Sensitivity Analysis for Outcome Tests with Binary Data

Elisha Cohen[*]

September 13, 2021

Abstract

Outcome tests are a method comparing rates of observed outcomes across selected groups to evaluate bias in decision making processes. Building on the lower bound estimand from Knox et al. (2020), I derive a lower bound in terms of relative risks and develop a sensitivity analysis to weaken the selection-on-observables assumption. Additionally I develop a covariate adjusted sensitivity analysis to assess sensitivity to unmeasured covariates. I am able to estimate a bias adjusted outcome test robust to both measured and unmeasured confounders. Applying this outcome test and sensitivity analysis to data from the Chicago Police Department (1985-2016), I find evidence for gender bias in hiring. I estimate at least 7.4% of men would not have been hired had they been women.

[*] PhD Candidate, Emory University

1 Introduction: Evaluating Bias in Decision Making Processes

From police stops to elections, there are many different decision making processes that we want to know are free from bias and discrimination. A common approach, known as benchmarking, assesses bias by looking at the selection rates across groups. Alternatively, outcome tests, suggested by Becker (1993), are an approach to directly evaluate discrimination in these decision making processes. Outcome tests compare occurrence rates of outcomes across groups to evaluate bias in the selection process. Traditionally, outcome tests are useful as suggestive evidence of bias but provide no measure of magnitude or quantity. As with benchmarking, outcome tests are used in many settings we care about – mortgage lending (Simoiu et al., 2017), pedestrian stops by police (Gelman et al., 2007) and legislator effectiveness (Anzia and Berry, 2011; Cohen and Glynn, 2021). Using the example of mortgage lending, benchmarking looks at whether Black applicants are denied loans at higher rates than white applicants. Outcome tests evaluate outcomes after the loans have been granted and compare default rates across Black and white loan recipients. If Black applicants are more frequently denied loans but on average default at lower rates, an outcome test would suggest this is evidence of bias in the decision making process to give mortgage loans (Simoiu et al., 2017).

The main limitation in traditional outcome tests is that they can suggest a process is discriminatory but do not quantify to what extent. To solve this problem Knox et al. (2020) use a principal stratification framework to derive a lower bound for the outcome test and apply it to racial bias in policing. The prominent drawback to the approach by Knox et al. (2020) is that it requires that within principal strata, groups are comparable after controlling for measured covariates. However, in many situations we may not be able to successfully measure all necessary variables to support this assumption. In this paper I weaken this assumption and develop a sensitivity analysis, building on the E-value approach from VanderWeele and Ding (2017), to evaluate the sensitivity of the lower bound estimand to unobserved confounders. Additionally, I advance an approach to adjust the lower bound estimate based on unmeasured confounders as large as important measured covariates. In Section 2 I discuss the assumptions required and derive the parameter of interest, in Section 3 I discuss estimation, Section 4 describes the sensitivity analysis and Section 5 empirically evaluates gender bias in hiring in the Chicago Police Department.

2 Outcome Test using Potential Outcomes

We will consider two groups in this selection process – a discriminated against group and non-discriminated against group. Let $\rho_i \in \{0, 1\}$ indicate person i's group status. $\rho_i = 0$ signifies person i is the non-discriminated against group and $\rho_i = 1$ signifies person i is in the discriminated against group. The selection process for person i is defined as $S_i \in \{0, 1\}$. The outcome is $Y_i \in \{0, 1\}$[1]. Using potential outcome notation we can define the potential selection of person i as $S_i(0)$ if person i is from group $\rho = 0$ and $S_i(1)$ if person i is from group $\rho = 1$. Using a principal stratification framework we can define the rate of the outcome under each selection strata. The population rate of the outcome by group ρ, is defined as follows:

Definition 1 (Population average among those selected from group ρ)**.**

$$E[Y_i|S_i(\rho) = 1, \rho_i]$$

Definition 1A (Sample average among those selected from group ρ)**.**

$$\overline{Y}_\rho \equiv \frac{\sum_{i=1}^{N_\rho} Y_i(\rho)}{\sum_{i=1}^{N_\rho} \rho_i}$$

where N_ρ are the number of observations in group ρ. This average can separately be defined

[1] This extends to the case of a continuous outcome $Y_i \in [0, \infty)$ see Cohen and Glynn (2021)

for each of the four strata defined by the selection process we are concerned with (see Table 1).

Definition 2 (Population Average Among Helped)**.** *Population average outcome among those with $\rho = 0$ that would not have been selected if $\rho = 1$.*

$$E[Y_i | S_i(0) = 1, S_i(1) = 0, \rho_i = 0]$$

Definition 2A (Sample Average Among Helped)**.**

$$\overline{Y}_{S_i(0)=1,S_i(1)=0,\rho_i=0} \equiv \frac{\sum_{i=1}^{N_{\rho=0}} Y_i(0)(S_i(0) = 1, S_i(1) = 0, \rho_i = 0)}{\sum_{i=1}^{N_{\rho=0}} \rho_i = 0}$$

Definition 3 (Population Average Among Always for Non-discriminated Group)**.** *Population average outcome among those with $\rho = 0$ that would have also been selected if $\rho = 1$.*

$$E[Y_i | S_i(0) = 1, S_i(1) = 1, \rho_i = 0] \equiv \frac{\sum_{i=1}^{N_{\rho=0}} Y_i(0)(S_i(0) = 1, S_i(1) = 1, \rho_i = 0)}{\sum_{i=1}^{N_{\rho=0}} \rho_i = 0}$$

Definition 3A (Sample Average Among Always for Non-discriminated Group)**.**

$$\overline{Y}_{S_i(0)=1,S_i(1)=1,\rho_i=0} \equiv \frac{\sum_{i=1}^{N_{\rho=0}} Y_i(0)(S_i(0) = 1, S_i(1) = 1, \rho_i = 0)}{\sum_{i=1}^{N_{\rho=0}} \rho_i = 0}$$

Definition 4 (Population Average Among Always for Discriminated Group)**.** *Population average outcome among those with $\rho = 1$ that would have also been selected if $\rho = 0$.*

$$E[Y_i | S_i(0) = 1, S_i(1) = 1, \rho_i = 1]$$

Definition 4A (Sample Average Among Always for Discriminated Group)**.**

$$\overline{Y}_{S_i(0)=1,S_i(1)=1,\rho_i=1} \equiv \frac{\sum_{i=1}^{N_{\rho=1}} Y_i(1)(S_i(0) = 1, S_i(1) = 1, \rho_i = 1)}{\sum_{i=1}^{N_{\rho=1}} \rho_i = 1}$$

The top left quadrant in Table 1, $S_i(1) = 0$ and $S_i(0) = 0$, occurs when a member of either group will not make it through the selection process. Usually we do not observe this group. The top right quadrant where $S_i(1) = 1$ and $S_i(0) = 0$ occurs when a person from the discriminated group $\rho_i = 1$ would be selected but would not be selected if from the non-discriminated against group $\rho_i = 0$. Given we are thinking about this process from the viewpoint of group $\rho = 0$, I will label this quadrant as the "hurt" group.

Table 1: Principal Strata for S and ρ

		$\rho_i = 1$	
		$S_i(1) = 0$	$S_i(1) = 1$
$\rho_i = 0$	$S_i(0) = 0$	Never $S_i(0) = 0, S_i(1) = 0$	Hurt $S_i(0) < S_i(1)$
	$S_i(0) = 1$	Helped $S_i(0) > S_i(1)$	Always $S_i(0) = S_i(1) = 1$

Note: The parameter of interest is trying to estimate the proportion of group $\rho_i = 0$ that falls in the grey quadrant, $S_i(0) > S_i(1)$.

