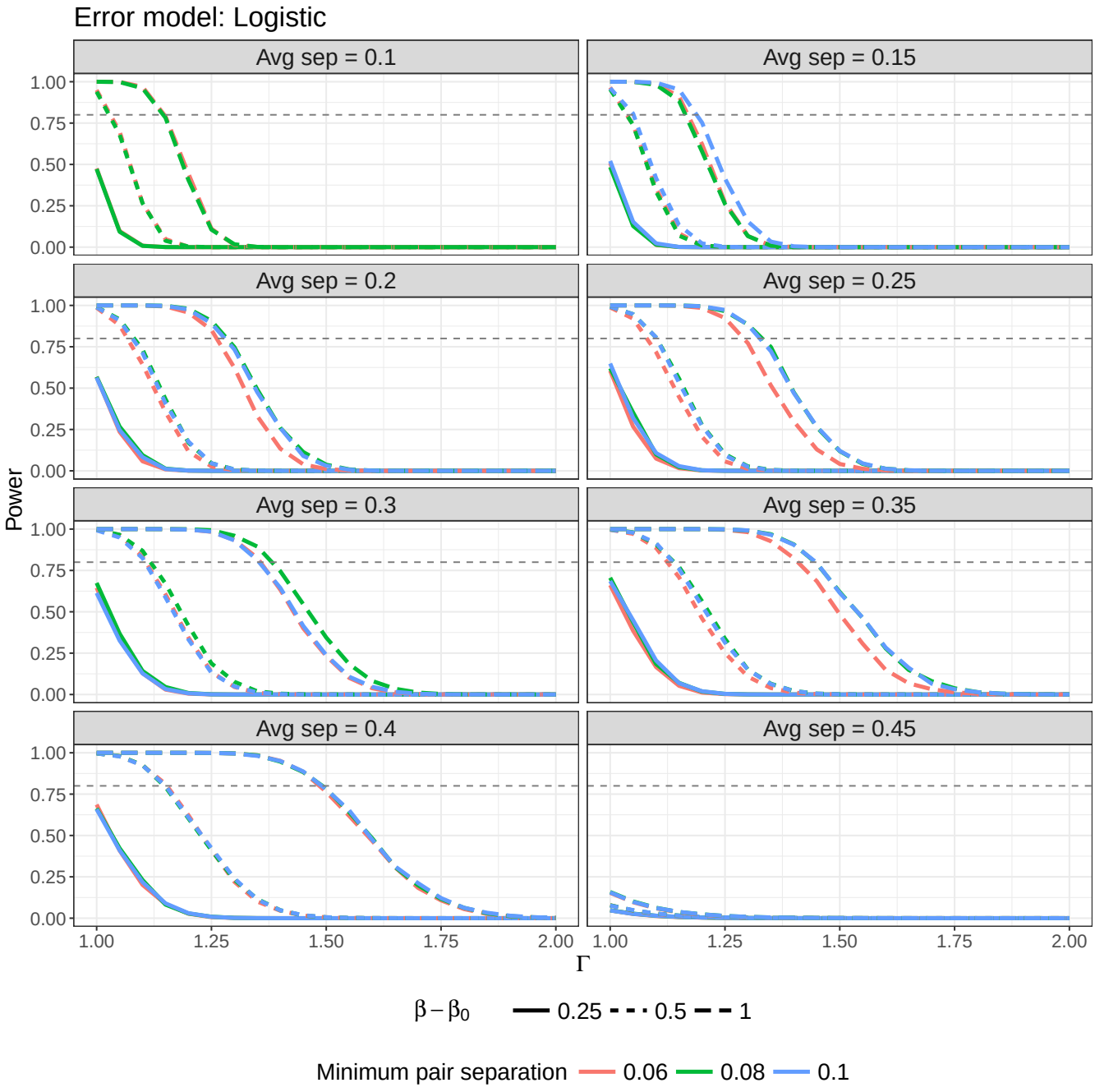


Figure 2.13: Power of a one-sided level-0.025 sensitivity analysis plotted against Γ for the Logistic error model. Each line represents a matched set, with the color representing the estimated proportion of compliers. The rows are for different values of the minimum pair separation and the columns are for the degrees of departure from $H_0 : \beta = \beta_0$. The dashed horizontal line denotes where power = 0.8.

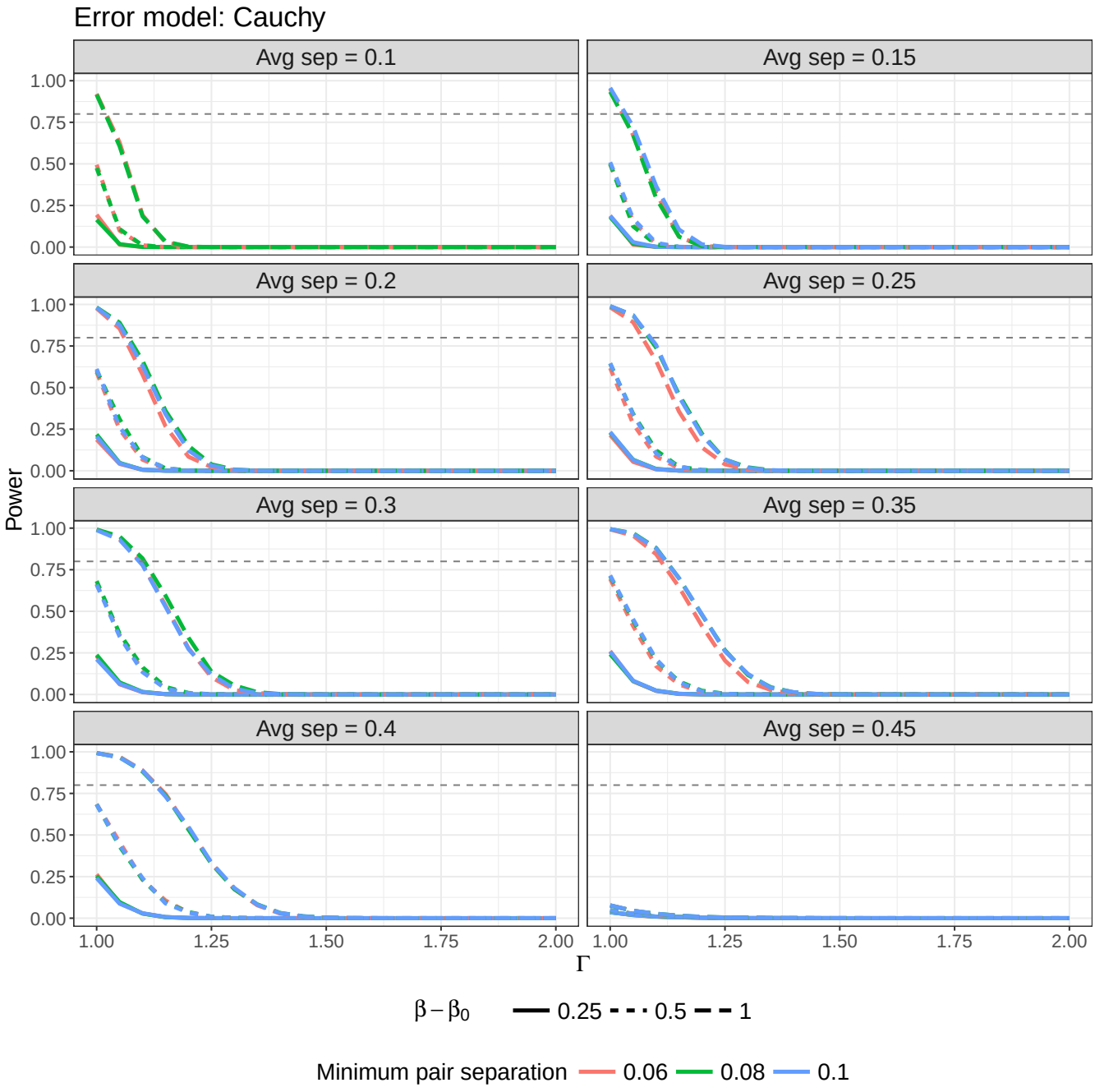


Figure 2.14: Power of a one-sided level-0.025 sensitivity analysis plotted against Γ for the Cauchy error model. Each line represents a matched set, with the color representing the estimated proportion of compliers. The rows are for different values of the minimum pair separation and the columns are for the degrees of departure from $H_0 : \beta = \beta_0$. The dashed horizontal line denotes where power = 0.8.

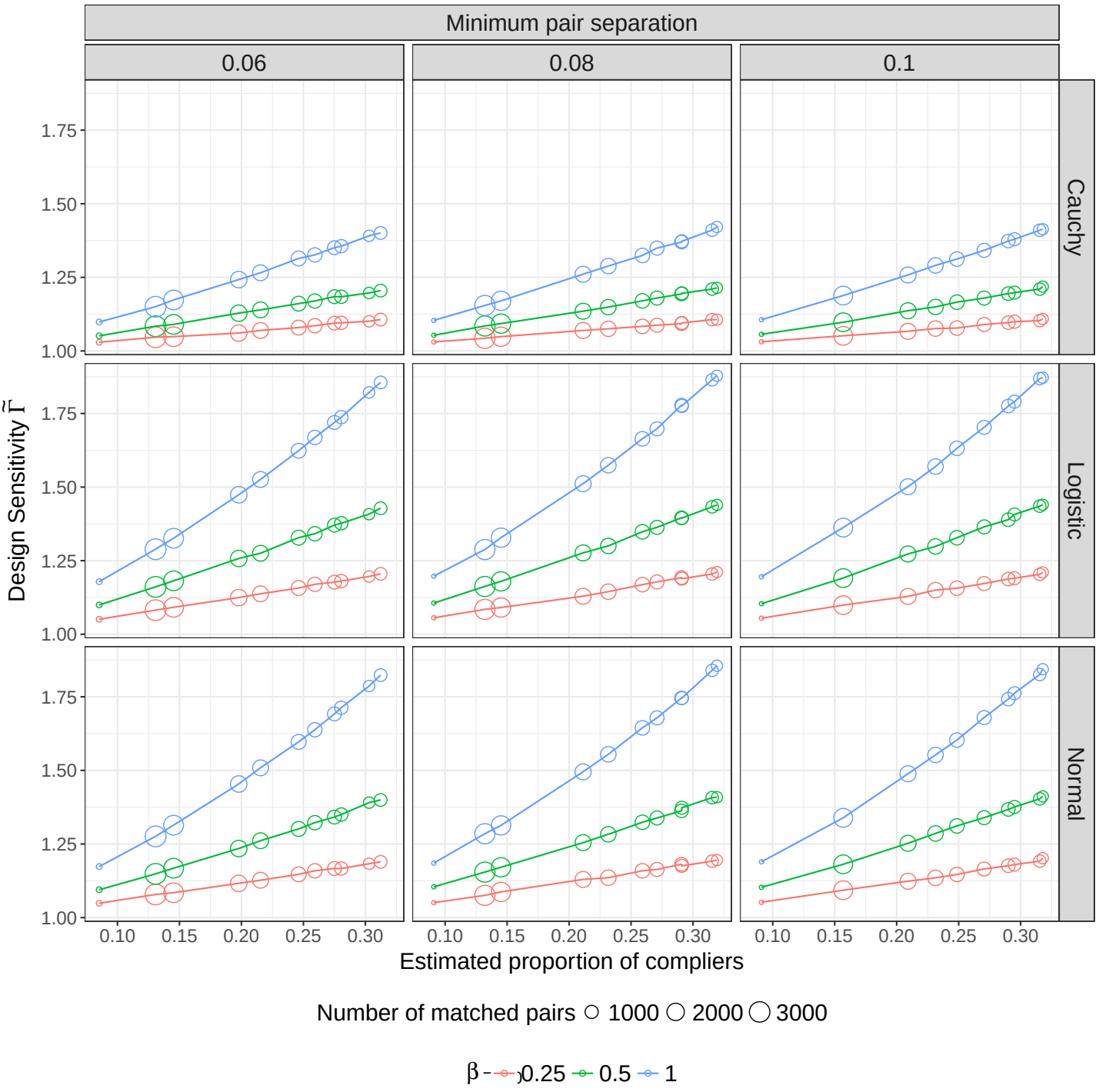


Figure 2.15: Design sensitivity $\tilde{\Gamma}$ plotted against instrument strength, measured as the difference in incarceration rates. Each circle represents a different match, with the size of the circle proportional to the number of pairs in the match.

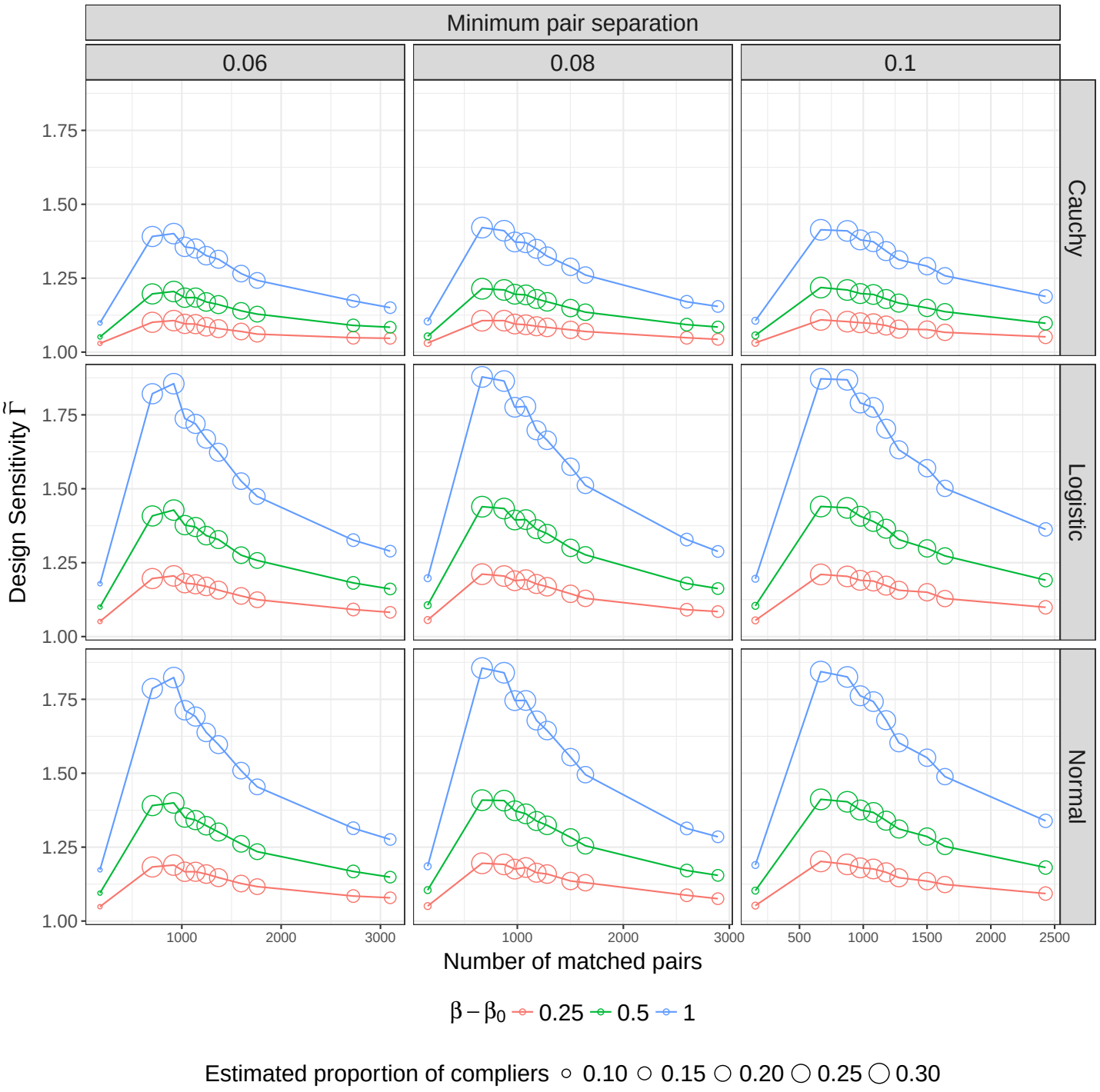


Figure 2.16: Design sensitivity $\tilde{\Gamma}$ plotted against the number of matched pairs. Each circle represents a different match, with the size of the circle proportional to the estimated proportion of compliers, which is the difference in incarceration rates in the match.

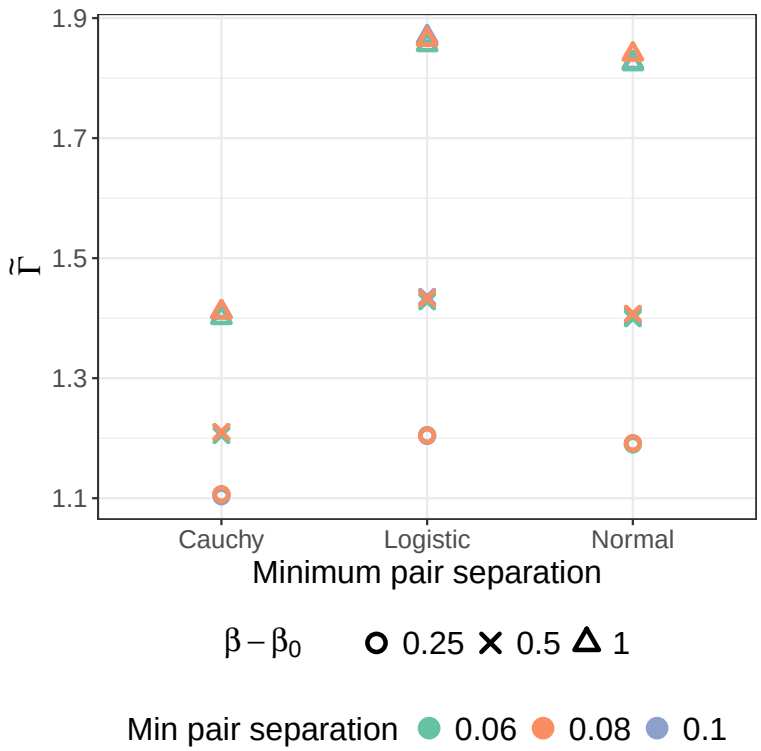


Figure 2.17: Design sensitivity $\tilde{\Gamma}$ plotted against three values of minimum pair separation for matches with average separation $\omega = 0.4$. The color of each point represents the departure from $H_0 : \beta - \beta_0$, while the shape represents the error model.

2.5.4 The Selected Match

Using the simulated power and design sensitivity, along with considerations for balance of important covariates, we select the match formed by requiring the minimum pair separation to be $\kappa = 0.08$ and the average separation to be $\omega = 0.4$. This match has $I = 875$ pairs and an estimated proportion of compliers $\widehat{\overline{\eta}} = 0.32$. The unmatched data have a total of $N = 53\,318$ subjects, from whom we could construct at most 18 874 matched pairs if we were to match only within county. In the unmatched data with $N = 53\,318$ offenders, the estimated proportion of compliers is 0.04. Though we have reduced our sample size by a factor of $53\,318/(2 * 875) = 30$, the instrument is now $0.32/0.04 = 8$ times stronger. Our simulations of the power of a sensitivity analysis and design sensitivity indicate that the improvements in instrument strength are worth this tradeoff. In addition, Ertefaie *et al.* show via simulation that, with a perfect instrument, a 5-fold loss in instrument strength requires a 25-fold increase in sample size to maintain the same power, and the required increase in sample size is even larger for an imperfect instrument.

Figure 2.18 shows the absolute standardized differences in means before (blue) and after (red) matching. While matching has made the balance on year (1998-2000) slightly worse, the difference is not above the conventional threshold of 0.1. Since our data only cover three years during which sentencing guidelines did not change, this slight imbalance is not concerning. The balance in both PRS and OGS is has improved after matching, and while age is slightly less balanced, the difference in means is still quite low at 0.025; see Figure 2.19. Because we have exactly matched for county, sex, race, and felony charges, those variables are perfectly balanced between the encouraged and unencouraged groups.

