The Impact of Differential Feature Under-reporting on Algorithmic Fairness

Nil-Jana Akpinar*
nakpinar@amazon.com
Amazon AWS AI/ML
USA

Zachary C. Lipton
Carnegie Mellon University
USA

Alexandra Chouldechova
Carnegie Mellon University
USA

ABSTRACT

Predictive risk models in the public sector are commonly developed using administrative data that is more complete for subpopulations that more greatly rely on public services. In the United States, for instance, information on health care utilization is routinely available to government agencies for individuals supported by Medicaid and Medicare, but not for the privately insured. Critiques of public sector algorithms have identified such "differential feature under-reporting" as a driver of disparities in algorithmic decision-making. Yet this form of data bias remains understudied from a technical viewpoint. While prior work has examined the fairness impacts of additive feature noise and features that are clearly marked as missing, little is known about the setting of data missingness absent indicators (i.e. differential feature under-reporting). In this work, we study an analytically tractable model of differential feature under-reporting to characterizethe impact of under-report on algorithmic fairness. We demonstrate how standard missing data methods typically fail to mitigate bias in this setting, and propose a new set of augmented loss and imputation methods. Our results show that, in real world data settings, under-reporting typically exacerbates disparities. The proposed solution methods show some success in mitigating disparities attributable to feature under-reporting.

1 INTRODUCTION

Regional and local governments around the world are using their increasingly digitized data systems to develop AI-driven decision-support technologies. The hope is that these tools improve decision quality, reduce inefficiencies, eliminate fraud, and improve outcomes for their citizens [25, 42]. Often, these technologies take the form of predictive risk models that are trained on administrative

---

*Work done while at Carnegie Mellon University.

data to assess the likelihood that a case will go on to have poor outcomes. Such models have been developed and deployed in criminal justice [9], child welfare [57], welfare fraud detection [58], federal tax audits [11, 37], homelessness services [41], health care [44], and many other settings.

Predictive risk models in the public sector have come under criticism over concerns that they are trained on biased data [10, 17, 43, 53]. In this paper, we consider a specific form of bias. We use the term 'differential feature under-reporting' to describe the phenomenon whereby administrative data records are more complete for individuals who have more greatly relied on public services. In the United States, for instance, administrative records often contain medical claims data for those who receive services through public insurance programs (Medicaid / Medicare), but lack information on physical, mental and behavioral healthcare utilization for the privately insured. A lack of *recorded* medical claims for an individual in this context is often indistinguishable from instances in which no medical claims have been made. In her critique of the Allegheny Family Screening Tool (AFST) used in screening child maltreatment referrals, Eubanks [26] writes, "by relying on data that is only collected on families using public resources, the AFST unfairly targets low-income families for child welfare scrutiny."

We provide a technical analysis of this problem. First, we introduce a statistical model of data collection with differential feature under-reporting. We then present theoretical results that characterize the impact of under-reporting on disparities in selection rates across groups. We describe why standard missing data methods generally fail to mitigate unfairness and, instead, propose new methods based on augmented loss estimation and optimal prediction imputation that are tailored to the under-reporting setting. Lastly, we present empirical results on semi-synthetic and real world data. Our results show that, while in theory under-reporting can decrease disparities, in practice, under-reporting usually leads to increasing disparities and our proposed mitigation methods alleviate this increase.

2 BACKGROUND AND RELATED WORK

Under-reporting vs. missingness. The problem of differential feature under-reporting is illustrated in Figure 1. An individual's risk prediction $\hat{Y}$ is formed based on observed administrative data features X which are a mismeasured version of a "true" latent feature vector Z. We assume that certain features in Z, such as demographic information, are correctly observed, whereas others, such as use of mental health services, are only correct for individuals who rely on publicly funded services. Problematically, we generally lack indicators on who is privately or publicly funded and for which services. For indicators and count features, e.g. the number of episodes in