2.0.1 Assumptions

Assumption 1 (Monotonicity)**.** *The probability that a selected person with $\rho = 1$, would not have been selected had they had $\rho = 0$.*

$$Pr[S(1) = 1, S(0) = 0)] = 0$$

The Monotonicity assumption 1, assumes the "hurt" group does not exist. We may be concerned that this assumption is difficult to support in practice if there are affirmative action type mandates in place aimed at hiring people from a historically marginalized group. For example, in hiring in police departments, if the physical assessment required to be hired is different for women than it is for men it is possible the "hurt" category does exist. However, as long as this group characteristic is associated only with the selection mechanism and is not associated with the outcome there is no violation of monotonicity. The bottom left quadrant describes the situation when $S_i(1) = 0$ and $S_i(0) = 1$. This occurs when a person from the non-discriminated group is selected but they would not have been selected if they were from the discriminated group. We will refer to this as the "helped" group as they are helped by the fact that there is bias in the decision making process. The bottom right quadrant occurs when $S_i(1) = 1$ and $S_i(0) = 1$. This is the "always" selected group because person i is selected no matter which group they belong to.

Assumption 2 (Comparability)**.** *Group $\rho = 0$ are comparable to the "always" select from group $\rho = 1$*

$$E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 1] = E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 0]$$

Comparability assumption 2 requires that the average rate of the outcome for the "always" members of for group $\rho = 1$ equal the average of the "always" members of group $\rho = 0$. Given that this comparability assumption is unlikely to hold except in cases of random sampling we will focus on the following assumption that incorporates covariates:

Assumption 2′ (Comparability within levels of X)**.**

$$E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 1, X_i = x] = E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 0, X_i = x] \quad \forall x \in X$$

Corollary 2.1.

$$\sum_x E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 1, X_i = x]Pr[X_i = x] =$$

$$\sum_x E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 0, X_i = x]Pr[X_i = x]$$

Assumption $2'$ states that within specified levels of X, the average outcome of the "always" observations in group $\rho = 0$ must equal the average outcome of the "always" observations in group $\rho = 1$. This assumption directly provides Corollary 2.1 that the weighted averages across levels of X for the "always" observations are equal. This comparability assumption combined with the monotonicity assumption that all members of group $\rho = 1$ are in the "always" group are necessary to estimate the lower bound of the outcome test.

2.1 Parameter of Interest

The key parameter of interest is the proportion of the non-discriminated group, $\rho = 0$, who would not have been selected if they were from the discriminated group $\rho = 1$. The non-selected group by definition is not observed and I will assume the "hurt" group does not exist[2]. I can rewrite $E[Y_i|\rho_i = 0]$ in terms of the bottom strata of Table 1.

$$E[Y_i|\rho_i = 0] = \pi \cdot E[Y_i|S_i(0) > S_i(1), \rho_i = 0] + (1 - \pi) \cdot E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 0] \quad (1)$$

To derive the lower bound I rearrange Equation 1 and solve for π (see Appendix A for details).

$$\pi = \frac{E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 0] - E[Y_i|\rho_i = 0]}{E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 0] - E[Y_i|S_i(0) > S_i(1), \rho_i = 0]} \quad (2)$$

[2] A violation of this would occur if selected members of the discriminated group would not have been selected had they been from the non-discriminated group.

We will estimate this using sample averages so we can re-write this as:

$$\pi = \frac{\overline{Y}_{\rho_i=0,S_i(0)=S_i(1)=1} - \overline{Y}_{\rho_i=0}}{\overline{Y}_{\rho_i=0,S_i(0)=S_i(1)=1} - \overline{Y}_{\rho_i=0,S_i(0)>S_i(1)}} \tag{3}$$

The smallest value possible for $\overline{Y}_{\rho_i=0,S_i(0)>S_i(1)}$ is zero, so I will substitute into Equation 3 to get a lower bound for π:

$$\pi \geq \frac{\overline{Y}_{\rho=0,s(0)=s(1)} - \overline{Y}_{\rho=0}}{\overline{Y}_{\rho_i=0,S_i(0)=S_i(1)}} \tag{4}$$

By assuming comparability holds, the "always" strata of group $\rho = 0$ can be estimated using the "always" strata of group $\rho = 1$. Assuming monotonicity means I can estimate this using the observed average rate of outcome for the the discriminated group $\rho = 1$.

$$\pi \geq \frac{\overline{Y}_{\rho_i=1} - \overline{Y}_{\rho_i=0}}{\overline{Y}_{\rho_i=1}} \tag{5}$$

Equation 5 is used to estimate the lower bound on the proportion of group $\rho = 0$ who would not have been selected had they been from group $\rho = 1$.

Given the properties of a binary outcome, we can instead write Equation 5 in terms of probabilities:

$$\begin{align}
\pi &\geq \frac{\overline{Y}_{\rho_i=1} - \overline{Y}_{\rho_i=0}}{\overline{Y}_{\rho_i=1}} \nonumber \\
&= \frac{Pr[Y_i = 1|\rho_i = 1] - Pr[Y_i = 1|\rho_i = 0]}{Pr[Y_i = 1|\rho_i = 1]} \nonumber \\
&= 1 - \frac{Pr[Y_i = 1|\rho_i = 0]}{Pr[Y_i = 1|\rho_i = 1]} \nonumber \\
&= 1 - \frac{1}{\frac{E[Pr[Y_i=1|\rho_i=1]]}{E[Pr[Y_i=1|\rho_i=0]]}} \tag{6} \\
&= 1 - \frac{1}{RR_{\rho Y}} \tag{7}
\end{align}$$

Where $RR_{\rho Y}$ is the relative risk of group ρ on the outcome Y.

3 Lower Bound Estimation

It is straightforward to estimate π using the relative risk as shown in Equation 7 comparing the probability of the outcome for groups $\rho = 0$ and $\rho = 1$ if no standardization is necessary. A major concern with this estimation procedure is that it requires the strong assumption 2, comparability. The contribution of this paper is to estimate π after weakening this assumption. First, to better support the assumption of comparability the probabilities and relative risks can be estimated conditioning on measured covariates $\mathbf{X}$. In practical terms the relative risks will use a regression-based approach to estimate these probabilities:

$$Pr[Y_i = 1|\rho_i] = h(\rho_i, \mathbf{X}_i, \beta) \tag{8}$$

Where $h(\cdot)^{-1}$ is a specified link function, ρ_i is group membership, $\mathbf{X}_i$ is a vector of explanatory variables and β is a $K \times 1$ vector of parameters. Conditioning on $\mathbf{X}$ helps support the assumption that the "always" strata of group $\rho = 1$ is comparable to the "always" strata of group $\rho = 0$. The relative risk $RR_{\rho Y|\mathbf{X}}$ is then the relative risk of group $\rho = 1$ compared to group $\rho = 0$ on the outcome Y marginalized over $\mathbf{X}$:

$$RR_{\rho Y|\mathbf{X}} = \frac{E[Pr[Y_i = 1|\rho_i = 1, \mathbf{X} = x]]}{E[Pr[Y_i = 1|\rho_i = 0, \mathbf{X} = x]]} \tag{9}$$

In the main empirical analysis in this paper I will use a Poisson model with a log-link (Greenland, 2004). The Poisson model allows us to directly estimate the relative risk and is constant across all values of $\mathbf{X} = \mathbf{x}$. The parameter on the ρ variable, β_ρ, is the log of the relative risk of the outcome when $\rho = 1$ compared to $\rho = 0$. Whenever a log-link is used the relative risk is directly estimated. With a logistic-link the probabilities are first separately estimated and then the ratio forms the relative risk (see Appendix B for estimation using a logistic regression approach). Using the relative risk, the lower bound on the proportion of group $\rho = 0$ who would not have been selected had they been group $\rho = 1$, marginalized across $\mathbf{X}$, is estimated as:

$$\hat{\pi}_{\rho,\mathbf{X}} \geq 1 - \frac{1}{\widehat{RR}_{\rho Y|\mathbf{X}}} \tag{10}$$