The matched dataset consists of five counties: Bucks, Dauphin, Lehigh, Mercer, and Philadelphia. This is in contrast to the 54 counties in the unmatched data and 67 counties in Pennsylvania as a whole. These counties include some of the largest cities in Pennsylvania (Philadelphia, Allentown), and Dauphin county includes the state

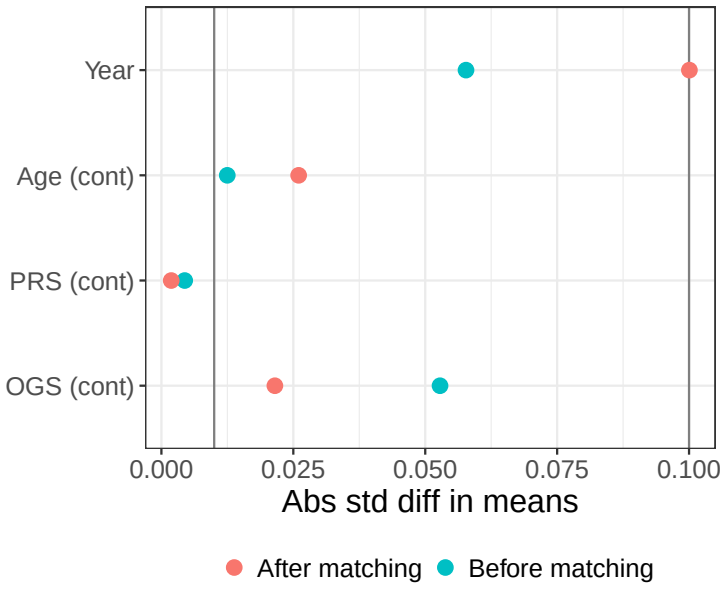


Figure 2.18: Absolute standardized differences in means for the selected match before (blue) and after (red) matching. Vertical grey lines are drawn at 0.01 and 0.1 to denote more and less conservative thresholds for balance.

capital of Harrisburg. The range of population sizes these five counties is quite large: Philadelphia is the largest county in the state, with a population of over 1.5 million in 2010, while Bucks county has 625,000 residents, Lehigh has 350,000, Dauphin has 268,000, and Mercer has 117,000. Despite the small number of counties in the match, they represent the geographic variation in Pennsylvania with a large urban county like Philadelphia, a small rural county like Mercer County, and counties with dense suburbs like Bucks County.

The matched data do not greatly differ from the unmatched data in terms of observed covariates. Table 2.1 shows the compositions of the matched and unmatched data in terms of the proportions of offenders who are male, white, and had felony charges, as well as the mean OGS, PRS, and age at date of offense. The matched data have slightly more males and fewer white offenders than the unmatched data, which is not surprising given that the population of Philadelphia County is 45% white while

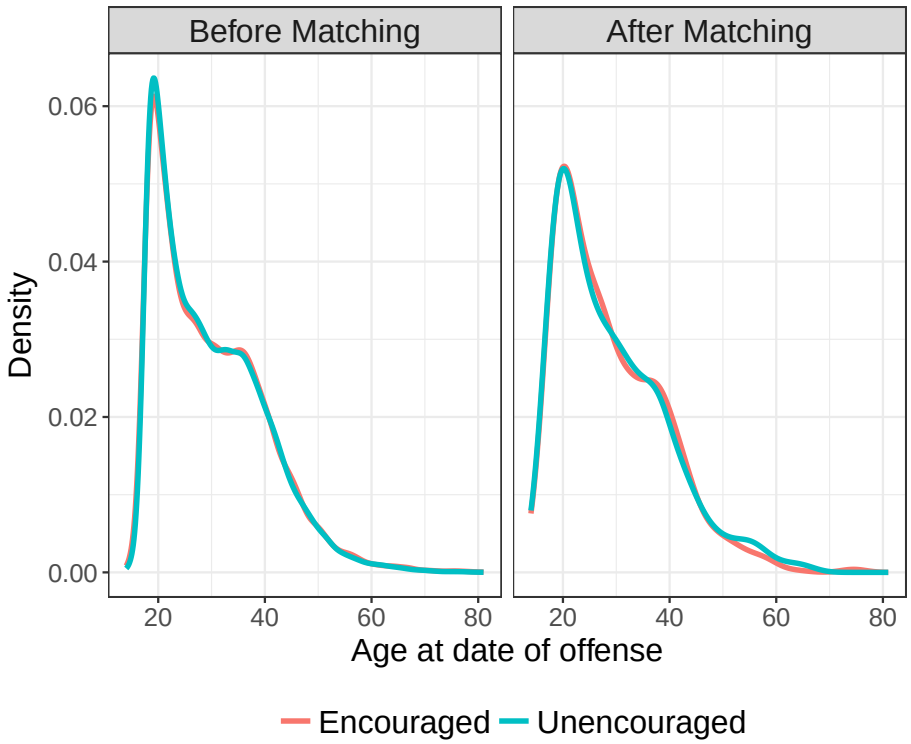


Figure 2.19: Density plot of age at date of offense for the selected match before (blue) and after (red) matching.

Covariate	Matched data	Unmatched data
Male	0.88	0.82
White	0.44	0.56
Felony charges	0.42	0.41
OGS	4.18	3.95
PRS	1.41	1.28
Age	29.02	29.65

Table 2.1: Compositions of the selected match and the unmatched data in terms of observed covariates. Values for the first three rows are the proportion of offenders falling into the given category, while the last three rows are means.

the other counties range from 72% to 93% white. The proportion of offenders facing felony charges is nearly equal. The OGS and PRS values are also slightly higher for the matched data, while age at date of offense is about six months older in the unmatched data. In removing observations from the data, we change the estimand and the population we are studying, but to the extent that this more restricted dataset better represents the population of compliers, we will obtain improved estimates of the causal effect of interest.

2.6 Effect Modification of the ITT

2.6.1 Methods for Discovering Effect Modification

Effect modification is an interaction between a treatment and a pretreatment covariate that affects the magnitude or stability of the treatment effect. If a treatment effect is smaller or more variable in some groups than in others, it is likely to be more sensitive to unobserved covariates. On the other hand, a large treatment effect is more difficult to explain away as the product of bias in treatment assignment. (Rosenbaum, 2004,

2005, 2010). Examining effect modification can uncover groups in the data in which estimates of treatments effects are less easily explained away by bias due to unobserved covariates.

We investigate effect modification for the intention-to-treat (ITT) effect, a relevant intermediate quantity for our analysis. The ITT is the causal effect of the encouragement on the outcome of interest. In our case, this is the effect of being assigned to a harsh judge on recidivism, and we denote it by $\overline{\delta} = (2I)^{-1} \sum_{i.j} \delta_{ij}$, where $\delta_{ij} = r_{Tij} - r_{Cij}$. The ITT ignores compliance, comparing recidivism in those assigned to a harsh judge with those assigned to a lenient one. Of course, we are most interested in the causal effect of the treatment, incarceration, but the ITT is simpler to handle because it focuses on the randomized encouragement assignment. Our investigation of effect modification of the ITT uses recently developed methods that allow for data-based discovery of covariates that may interact with the treatment. While subject-matter knowledge can often help guide the search for covariates that are likely to be effect modifiers, a data-driven method for discovering subgroups of the data or subpoplations that exhibit evidence of effect modification is appealing; not only can such a procedure save time, it also lessens researcher degrees of freedom. These methods are both exploratory, in that they use the data to discover which subgroups of the data show evidence of effect modification, and confirmatory, in that the familywise error rates arising from testing multiple hypotheses with correlated test statistics and the resulting sensitivity analyses are controlled at a prespecified level.

We use three recently developed methods that strongly control the familywise error rate of sensitivity analyses. All use pair-matched data that are exactly matched for a set of covariates $\mathbf{x}_{ij}$ so that $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ in all pairs i. The first method, proposed by Hsu *et al.* (2013), creates a classification and regression tree (CART) (Breiman *et al.*, 1984) to predict the ranks of $|Y_i|$, the absolute value of the observed encouraged-minus-unencouraged pair difference in outcomes, $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$, using the

covariates $\mathbf{x}_i$. The leaves of the regression tree define $G \geq 1$ groups $\mathcal{G} = \{s_1, \ldots, s_G\}$, where each s_g is a subset of the pair indices $i = 1, \ldots : L$, $s_g \subseteq \{1, \ldots, I\}$. It is these G groups in which we test for the presence of effect modification.

Using $|Y_i|$ instead of Y_i is crucial here. Consider testing Fisher's sharp null hypothesis H_0 of no effect, in which case $r_{Tij} = r_{Cij}$ for all ij. Then $|Y_i| = |r_{i1} - r_{i2}| = |r_{Ci1} - r_{Ci2}|$ is a function of $\mathcal{F}$, so the tree is fixed regardless of the distribution of encouragement assignments $\mathbf{Z}$, and a test statistic calculated using the observed values of Z_{ij} and Y_i for $i \in s_g$ can be bounded under H_0 as in (2.8) (Hsu *et al.*, 2013, 2015). As summarized by Hsu *et al.* (2015), Hsu *et al.* (2013) calculate p-values for the bounds in (2.8) for each of the G subgroups and combine them using the truncated product of Zaykin *et al.* (2002) for an overall test of Fisher's sharp null H_0. In this way, they pool evidence from the subgroups to test the null hypothesis of no effect in any of the G subgroups.

Of course, if H_0 is rejected, the next natural question is for which subgroups s_g the null hypothesis H_{0s_g} of no effect can be rejected. The problem here is that since the subgroups were generated from a regression tree constructed from the specific data at hand, so were the subgroup null hypotheses we wish to test. In other words, "[w]hat does it mean to speak about the probability of falsely rejecting $[H_{0s_g}]$ if most datasets would not lead us to test $[H_{0s_g}]$?" (Lee *et al.*, 2017a). Hsu *et al.* (2015) extends the work of Hsu *et al.* (2013) by proving that 1) if Fisher's strong null of no effect holds within subgroup g (i.e. H_{0s_g} is true), then the randomization distribution of treatment (encouragement) assignments within s_g has its usual null distribution, conditional on the groups $\mathcal{G}$ and $(\mathcal{F}, \mathcal{Z})$; and 2) the familywise error rate arising from testing null hypotheses within subgroups can be controlled at a prespecified level α if the test used would control the familywise error rate with groups that were fixed *a priori* instead of data-driven (Lee *et al.*, 2017a). Hsu *et al.* (2015) show that the control of the familywise error rate also applies in the case of a sensitivity analysis with bias of at most $\Gamma \geq 1$. They achieve this strong control by using the closed testing

procedure of Marcus *et al.* (1976). This procedure considers testing a hypothesis $H_{\mathcal{K}}$, where $\mathcal{K} \subseteq \{1, \ldots, G\}$, which states that there is no effect in any of the pairs in any of the subgroups $\bigcup_{g \in \mathcal{K}} s_g$. As described in Hsu *et al.* (2015), closed testing rejects $H_{\mathcal{K}}$ at level α if and only if it rejects at level α the null hypotheses $H_{\mathcal{L}}$ for all $\mathcal{L}$, where the set $\mathcal{L}$ is the set of all subsets of $\{1, \ldots, G\}$ that contain $\mathcal{K}$: $\mathcal{K} \subseteq \mathcal{L} \subseteq \{1, \ldots, G\}$.

The third method is the submax method and was proposed by Lee *et al.* (2017b). Its advantage is that its power and design sensitivity can be calculated via analytical formulas, while the above two CART-based methods require simulation. While it does not require matched pairs, we outline the method assuming that we have I pairs matched exactly for categorical covariates $\mathbf{x}_i$. Consider the G groups formed by the interactions of the covariate categories; with two categorical covariates with three levels each, there are $G = 3^2 = 9$ subgroups, and with L binary covariates, there are 2^L subgroups. In either case, the interaction subgroups quickly get sparse in terms of data as the number of covariates gets larger. Suppose we have L binary covariates. The submax method considers each of the $2L$ subgroups formed by splitting the population in two for each covariate separately. It does one overall test and $2L$ subgroup tests for a total of $2L + 1$ tests, which are highly correlated because the same data are used in many of the tests. For example, in our data we have binary covariates for sex, race, and whether the offense was a felony or not. Instead of $G = 2^3 = 8$ subgroup tests, we do $2L + 1 = 7$ tests. The tests for no effect among men and women are independent because separate portions of the data are used for each test. On the other hand, the two test statistics for women and men are correlated with the test statistics for race (white/not white) and felony (felony/not felony) because those statistics use the same people. Lee *et al.* (2017b) point out that the correction for multiple testing in this context turns out to be small because of the high correlation among the test statistics. In addition to testing Fisher's sharp null hypothesis of no effect for any individual, Lee *et al.* (2017b) further describe how to use closed testing to test subgroup-specific null hypotheses of no effect while strongly

controlling the familywise error rate at α.

For all of these methods, we use a test statistic that is a form of Maritz (1979)'s version of a Huber M-statistic as studied by Rosenbaum (2013). Specifically, we use

$$\sum_{i=1}^{I} \text{sign}(Y_i)\psi(|Y_i|/s), \tag{2.10}$$

where $\text{sign}(y) = 1$ if $y > 0$, $\text{sign}(y) = -1$ if $y < 0$, and $\text{sign}(y) = 0$ if $y = 0$; s is the median of $|Y_i|$; and $\psi(u)$ is given by

$$\psi(u) = (4/3)\text{sign}(u)\max(0, \min(h, |u|) - \iota), \tag{2.11}$$

where $h = 2$ and $\iota = 1/2$. This is an m-statistic that is similar to a trimmed mean and levels off at $h = 2$ times the median; see Rosenbaum (2013) for further discussion.

2.6.2 Results

We apply the above methods to the match discussed in Section 2.5.4, for estimating the causal effect of encouragement, in the form of being assigned to a harsh judge, on the number of arrests in the three years after sentencing. Our Figure 2.20 displays a density plot of the encouraged-minus-unencouraged number of arrests in the three years after sentencing. The density is highly peaked at zero, with long tails to the right and left. However, maybe we can discover subgroups in the data for which we can reject the null hypothesis that the ITT is zero.

Figure 2.21 shows the regression tree from the CART-based methods using the default complexity parameter of 0.05. Each node displays the mean value of $|Y_i|$ at that node, with the percentage of observations at that node shown below. The regression tree for the number of arrests identifies a prior record score of 0 as the most imortant variable to split on. For offenders with a prior record score greater than zero, the tree then splits on whether the offender is under age 23 or not. Offenders who have no prior record are split by sex.

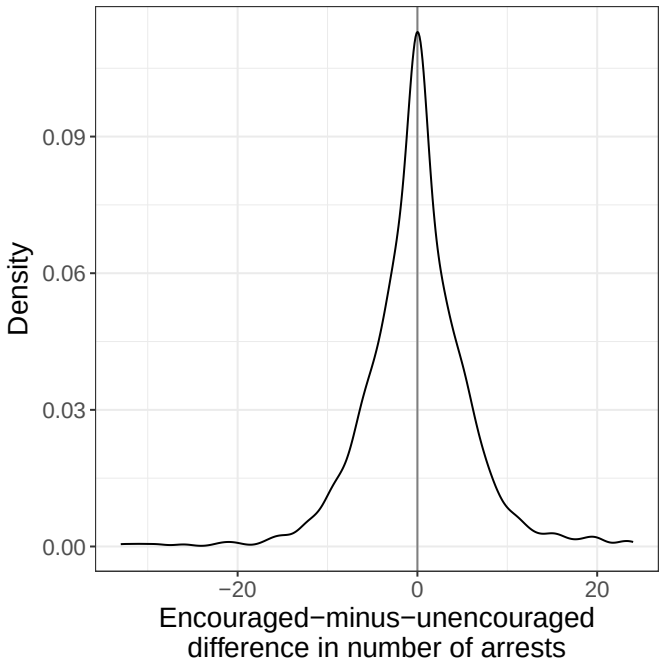


Figure 2.20: Density plot of the encouraged-minus-unencouraged difference in the number of arrests three years after sentencing.

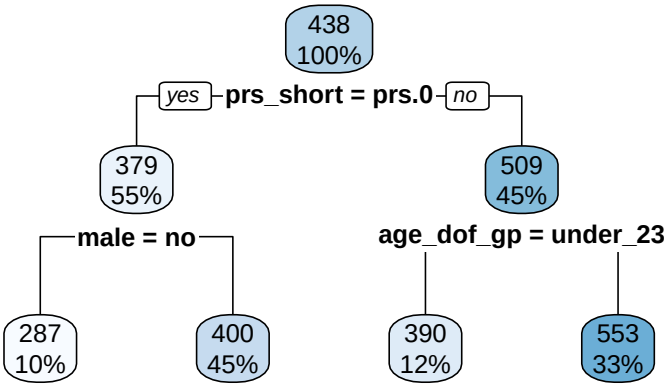


Figure 2.21: Regression tree for number of arrests in the three years after sentencing. Each leaf shows the predicted rank of $|Y_i|$ and the percentage of observations in that leaf.

	Group 1	Group 2	Group 3	Group 4	Pooled
PRS = 0	Yes	Yes	No	No	
Male	No	Yes	Both	Both	
Under 23	Both	Both	Yes	No	
Number of pairs	87	108	391	289	875
Arrests (mean), encouraged	1.69	5.18	2.71	6.73	4.24
Arrests (mean), unencouraged	1.75	4.98	2.93	6.82	4.35
Difference	-0.06	0.19	-0.21	-0.09	-0.11

Table 2.2: Number of arrests in three years after sentencing for each group defined by the leaves of the tree in Figure 2.21.

We display summary statistics for the groups identified by the regression tree in Table 2.2. The first three rows are for the variables that define the leaves of the tree. For example, the first group (leftmost leaf in Figure 2.21) is for women with PRS = 0, regardless of age since that variable was not used to define the first group. The fourth row shows the number of pairs in each group, followed by the mean number of arrests in the three years after sentencing and the encouraged-minus-unencouraged difference. The differences are small, less than one arrest in magnitude, and vary from slightly positive to slightly negative. We display the distributions of the differences in the number of arrests by group in Figure 2.22. For all groups, the differences are centered around zero and quite variable, particularly for the

For all three methods of discovering effect modification described above, we were not able to reject the null hypothesis of no effect, even under the assumption of a perfect instrument, $\Gamma = 1$. This is true for the global null hypothesis of no effect in any subgroup, as well as the subgroup-specific null hypotheses. Tables 2.3 to 2.6 are modeled after Table 1 in Hsu *et al.* (2015). Each table displays the upper bounds on one-sided p-values for the test of no effect for different values of Γ. In Table 2.3, the column labeled "Overall" calculates the p-value for the entire matched sample

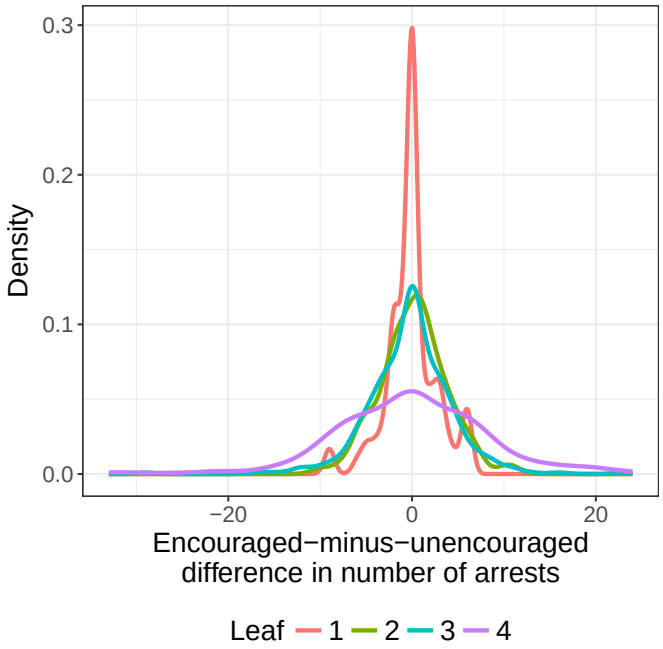


Figure 2.22: Density plot of the encouraged-minus-unencouraged difference in the number of arrests three years after sentencing for each leaf identified in the regression tree. The leaf numbers correspond to the leaves in Figure 2.21 from left to right.