(a) General graph

G: High vs. low income group
Z: Number of doctor visits in the past year
ξ: Publicly insured ($\xi = 1$) or privately insured ($\xi = 0$)
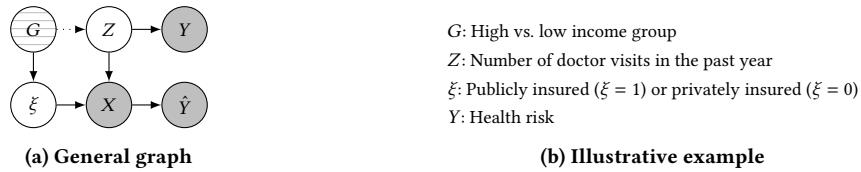Y: Health risk

(b) Illustrative example

Figure 1: We study a prediction model on feature vectors with differential under-reporting X where true outcomes Y are a function of the latent 'true' features Z. Missingness ξ is influenced by group membership G. We consider both cases in which feature distributions vary by group membership and cases with $G \perp Z$. In our setting, missingness indicators ξ are unobserved and group membership G is only used for model evaluation and not as a feature. The graph reflects the dependencies at prediction time.

inpatient mental health treatment in the past year, the mismeasured feature will simply show the value 0 for individuals who received those services through privately funded mechanisms. This means that when we observe $X_j = 0$, we do not know whether $Z_j = 0$, or if $Z_j \neq 0$ and the feature has been mismeasured. In the graph, the *unobserved* missingness indicators are denoted by ξ. This distinguishes the under-reporting setting from standard missingness, wherein ξ is assumed to be fully observed.

Missing data literature distinguishes three types of mechanisms: (1) Missing Completely At Random (MCAR) where missing values are independent of both observed and unobserved data, (2) Missing At Random (MAR) where missingness depends on observed variables, and (3) Missing Not At Random (MNAR) where missing values depend on unobserved data [54]. In administrative data, records are more available for individuals who rely more greatly on public services which often correlates with demographic attributes excluded from modeling. This implies an MNAR setting.

Under-reporting in real-world applications. The problem of feature under-reporting extends beyond the administrative data context. In health applications, Electronic Health Record (EHR) data is often under-reported at different levels for different population sub-groups [19, 33, 35, 51, 60]. Socioeconomically disadvantaged patients may be missing more diagnostic tests due to limited health care access [2, 7]. Reliance on clinical decision support systems trained on EHR data could exacerbate already existing health care disparities [15, 33, 38]. Similarly, the extent of under-reporting often varies across domains (e.g. hospitals) which has been studied by Zhou et al. [62]. While Zhou et al. [62] consider model adaptation when shifting to unlabeled target data with different level of under-reporting, we focus on a single domain with varying levels of under-reporting across groups and study fairness implications.

Under-reporting in biomedical research. In epidemiological surveys, social stigma can lead participants to provide false negative responses (e.g. true maternal smoking status Z vs. reported status X) [34, 40, 45, 55]. Researchers have proposed various methods to estimate association between Z and outcome Y by leveraging X including correction factors for independence tests [13, 55], adjusted mutual information [55], odds-ratios [18, 22, 23], risk-ratios [12, 50], and full likelihood approaches [1]. These methods are typically limited to binary features and outcomes, and focus on inferring the relationship between Y and some Z rather than on prediction. In the study of geographical disease counts, [8, 20, 32, 56] Bayesian

methods have been used to make inference in the under-reporting setting of $X \leq Z$. Such methods require a host of parametric and distributional assumptions as well as informative priors. In single-cell RNA sequencing, under-report arises as 'zero-inflation,' which refers to genes going undetected despite being expressed in a cell due to low levels of RNA. Methods correcting for zero-inflation are often Bayesian and highly specialized for the single-cell RNA sequencing task. Note that even if some of the above approaches were applicable in our setting to learn a correctly specified model $f(z) = \mathbb{E}\left[Y \mid Z = z\right]$ from observations of X and Y, it is unclear how to use such models *for prediction* when only the under-reported features X are available at prediction time.

Additive noise and fairness. The algorithmic fairness literature has studied various types of feature mismeasurement as summarized in Table 1. A commonly studied setting is additive feature noise where, instead of a feature z_1, we observe a noisy version $x_1 = z_1 + \varepsilon$. The random noise ε is often assumed to be zero-mean, of small variance, and independent of other variables. This implies that, while some of the information in the feature is diluted, large portions of the encoded information remains intact. Khani and Liang [39] show that adding the same amount of feature noise to a group-blind model can introduce statistical loss discrepancy. This is in line with earlier observations from the statistical discrimination literature [3, 48]. Chen et al. [16] propose data collection strategies targeted at decreasing discrepancy and come to the conclusion that overcoming differential noise across protected groups may require collection of additional data. In contrast to additive noise, under-reporting removes all information from impacted feature entries and typically biases the feature mean. Some works [e.g. 39] suggest that feature missingness can be modeled as a special case of additive noise by selecting noise terms with very high variance. However, this only covers a special case of under-reporting in which feature entries are missing for *all* observations.