4 Sensitivity Analysis

A major concern we may have when estimating bias in a decision making process is that there are omitted variables associated with both group membership and the outcome. As seen in Figure 1, the unobserved confounder U could be associated with both group membership ρ and the outcome Y. I assume $\mathbf{X}$ and U are independent by conceptualizing U (or the linear combination of many Us) as the part of the outcome orthogonal to $\mathbf{X}$. An omitted variable having this relationship with group membership and the outcome would violate the assumption of comparability and mean that the true model that I would like to estimate is as follows in Equation 11:

$$Pr[Y_i = 1|\rho_i] = h(\rho_i, \mathbf{X}_i, U_i, \beta) \tag{11}$$

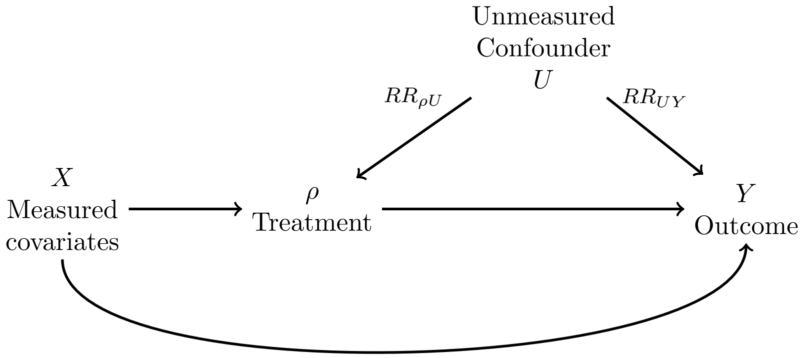

Figure 1: DAG with unmeasured confounder

The original specification from Equation 8 was then actually estimating a restricted model where the true model and observed (restricted) model are related in the following way:

$$RR_{\rho Y|\mathbf{X},U} \geq \frac{RR_{\rho Y|\mathbf{X}}}{BF_{U,\mathbf{X}}} \tag{12}$$

where $BF_{U,\mathbf{X}}$ is the bias factor

$$BF_{U,\mathbf{X}} = \frac{RR_{UY|\rho,\mathbf{X}} RR_{\rho U|\mathbf{X}}}{RR_{UY|\rho,\mathbf{X}} + RR_{\rho U|\mathbf{X}} - 1} \tag{13}$$

4.0.1 Binary confounder

The bias factor is the amount of estimation bias due to omitting a confounder and is determined by both the association of the confounder with the outcome and the treatment with the confounder (Ding and VanderWeele, 2016). $RR_{UY|\rho,\mathbf{X}}$ is the associated relative risk of the unobserved confounder U with the outcome Y for $\rho = \rho$ and $\mathbf{X} = x$. This is a measure of how important the association of the confounder is with the outcome. For a binary confounder, continuing to marginalize across $\mathbf{X}$, we have

$$RR_{UY|\mathbf{X}} = \frac{E\left[Pr[Y=1|U=1,\rho,\mathbf{X}=x]\right]}{E\left[Pr[Y=1|U=0,\rho,\mathbf{X}=x]\right]}$$

For the relative risk for group membership on the unobserved confounder we have $RR_{\rho U|\mathbf{X}}$ as

$$RR_{\rho U|\mathbf{X}} = \frac{E\left[Pr[U=1|\rho=1,\mathbf{X}=x]\right]}{E\left[Pr[U=1|\rho=0,\mathbf{X}=x]\right]}$$

4.0.2 Non-binary confounder

The confounder need not be binary and if this is the case $RR_{\rho U|\mathbf{X}}$ denotes the maximum relative risk $U = k$ for all $k = 0, 1, \ldots K-1$ and $U = l$ for all $l = 0, 1, \ldots, L-1$ marginalized across $\mathbf{X}$. In each of the following we choose the levels of k and l to maximize the ratio such that

$$RR_{UY|\rho,\mathbf{X}} = \frac{E\left[max_k Pr[Y=1|U=k,\rho,\mathbf{X}=x]\right]}{E\left[min_l Pr[Y=1|U=l,\rho,\mathbf{X}=x]\right]}$$

Similarly, $RR_{\rho U|\mathbf{X}}$ is the maximum relative risk for group membership on the unobserved confounder where k is the level at which the relative risk of treatment on the outcome is the largest and marginalized across $\mathbf{X}$.

$$\underset{max\,\rho U|\mathbf{X}}{RR} = \frac{E\left[Pr[U=k|\rho=1,\mathbf{X}=x]\right]}{E\left[Pr[U=k|\rho=0,\mathbf{X}=x]\right]}$$

Given knowledge about the magnitude of these confounder relative risks it would be straightforward to calculate the confounder bias $BF_{U,\mathbf{X}}$ (Equation 13), the relative risk $RR_{\rho Y|\mathbf{X},U}$(Equation 12) and finally $\pi_{\rho,\mathbf{X},U}$

$$\pi_{\rho,\mathbf{X},U} = 1 - \frac{1}{\frac{RR_{\rho Y|\mathbf{X}}}{BF_{U,\mathbf{X}}}} \tag{14}$$

$$\pi_{\rho,\mathbf{X},U} = 1 - \frac{1}{RR_{\rho Y|\mathbf{X},U}} \tag{15}$$

where $\pi_{\rho,\mathbf{X},U}$ is the proportion of members in group $\rho = 0$ who would not have been selected had they been in group $\rho = 1$ marginalized across observed $\mathbf{X}$ and unobserved U. In rare cases with domain expertise, there may be a reason why we are unable to observe U directly to include in our main estimation but we are sufficiently knowledgable about the magnitudes of the associated relative risks. For any pair of values for the associated relative risks we could carryout the above process for estimating our quantity of interest. Without substantiated knowledge about the magnitudes of the unobserved confounder we need a systematic approach for contextualizing the robustness of our findings.

4.1 Explaining away the entire effect

Given the difficulty in knowing appropriate values for these parameters it is common to instead focus on what size of relative risks would give us a null effect. We see this going back to Cornfield et al. (1959) and their evaluation that a confounder would have to be 9 times the risk of lung-cancer for a non-smoker to believe the confounder, and not smoking, leads to lung cancer. More recently, this has been formalized by the E-value as the minimum association of an unmeasured confounder that is necessary to fully explain away the treatment-outcome effect (VanderWeele and Ding, 2017). Thinking about the magnitude of an unmeasured confounder in this way helps us to evaluate the feasibility and likelihood of possible confounders that could substantively alter our findings. The two thresholds that must be met to fully explain away the effect are discussed below.

4.1.1 Low and high thresholds for bounding the confounder

The general Cornfield conditions state that in order to completely wipe out the effect of treatment with the outcome the threshold that must be met is $min(RR_{\rho U|\mathbf{X}}, RR_{UY|\rho,\mathbf{X}}) > RR_{\rho Y|\mathbf{X}}$ (Cornfield et al., 1959). For this to hold, *both* the confounder relatives risks must be at least as large as the main overall relative risk of treatment on outcome. This is shown as the red dashed lines in Figure 5. If the main treatment-outcome effect is $RR_{\rho Y|\mathbf{X}}$ then both confounder relative risks must be at least this large for the confounder to completely wipe out the effect. Additionally, there is a high threshold condition that must also be satisfied in order to explain away the effect – the contour line. This contour comes from the combination of confounder relative risks such that the bias factor would equal that of observed $RR_{\rho Y|\mathbf{X}}$. The point labelled “E-value” is the E-value calculation from VanderWeele and Ding (2017) where the two confounder relative risks are set equal to each other. To completely wipe out the main effect the pair of relative risks must be on or above this contour line.