Γ	Overall	Truncated
1	0.70	1
1.1	0.70	1
1.2	0.69	1

Table 2.3: Sensitivity analysis results for the ITT. For each value of Γ, the table shows the upper bound on a one-sided p-value. In the column "Overall", we calculate the p-value for the entire matched sample; here were are not looking for effect modification. In the column "Truncated", we calculate the p-values within each subgroup and pool them with the truncated product method from Zaykin *et al.* (2002).

using the Huber-Maritz test statistic in (2.10) as if we were not looking for effect modification. The column labeled "Truncated" calculates the p-value in each of the four groups defined by the leaves of the regression tree and combines them using the truncated product method of Zaykin *et al.* (2002); this column corresponds to the first method of discovering effect modification described in Section 2.6.1. The p-values are all well above 0.05, so we cannot reject the null hypothesis of no effect in the sample as a whole.

Tables 2.4 to 2.6 show the same p-value upper bounds, this time calculated separately for each comination of the four groups. Table 2.4 is for the groups taken two at a time, Table 2.5 is for the groups taken three at a time, and Table 2.6 is each group separately. The p-values are again large enough in each case that we cannot reject the null hypothesis of no effect. These tables correspond to the second method described in Section 2.6.1, and Propositions 1 and 2 of Hsu *et al.* (2015) ensure that the familywise error rate is controlled at level $\alpha = 0.05$, even though the groups we use here were not defined a priori and were instead discovered empirically from the data.

We also find no evidence of effect modification using the submax method, the third method described in Section 2.6.1. Here we test the null hypothesis of no effect in

Two groups

Γ	1,2	1,3	1,4	2,3	2,4	3,4
1	1	1	1	1	1	1
1.1	1	1	1	1	1	1
1.2	1	1	1	1	1	1

Table 2.4: Sensitivity analysis results for the ITT using the closed testing method of Hsu *et al.* (2015) for each combination of two groups. For each value of Γ, the table shows the upper bound on a one-sided p-value. Groups: 1 = women with PRS = 0; 2 = men with PRS = 0; 3 = offenders under 23 with PRS > 0; 4 = offenders over 23 with PRS > 0.

Three groups

Γ	1,2,3	1,2,4	1,3,4	2,3,4
1	1	1	1	1
1.1	1	1	1	1
1.2	1	1	1	1

Table 2.5: Sensitivity analysis results for the ITT using the closed testing method of Hsu *et al.* (2015) for each combination of three groups. For each value of Γ, the table shows the upper bound on a one-sided p-value. Groups: 1 = women with PRS = 0; 2 = men with PRS = 0; 3 = offenders under 23 with PRS > 0; 4 = offenders over 23 with PRS > 0.

	Individual groups			
Γ	1	2	3	4
1	0.68	0.34	0.83	0.60
1.1	0.67	0.35	0.82	0.59
1.2	0.66	0.35	0.81	0.59

Table 2.6: Sensitivity analysis results for the ITT using the closed testing method of Hsu *et al.* (2015) for each group separately. For each value of Γ, the table shows the upper bound on a one-sided p-value. Groups: 1 = women with PRS = 0; 2 = men with PRS = 0; 3 = offenders under 23 with PRS > 0; 4 = offenders over 23 with PRS > 0.

each of the eight populations defined by treating the four groups as binary variables. For example, group 1 is women with PRS = 0, so we test for no effect in women with PRS = 0 and in everyone else, namely all men and women with PRS > 0. Similarly, Group 2 is men with PRS = 0, so we test for no effect in men with PRS = 0 and in all women and men with PRS > 0. Because the data for each offender is used in exactly four of the eight test statistics, the test statistics are highly correlated. We apply the correction developed in Lee *et al.* (2017b) to the upper bound on the p-value, which yields a p-value for testing the global null hypothesis of no effect, and we also use closed testing to test the null hypothesis in each of the eight subpopulations. For $\Gamma = 1$, the p-value for the global null hypothesis of no effect is 0.70. For testing the null within each of the eight subpopulations, the critical value for $\Gamma = 1$ is $d_\alpha = 2.37$ for $\alpha = 0.05$, but the maximum deviate among the eight subpopulations is 0.46, so we cannot reject the null in any of the eight.

2.7 Estimating the Effect Ratio

Having looked for evidence of effect modification in the ITT, we now turn to estimating the effect ratio λ as defined i (2.1) in Section 2.4.2. We use the testing procedure developed in Baiocchi *et al.* (2010) to obtain point estimates and confidence intervals for the effect of incarceration on recidivism in terms of number of arrests in the three years after sentencing. The test statistic for testing $H_0 : \lambda = \lambda_0$ is $T(\lambda_0)/S(\lambda_0)$, where

$$T(\lambda_0) = \frac{1}{I}\sum_{i=1}^{I}\left(\sum_{j=1}^{2} Z_{ij}(R_{ij} - \lambda_0 D_{ij}) - \sum_{j=1}^{2}(1 - Z_{ij})(R_{ij} - \lambda_0 D_{ij})\right) = \frac{1}{I}\sum_{i=1}^{I} V_i(\lambda_0)$$

and

$$S^2(\lambda_0) = \frac{1}{I(I-1)}\sum_{i=1}^{I}\left(V_i(\lambda_0) - T(\lambda_0)\right)^2.$$

Specifically, we test $H_0 : \lambda = \lambda_0$ by comparing $T(\lambda_0)/S^2(\lambda_0)$ to the standard normal distribution. We obtain 95% confidence intervals and point estimates for λ by solving for $T(\lambda_0)/S(\lambda_0) = \pm 1.96$ and $T(\lambda_0)/S(\lambda_0) = 0$, respectively (Baiocchi *et al.*, 2010).

Using this method, the point estimate of λ is -0.37, with a 95% confidence interval of $[-1.75, 0.96]$. Thus, while the point estimate is negative and suggests that incarceration reduces the number of arrests in a three-year window after sentencing, our data are consistent with both positive and negative point estimates. As Baiocchi *et al.* (2010) describe, we can interpret λ as the ratio of two average treatment effects, that of the encouragement on the outcome to that of the encouragement on the treatment. A point estimate of -0.37 suggests that for every 100 offenders "encouraged" to be incarcerated by a harsh judge, there are 37 fewer arrests in the three years following sentencing. If the exclusion restriction and monotonicity assumptions hold, then we can interpret λ as the change in the number of arrests *caused by* being sentenced to either prison or jail.

2.8 Discussion

Instrumental variables are a popular method for causal inference, particularly in the context of natural experiments. The pitfalls of weak instruments are well known, but their potential utility in studying important causal effects that can only be investigated via observational studies means they are a staple of the social scientist's toolbox. This work furthers the utility of instrumental variables in observational studies in several ways. First, we demonstrate the use of recent advances in matching methods to increase instrument strength by enforcing restrictions on both the minimum within-pair difference κ and average across-pair difference ω in the value of the continuous instrument, here judge harshness. These methods allow for many forms of balance constraints, and while we only employ exact matching in this study, other balance requirements like mean balance and fine balance are straightforward to implement. By varying both κ and ω, we obtain multiple matched datasets with varying degrees of instrument strength, as measured by the estimated proportion of compliers: the difference in incarceration rates between encouraged and unencouraged offenders. Because nothing in the matching procedure uses the outcome, we are free to select the best match without fear of introducing bias into our outcome analysis.

Given the resulting set of matches, a researcher then has to choose one to use in her analysis. Balance on observed covariates is an important and standard criterion, but choosing between matches that have equally good balance but differ in terms of sample size and instrument strength is more difficult. To aid in this choice, we propose using simulation to examine the power of a sensitivity analysis for each match under different assumptions about the degree of unobserved bias, the error model and the effect size. Graphically exploring the relationship between power, sample size, and instrument strength is a useful way to weigh the tradeoff between the latter two factors in terms of the first. We also propose calculating the design sensitivity. As an asymptotic quantity, the design sensitivity does not help us with choosing the sample size, and as a monotonically increasing function of instrument strength (at

least in the current application), maximizing design sensitivity will always lead us to choose the match with the largest estimated instrument strength. However, the design sensitivity is another factor we can weigh along with balance on observed covariates and the power of a sensitivity analysis in choosing a match.

In strengthening the instrument, some of the matches resulted in worsened covariate balance compared to before matching. This is in part due to our matching strategy, which matches exactly for six categorical covariates: binary indicators for sex (male/female), race (white/not white), whether the offense was a felony, three age groups, three PRS groups, and three OGS groups. We used exact matching without enforcing other forms of balance like mean balance or fine balance because the methods we use to detect effect modification in the ITT require candidate covariates to be exactly matched. With a focus strictly on the effect ratio, we can employ a matching strategy that employs additional forms of balancing in hopes of better estimating that quantity.

The match we selected has a significantly smaller sample size than the unmatched data ($2 \times 875 = 1750$ offenders vs 53 318 offenders), but a much stronger instrument (difference in incarceration rates of 0.32 vs 0.04). The counties in the match are representative of the geographic and demographic diversity in Pennsylvania, including the large urban county of Philadelphia, the medium-sized suburban Bucks county, and the small rural Mercer county. In addition, the composition of the matched data in terms of observed covariates is close to that of the unmatched data. This match is ideal in terms of its balance on observed covariates and the power that it achieves in our simulations.

In addition to demonstrating methods for strengthening the instrument and choosing a match, we also demonstrate the application of three recently developed methods for effect modification. These methods are innovative in that they allow for the discovery of effect modification in subgroups discovered empirically in the data, instead of subgroups that are defined a priori, while maintaining strong control of the fam-

ilywise error rate. In addition, the submax method has the attractive property of having analytical expressions for its large-sample power, negating the need to study power via simulation.

We found no evidence of effect modification in the ITT, and indeed could not reject the null that the ITT is zero despite having chosen a match with a strong instrument. It is possible that the causal effect of being assigned to a harsh judge is indistinguishable from zero, or that it works in one direction for some offenders and the opposite for others. It may also be the case that there is some important covariate that we have not observed that is biasing the results toward zero. The estimate of the effect ratio yielded a negative point estimate with a 95% confidence interval that includes zero. As such, we cannot conclusively determine whether incarceration reduces reoffending in the first three years after arrest; assuming that all of the instrumental variable assumptions hold, our data are consistent with both increases and decreases in reoffending caused by incarceration.

One potential criticism of our approach is that it is subject to researcher degrees of freedom and the idea of the "Garden of Forking Paths" (**?**), which raises concerns about multiple comparisons even with just one analysis and asks whether the exact same data cleaning and analysis decisions would have been made with a different dataset. To the extent that our data cleaning process removes noise and bias that would otherwise adversely affect our analyses, it helps us achieve better estimates. At the same time, it is true that many of our data cleaning choices were motivated by the data at hand, and we may conceivably have made different choices with a different dataset.

As with any observational study, our ability to make causal claims about the effects of interest depend on how well the underlying assumptions are met. We have described these assumptions in detail and explained why we believe our data meet them and the ways in which they could be violated. As is often the case particularly with observational data, we cannot conclusively prove that the assumptions are met,

but our extensive graphical checks of the data support our claims that they are.

This work can be extended in several interesting and useful ways. First, while the number of arrests in the three years after sentencing is a common measure of recidivism, it would also be useful to investigate the time until the first rearrest after sentencing, or the probability of being rearrested within a certain time frame. Incarceration could conceivably have different effects on each of these outcomes, as well as different effects over different time periods. Analyzing different outcomes with the methods used here, possibly with extensions to control for multiple comparisons, would give a more nuanced picture of the effect of incarceration on recidivism.

Second, we can treat incarceration as multi-valued treatment prison, jail, and time served considered to be separate treatment options. This approach would help differentiate between the effects of being confined to state prison, which is generally for sentences longer than two years, and to county jail, which is for shorter sentences. In addition, because many offenders are jailed during their trial due to inability to make bail, they do experience incarceration even if not sentenced to additional prison or jail. Because many offenders, especially those who cannot make bail, are imprisoned for the duration of their trial, understanding the causal effects of time served is important.

Third, we have data on the duration of prison sentences for those offenders sentenced to state prison. Because we consider the treatment to be a sentence of either prison or jail and do not have information on the exact duration of jail sentences, we did not incorporate the duration of the prison sentence into this analysis. However, investigating the causal effects of the duration of a prison sentence is another important avenue of future research.

Chapter 3

Measurement Error in Hierarchical Models

3.1 Introduction

Many public health and social science research questions involve investigating the relationships between individual outcomes and variables at the neighborhood, community, and environmental levels (Diez-Roux, 2000; Kawachi and Berkman, 2003; Pickett and Pearl, 2001). Accurate measurement of neighborhood-level variables is critical for valid inference about these relationships (Diez-Roux, 2008), but often these variables are not measured or observed directly. Instead, they must be estimated from individual-level responses collected in sample surveys or measured via proxies from other sources. When neighborhood-level sample sizes (or the neighborhoods themselves) are small, or when the proxies do not accurately capture the construct or concept of interest, measurement error becomes a concern.

For example, the Demographic and Health Surveys (DHS) are household surveys conducted in lower-income countries and constitute an integral source of information key maternal and child health indicators. These surveys are conducted in multiple stages, with villages or census blocks of 100-200 households serving as primary sampling units (PSUs), from which roughly 30 households are sampled (Rutstein and Rojas, 2006). Some PSU-level covariates, like the proportion of households with an improved water source, can be estimated from variables measured for every sampled household, but other coviariates like the rate of school attendance among children aged 6-15 or the proportion of children with diarrhea in the last two weeks can only be measured for the subset of households with individuals in the relevant demographic groups. In this case, the reliability of PSU-level means may be lower than is desirable.

In other cases, neighborhood-level variables may not be directly measurable or observable, forcing researchers to rely on proxies that may be highly noisy or imperfectly capture the feature of interest. Diez-Roux *et al.* (2001) study the relationship between neighborhood of residence and incidence of coronary heart disease. The neighborhood-level covariate of interest is its socioeconomic environment. However, because this covariate is not directly observable, the authors generate a proxy using a

summary score consisting of a set of socioeconomic indicators. The reliability of the subsequent scientific conclusions then depend, in large part, on how well this proxy corresponds to the true underlying socioeconomic position. Mujahid *et al.* (2007) construct measures of neighborhood socioeconomic position aggregating individual-level survey responses to the neighborhood level and assess their psychometric (within-person) and ecometric (within-neighborhood) properties. While their results indicate that their constructed measures have good psychometric and ecometric properties, they note that some neighborhoods had small sample sizes and "future research is needed to examine the consequences of using simple means or empirical Bayes estimates as predictors of health outcomes."

As noted by Muff and Keller (2015), measurement error has a long history in statistics, dating back to Pearson (1902) and Wald (1940), and ignoring measurement error can lead to biased estimates and misleading confidence intervals (Fuller, 2009). Not only can measurement error attenuate the estimated coefficient of the mismeasured variable, it can also affect the coefficients of other variables in the model, and the degree and direction of the impact of the mismeasured variable on other coefficients depends, among other things, on the correlation between the mismeasured variable and the other variables in the model (Carroll *et al.*, 1985; Gustafson, 2003). In a hierarchical model, where we may have measurement error at one level and coefficients of particular interest at another, this interplay can be particularly important.

In this chapter, we consider measurement error in the context of hierarchical models. Specifically, we investigate the consequences of ignoring measurement error in a group-level covariate, where the error is due either to sampling or to classical measurement error, and demonstrate that ignoring this error leads to biased and inefficient estimates. We present a single Bayesian framework for explicitly incorporating both types of measurement error into the outcome model and conduct a simulation study to demonstrate that accounting for measurement error in this way leads to greatly improved inference. We implement our methods in Stan (Stan Development Team,

2016b), a probabilistic programming language that allows for fully Bayesian inference via MCMC sampling using Hamiltonian Monte Carlo (HMC) (Stan Development Team, 2016b). Other Bayesian approaches have used Gibbs sampling (Richardson and Gilks, 1993; Bernardinelli *et al.*, 1997) and integrated nested Laplace approximations (INLA) (Muff *et al.*, 2015), but they lack the flexibility and intuitive model specification framework that makes incorporating the measurement error process into outcome modeling simple and straightforward.

3.2 Methods

3.2.1 Sampling-induced measurement error

In the context of survey sampling, measurement error occurs when, for example, we estimate group- or area-level variables using sample averages of unit-level variables. Consider a population consisting of J primary sampling units (PSUs), such as villages or census enumeration blocks. Each of the $j = 1, \ldots, J$ PSUs consists of N_j units (e.g. individuals or households), with a total population size of $N = \sum_{j=1}^{J} N_j$. Suppose we have a sample from this population taken under a two-stage sampling design, with J_s PSUs sampled in the first stage and n_j units within each selected PSU sampled in the second stage. The total sample size is then $n = \sum_{j=1}^{J_s} n_j$ out of a total population size of $N = \sum_{j=1}^{J} N_j$. Let S_j denote the set of all individuals i in PSU j and let s_j denote the set of sampled individuals i in PSU j, so that $|S_j| = N_j$ and $|s_j| = n_j$.

Suppose the survey collects information on a binary unit-level covariate z_i and a unit-level outcome y_i. We wish to understand the relationship between y_i and the unit-level covariate z_i and a PSU-level characteristic $\theta_j \in [0, 1]$, where $z_i \sim \text{Bern}(\theta_{j[i]})$. Here θ_j acts as a latent prevalence of (or propensity for) $z_i = 1$, with a realized PSU-level prevalence of $p_j^* = 1/N_j \sum_{i \in S_j} z_i$.[1] Since we do not have data on all units in PSU j, we cannot calculate p_j^*. The sample proportion $p_j = 1/n_j \sum_{i \in s_j} z_i$ is the maximum

[1] In many public health and social science applications, the more relevant PSU-level characteristic

likelihood estimator for θ_j, and from the Central Limit Theorem, we know that its standard error approaches $\sqrt{p_j(1-p_j)/n_j}$, which has a maximum of $0.5/\sqrt{n_j}$. For $n_j = 30$, a typical within-PSU sample size in many large-scale household surveys like the Demographic and Health Surveys (DHS), $0.5/\sqrt{n_j} = 0.09$.

However, in many cases, the within-PSU sample size may be much smaller, particularly if the PSU-level characteristic of interest is measured for only a subset of the sampled units. For example, in a survey where PSUs are villages and units are households, if the PSU-level variable of interest is measured for a specific demographic (e.g. children under five or women of childbearing age), the sample size will be smaller if not every household has respondents in that demographic. In these cases, p_j may be a poor estimator for θ_j.

3.2.2 Classical measurement error

Suppose we have information on an outcome y_i and covariate z_i for individuals i who are nested in neighborhoods j. Suppose we wish to relate y_i and z_i to a neighborhood-level variable u_j such as socioeconomic environment or social cohesion. The true variable u_j is unobserved, perhaps because it is inherently unobservable or because existing instruments are imperfect, and in its place we have the noisy proxy w_j. In many applications, it is reasonable to assume that we have repeated measurements of w for each u, so that

$$w_{kj} = u_j + \epsilon_{kj}, \quad k = 1, \ldots, m_j, \quad j = 1, \ldots, J,$$

where ϵ_{jk} denotes the error term and m_j the number of repeated measurements for unit j. However, here we consider the case of $m_j = 1$, so there is only one measurement of the proxy w_j for each true u_j. We assume that the measurement errors are independent and follow a normal distribution with mean zero and constant variance

may be the realized prevalence p_j^*. However, it differs negligibly from θ_j for large enough N_j, so we ignore this distinction and focus on θ_j.