Missing data methods and fairness. Feature missingness has been studied in the statistical literature for several decades [e.g. 54]. This line of work generally assumes that we observe missingness indicators or, equivalently, missing values are clearly marked (e.g. NaN). Various methods with different fairness implications have been proposed (Table 1).

(1) *Complete case analysis and reweighing.* In some cases, it may be desirable to remove incomplete rows which can lead to significant biases [54]. Various reweighing procedures have

	Complete data				Additive noise				Missing with indicator					Under-reporting			
	g	z_1	z_2	y	g	x_1	z_2	y	g	x_1	r	z_2	y	g	x_1	z_2	y
Setting	0	**10**	2	1	0	**10.2**	2	1	0	**10**	1	2	1	0	**10**	2	1
	0	**7**	1	0	0	**6.5**	1	0	0	m	0	1	0	0	m	1	0
	1	**0**	3	1	1	**0.8**	3	1	1	m	0	3	1	1	m	3	1
	Features fully observed				Feature values with added random noise ε				Some feature values take default value m; r indicates which values are observed					Some feature values take default value m; No indicators for missingness			
Previous fairness work	No feature mismeasurement				Khani and Liang [39], Phelps [48], Aigner and Cain [3], Chen et al. [16]				Zhang and Long [61], Wang and Singh [59], Jeanselme et al. [38], Fernando et al. [27], Fricke [30], Ahmad et al. [2]					**This work**, Eubanks [26]			

Table 1: Different types of feature mismeasurement and previous work addressing fairness implications. In the data examples, mismeasured features are denoted with x while correctly observed features are denoted by z. Column g encodes group membership.

been proposed to deal with this problem. Zhang and Long [61] suggest learning only from complete observations while employing an importance sampling procedure. Wang and Singh [59] suggest that reweighing and resampling methods in the context of categorical data can lead to considerable fairness improvements over learning with missing data directly.

(2) *Imputation.* Jeanselme et al. [38] compare different imputation strategies under clinical presence and find that there is no imputation strategy that reliably outperforms other imputation methods in terms of fairness. Fernando et al. [27] and Fricke [30] compare feature imputation to complete case analysis and find that rows with missing values can contribute to fairer outcomes via observed columns.

In contrast to the feature missingness setting, we do not observe indicators for missing entries in this work. Despite this difficulty, we experiment with row omission and imputation in Section 7.

3 PROBLEM SETUP

Setting. We study the effect of feature under-reporting on algorithmic fairness through the lens of the regression setting displayed in Figure 1. Assume latent feature vectors $z \in \mathbb{R}^d$ and group information $g \in \{0, 1\}$. We assume a noiseless regression setting in which the outcome y is a linear function of z, i.e. $y = \alpha + \beta^T z$ with $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}^d_{\neq 0}$. Instead of the true features z, we observe a mismeasured vector x in which entries default to 0 with group-dependent probabilities. That is, we set $x = z \odot \xi^g$ where $\odot$ denotes element-wise multiplication, $\xi^g \sim \text{Bern}(m^g)$, and $(1 - m^0), (1 - m^1) \in (0, 1]^d$ are under-reporting rates in the two groups. More formally, we have a group random variable $G \sim \text{Bern}(r)$ and a random feature vector Z. The vector of under-reported features X can be written as $X = Z \odot \xi$ where $\xi = G\xi^1 + (1 - G)\xi^0$. This setting allows for different dependence structures depending on whether we assume $G \perp Z$.

Two-step bias. Differential under-reporting introduces bias in two ways: (1) Under-reporting in training data influences estimation of the prediction model *(estimation step)*, and (2) input data with under-reporting leads to biased predictions at test time *(prediction step)*. It is generally not sufficient to recover the true model parameters α, β as only biased features are available at prediction time. In fact, our experiments in Section 8 demonstrate that using

true parameters for prediction can lead to worse fairness outcomes than using a model estimated with biased data.