Therefore, the first step in the sensitivity analysis, is to estimate these thresholds and consider the likelihood of an unmeasured confounder meeting them. Thresholds that are

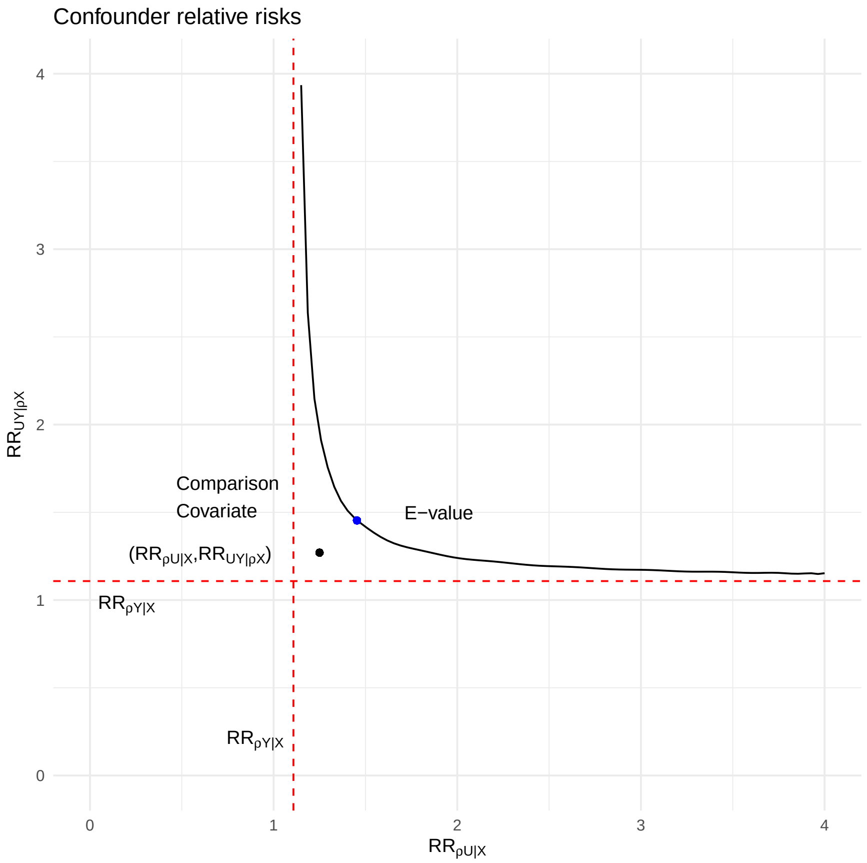

Figure 2: Low and high thresholds for completely exlaining away the main effect. The red dashed lines represent the low threshold requirement. The confounder relative risks must be on or above the contour line to completely wipe out the main effect.

large give suggestive evidence that the main relative risk of interest is robust to unobserved confounders. Understanding how robust our main relative risk is to confounding is important, but is a quite conservative approach viewing these estimates in an all or nothing manner. A more practical approach may be to consider how a reasonable amount of confounder bias could change – but not wipe out completely – our results. In this paper I propose two additional steps. First, is to use an observed covariate as a comparison metric to better evaluate the feasibility of meeting these thresholds and fully wiping out the observed treatment-outcome effect. Second, if it is unlikely that the both thresholds are met, I make a confounder bias adjustment to the main relative risk and calculate an adjusted lower bound on the outcome test.

4.2 Comparison covariate

Given that I do not observe U and therefore cannot directly estimate the bias from leaving U out of the model, I will conduct a sensitivity analysis using a comparison covariate in order to weaken the selection-on-observables assumption. I will choose a comparison covariate, X_j, that is both theoretically and substantially associated with group membership and with the outcome. To perform the sensitivity analysis I will use $RR_{X_jY|\rho,\mathbf{X}_{-j}}$ in place of $RR_{UY|\rho,\mathbf{X}}$ and $RR_{\rho X_j|\mathbf{X}_{-j}}$ in place of $RR_{\rho U|\mathbf{X}}$. I first evaluate whether a confounder the same strength as the comparison covariate could potentially wipe out the entire effect of group membership on the outcome. That is, does the comparison covariate meet the low and high thresholds discussed in subsubsection 4.1.1. If that threshold is not met I can then estimate the bias factor $BF_{X_j,X_{-j}}$, the adjusted overall relative risk (Equation 17) and outcome test lower bound for including a confounder of similar strength as the comparison. Using the comparison covariate the bias factor will be

$$BF_{X_j,X_{-j}} = \frac{RR_{X_jY|\rho,\mathbf{X}_{-j}} RR_{\rho X_j|\mathbf{X}_{-j}}}{RR_{X_jY|\rho,\mathbf{X}_{-j}} + RR_{\rho X_j|\mathbf{X}_{-j}} - 1} \tag{16}$$

the adjusted overall relative risk is then

$$\widehat{RR}^{adj.}_{\rho Y|\mathbf{X}} = \frac{\widehat{RR}^{obs}_{\rho Y|\mathbf{X}}}{BF_{X_j,X_{-j}}} \tag{17}$$

and the adjusted outcome test lower bound is

$$\hat{\pi}^{adj}_{\rho,\mathbf{X}} \geq 1 - \frac{1}{\widehat{RR}^{adj}_{\rho Y|\mathbf{X}}} \tag{18}$$

It is important to note that all relative risks discussed have been conditional on observed covariates $\mathbf{X}$. Therefore any results such as the adjusted relative risk $\widehat{RR}^{adj}_{\rho Y|\mathbf{X}}$ or the adjusted parameter $\hat{\pi}^{adj}_{\rho,\mathbf{X}}$ hold within a given stratum of $\mathbf{X}$. In more practical terms, $\hat{\pi}^{adj}_{\rho,\mathbf{X}}$ will be averaged across all strata $\mathbf{X}$. Next, I will apply this new methodological approach to evaluate gender bias in hiring of women in policing.

5 Empirical Application: Women in Policing

In 2001, the National Center for Women & Policing conducted their fifth annual status report on women in law enforcement in the United States. They report a discouragingly small proportion of all sworn law enforcement positions held by women – 12.7% for agencies with 100 or more sworn officers (8.1% for small and rural agencies). In these large agencies, women also hold few advanced positions – 7.3% in top command positions and 9.6% in supervisory roles. The report also highlights that not only has the proportion of women increased only 4 percentage points from 1990 to 2001 but that this change has stalled or even decreased since the late 1990s (Lonsway et al., 2002).

More recent reports offer a similar story. From the NIBRS 2017 report of 12.5% proportion of female police officers, we see this stagnating trend has not improved. Notably, the 2001 report is now 20 years old because this organization, dedicated to supporting and studying women in policing, has stopped being funded. This highlights the lack of attention given to the gender inequity in policing in the United States. The status of women in policing has not demonstrably changed in the last two decades and while there is a growing focus in political science on race and ethnicity in policing, little attention has been given to gender.

The police are a political and social institution that have direct and often frequent contact in people's lives. Street-level bureaucrats like the police, social workers, etc. are the agents of the government that people most frequently come into contact with and have a high level of discretion over the allocation of public goods, such as welfare benefits or public housing as well as the ability to confer status such as "criminal". "Moreover, when taken together the individual decisions of these workers become, or add up to, agency policy" (Lipsky, 2010). The lack of descriptive representation in the police can directly affect outcomes through the individual behaviors of officers as even with a strong police culture, individual group identities may matter (LeCount, 2017). Additionally, in government, increased collective descriptive representation is associated with higher perceived responsiveness and legitimacy by its citizens (Atkeson and Carrillo, 2007) and so too can the perception of responsiveness and legitimacy of the police be threatened when the officers seem far removed on gender, race and ethnicity form the people they serve (Lipsky, 2010). This could be especially important in disadvantaged neighborhoods that are subject to aggressive policing but have slow responses to citizen complaints and citizen initiated interactions. Lerman and Weaver (2014, page 205) write "[B]ecause citizens derive evaluations of authorities from their personal and vicarious contact, both the concentration and character of policing can have a powerful influence on resulting attitudes about law enforcement." The lack of women in policing due to gender bias has implications for the efficacy, repsonsiveness and legitimacy of the police.

5.1 Women in policing bias

The raw numbers of women in policing are very low compared to the population. However, this alone is not evidence of gender bias. If women have different preferences and choose to apply to become police officers at much lower rates, than even a completely unbiased recruiting and hiring process could result in disproportionate group sizes. An outcome test can be used to reveal bias in policing if on average women officers have higher performance than men officers. There may be gender bias in the selection process from two mechanisms. First, women may perceive there to be gender bias in recruiting and hiring. If this is the case only the most qualified and capable women will even apply resulting in higher performance on average. Second, there may be gender bias by police in the recruiting and hiring of women. Therefore only the most qualified women applicants will be hired leading to, on average, higher performance by women than men.