σ_ϵ^2: $\epsilon_j \sim N(0, \sigma_\epsilon)$. This assumption then implies that the measurement error is non-differential, meaning that ϵ_j gives no additional information about the outcome y, conditional on the noisy proxy w and other relevant (accurately-measured) covariates (Carroll *et al.*, 2006).

3.2.3 Measurement error in a Bayesian framework

In epidemiology, measurement error models are often broken down into three sub-models (Richardson and Gilks, 1993). The first is a disease model that describes the relationship between the outcome y and the true risk factor r, and possibly other accurately measured covariates x. Next is a measurement model that relates the true risk factor r to the mismeasured surrogate s, and last is the exposure model that describes the distribution of the true risk factor r in the population. For the sampling-induced measurement error described in Section 3.2.1, suppose the disease, measurement, and exposure models are

$$\begin{aligned}
\text{disease model:} \quad & y_i \sim \mathrm{N}(\beta_{0j[i]} + \beta_1 z_i, \sigma_y^2) \\
& \beta_{0j} \sim \mathrm{N}(\alpha_0 + \gamma_0 \theta_j, \sigma_{\beta_0}^2) \\
& z_i \sim \mathrm{Bern}(\theta_{j[i]}) \\
\text{measurement model:} \quad & \sum_{i \in s_j} z_i \sim \mathrm{Bin}(n_j, \theta_j) \\
\text{exposure model:} \quad & \mathrm{logit}(\theta_j) \sim \mathrm{N}(\mu, \tau^2)
\end{aligned}$$

For classical measurement error as described in Section 3.2.2, the three models would be

$$\begin{aligned}
\text{disease model:} \quad & y_i \sim \mathrm{N}(\beta_{0j[i]} + \beta_1 z_i, \sigma_y^2) \\
& \beta_{0j} \sim \mathrm{N}(\alpha_0 + \gamma_0 u_j, \sigma_{\beta_0}^2) \\
\text{measurement model:} \quad & w_j \sim N(u_j, \sigma_\epsilon) \\
\text{exposure model:} \quad & u_j \sim N(0, \sigma_u).
\end{aligned}$$

We fit the models in Stan, a probabilistic programming language that allows for fully Bayesian inference. The straightforward way of specifying statistical models in Stan means that the disease, measurement, and exposure models can be coded almost exactly as written above. We include weakly informative priors on the regression, variance, and measurement error parameters, but we can easily incorporate additional information on the measurement error process into the prior.

3.3 Simulation study

We conduct a simulation study to illustrate the effects of measurement error in a group-level covariate in a hierarchical model with both unit- and group-level predictors and investigate the performance of our proposed methods in comparison to a naive model that ignores measurement error. The first simulation scenario considers the case of measurement error in a cluster-level covariate induced by sampling as described in Section 3.2.1, while the second considers classical measurement error as in Section 3.2.2.

We implement our proposed models in Stan, which generates posterior samples in a fully Bayesian framework. We carefully monitor the detailed warnings and diagnostics that Stan provides to detect when posterior inferences may be unreliable due to difficulties in sampling. Divergent transitions indicate that the sampler is unable to explore a portion of the parameter space, which can lead to significant bias in the resulting posterior distribution and ultimately unreliable inferences (Stan Development Team, 2016c). Stan reports the number of divergent transitions for each run, and even one divergent transition indicates that the results may be suspect. If divergent transitions occur, we follow the recommendation of Stan developers and iteratively increase the target acceptance rate `adapt_delta` (Stan Development Team, 2016a). We also monitor the estimated potential scale reduction factor $\widehat{R}$, a diagnostic that assesses the mixing of the chains; at convergence, $\widehat{R} = 1$. If $\widehat{R} \geq 1.1$ for any pa-

rameter, we increase the number of iterations by 1000 until all values of $\widehat{R}$ are less than 1.1, up to 7000 iterations. If values of $\widehat{R} \geq 1.1$ remain with 7000 iterations, we discard the simulation. The results presented here are based on a minimum of 75 simulations for each scenario.

3.3.1 Measurement error from sampling

This simulation considers a simple linear model with a group-level varying intercept:

$$\begin{aligned} y_i &\sim \mathrm{N}(\beta_{0j[i]} + \beta_1 z_i, \sigma_y^2) \\ \beta_{0j} &\sim \mathrm{N}(\alpha_0 + \gamma_0 \theta_j, \sigma_{\beta_{0j}}^2) \\ z_i &\sim \mathrm{Bern}(\theta_{j[i]}) \\ \mathrm{logit}(\theta_j) &\sim \mathrm{N}(0, 1). \end{aligned} \tag{3.1}$$

We use this data-generating model to create a fixed population as follows. We create $J = 3850$ PSUs consisting of N_j units each, where N_j is sampled uniformly from the integers 100 to 300, $j = 1, \ldots, J$. We draw the latent cluster-level variable θ_j as $\mathrm{logit}(\theta_j) \sim \mathrm{N}(0, 1)$ and set the true regression parameters to $\alpha_0 = 0.2$, $\gamma_0 = 2$, $\sigma_{\beta_0} = 0.2$, $\beta_1 = 0.75$ and $\sigma_y = 0.05$; we also create a binary outcome y_i as $\Pr(y_i = 1) = \mathrm{logit}^{-1}(\beta_{0j[i]} + \beta_1 z_i)$. We then repeatedly take samples from this fixed population using a two-stage sampling scheme. In the first stage, we sample $J_s < J$ clusters via simple random sampling (SRS), and then sample n_j units from each selected cluster, with $J_s \in \{5, 15, 50\}$ and $n_j \in \{10, 30, 60\}$.

We fit two models to each sample: the true model given in (3.1) and a naive model that uses the observed proportion p_j as the PSU-level predictor:

$$\begin{aligned} y_i &\sim \mathrm{N}(\beta_{0j[i]} + \beta_1 z_i, \sigma_y^2) \\ \beta_{0j} &\sim \mathrm{N}(\alpha_0 + \gamma_0 p_j, \sigma_{\beta_{0j}}^2), \end{aligned} \tag{3.2}$$

We also consider the case of binary y. Here the naive and full models are analogous to (3.1) and (3.2), and the outcome model is given by

$$\Pr(y_i = 1) = \mathrm{logit}^{-1}(\beta_{0j[i]} + \beta_1 z_i).$$

In all cases, we use weakly informative prior distributions for the regression parameters Gelman (2006):

$$\alpha_0, \gamma_0, \beta_1 \overset{ind}{\sim} \mathrm{N}(0, 10^2)$$
$$\sigma_{\beta_0}, \sigma_{\beta_1}, \sigma_y \overset{ind}{\sim} \mathrm{Cauchy}^+(0, 2.5),$$

where $\mathrm{Cauchy}^+(0, 2.5)$ denotes a Cauchy distribution with location 0 and scale 2.5 restricted to positive values.

3.3.2 Classical measurement error

This scenario considers classical measurement error in a group-level predictor. The true group-level predictor is u_j, generated as $u_j \sim N(0, 1)$, but we only observe the noisy proxy w_j, where $w_j \sim \mathrm{N}(u_j, \sigma_\epsilon)$ and $\epsilon = 0.5$. We generate a population of $J = 200$ clusters of size N_j, where N_j is generated as $N_j \sim \mathrm{NegBin}(\mu, \phi) + 5$ with the mean parameter $\mu = 30$ and the dispersion parameter $\phi = 8$ (we add 5 to avoid $N_j = 0$). The clusters therefore have an average size of 35 with a standard deviation of $\sqrt{(}\mu + \mu^2/\phi) = 11.9$. The data-generating model is similar to that in (3.1):

$$\begin{aligned} y_i &\sim \mathrm{N}(\beta_{0j[i]} + \beta_1 z_i, \sigma_y^2) \\ \beta_{0j} &\sim \mathrm{N}(\alpha_0 + \gamma_0 u_j, \sigma_{\beta_{0j}}^2) \\ w_j &\sim \mathrm{N}(u_j, \sigma_\epsilon). \end{aligned} \tag{3.3}$$

The true parameter values are: $\alpha_0 = -1.5$, $\gamma_0 = 0.8$, $\sigma_{\beta_0} = 0.1$, $\beta_1 = 0.5$, and $\sigma_y = 0.5$. The naive model uses the observed proxy w_j in place of u_j as the group-level predictor:

$$\begin{aligned} y_i &\sim \mathrm{N}(\beta_{0j[i]} + \beta_1 z_i, \sigma_y^2) \\ \beta_{0j} &\sim \mathrm{N}(\alpha_0 + \gamma_0 w_j, \sigma_{\beta_{0j}}^2), \end{aligned} \tag{3.4}$$

We also consider a binary outcome y, with $\Pr(y_i = 1) = \mathrm{logit}^{-1}(\beta_{0j[i]} + \beta_1 z_i)$ as before. The priors are uninformative as described above.

In contrast to the previous simulation scenario, in which we subsample from a fixed population, in this scenario we generate a new population for each simulation

and fit the model to that entire population. We do not include the additional step of sampling from the population because our goal is to evaluate how well the full model adjusts for measurement error in w_j, and including sampling would simply add unnecessary noise.

We also expand the models in (3.3) and (3.4) to include varying slopes in addition to varying intercepts. The data-generating model is then given by

$$\begin{aligned} y_i &\sim \mathrm{N}(\beta_{0j[i]} + \beta_{1j[i]} z_i, \sigma_y^2) \\ \begin{pmatrix} \beta_{0j} \\ \beta_{1j} \end{pmatrix} &\sim \mathrm{N}\left(\begin{pmatrix} \alpha_0 + \gamma_0 u_j \\ \alpha_1 + \gamma_1 u_j \end{pmatrix}, \Sigma \right) \\ w_j &\sim \mathrm{N}(u_j, \sigma_\epsilon), \end{aligned} \tag{3.5}$$

where the covariance matrix Σ is

$$\Sigma = \begin{pmatrix} \sigma_{\beta_0}^2 & \rho\sigma_{\beta_0}\sigma_{\beta_1} \\ \rho\sigma_{\beta_0}\sigma_{\beta_1} & \sigma_{\beta_1}^2 \end{pmatrix}$$

and ρ denotes the correlation between β_{0j} and β_{1j}. Here the true parameter values are $\alpha_0 = -1.5$, $\gamma_0 = 0.8$, $\alpha_1 = -0.7$, $\gamma_1 = -0.5$, $\sigma_{\beta_0} = 0.1$, $\sigma_{\beta_1} = 0.2$, $\rho = 0.4$, $\sigma_y = 0.5$. The naive model is analogous to (3.4):

$$\begin{aligned} y_i &\sim \mathrm{N}(\beta_{0j[i]} + \beta_{1j[i]} z_i, \sigma_y^2) \\ \begin{pmatrix} \beta_{0j} \\ \beta_{1j} \end{pmatrix} &\sim \mathrm{N}\left(\begin{pmatrix} \alpha_0 + \gamma_0 w_j \\ \alpha_1 + \gamma_1 w_j \end{pmatrix}, \Sigma \right) \end{aligned} \tag{3.6}$$

3.4 Results

3.4.1 Measurement error from sampling

Figures 3.1 and 3.2 show simulation results for the group-level parameters α_0, γ_0, and σ_{β_0} for continuous and binary y, respectively. We focus on the group-level parameters

because the individual-level parameters β_1 and σ_y are very accurately estimated under both models, so we do not discuss them further here.

In both figures, the six panels display relative bias, relative root mean squared error (RRMSE), coverage of 50% and 95% uncertainty intervals, and the average relative widths of the 50% and 95% uncertainty intervals. The relative bias is calculated as $\frac{1}{L}\sum_{\ell=1}^{L}\frac{\widehat{\theta}_\ell-\theta}{\theta}$, where θ is the true parameter value, $\widehat{\theta}_\ell$ is the estimated value from the ℓ-th simulation, and L is the number of simulations ($L \geq 75$). Ppositive values of relative bias thus indicate that the coefficient estimates are inflated compared to the truth, while negative values of relative bias indicate attenuation in the coefficient estimates. RRMSE is calculated as $\sqrt{\frac{1}{L}\sum_{\ell=1}^{L}\left(\frac{\widehat{\theta}_\ell-\theta}{\theta}\right)^2}$. We calculate the 50% (95%) intervals from the 25th and 75th (2.5th and 97.5th) percentiles of the posterior predictive distribution for each parameter. The relative widths of the uncertainty intervals are calculated by dividing the width of the uncertainty interval by the true parameter value and averaging across the L simulations. In each plot, y-axis denotes the number of clusters sampled ($J_s \in \{5, 15, 50\}$), the color represents the two models (full or naive), and the symbols represent the number of units sampled ($n_j \in \{10, 30, 60\}$).

For continuous y, the full model (red) outperforms the naive one (blue) in terms of relative bias in every sampling scenario as seen in the top left panel of Figure 3.1. Even with $J_s = 50$ clusters and 30 units sampled per cluster, the level of relative bias remains high under the naive model: it inflates the true value of the intercept term α_0 by nearly 50% and attenuates γ_0 by nearly 10%. Going from $J_s = 5$ to $J_s = 50$ clusters does little to reduce bias under the naive model, except in estimating the group-level variance σ_{β_0}. In contrast, the bias in α_0 and γ_0 under the full model with only 10 units sampled per cluster is comparable to that under the naive model with 60 units per cluster. In general, the naive model inflates the estimates of α_0 and attenuates those of γ_0, while the full model does the opposite but at much smaller magnitudes.

For binary y, we again see that the naive model results in inflation of α_0 and

attenuation in γ_0 (top left panel of Figure 3.2). The relative bias in α_0 is smaller under the full model than under the naive model in all sampling scenarios, but for γ_0, the full model overcorrects the attenuation and results in higher relative bias than the naive model until $J_s = 50$ clusters are sampled.

The differences in RRMSE are less dramatic than in relative bias. For $J_s = 5$ clusters, the RRMSE is often slightly higher for the full model than the naive model under both continuous and binary y. However, with $J_s = 50$ clusters, the RRMSEs for the two models are identical under binary y. For continuous y, the full model is particularly advantageous in the case of a large sample of clusters ($J_s = 50$) and a small within-cluster sample size ($n_j = 10$), with RRMSEs under the full model nearly half of those from the naive model.

In addition to reduced bias, another clear advantage of the full model comes in uncertainty intervals (UIs) that achieve nominal coverage levels (middle row of plots in Figures 3.1 and 3.2). Even with only $J_s = 5$ sampled clusters, UIs from the full model are at or above the nominal levels (the one exception is σ_{β_0} under binary y with 10 or 30 units per cluster). In contrast, the coverage rates of the UIs for the naive model are often well below the nominal levels and even sometimes zero and actually decrease as the number of sampled clusters increases.

The price of improved coverage is wider UIs under the full model than the naive model, most dramatically so when the number of clusters sampled is small (bottom row of plots in Figures 3.1 and 3.2). However, once we sample at least $J_s = 15$ clusters, the difference in UI lengths between the naive and full models is generally negligible.

3.4.2 Classical measurement error

Figure 3.3 shows the results for estimating the group-level coefficients α_0, γ_0, and σ_{β_0} for the case of classical measurement error in a group-level predictor as described in Section 3.2.2. Hollow circles denote binary y and crossed denote continuous y. To

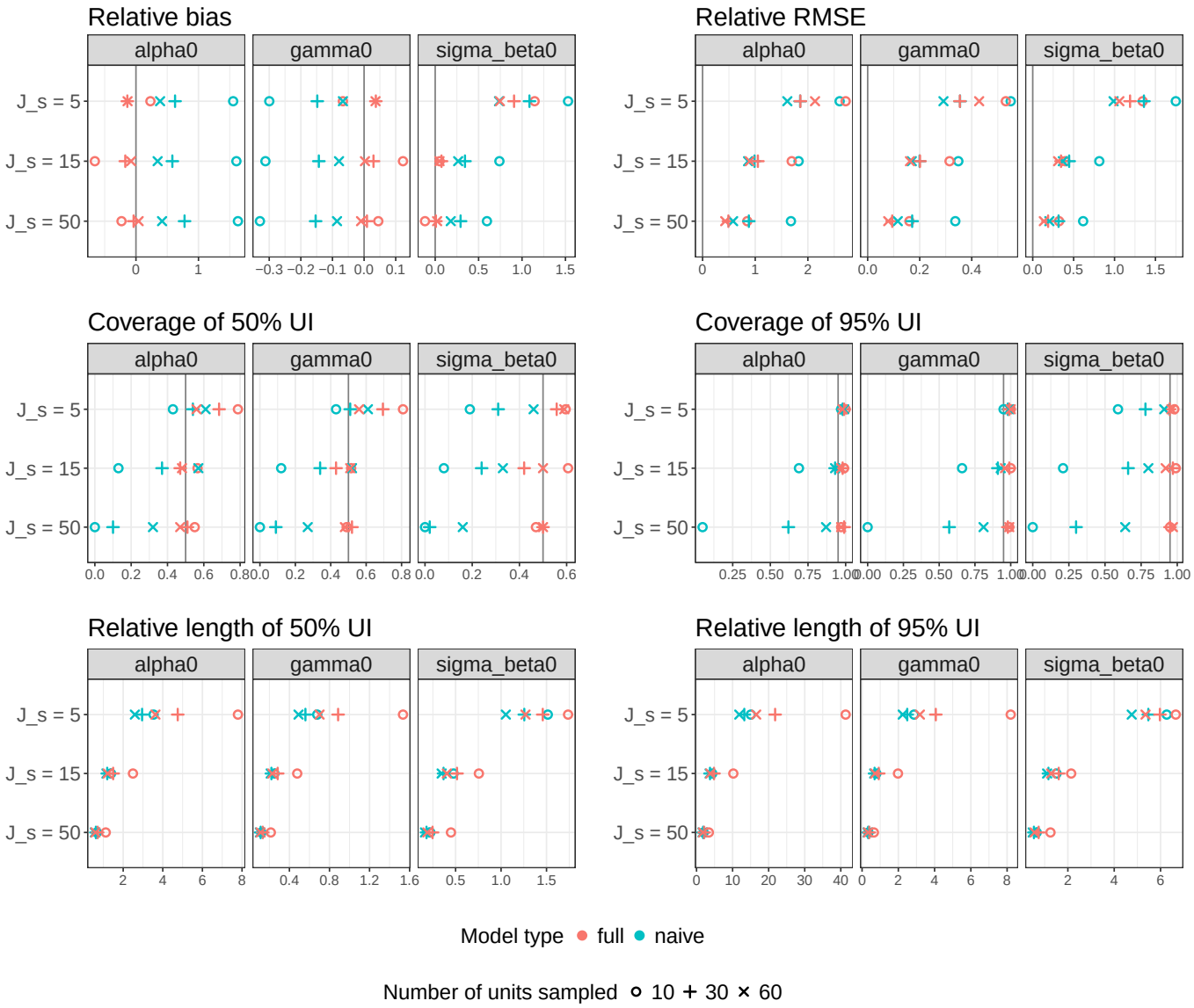


Figure 3.1: Simulation results for continuous y for the group-level parameters α_0, γ_0, and σ_{β_0} for the scenario described in Section 3.3.1. In each of the six panels, the x-axis is the value of the metric being plotted and the y-axis is the number of sampled clusters ($J_s \in \{5, 15, 50\}$). The color of the symbol denotes the model (full vs naive), and the shape of the symbol denotes the number of sampled units (10, 30, or 60).