5.2 Chicago Police Department Data

The Invisible Institute, mainly through FOIAs, have collected data from the Chicago Police Department (CPD) through what they call the “Citizens Police Data Project” (Invisible Institute, 2019). Their goal is to collect and release different forms of interactions between civilians and police officers such as citizen complaints and use of force reports as well as salary and roster information about the officers. The data has been released via a GitHub repository (see CPD (2019)) to support transparency. Using the matching done by Invisible Institute (2019), I am able to link police officer roster data with data on monetary settlements paid by the CPD to settle police misconduct lawsuits.

The roster data goes back until 1946 but the reliability is questionable. Women were only first assigned to patrol duties in Chicago in 1974 ((Chicago Police Department, 2019)). Given that the outcome I am focusing on is settlements, which include incidents involving excessive force and unlawful arrest, it does not make sense to use data prior from when women could have been involved in these types of incidents. Additionally, I have Law Enforcement Management and Administrative Statistics (LEMAS) data from 1985 which gives aggregate counts of sworn male and female police officers totaling 12,478 people (10.64% are women)(Bureau of Justice Statistics., 1985). From the aggregate numbers provided by the LEMAS data I am able to verify the stock total of officers in the CPD data in 1985 (note that many officers were appointed before 1985).

Therefore, I will use CPD data from 1985 until 2016 to conduct the empirical analyses. Every officer is assigned a 0 or 1 value for the settlement outcome where no settlement is a 1. From 1985 to 2016, 4.3% of women and 7.9% of men have had a settlement. In the next section I will use this data and apply the outcome test methodology to estimate gender bias in hiring.

5.3 Outcome Test Results on Chicago Police Department Data

Using the CPD data, group membership is $\rho = 0$ for men and $\rho = 1$ for women. In this application men are the non-discriminated group and women are the discriminated against group. The goal of the lower bound estimation $\hat{\pi}_{\rho}$, is to estimate the proportion of men who would not have been hired had they been women. No settlement is the outcome Y measuring one aspect of job performance. Given the observed rate of no settlement for women is 95.6% and 92.0% for men, the estimated lower bound of the proportion of men who would not have been hired had they been women is 3.8%. With no adjustment of covariates, 3.8% of men (917 men) would not have been hired had they been women.

In order to better support Assumption 2, comparability of groups in the "always" strata, I use Poisson regression specifications to estimate relative risks for the parameter $\hat{\pi}_{\rho,\mathbf{X}}$ conditioned on covariates. More comparable samples help uphold the assumption that the "always" hired are the same on average for both men and women. For the regression specification, the "no settlement" outcome is regressed on the main variable of interest, gender. Additional covariates for race and appointed year are included in the main specification. Table 2 shows the Poisson regression results using a log link. The exponentiated coefficients give us the relative risks. Looking at conditioning set 2, women have a 9% increase in the likelihood of no settlement as compared to men. A limitation of this data is that the roster of complete information is from a single point in time, 2017. Therefore, appointed year (which I have binned into categories of five years), captures both time trends related to policing and the behaviors that may lead to lawsuits, as well as experience of the officer. Figure 3 shows the total number of settlements by year of the incident. The number of incidents occurring peaked in 2011. Additionally, we can see from Figure 4 that officers hired in the early 2000s were the cohort most heavily involved in incidents leading to settlements. Including appointed year as a binned variable in the model specification helps parse out the different associations of outcome with gender versus these trends in policing. Using conditioning sets 1 and 2, I estimate that at least 9.6% (8.3%) of men would not have been hired had they been women (Table 3).

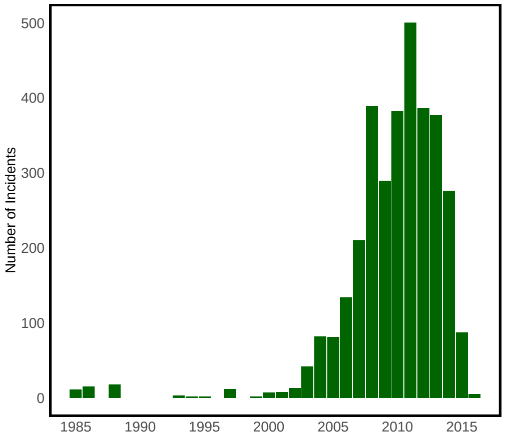

Figure 3: Total number of settlements by year of incident

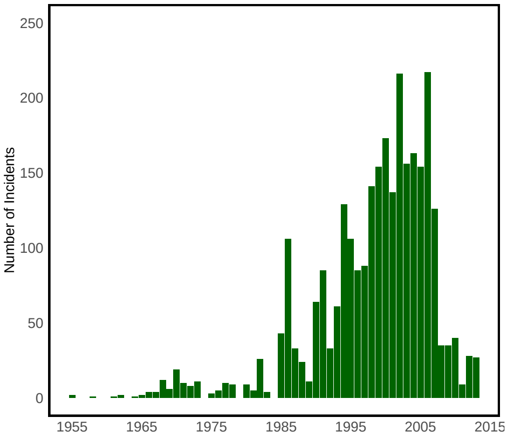

Figure 4: Total number of settlements by appointed year of officer involved

5.4 Sensitivity analysis adjustment

The sensitivity analysis is necessary to evaluate how strong an unobserved confounder U must be to potentially reduce the estimated lower bound on the parameter of interest to zero. In order to adjust my estimates to be robust to possible unmeasured confounders, I choose an observed covariate with a strong association with the outcome to use as a comparison. I use the cohort of officers appointed in 2000-2005, a time period with heavily enforced broken windows policing compared to the cohort from 1985-1990. From Table 2 we can see that the poisson coefficients on the 1985-1990 cohort as compared to the 2000-2005 are 0.20 and 0.19 on the log scale making the relative risks 1.11 and 1.09.

I separately estimate the $RR_{X_JY|\mathbf{X}_{-j}}$ for the association of this cohort with the outcome and $RR_{\rho X_j|\mathbf{X}_{-j}}$ for the relative risk of gender with this cohort. Using the process discussed in subsubsection B.0.2 I choose the maxima of the comparisons between the levels for appointed year by group membership. When using a Poisson regression to estimate these relative risks I do not have to separately estimate for group $\rho = 1$ and $\rho = 0$ since they are equivalent. The largest ratio is between the 2000-2005 and 1985-1990 comparison as follows:

	No Settlement	No Settlement
	(1)	(2)
Woman ($\rho = 1$)	0.10^{***}	0.09^{***}
	(0.01)	(0.01)
Black	0.04^{***}	0.04^{***}
	(0.01)	(0.01)
Hispanic	0.01	0.01
	(0.01)	(0.01)
Asian	0.01	-0.01
	(0.02)	(0.02)
Native American	-0.02	-0.05
	(0.07)	(0.07)
Appointed Year (1985,1990]	0.20^{***}	0.19^{***}
	(0.01)	(0.01)
Appointed Year (1990,1995]	0.18^{***}	0.17^{***}
	(0.01)	(0.01)
Appointed Year (1995,2000]	0.13^{***}	0.12^{***}
	(0.01)	(0.01)
Appointed Year (2005,2010]	0.02	0.02
	(0.02)	(0.02)
Appointed Year (2010,2015]	0.25^{***}	0.24^{***}
	(0.01)	(0.01)
Intercept	-0.31^{***}	-0.66^{*}
	(0.01)	(0.27)
Current Unit	No	Yes
Current Rank	No	Yes
AIC	31568.53	31795.99
BIC	31653.00	33093.64
Log Likelihood	-15773.27	-15728.99
Deviance	3600.53	3511.99
Num. obs.	15969	15969

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Poisson regression results used for estimating risk ratios. Robust standard errors shown.

		Covariate Set 1		Covariate Set 2	
		Est.	CI	Est.	Lower CI
Main	$\widehat{RR}_{\rho Y\|\mathbf{X}}$	1.107	[1.095,∞]	1.091	1.079
estimates	$\hat{\pi}_{\rho,\mathbf{X}}$	9.6%	[8.7%,100%]	8.3%	[7.3%,100%]
	BF	1.014	[1.005,∞]	1.01	[1.005,∞]
U adjusted	$\widehat{RR}^{adj}_{\rho Y\|\mathbf{X}}$	1.091	[1.09,∞]	1.08	[1.078,∞]
estimates	$\hat{\pi}^{adj}_{\rho,\mathbf{X}}$	8.3%	[8.2%,100%]	7.4%	[7.2%,100%]

Table 3: Estimated RR and outcome test lower bound conditional on covariates using Poisson regression. Bottom two rows are adjusted for unobserved confounding as large as an observed comparison covariate by first estimating the bias factor and then the relative risk and lower bound using Equation 17 and Equation 18. All confidence interval estimates use robust standard errors (Zou, 2004).

$$\underset{max\, X_jY|\rho,\mathbf{X}_{-j}}{RR} = \frac{max_{1985-1990}Pr[Y=1|\rho, X_j = Appointment_{1985-1990}, \mathbf{X}_{-j}]}{min_{2000-2005}Pr[Y=1|\rho, X_j = Appointment_{2000-2005}, \mathbf{X}_{-j}]}$$

Similarly, the largest ratio for the model of confounder regressed on treatement is:

$$\underset{max\, \rho X_j|\mathbf{X}_{-j}}{RR} = \frac{Pr[X_j = Appointment_{1985-1990}|\rho = 1, \mathbf{X}_{-j}]}{Pr[X_j = Appointment_{1985-1990}|\rho = 0, \mathbf{X}_{-j}]}$$

For the first conditioning set used in model 1, $\underset{max\, X_jY|\rho,\mathbf{X}_{-j}}{RR} = 1.22$ and $\underset{max\, \rho X_j|\mathbf{X}_{-j}}{RR}$ is estimated to be 1.09. Given the estimated main effect of 1.11, the estimated confounder relative risks do not satisfy the low threshold requirements and an unobserved confounder the same strength as the comparison covariate would not be strong enough to wipe out the entire effect. Using 1.22 and 1.09 in the bias factor equation I can estimate an adjusted risk ratio and then an adjusted quantity of interest. After controlling for the conditioning set in model 1 and adjusting for an unobserved confounder the same strength as the comparison covariate, I find that 8.3% of men would not have been hired had they been women. Repeating this process on conditioning set 2 I estimate a $\underset{max\, X_jY|\rho,\mathbf{X}_{-j}}{RR} = 1.21$ and $RR_{\rho X_j|\mathbf{X}_{-j}} = 1.06$. Given the main relative risk of 1.091 the confounder relative risks are not large enough to completely wipe out the effect. With conditioning set 2 the adjusted risk ratio is 1.08 telling us that

7.4% of men would not have been hired had they been women. [3]

6 Discussion/Conclusion

My estimation of π, the proportion of men who would not have been hired had they been women, is robust and substantively meaningful. Conditioning on covariates actually increased the estimate of π from 3.8% to 9.6% (8.3%). Conditioning on covariates allows us to parse out how race and appointed year (as well as current unit and rank) may be associated with job performance separately from gender. This allows us to be more confident that we are comparing men police officers with similar women police officers. The sensitivity adjusted estimate π of 7.4% continues to be remarkably robust even after adjusting for a confounder as strong as appointed year further strengthening our conclusions that there is gender bias in hiring in the Chicago Police Department. This gives us confidence that we are measuring a true, but likely conservative, estimate of gender bias in selection. Compared to most U.S. cities Chicago is both large in population and high in yearly police department settlements. Future analyses of cities across the United States would be informative. Most likely the CPD is not an outlier in terms of gender bias in the hiring process and a national analysis could shed light on city differences in hiring.

[3] The use of a comparison covariate to contextualize the potential bias from an unobserved confounder is a "naive" benchmark as discussed in Cinelli and Hazlett (2020). This approach does not adjust for potential collider bias and therefore could be underestimating the bias (VanderWeele et al., 2019). However, given that using the covariate of cohort as a confounder comparison only minimally reduces the effect of group membership on the outcome it is very unlikely that collider bias could significantly alter the results.

7 References

References

(2019). GitHub Repository Chicago Police Data.

Anzia, S. F. and Berry, C. R. (2011). The jackie (and jill) robinson effect: why do congresswomen outperform congressmen? *American Journal of Political Science*, 55(3):478–493.

Atkeson, L. R. and Carrillo, N. (2007). More is better: The influence of collective female descriptive representation on external efficacy. *Politics & Gender*, 3(1):79–101.

Becker, G. S. (1993). Nobel lecture: The economic way of looking at behavior. *Journal of political economy*, 101(3):385–409.

Bureau of Justice Statistics. (1985). Law Enforcement Management and Administrative Statistics (LEMAS). Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor].

Chicago Police Department (2019).

Cinelli, C. and Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of Royal Statistical Society Series B*, 82(1):39–67.

Cohen, E. and Glynn, A. (2021). Estimating bounds on selection bias with outcome tests. Working Paper.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22(1):173–203.

Ding, P. and VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology*, 27(3):368.

Gelman, A., Fagan, J., and Kiss, A. (2007). An analysis of the new york city police department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American Statistical Association*, 102(479):813–823.

Greenland, S. (2004). Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *American journal of epidemiology*, 160(4):301–305.

Invisible Institute (2019). Citizens police data project.

Knox, D., Lowe, W., and Mummolo, J. (2020). Administrative records mask racially biased policing. *American Political Science Review*, 114(3):619–637.

LeCount, R. J. (2017). More black than blue? comparing the racial attitudes of police to citizens. In *Sociological Forum*, volume 32, pages 1051–1072. Wiley Online Library.

Lerman, A. E. and Weaver, V. (2014). Staying out of sight? concentrated policing and local political action. *The ANNALS of the American Academy of Political and Social Science*, 651(1):202–219.

Lipsky, M. (2010). *Street-level bureaucracy dilemmas of the individual in public services.* UPCC book collections on Project MUSE. Russell Sage Foundation, New York, 30th anniversary expanded edition.. edition.

Lonsway, K., Carrington, S., Aguirre, P., Wood, M., Moore, M., Harrington, P., Smeal, E., Spillar, K., et al. (2002). Equality denied: The status of women in policing: 2001. *National Center for Women & Policing, a Division of the Feminist Majority Foundation.*

Simoiu, C., Corbett-Davies, S., Goel, S., et al. (2017). The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216.

VanderWeele, T. J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the e-value. *Annals of internal medicine.*

VanderWeele, T. J., Ding, P., and Mathur, M. (2019). Technical considerations in the use of the e-value. *Journal of Causal Inference*, 7(2).

Zou, G. (2004). A modified poisson regression approach to prospective studies with binary data. *American journal of epidemiology*, 159(7):702–706.

Appendix

A Derivation of π from Equation 1

$$\begin{aligned}
&E[Y_i|\rho_i = 0] = \\
&\qquad \pi \cdot E[Y_i|S_i(0) > S_i(1), \rho_i = 0] + (1-\pi) \cdot E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 0] \\
&E[Y_i|\rho_i = 0] = \pi \cdot E[Y_i|S_i(0) > S_i(1), \rho_i = 0] \\
&\qquad + E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 0] - \pi \cdot E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 0] \\
&\pi \cdot E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 0] - \pi \cdot E[Y_i|S_i(0) > S_i(1), \rho_i = 0] = \\
&\qquad E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 0] - E[Y_i|\rho_i = 0] \\
&\pi(E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 0] - E[Y_i|S_i(0) > S_i(1), \rho_i = 0]) = \\
&\qquad E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 0] - E[Y_i|\rho_i = 0] \\
&\qquad \pi = \frac{E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 0] - E[Y_i|\rho_i = 0]}{E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 0] - E[Y_i|S_i(0) > S_i(1), \rho_i = 0]}
\end{aligned}$$

Given Assumption 2, $E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 1]$ can be substituted in for $E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 0]$

$$\pi = \frac{E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 1] - E[Y_i|\rho_i = 0]}{E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 1] - E[Y_i|S_i(0) > S_i(1), \rho_i = 0]}$$