improve readability, we divide the values of relative bias and RRMSE for σ_{β_0} by 10. While relative bias in α_0 is very similar between the naive and full models, the naive

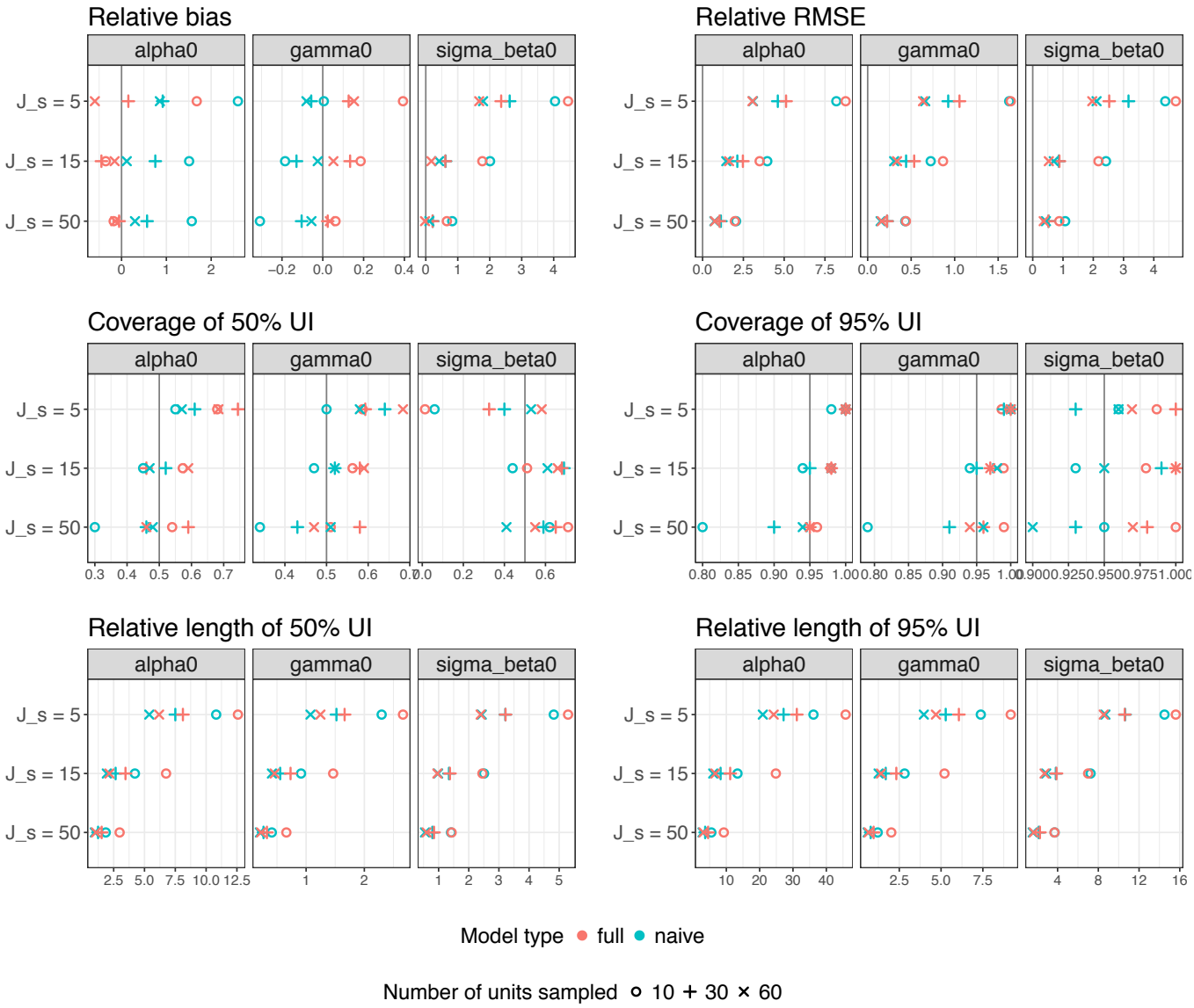


Figure 3.2: Simulation results for binary y for the group-level parameters α_0, γ_0, and σ_{β_0} for the scenario described in Section 3.3.1. In each of the six panels, the x-axis is the value of the metric being plotted and the y-axis is the number of sampled clusters ($J_s \in \{5, 15, 50\}$). The color of the symbol denotes the model (full vs naive), and the shape of the symbol denotes the number of sampled units (10, 30, or 60).

model attenuates the value of γ_0 by an average of 20%. In contrast, the relative bias in γ_0 under the full model is much smaller: 7% for binary y and 1% for continuous y.

Similarly, the relative bias of σ_{β_0} is twice as small under the full model as under the naive model for binary y and five times smaller for continuous y.

The naive model also yields estimates that are much more variable. The RRMSEs of γ_0 is twice as large under the naive model compared to the full model for binary y and four times larger for continuous y; the pattern for σ_{β_0} is similar, while the differences are negligible for α_0.

As in the previous simulation scenario, the full model yields uncertainty intervals whose coverage rates are much closer to nominal levels than those from the naive model. However, the 50% UIs from the full model fall short of nominal coverage rates particularly for binary y. The UI lengths for α_0 and γ_0 are nearly identical between the full and naive models, but for σ_{β_0} they are two to three times as long under the full model.

The results for the case of varying slopes and intercepts are in Figures 3.4 and 3.5. Figure 3.4 shows results for the regression parameters α_0, γ_0, α_1, and γ_1, and Figure 3.5 for the variance/covariance parameters σ_{β_0}, σ_{β_1}, and ρ.

In Figure 3.4, we see that ignoring measurement error in the naive model with binary y leads to high bias: over 150% for α_0 and nearly 90% for the other regression parameters. For continuous y, both the naive and full models yield nearly unbiased estimates of α_0 and α_1, but for γ_0 and γ_1, the estimates from the naive model are attenuated by about 20%. We see a similar pattern in RRMSE, where the differences between the naive and continuous model are large for binary y and much smaller for continuous y.

The high bias of the naive model estimates with binary y lead to UIs that do not cover the true value; for continuous y, the UIs for α_0 and α_1 are close to the nominal levels, but those for γ_0 and γ_1 fail to contain the true value. In contrast, the coverage rates of the UIs from the full model are close to or exceed the nominal levels. The lengths of the UIs between the full and naive models are very similar for continuous y. For binary y, the UIs from the full model are much longer than those from the

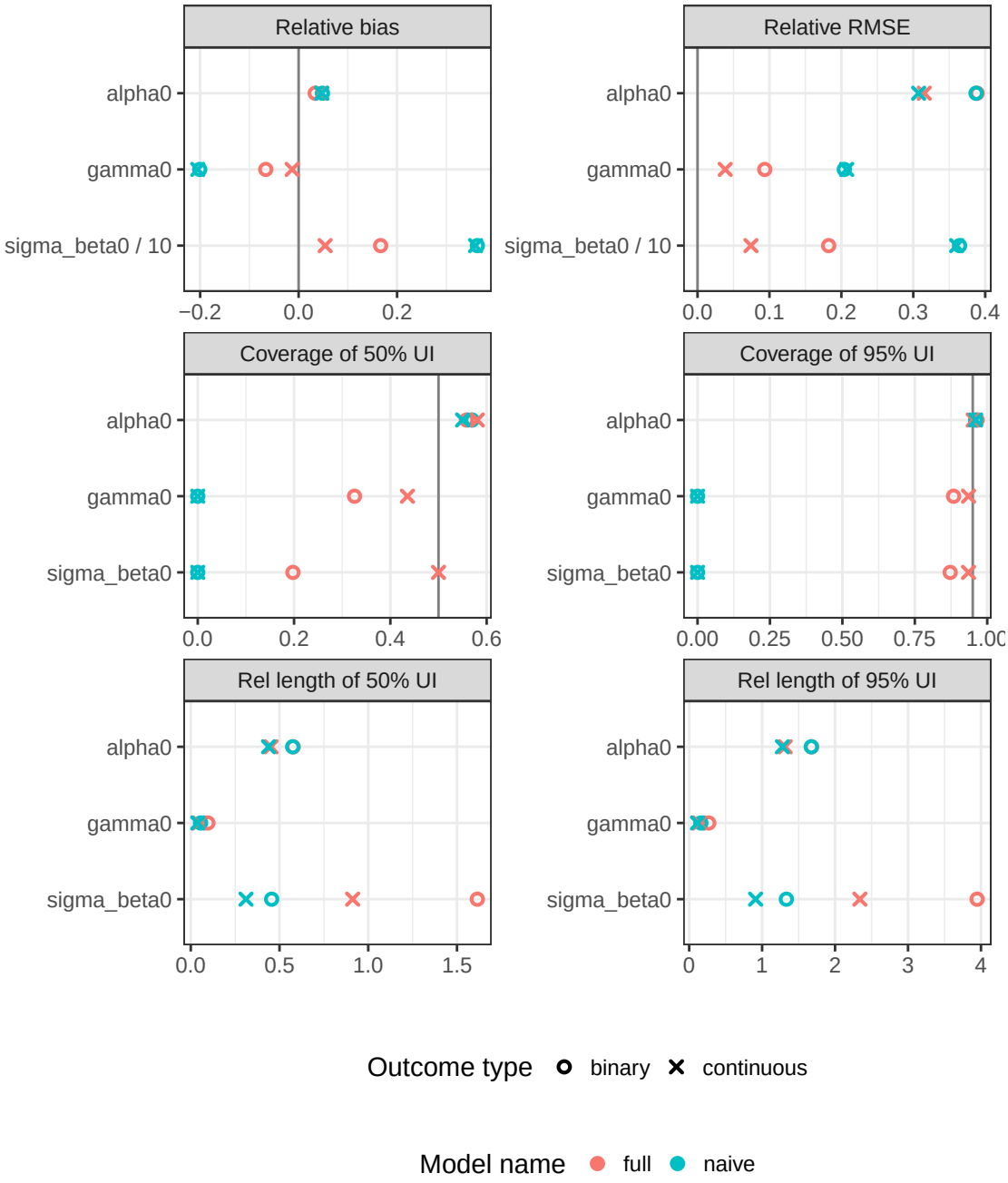


Figure 3.3: Results for binary y (hollow circles) and continuous y (crosses) with classical measurement error and a group-varying intercept only as described in Section 3.3.2. Red points denote the full model and blue points the naive model. Note that to improve readability, we divide the values of relative bias and RRMSE for σ_{β_0} by 10.

naive model.

The full model leads to better estimates of the variance parameters as well, as seen in Figure 3.5. The relative bias in σ_{β_0} and σ_{β_1} is much lower under the full model, particularly for continuous y. Both the full and naive models underestimate the correlation parameter ρ, but the bias under the naive model is twice as large for binary y and over six times larger for continuous y. The full model also leads to reduced RRMSEs in most cases.

In Figures 3.3, 3.4, and 3.5, the coverage rates of the UIs from the naive model are essentially zero, while those from the full model are near or above the nominal levels. Figure 3.6 helps illustrate why this is the case. In each panel, we plot the distribution of posterior means across the L simulations ($L \geq 75$) from the full model (red) and naive model (blue) for both binary y (solid line) and continuous y. The vertical lines denote the true parameter values. We see that the posterior means from the naive model tend to be highly peaked around a biased value, leading to the large relative bias and RRMSE and low coverage rates and UI lengths in Figures 3.3 to 3.5. The posterior means from the full model, on the other hand, have a much larger spread but are generally centered around the true value. In this way, the inferences from the full model more completely reflect the uncertainty in the parameter estimates, while those from the naive model are falsely precise and highly biased.

3.5 Discussion

We propose a simple method to account for measurement error in a group-level covariate in the context of Bayesian hierarchical models by explicitly including the measurement process in the outcome model. Our simulation results demonstrate the well-known pitfalls of relying on naive models that do not account for measurement error – point estimates that are far from their true values and uncertainty intervals that are misleadingly short with poor coverage rates – and the improved inference

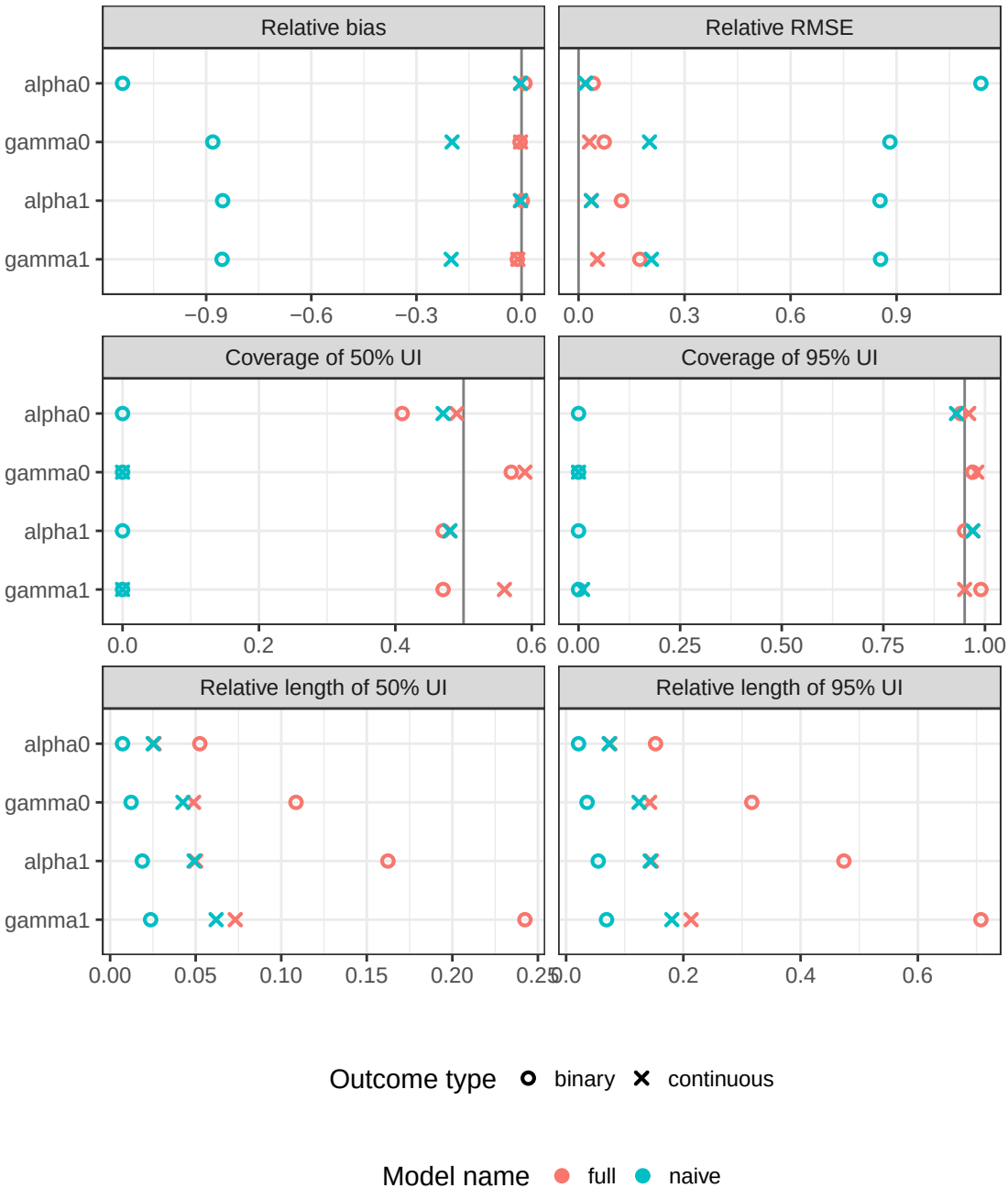


Figure 3.4: Results for group-level regression parameters α_0, γ_0, α_1, and γ_1 under binary y (hollow circles) and continuous y (crosses) with classical measurement error and group-varying slopes and intercepts as described in Section 3.3.2. Red points denote the full model and blue points the naive model.

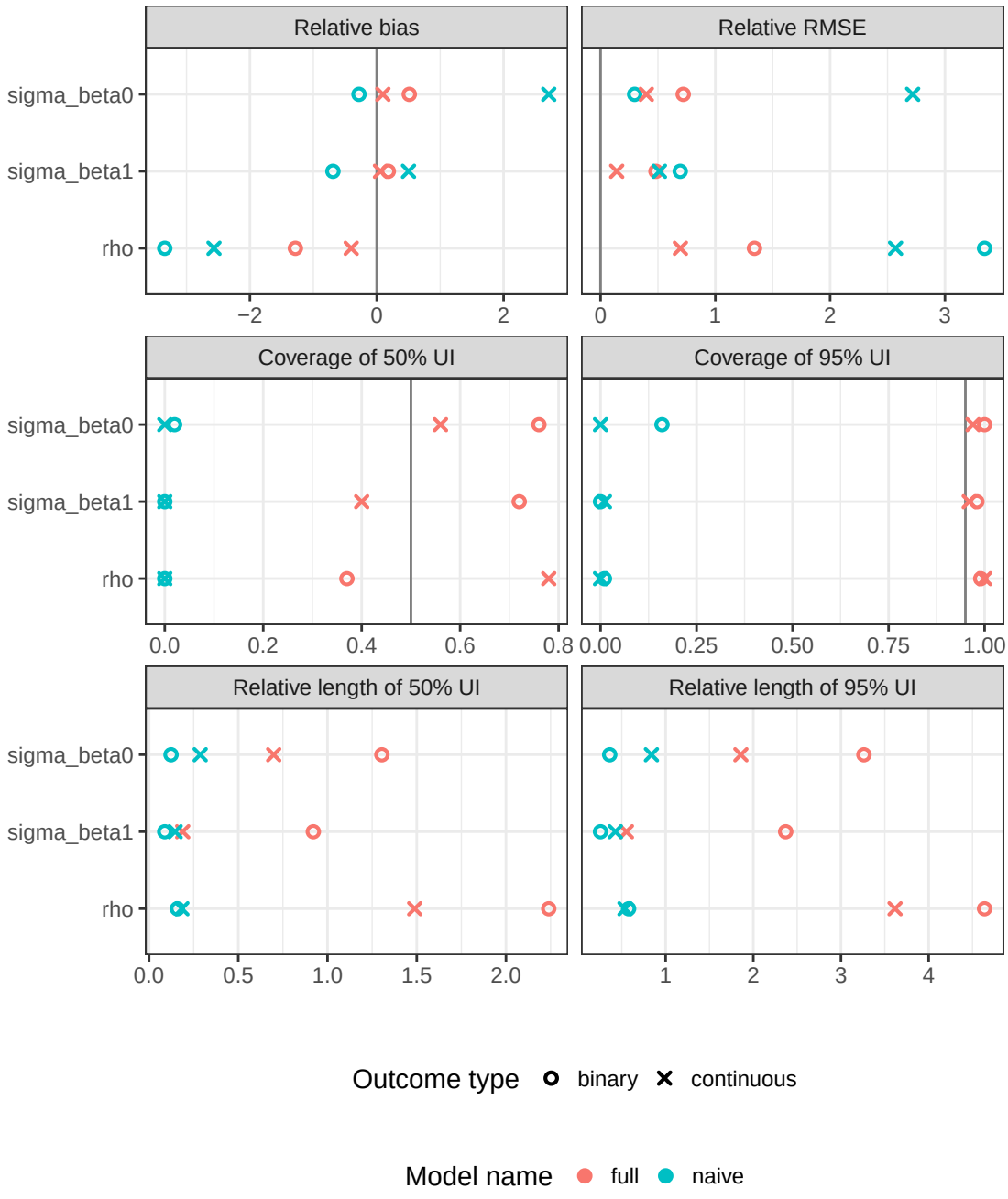


Figure 3.5: Results for group-level variance/covariance parameters σ_{β_0}, σ_{β_1}, and ρ under binary y (hollow circles) and continuous y (crosses) with classical measurement error and group-varying slopes and intercepts as described in Section 3.3.2. Red points denote the full model and blue points the naive model.

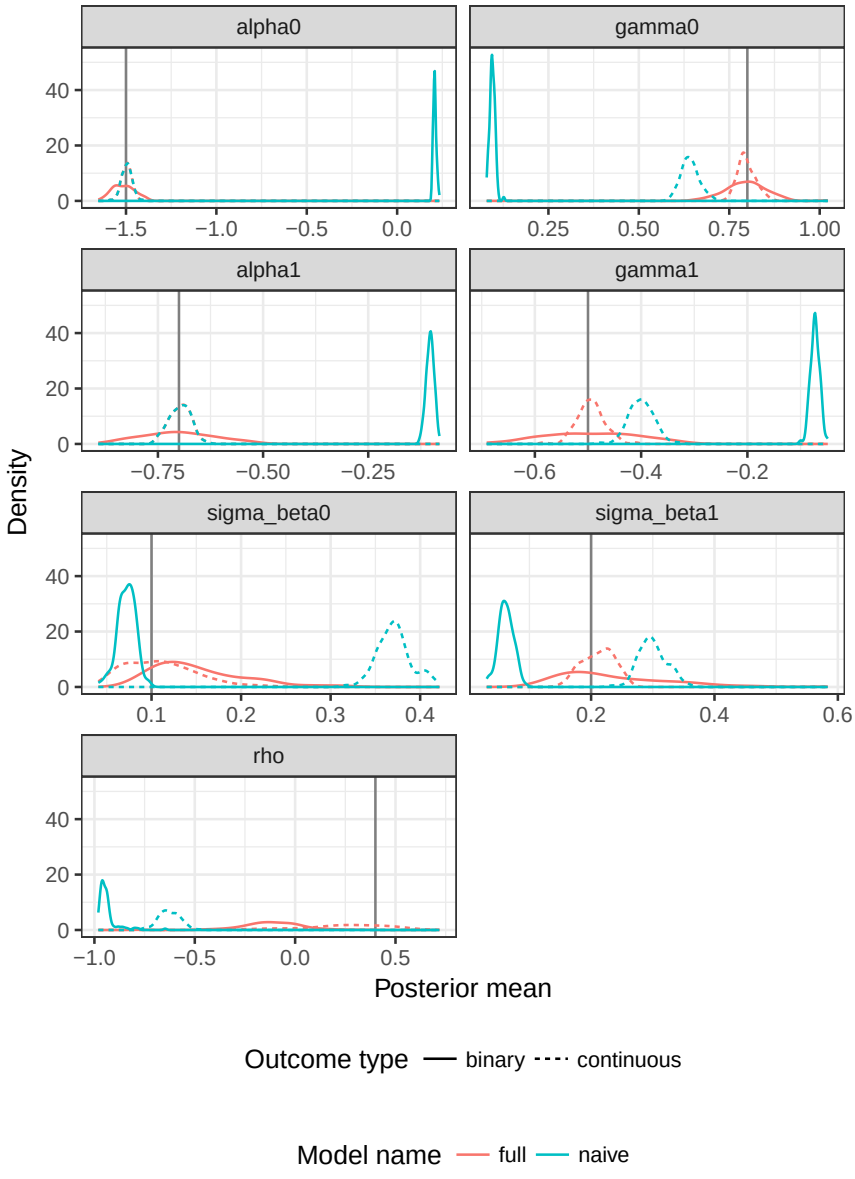


Figure 3.6: Distribution of posterior means across simulations for binary y (solid line) and continuous y (dashed line) under classical measurement error in the group-level predictor with both group-varying slopes and intercepts as described in Section 3.3.2.

that can be achieved by explicitly modeling the measurement error process.

Our simulations show that accounting for measurement error is critical, especially when creating area-level covariates by averaging unit-level ones. They demonstrate that our proposed method leads to reduced bias and yields uncertainty intervals with appropriate coverage levels. This is particularly important when the research question of interest centers on the relationship between group-level covariates and individual-level outcomes.

In our simulations of classical measurement error, we assume that the error is nondifferential and homoskedastic. Recent work (Muff *et al.*, 2015) has demonstrated the importance of heterogeneity in measurement error, particularly in the context of estimating interaction parameters. Allowing for measurement error to differ between $z_i = 1$ and $z_i = 0$ would be a straightforward extension of our proposed model. We have also not considered spatial correlation in neighborhood-level variables or their errors, which has been shown to be important particularly in epidemiology (Xia and Carlin, 1998; Bernardinelli *et al.*, 1997). Incorporating spatial correlation in the latent neighborhood-level variable or the measurement error process is straightforward in Stan (Morris, 2018).

Advances in statistical computation like the Stan probabilistic programming language make it easy to seamlessly and explicitly incorporate measurement error into fully Bayesian models. Previous Bayesian approaches like Gibbs sampling often face difficulties in practice like poor mixing (Gryparis *et al.*, 2009), while INLA requires that the model for the (latent) true covariate be Gaussian (Muff *et al.*, 2015). In contrast, the Stan language is flexible and enables users to specify models in a highly intuitive way without restrictions on the specific functional or distributional form of the measurement model. It allows for fine tuning of the HMC sampler and provides detailed warnings about and diagnostics for lack of convergence and mixing, ensuring that the user is aware when inferences may not be reliable and enabling them to modify sampling parameters as necessary. Stan is well-equipped to handle measure-

ment error models more complex than those considered here, opening the door to ever-improved inference in the presence of measurement error.

Bibliography

Anna Aizer and Joseph J. Doyle, Jr. Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges. *The Quarterly Journal of Economics*, 130(2):759–803, 2015.

Rebecca R. Andridge. Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical Journal*, 53(1):57–74, 2011.

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):pp. 444–455, 1996.

Mike Baiocchi, Dylan S. Small, Scott Lorch, and Paul R. Rosenbaum. Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association*, 105(492):1285–1296, 2010.

Michael Baiocchi, Jing Cheng, and Dylan S. Small. Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340, 2014.

L Bernardinelli, C Pascutto, N G Best, and W R Gilks. Disease mapping with errors in covariates. *Stat Med*, 16(7):741–752, Apr 1997.

M. J. Betancourt and Mark Girolami. Hamiltonian monte carlo for hierarchical models, 2013.

John Bound, David A. Jaeger, and Regina M. Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450, 1995.

L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

Barbara Lepidus Carlson. Fragile families & child wellbeing study: Methodology for constructing mother, father, and couple weights for core telephone surveys waves 1-4. Technical report, Mathematica Policy Research, 2008.

Raymond J. Carroll, Paul Gallo, and Leon Jay Gleser. Comparison of least squares and errors-in-variables regression, with special reference to randomized analysis of covariance. *Journal of the American Statistical Association*, 80(392):929–932, 1985.

R.J. Carroll, D. Ruppert, L.A. Stefanski, and C.M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2006.

Q.X. Chen, Michael R. Elliott, and R.J. Little. Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Survey Methodology*, 36:23–34, 2010.

Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382, 1992.

A V Diez-Roux, Sharon Stein Merkin, Donna Arnett, Lloyd Chambless, Mark Massing, F. Javier Nieto, Paul Sorlie, Moyses Szklo, Herman A. Tyroler, and Robert L. Watson. Neighborhood of residence and incidence of coronary heart disease. *New England Journal of Medicine*, 345(2):99–106, 2001. PMID: 11450679.

A V Diez-Roux. Multilevel analysis in public health research. *Annual Review of Public Health*, 21(1):171–192, 2000. PMID: 10884951.

A V Diez-Roux. Next steps in understanding the multilevel determinants of health. *Journal of Epidemiology & Community Health*, 62(11):957 – 959, 2008.

Ashkan Ertefaie, Dylan S. Small, and Paul R. Rosenbaum. Quantitative evaluation of the trade-off of strengthened instruments and sample size in observational studies. *Journal of the American Statistical Association*, page to appear.

W.A. Fuller. *Measurement Error Models*. Wiley Series in Probability and Statistics. Wiley, 2009.

Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. *Bayesian Data Analysis, Third Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, London, third edition, November 2013.

Andrew Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1:1–19, 2006.

Donald P. Green and Daniel Winik. Using random judge assignments to estimate the effects of incarceration and probation on recidivism among drug offenders. *Criminology*, 48(2):357–387, 2010.

Alexandros Gryparis, Christopher J Paciorek, Ariana Zeka, Joel Schwartz, and Brent A Coull. Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*, 10(2):258–274, Apr 2009.

Paul Gustafson. *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. CRC Press, 2003.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

Jesse Y. Hsu, Dylan S. Small, and Paul R. Rosenbaum. Effect modification and design sensitivity in observational studies. *Journal of the American Statistical Association*, 108(501):135–148, 2013.

Jesse Y. Hsu, José R. Zubizarreta, Dylan S. Small, and Paul R. Rosenbaum. Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika*, 102(4):767–782, 2015.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

Guido W. Imbens and Paul R. Rosenbaum. Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1):109–126, 2005.

Ichiro Kawachi and Lisa F. Berkman. *Neighborhoods and Health*. Oxford University Press, 2003.

Luke Keele and Jason W. Morgan. How strong is strong enough? strengthening instruments through matching and weak instrument tests. *Annals of Applied Statistics*, page forthcoming, 2016.

Kwonsang Lee, Dylan S. Small, Jesse Y. Hsu, Jeffrey H. Silber, and Paul R. Rosenbaum. Discovering effect modification in an observational study of surgical mortality at hospitals with superior nursing. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, page to appear, 2017.

Kwonsang Lee, Dylan S. Small, and Paul R. Rosenbaum. A new, powerful approach to the study of effect modification in observational studies. *ArXiv e-prints*, February 2017.

Roderick J.A. Little and H Zheng. The Bayesian approach to the analysis of finite population surveys. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,

D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 8*, pages 283–302 (with discussion and rejoinder). Oxford University Press, Oxford, 2007.

Roderick J. Little. To Model or Not To Model? Competing Modes of Inference for Finite Population Sampling. *Journal of the American Statistical Association*, 99:546–556, January 2004.

Charles E. Loeffler. Does imprisonment alter the life course? evidence on crime and employment from a natural experiment. *Criminology*, 51(1):137–166, 2013.

Thomas Lumley. Analysis of complex survey samples. *Journal of Statistical Software*, 9(1):1–19, 2004.

Ruth Marcus, Eric Peritz, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

J. S. Maritz. A note on exact robust confidence intervals for location. *Biometrika*, 66(1):163–166, 1979.

Glen Meeden. A non-informative Bayesian approach for two-stage cluster sampling. *Sankhya, Series B*, 61:133–144, 1999.

Mitzi Morris. Spatial models in stan: Intrinsic auto-regressive models for areal data. GitHub repository, 2018.

Michael Mueller-Smith. The criminal and labor market impacts of incarceration. Technical report, University of Michigan, 2015.

Stefanie Muff and Lukas F. Keller. Reverse attenuation in interaction terms due to covariate measurement error. *Biometrical Journal*, 57(6):1068–1083, 2015.

Stefanie Muff, Andrea Riebler, Leonhard Held, Håvard Rue, and Philippe Saner. Bayesian analysis of measurement error models using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(2):231–252, 2015.

Mahasin S. Mujahid, Ana V. Diez Roux, Jeffrey D. Morenoff, and Trivellore Raghunathan. Assessing the measurement properties of neighborhood scales: From psychometrics to ecometrics. *American Journal of Epidemiology*, 165(8):858–867, 2007.

Daniel S. Nagin and G. Matthew Snodgrass. The effect of incarceration on reoffending: Evidence from a natural experiment in pennsylvania. *Journal of Quantitative Criminology*, 29(4):601–642, 2013.

Paul Nieuwbeerta, Daniel S. Nagin, and Arjan A. J. Blokland. Assessing the impact of first-time imprisonment on offenders' subsequent criminal career development: A matched samples comparison. *Journal of Quantitative Criminology*, 25(3):227–257, 2009.

G.P. Patil and C.R. Rao. Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, 34(2):179–189, 1978.

Karl Pearson. V. on the methematical theory of errors of judgement, with special reference to the personal equation. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 198(300-311):235–299, 1902.

Pennsylvania Commission on Sentencing. Pennsylvania Sentencing Data. 1998–2000.

K E Pickett and M Pearl. Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review. *Journal of Epidemiology & Community Health*, 55(2):111–122, 2001.

Nancy E. Reichmann, Julien O. Teitler, Irwin Garfinkel, and Sara S. McLanahan. Fragile families: Sample and design. *Children and Youth Services Review*, 23(4/5):303–326, 2001.

Jerome P. Reiter, Trivellore E. Raghunathan, and S. K. Kinney. The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology*, 32:143–149, 2006.

Sylvia Richardson and Walter R. Gilks. Conditional independence models for epidemiological studies with covariate measurement error. *Statistics in Medicine*, 12(18):1703–1722, 1993.

Paul R. Rosenbaum. Identification of causal effects using instrumental variables: Comment. *Journal of the American Statistical Association*, 91(434):465–468, 1996.

Paul R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327, 2002.

P.R. Rosenbaum. *Observational Studies*. Springer Series in Statistics. Springer, 2002.

Paul R. Rosenbaum. Design sensitivity in observational studies. *Biometrika*, 91(1):153–164, 2004.

Paul R. Rosenbaum. Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *The American Statistician*, 59(2):147–152, 2005.

P.R. Rosenbaum. *Design of Observational Studies*. Springer Series in Statistics. Springer New York, 2010.

P. R. Rosenbaum. Impact of multiple matched controls on design sensitivity in observational studies. *Biometrics*, 69:118–127, 2013.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

D. B. Rubin. Comment on "An evaluation of model-dependent and probability-sampling inferences in sample surveys", by M. H. Hansen, W. G. Madow and B. J. Tepping. *Journal of the American Statistical Association*, 78:803–805, 1983.

Donald B Rubin. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990.

S.O. Rutstein and G Rojas. Guide to dhs statistics. 01 2006.

C.E. Särndal, B. Swensson, and J.H. Wretman. *Model Assisted Survey Sampling*. Springer series in statistics. Springer-Verlag, 1992.

J. L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London, 1997.

Yajuan Si, Natesh S. Pillai, and Andrew Gelman. Bayesian nonparametric weighted sampling inference. *Bayesian Anal.*, 10(3):605–625, 09 2015.

Dylan S Small and Paul R Rosenbaum. War and wages. *Journal of the American Statistical Association*, 103(483):924–933, 2008.

G. Matthew Snodgrass, Arjan A. J. Blokland, Amelia Haviland, Paul Nieuwbeerta, and Daniel S. Nagin. Does the time cause the crime? an examination of the relationship between time served and reoffending in the netherlands. *Criminology*, 49(4):1149–1194, 2011.

Stan Development Team. Brief guide to stan's warnings, 2016.

Stan Development Team. The stan c++ library, version 2.15.0, 2016.

Stan Development Team. Stan modeling language users guide and reference manual, version 2.15.0, 2016.

Peter Wagner and Bernadette Rabuy. Mass incarceration: The whole pie. Technical report, Prison Policy Initiative, https://www.prisonpolicy.org/reports/pie2017.html, 2017.

Peter Wagner and Alison Walsh. States of incarceration: The global context. Technical report, Prison Policy Initiative, https://www.prisonpolicy.org/global/2016.html, 2016.

Abraham Wald. The fitting of straight lines if both variables are subject to error. *Ann. Math. Statist.*, 11(3):284–300, 09 1940.

K. Wolter. *Introduction to Variance Estimation*. Springer-Verlag, New York, NY, 2007.

H. Xia and B. P. Carlin. Spatio-temporal models with errors in covariates: mapping Ohio lung cancer mortality. *Stat Med*, 17(18):2025–2043, Sep 1998.

Sahar Z. Zangeneh and Roderick J. A. Little. Bayesian inference for the finite population total from a heteroscedastic probability proportional to size sample. *Journal of Survey Statistics and Methodology*, 3(2):162–192, 2015.

Sahar Z. Zangeneh, Robert W. Keener, and Roderick J. A. Little. Bayesian nonparametric estimation of finite population quantities in absence of design information on nonsampled units. In *JSM Proceedings. Section on Survey Research Methods. Miami Beach, FL, USA. American Statistical Association-IMS*, pages 3429–3440, 2011.

Sahar Z. Zangeneh, Robert W. Keener, and Roderick J.A. Little. Bayesian nonparametric estimation of finite population quantities in absence of design information on nonsampled units. *Proceedings of the Joint Statistical Meetings*, 2011.

D.V. Zaykin, Lev A. Zhivotovsky, P.H. Westfall, and B.S. Weir. Truncated product method for combining p-values. *Genetic Epidemiology*, 22(2):170–185, 2002.

H Zheng and R.J. Little. Inference for the population total from probability-proportional-to-sizes samples based on predictions from a penalized spline non-parametric model. *Journal of Official Statistics*, 21(1):1–20, 2005.

Hanzhi Zhou, Michael R. Elliott, and Trivellore E. Raghunathan. Multiple imputation in two-stage cluster samples using the weighted finite population bayesian bootstrap. *Journal of Survey Statistics and Methodology*, 2016.

Jose R. Zubizarreta, Dylan S. Small, Neera K. Goyal, Scott Lorch, and Paul R. Rosenbaum. Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *The Annals of Applied Statistics*, 7(1):25–50, 2013.

Appendices

Appendix

Instrument Validity Figures

A.1 Instrument Validity Figures

In Chapter 2, we describe the assumptions required for an instrumental variable to yield valid causal inferences. One assumption is that the instrument assignment must be as-if random. In the context of recidivism, this assumption implies that the assignment of offenders to judges is random, and if this is so, there should be no relationship between the harshness of a judge and the types of offenders they see. We show graphical checks of this assumption in Figures 2.3 and 2.4 for offense severity in both the data from 1997, which we use to calculate the instrument itself, and in the offense data from 1998-2000. This Appendix includes figures for additional offender characteristics. In each figure, the x-axis is the (continuous) harshness of the judge in each county and the y-axis corresponds to the characteristic of interest, generally a proportion, mean, or median. The size of each circle is proportional to the number of cases the judge saw in that time period (1997 or 1998-2000); only judges who saw at least 30 cases in a county in each time period and counties with at least two such judges are shown. If cases are indeed randomly assigned, we would not see any strong relationship between harshness and these characteristics and all of the lines would be approximately horizontal. Overall, these figures indicate there is little relationship between harshness and the various offender and case characteristics.