Given Assumption 1, we assume that all members of $\rho = 1$ are in the strata of $E[Y_i|S_i(0) = S_i(1) = 1, \rho_i = 1]$ therefore this can be substituted with the observed rate of the outcome for group $\rho = 1$ and $E[Y_i|\rho_i = 0]$ continues to be the observed rate for group $\rho = 0$.

$$\pi = \frac{E[Y_i|\rho_i = 1] - E[Y_i|\rho_i = 0]}{E[Y_i|\rho_i = 1] - E[Y_i|S_i(0) > S_i(1), \rho_i = 0]}$$

Now the only unobserved quantity is $E[Y_i|S_i(0) > S_i(1), \rho_i = 0]$. Given that the outcome is defined as $Y_i \in \{0, 1\}$, the smallest value this rate could take is 0, making this a lower bound:

$$\pi = \frac{E[Y_i|\rho_i = 1] - E[Y_i|\rho_i = 0]}{E[Y_i|\rho_i = 1] - E[Y_i|S_i(0) > S_i(1), \rho_i = 0]} \geq \frac{E[Y_i|\rho_i = 1] - E[Y_i|\rho_i = 0]}{E[Y_i|\rho_i = 1]}$$

B Estimating with a logistic regression

Instead of using the average of the outcomes by group I can use logistic regression to estimate the risk ratio $\widehat{RR}_{\rho Y|\mathbf{X}}$. This allows us to condition on covariates to better support the assumption that the "always" strata of each group are comparable.
I can estimate the following model specification:

$$P[Y_i = 1|\rho_i, X_i] = \frac{exp^{(\beta_\rho \rho_i + \mathbf{X}_i^{'}\gamma)}}{1 + exp^{(\beta_\rho \rho_i + \mathbf{X}_i^{'}\gamma)}} \tag{19}$$

From Equation 19, the main coefficient of interest is β_ρ which is the estimated coefficient for the binary variable of group membership. $\mathbf{X}_i$ is a $n \times k$ matrix containing k observed covariates (including a constant). Conditioning on $\mathbf{X}_i$ helps support the assumption that the "always" strata of group $\rho = 1$ is comparable to the "always" strata of group $\rho = 0$. β_ρ represents the average difference in outcome, all else equal, between group $\rho = 1$ and group $\rho = 0$. Using the results from the logistic regression $\widehat{RR}_{\rho Y|\mathbf{X}}$ is estimated in the following manner:

$$\widehat{RR}_{\rho Y|\mathbf{X}} = \frac{E[P(\hat{Y} = 1|\rho = 1, X_i = x)]}{E[P(\hat{Y} = 1|\rho = 0, X_i = x)]} \tag{20}$$

$$P[Y_i = 1|\rho_i, X_i, U_i] = \frac{exp^{(\beta_\rho \rho_i + \mathbf{X}_i^{'}\gamma + \hat{\tau} U_i)}}{1 + exp^{(\beta_\rho \rho_i + \mathbf{X}_i^{'}\gamma + \hat{\tau} U_i)}} \tag{21}$$

I separately estimate the $RR_{X_J Y|\mathbf{X}_{-j}}$ for the association of this cohort with the outcome and $RR_{\rho X_j|\mathbf{X}_{-j}}$ for the relative risk of gender with this cohort. Using the process discussed in subsubsection B.0.2 I choose the maxima of the comparisons between the levels for appointed year by group membership. When using a Poisson regression to estimate these relative risks I do not have to separately estimate for group $\rho = 1$ and $rho = 0$ since they are equivalent. The largest ratio is between the 2000-2005 and 1985-1990 comparison as follows:

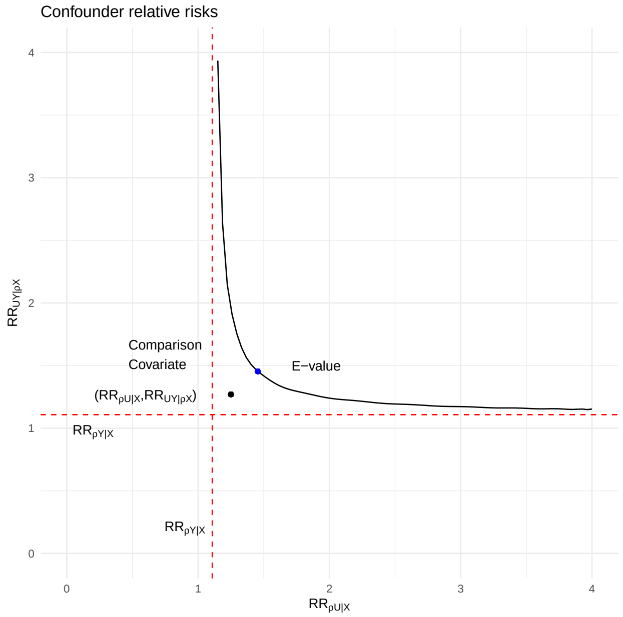

Figure 5: Generic thresholds

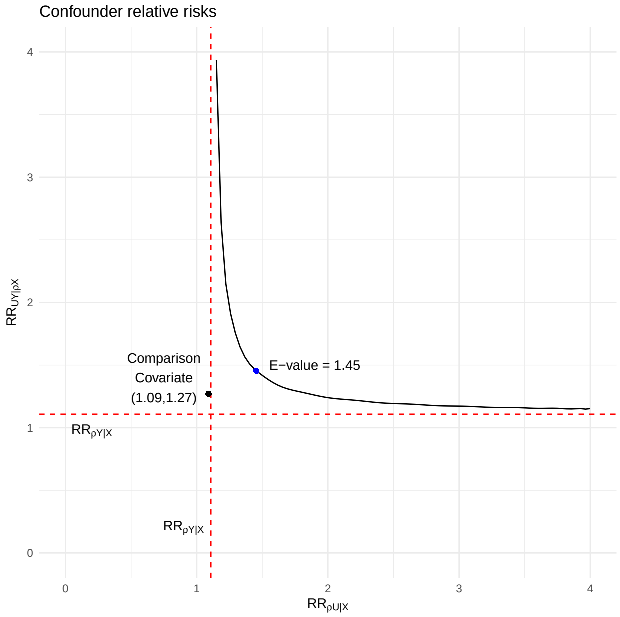

Figure 6: Thresholds with estimates from Model 1

$$RR_{X_jY|\rho=1,\mathbf{X}_{-j}} = \frac{max_{1985-1990}Pr[Y=1|\rho=1, X_j = Appointment_{1985-1990}, \mathbf{X}_{-j}]}{min_{2000-2005}Pr[Y=1|\rho=1, X_j = Appointment_{2000-2005}, \mathbf{X}_{-j}]}$$

$$RR_{X_jY|\rho=0,\mathbf{X}_{-j}} = \frac{max_{1985-1990}Pr[Y=1|\rho=0, X_j = Appointment_{1985-1990}, \mathbf{X}_{-j}]}{min_{2000-2005}Pr[Y=1|\rho=0, X_j = Appointment_{2000-2005}, \mathbf{X}_{-j}]}$$

Similarly this comparison also gives the largest ratio for

$$RR_{\rho X_j|\mathbf{X}_{-j}} = \frac{Pr[X_j = Appointment_{1985-1990}|\rho=1, \mathbf{X}_{-j}]}{Pr[X_j = Appointment_{2000-2005}|\rho=0, \mathbf{X}_{-j}]}$$

For the first conditioning set used in model 1, $RR_{X_jY|\rho,\mathbf{X}_{-j}} = max(RR_{UY|\rho=1,\mathbf{X}}, RR_{UY|\rho=0,\mathbf{X}}) = max(1.09, 1.27)$. Therefore 1.27 is used as the comparison covariate relative risk with the outcome to be subbed in for $RR_{X_jY|\rho,\mathbf{X}_{-j}}$ and $maxRR_{\rho X_j|\mathbf{X}_{-j}}$ is estimated to be 1.09. Given the estimated main effect of 1.11, the estimated confounder relative risks do not satisfy the low threshold requirements and an unobserved confounder the same strength as the comparison covariate would not be strong enough to wipe out the entire effect. Using 1.27 and 1.09 in the bias factor equation I can estimate an adjusted risk ratio and then an adjusted quantity of interest. After controlling for the conditioning set in model 1 and adjusting for an unobserved confounder the same strength as the comparison covariate, I find that 8.2% of men would not have been hired had they been women. Repeating this process on conditioning set 2 I estimate a $RR_{X_jY|\rho,\mathbf{X}_{-j}} = 1.25$ and $RR_{\rho X_j|\mathbf{X}_{-j}} = 1.06$. Given the main relative risk of 1.097 the confounder relative risks are not large enough to completely wipe out the effect. With conditioning set 2 the adjusted risk ratio is 1.086 telling us that 7.9% of men would not have been hired had they been women.