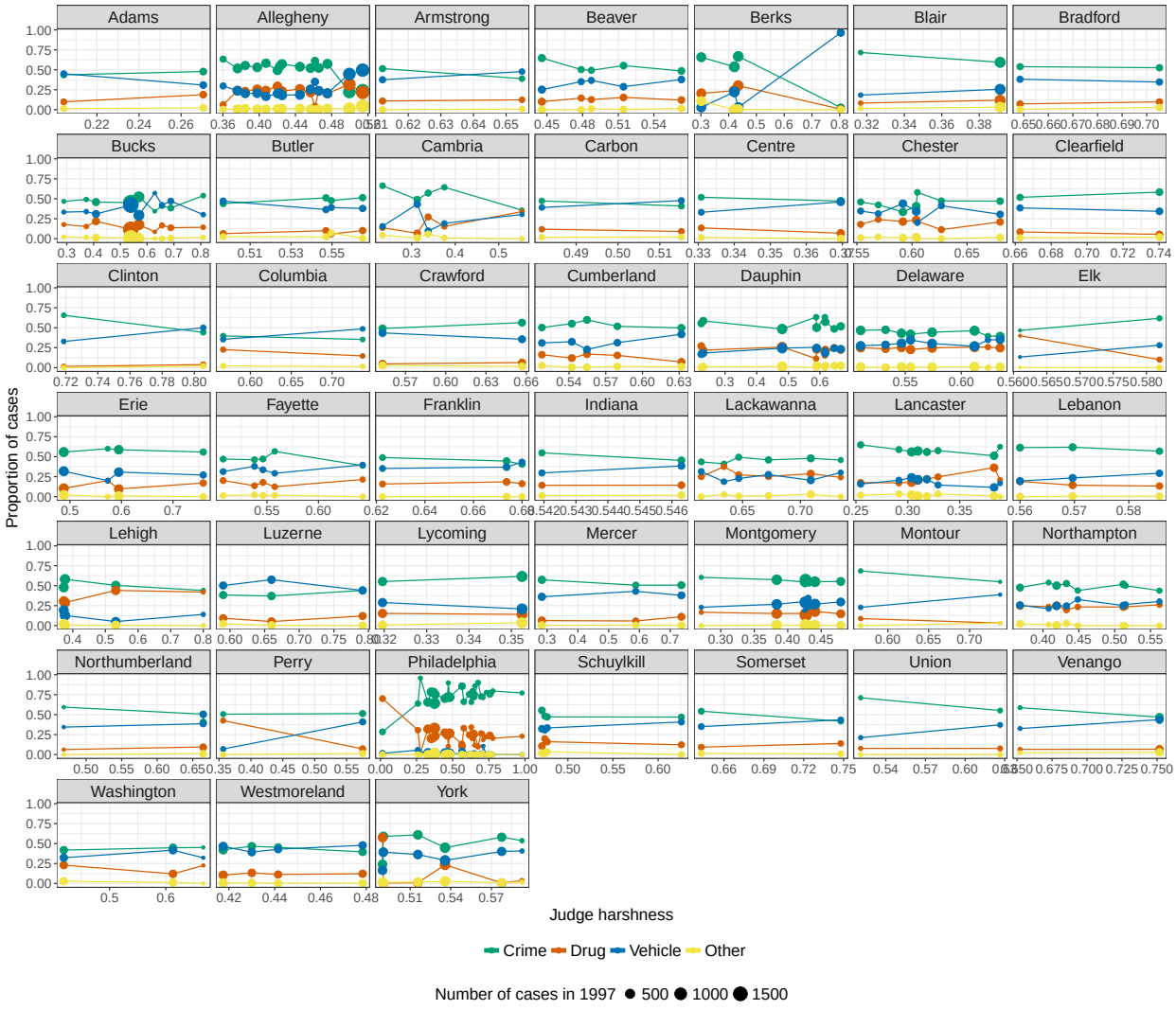


Figure A.1: Proportion of each judge's 1997 cases that fell into each of four major offense categories: crime (green), drugs (orange), vehicle (blue), and other (yellow), plotted against the harshness of the judge. If cases are, in fact, randomly assigned to judges, we should not see any strong relationship between offense category and harshness. Only counties with at least two judges who saw at least 30 cases in 1997 are shown.

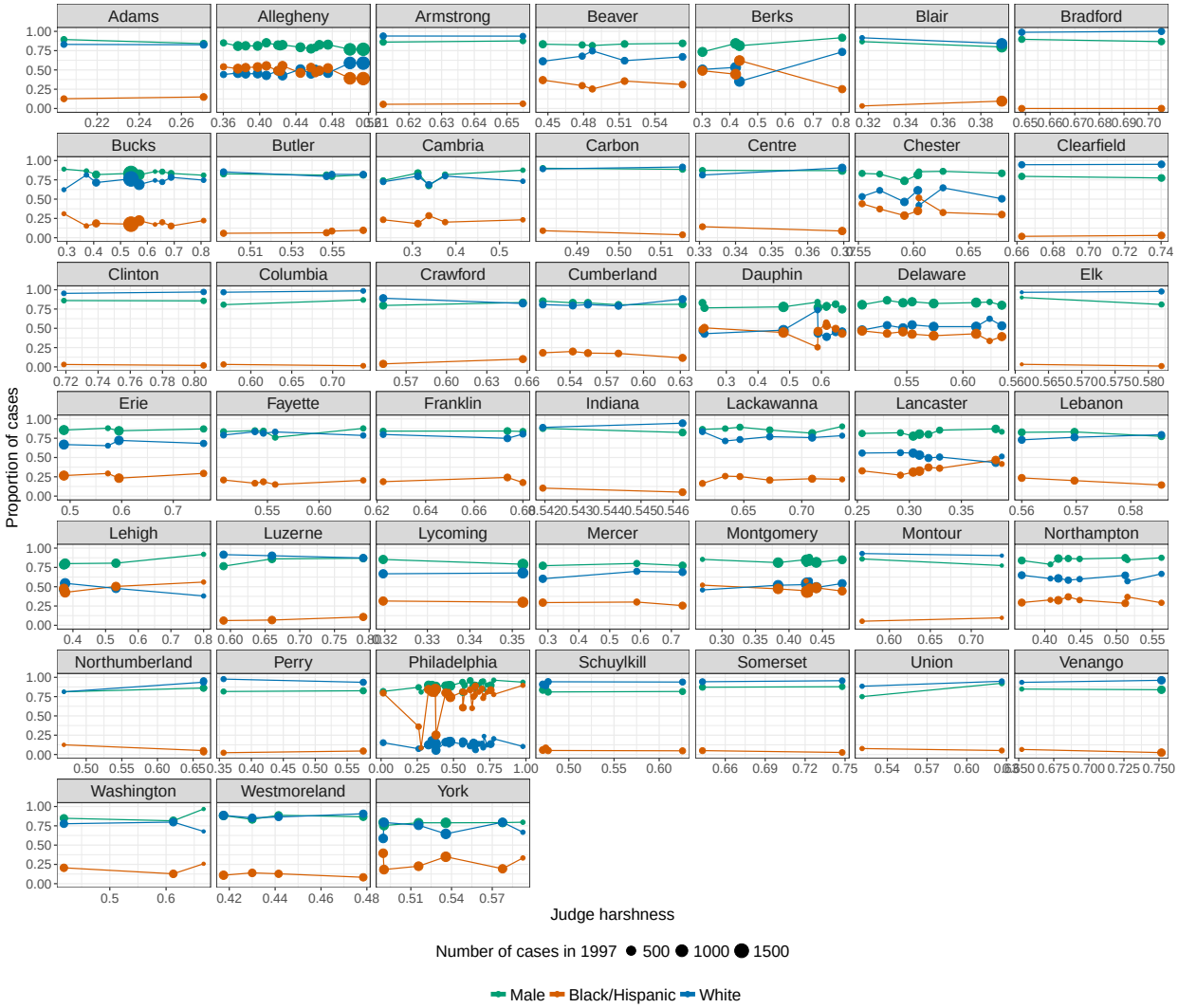


Figure A.2: Proportion of each judge's 1997 cases by sex and race, plotted against the harshness of the judge. If cases are, in fact, randomly assigned to judges, we should not see any strong relationship between the demographic categories and harshness. Only counties with at least two judges who saw at least 30 cases in 1997 are shown.

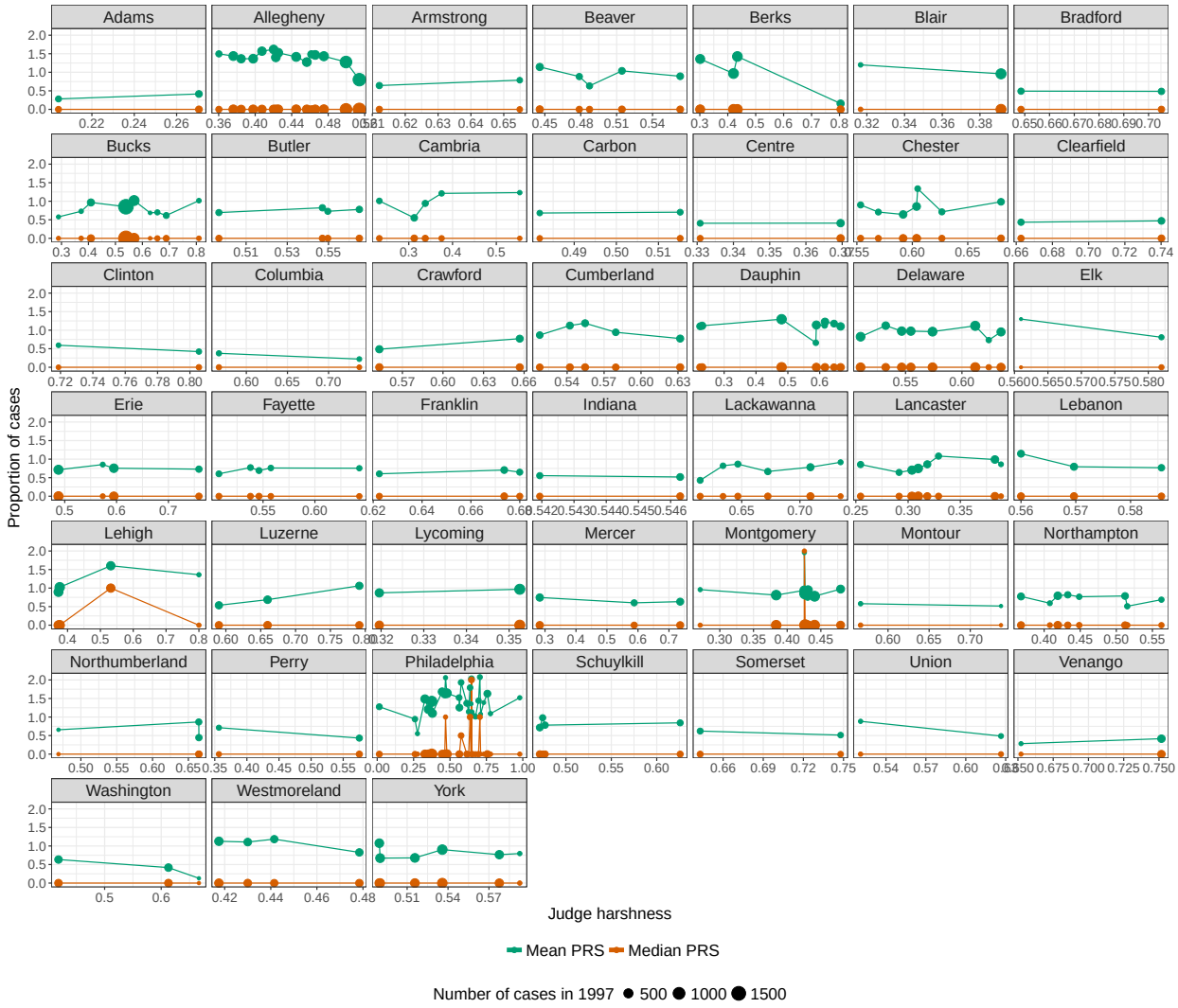


Figure A.3: Mean and median prior record scores of offenders in the cases seen by each judge in 1997, plotted against the harshness of the judge. If cases are, in fact, randomly assigned to judges, we should not see any strong relationship between prior record score and harshness. Only counties with at least two judges who saw at least 30 cases in 1997 are shown.

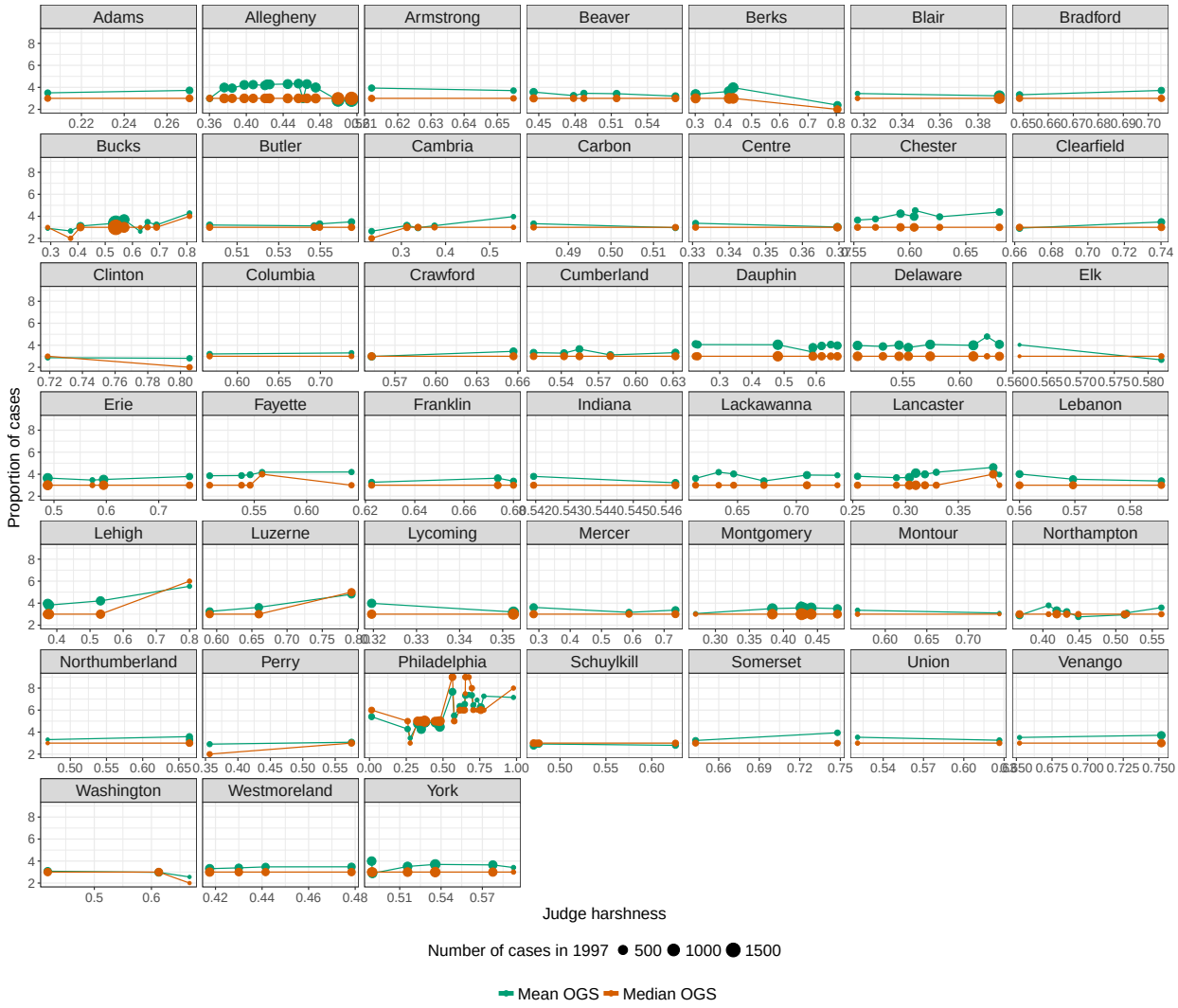


Figure A.4: Mean and median offense gravity scores of offenders in the cases seen by each judge in 1997, plotted against the harshness of the judge. If cases are, in fact, randomly assigned to judges, we should not see any strong relationship between offense gravity score and harshness. Only counties with at least two judges who saw at least 30 cases in 1997 are shown.

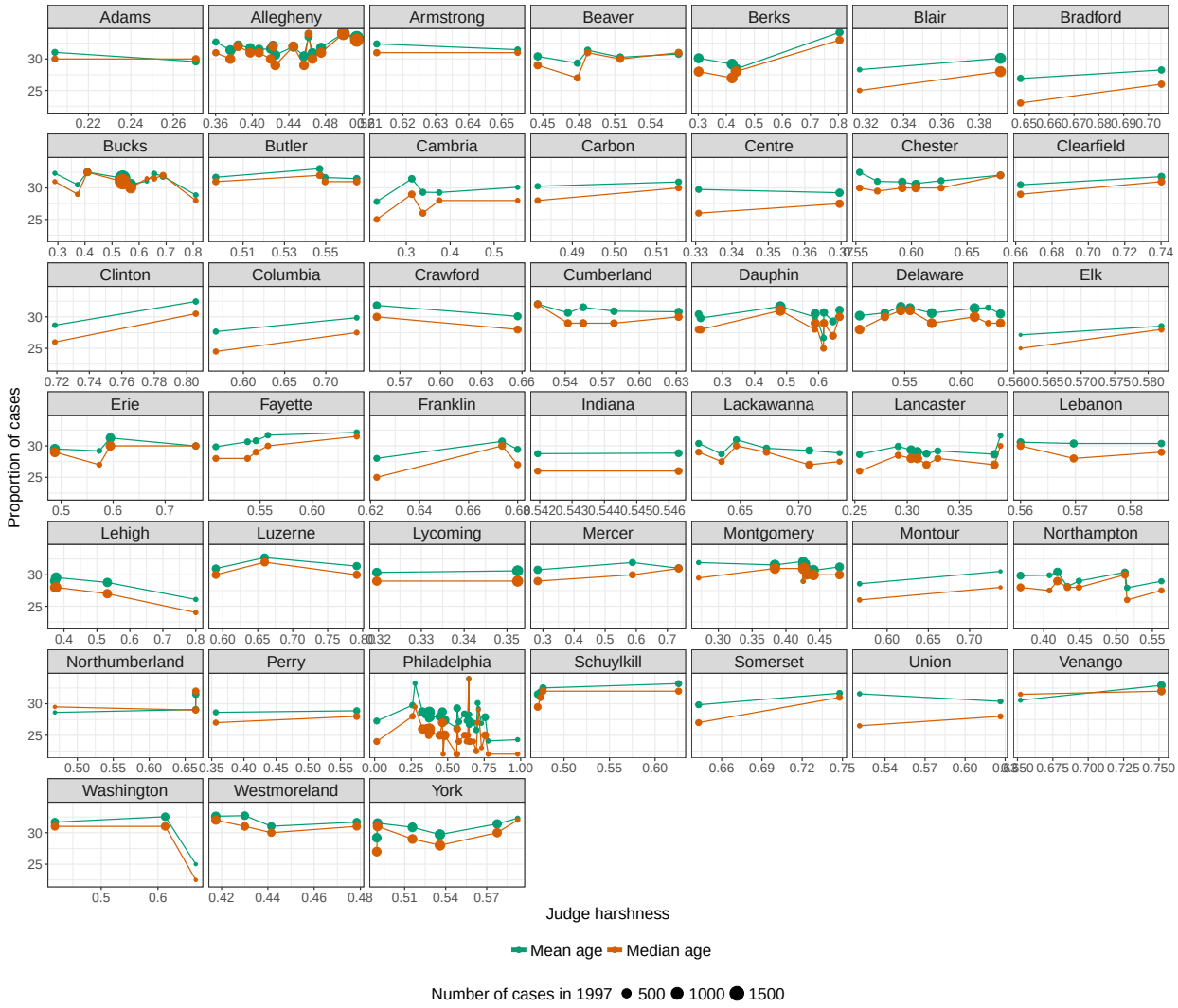


Figure A.5: Mean and median age at date of offense for offenders in the cases seen by each judge in 1997, plotted against the harshness of the judge. If cases are, in fact, randomly assigned to judges, we should not see any strong relationship between age and harshness. Only counties with at least two judges who saw at least 30 cases in 1997 are shown.