		Covariate Set 1 Est.	Covariate Set 2 Est.
main	$\widehat{RR}_{\rho Y\|\mathbf{X}}$	1.108	1.097
estimates	$\hat{\pi}_{\rho,\mathbf{X}}$	9.8%	8.9%
U adjusted	$\widehat{RR}^{adj}_{\rho Y\|\mathbf{X}}$	1.089	1.086
estimates	$\hat{\pi}^{adj}_{\rho,\mathbf{X}}$	8.2%	7.9%

Table 4: Estimated RR and outcome test lower bound conditional on covariates. Bottom two rows are adjusted for unobserved confounding as large as an observed comparison covariate by first estimating the bias factor and then the relative risk and lower bound using Equation 17 and Equation 18.

B.0.1 Binary confounder

The bias factor is the amount of estimation bias due to omitting a confounder and is determined by both the association of the confounder with the outcome and the treatment with the confounder (Ding and VanderWeele, 2016). $RR_{UY|\rho,\mathbf{X}}$ is the associated relative risk of the unobserved confounder U with the outcome Y within given levels of X, or how important the association of the confounder is with the outcome. For a binary confounder, continuing to marginalize across $\mathbf{X}$, we have both

$$RR_{UY|\rho=1,\mathbf{X}} = \frac{E\left[Pr[Y=1|\rho=1, U=1, \mathbf{X}=x]\right]}{E\left[Pr[Y=1|\rho=1, U=0, \mathbf{X}=x]\right]}$$

and

$$RR_{UY|\rho=0,\mathbf{X}} = \frac{E\left[Pr[Y=1|\rho=0, U=1, \mathbf{X}=x]\right]}{E\left[Pr[Y=1|\rho=0, U=0, \mathbf{X}=x]\right]}$$

.

In order for the bias factor to bound the possible bias we choose the relative risk that maximizes this such that $\underset{max}{RR}_{UY|\rho\mathbf{X}} = max(RR_{UY|\rho=1,\mathbf{X}}, RR_{UY|\rho=0,\mathbf{X}})$. If we are using a Poisson model to estimate these relative risks than this maximation is unneccessary as these relative risks are equal to each other. Therefore we only need to estimate

$$RR_{UY|\mathbf{X}} = \frac{E\left[Pr[Y=1|U=1, \rho, \mathbf{X}=x]\right]}{E\left[Pr[Y=1|U=0, \rho, \mathbf{X}=x]\right]}$$

For the relative risk for group membership on the unobserved confounder we have $RR_{\rho U|\mathbf{X}}$ as

$$RR_{\rho U|\mathbf{X}} = \frac{E\left[Pr[U=1|\rho=1, \mathbf{X}=x]\right]}{E\left[Pr[U=1|\rho=0, \mathbf{X}=x]\right]}$$

B.0.2 Non-binary confounder

The confounder need not be binary and if this is the case $RR_{\rho U|\mathbf{X}}$ denotes the maximum relative risk $U = k$ for all $k = 0, 1, \dots K-1$ and $U = l$ for all $l = 0, 1, \dots, L-1$ marginalized across $\mathbf{X}$. In each of the following we choose the levels of k and l to maximize the ratios such that

$$RR_{UY|\rho=1,\mathbf{X}} = \frac{E\left[max_k Pr[Y=1|\rho=1, U=k, \mathbf{X}=x]\right]}{E\left[min_l Pr[Y=1|\rho=1, U=l, \mathbf{X}=x]\right]}$$

and

$$RR_{UY|\rho=0,\mathbf{X}} = \frac{E\left[max_k Pr[Y=1|\rho=0, U=k, \mathbf{X}=x]\right]}{E\left[min_l Pr[Y=1|\rho=0, U=l, \mathbf{X}=x]\right]}$$

Then, as in the binary case, $\underset{max}{RR}_{UY|\rho\mathbf{X}} = max(RR_{UY|\rho=1,\mathbf{X}}, RR_{UY|\rho=0,\mathbf{X}})$. And with the Poisson we need only estimate

$$RR_{UY|\rho,\mathbf{X}} = \frac{E\left[max_k Pr[Y=1|\rho, U=k, \mathbf{X}=x]\right]}{E\left[min_l Pr[Y=1|\rho, U=l, \mathbf{X}=x]\right]}$$

Similarly, $RR_{\rho U|\mathbf{X}}$ is the maximum relative risk for group membership on the unobserved confounder where k is the level at which the relative risk of treatment on the outcome is the largest and marginalized across $\mathbf{X}$.

$$\underset{max}{RR}_{\rho U|\mathbf{X}} = \frac{E\left[Pr[U=k|\rho=1, \mathbf{X}=x]\right]}{E\left[Pr[U=k|\rho=0, \mathbf{X}=x]\right]}$$

Given knowledge about the magnitude of these confounder relative risks it would be straightforward to calculate the confounder bias $BF_{U,\mathbf{X}}$ (Equation 13), the relative risk $RR_{\rho Y|\mathbf{X},U}$(Equation 12) and finally $\pi_{\rho,\mathbf{X},U}$

$$\pi_{\rho,\mathbf{X},U} = 1 - \frac{1}{\frac{RR_{\rho Y|\mathbf{X}}}{BF_{U,\mathbf{X}}}} \tag{22}$$

$$\pi_{\rho,\mathbf{X},U} = 1 - \frac{1}{RR_{\rho Y|\mathbf{X},U}} \tag{23}$$

where $\pi_{\rho,\mathbf{X},U}$ is the proportion of members in group $\rho = 0$ who would not have been selected had they been in group $\rho = 1$ marginalized across observed $\mathbf{X}$ and unobserved U. In rare cases with domain expertise, there may be a reason why we are unable to observe U directly to include in our main estimation but we are sufficiently knowledgable about the magnitudes of the associated relative risks. For any pair of values for the associated relative risks we could carryout the above process for estimating our quantity of interest.

We can also interpret these on the odds-ratio scale as the 1985-1990 cohort has an increased odds of 4.3 (4.5) of having no settlement as compared to the 2000-2005 cohort. For the first conditioning set used in model 1, $\underset{max}{RR}_{X_jY|\rho,\mathbf{X}_{-j}} = max(RR_{UY|\rho,\mathbf{X}}, RR_{UY|\rho=0,\mathbf{X}}) = max(1.09, 1.27)$. Therefore 1.27 is used as the comparison covariate relative risk with the outcome to be subbed in for $RR_{X_jY|\rho,\mathbf{X}_{-j}}$ and $\underset{max}{RR}_{\rho X_j|\mathbf{X}_{-j}}$ is estimated to be 1.09. Given the estimated main effect of 1.11, the estimated confounder relative risks do not satisfy the low threshold requirements and an unobserved confounder the same strength as the comparison covariate would not be strong enough to wipe out the entire effect. Using 1.27 and 1.09 in the bias factor equation I can estimate an adjusted risk ratio and then an adjusted quantity of interest. After controlling for the conditioning set in model 1 and adjusting for an unobserved confounder the same strength as the comparison covariate, I find that 8.2% of men would not have been hired had they been women. Repeating this process on conditioning set 2 I estimate a $RR_{X_jY|\rho,\mathbf{X}_{-j}} = 1.25$ and $RR_{\rho X_j|\mathbf{X}_{-j}} = 1.06$. Given the main relative risk of 1.097 the confounder relative risks are not large enough to completely wipe out the effect. With conditioning set 2 the adjusted risk ratio is 1.086 telling us that 7.9% of men would not have been hired had they been women.