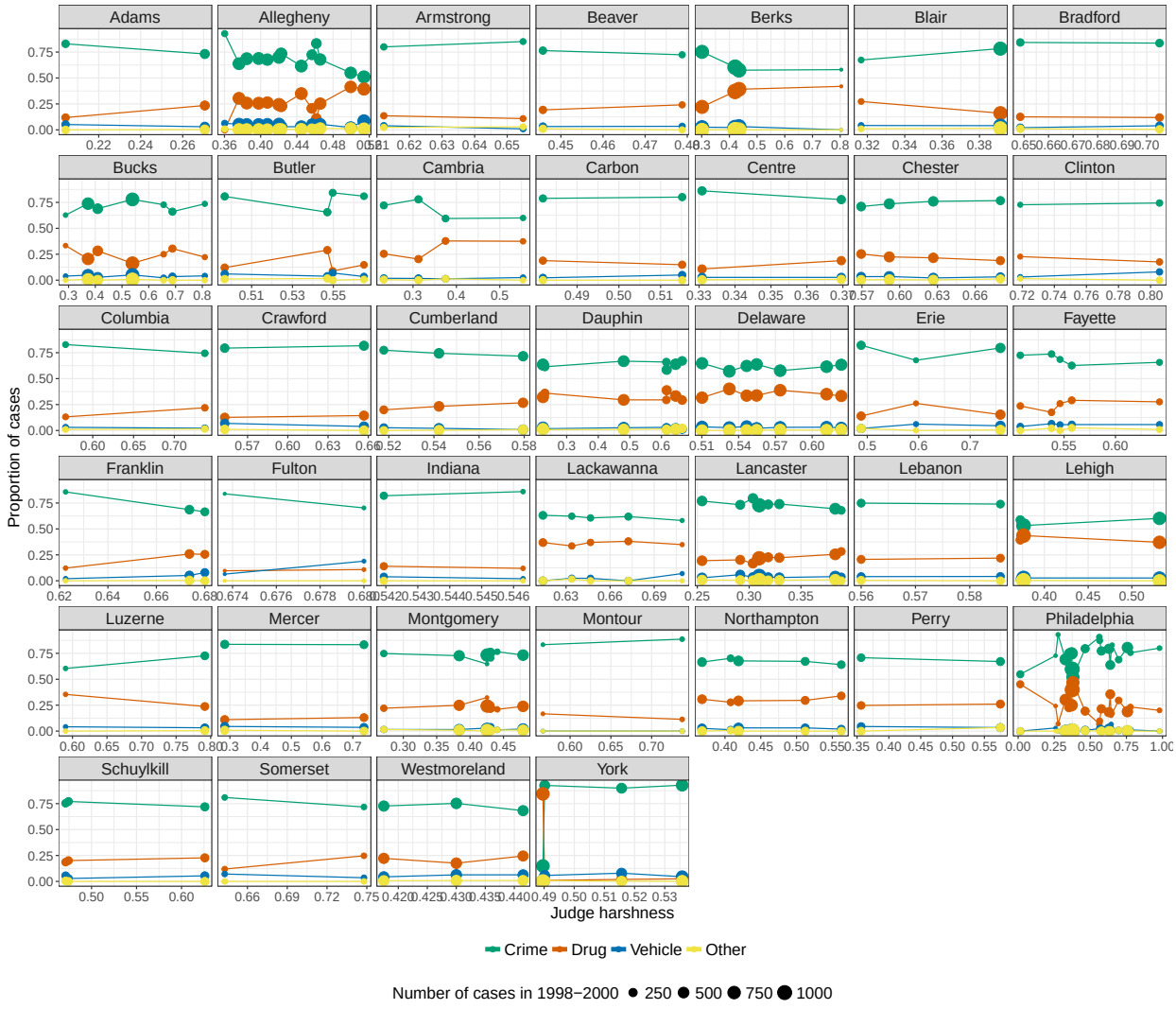


Figure A.6: Proportion of each judge's 1997 cases that fell into each of four major offense categories: crime (green), drugs (orange), vehicle (blue), and other (yellow), plotted against the harshness of the judge. If cases are, in fact, randomly assigned to judges, we should not see any strong relationship between offense category and harshness. Only counties with at least two judges who saw at least 30 cases in 1998-2000 are shown.

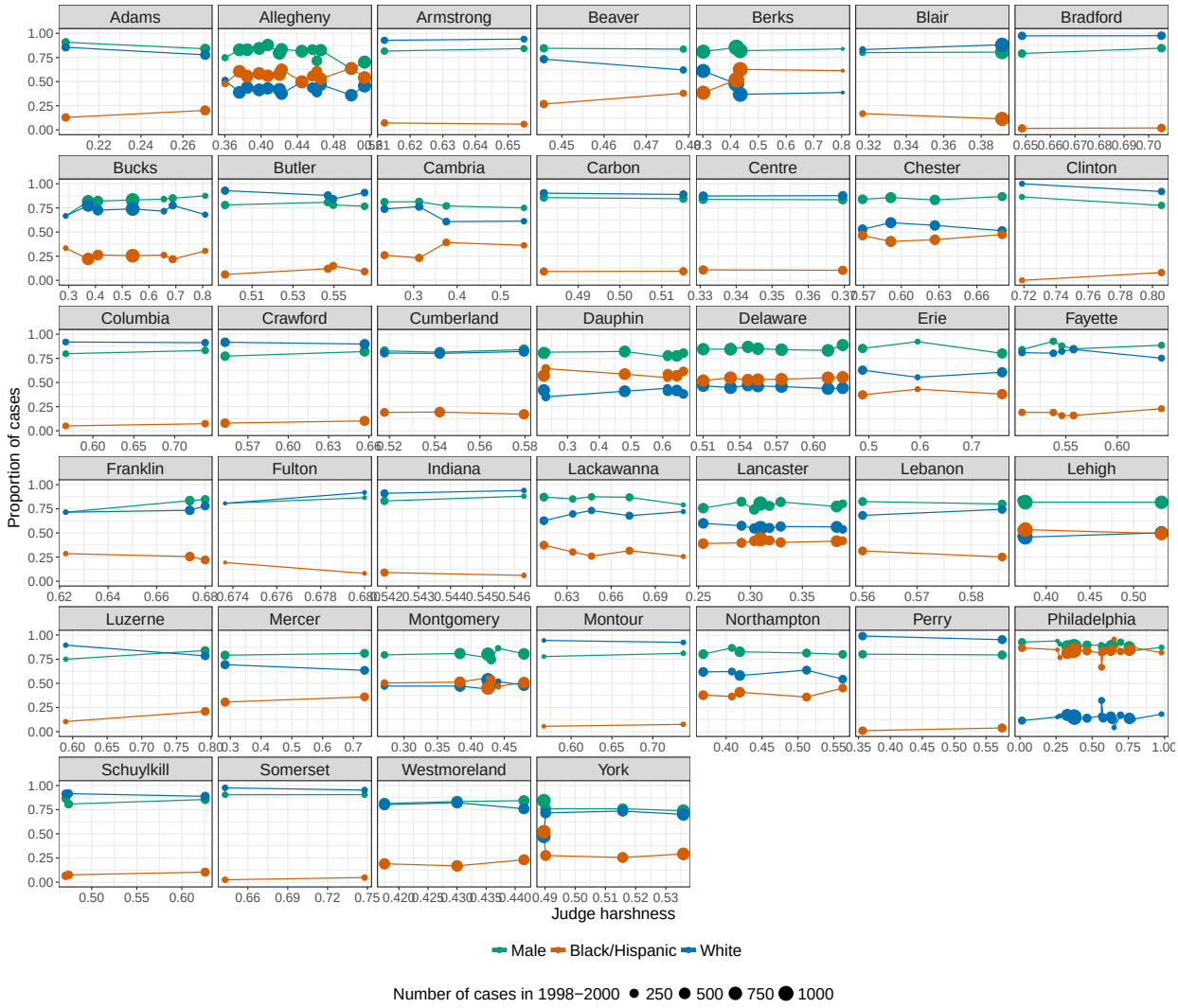


Figure A.7: Proportion of each judge's 1998-2000 cases by sex and race, plotted against the harshness of the judge. If cases are, in fact, randomly assigned to judges, we should not see any strong relationship between the demographic categories and harshness. Only counties with at least two judges who saw at least 30 cases in 1998-2000 are shown.

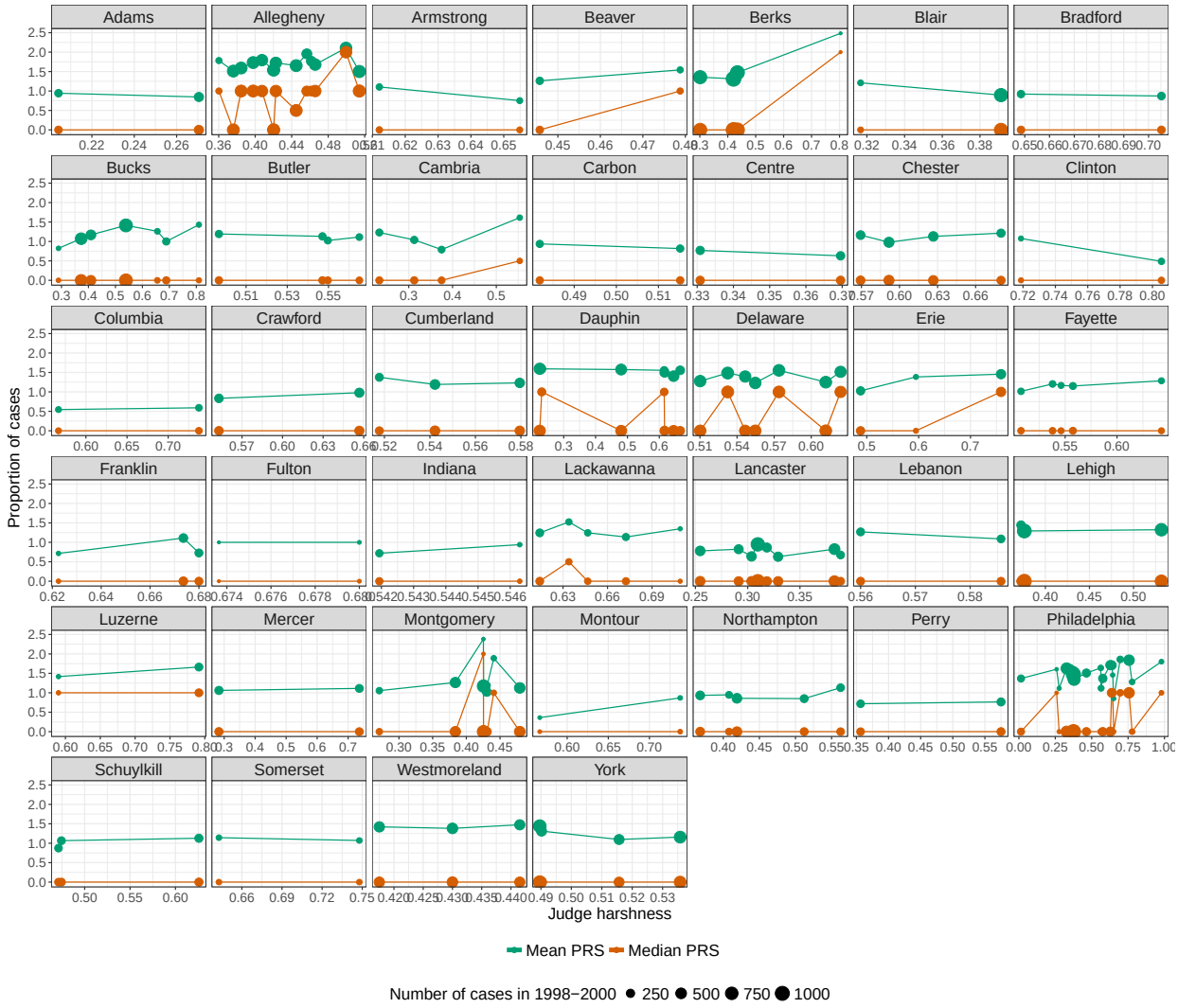


Figure A.8: Mean and median prior record scores of offenders in the cases seen by each judge in 1998-2000, plotted against the harshness of the judge. If cases are, in fact, randomly assigned to judges, we should not see any strong relationship between prior record score and harshness Only counties with at least two judges who saw at least 30 cases in 1998-2000 are shown.

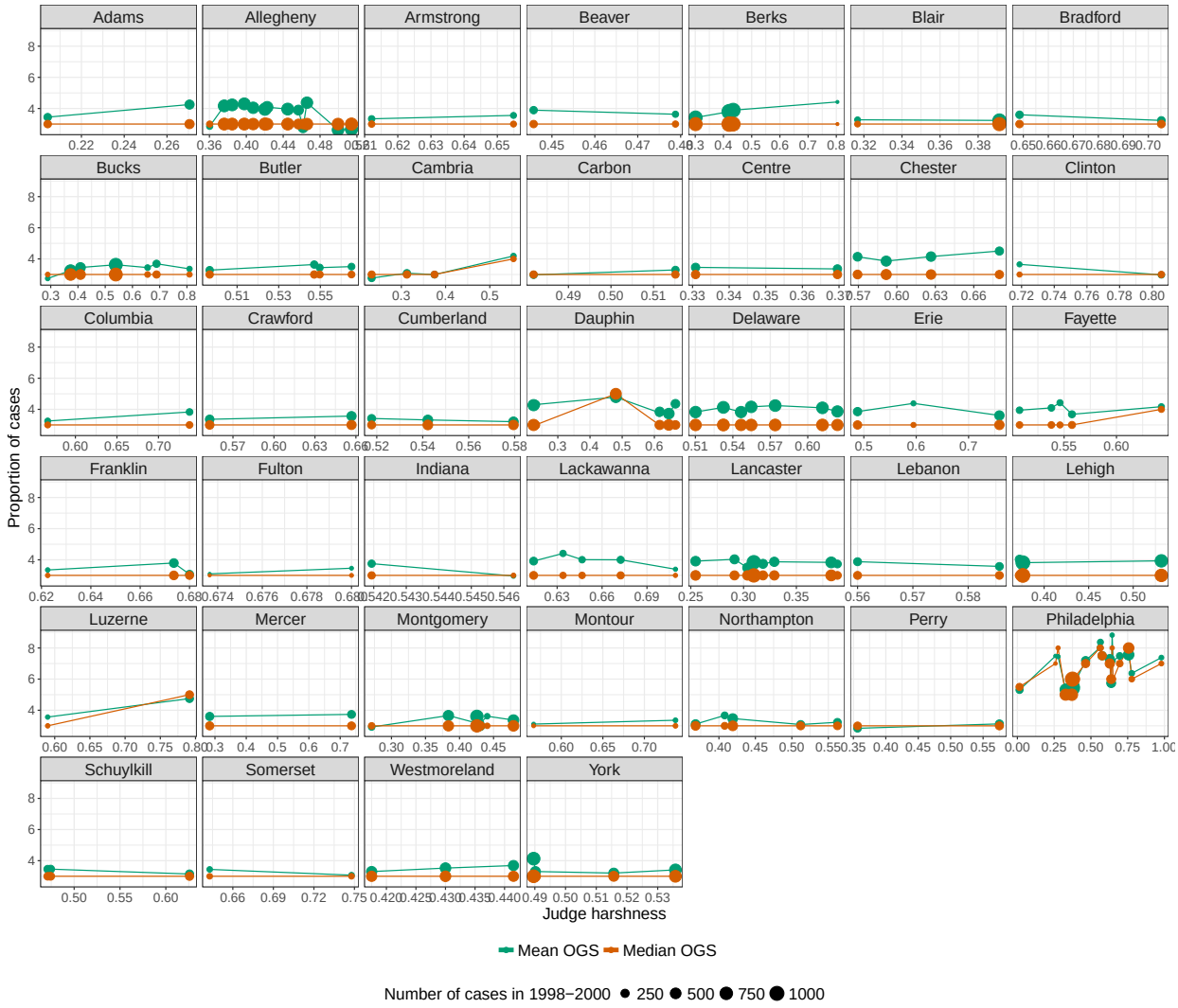


Figure A.9: Mean and median offense gravity scores of offenders in the cases seen by each judge in 1998-2000, plotted against the harshness of the judge. If cases are, in fact, randomly assigned to judges, we should not see any strong relationship between offense gravity score and harshness. Only counties with at least two judges who saw at least 30 cases in 1998-2000 are shown.

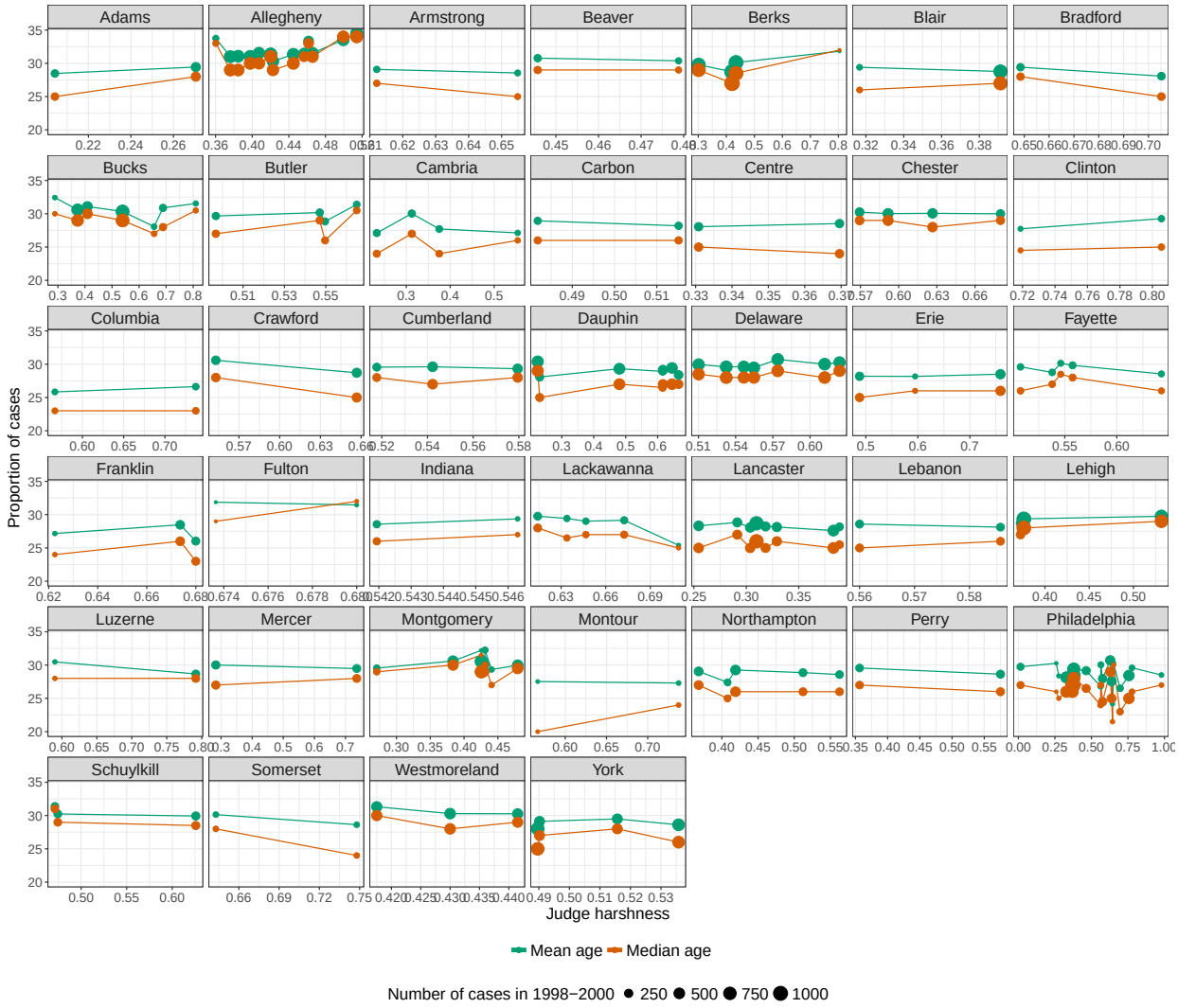


Figure A.10: Mean and median age at date of offense for offenders in the cases seen by each judge in 1998-2000, plotted against the harshness of the judge. If cases are, in fact, randomly assigned to judges, we should not see any strong relationship between age and harshness. Only counties with at least two judges who saw at least 30 cases in 1998-2000 are shown.