Thresholded prediction. We assume a thresholded prediction setting reminiscent of predictive risk modeling in the public sector. A predictor f is fit on (X, Y) to produce predictions $\hat{Y} = f(X) = f(Z \odot \xi^G)$. We consider group-wise shares of predictions above a given threshold $\tilde{y}$: $P(\hat{Y} \geq \tilde{y} \mid G = g)$ which we refer to as *selection rates* at threshold $\tilde{y}$. This implicitly assumes a setting in which the *highest* risk individuals are selected (e.g. child welfare screenings, fraud detection, federal tax audits). However, it is straightforward to reverse the analysis for scenarios in which *low* risk leads to selection (e.g. bail decisions in criminal risk assessment). In addition, we assume that being selected is *undesirable*. Crucially, this assumption is only made to simplify interpretation and we could easily consider the opposite case.

Excess selection rates. We assume that the threshold on predictions $\hat{Y}$ is set to achieve a desired overall *selection rate* $P(\hat{Y} \geq \tilde{y}) = C \in [0, 1]$. Given the cumulative distribution function of predictions $F_{\hat{Y}}$, the percentile threshold C implies an absolute threshold $\tilde{y} = F_{\hat{Y}}^{-1}(1 - C)$ such that the selection rate for group g can be written as $P(\hat{Y} \geq \tilde{y} \mid G = g)$. In order to isolate the effect of under-reporting, we need to account for a ground truth difference in selection rates. Let $\hat{Y}_X$ denote the predictions of a model trained on (X, Y) and $\hat{Y}_Z$ the predictions of a model trained on (Z, Y). When distributions of $\hat{Y}_Z$ and $\hat{Y}_X$ differ, the predictions imply different thresholds $y' = F_{\hat{Y}_Z}^{-1}(1 - C)$ and $\tilde{y} = F_{\hat{Y}_X}^{-1}(1 - C)$. With this notation, we define a metric for impact of differential feature under-reporting on disparities in selection rates.

DEFINITION 1 (EXCESS SELECTION RATE DUE TO UNDER-REPORTING). *The excess selection rate for group $g \in \{0, 1\}$ at overall selection rate $C \in [0, 1]$*

$$\Delta(g, C) := P(\hat{Y}_X \geq \tilde{y} \mid G = g) - P(\hat{Y}_Z \geq y' \mid G = g),$$

is the difference in selection rates when ranking according to a model trained on X compared to a model trained on Z. We say that group g is over-selected *at level C if $\Delta(g, C) > 0$. If $\Delta(g, C) < 0$, we say that g is* under-selected.

In principle, we could directly consider a "difference in difference": the difference in selection rates between groups $g = 0, 1$ when selection occurs according to the model $\hat{Y}_X$ versus the unbiased predictions $\hat{Y}_Z$. However, since we select a fixed share of

the population C, an increase of the selection rate of group g when moving from $\hat{Y}_Z$ to $\hat{Y}_X$ already implies a decrease for group $1 - g$. It is generally difficult to argue about the excess selection rate $\Delta(g, C)$ analytically. Even in a simple setting with group-dependent Gaussian features $Z \mid G \sim \mathcal{N}(\mu_G, \Sigma_G)$, there is no closed-form expression for the quantile $y' = F^{-1}_{\hat{Y}_Z}(1 - C)$ and determining the sign of $\Delta(g, C)$ requires analysis of a difference in cdfs which is often intractable. Instead, we simplify the setting and assume that Z follows the same distribution across groups. In this case, the selection rates on the true outcome Y are the same in both groups at every threshold, and we can simplify.

DEFINITION 2 (EXCESS SELECTION RATE DUE TO UNDER-REPORT-ING, INDEPENDENT CASE). *If $Z \perp G$, we say that group $g \in \{0, 1\}$ is* over-selected *at threshold $C \in [0, 1]$ if $P(\hat{Y}_X \geq \tilde{y} \mid G = g) > P(\hat{Y}_X \geq \tilde{y} \mid G = 1 - g)$, and* under-selected *if the inequality is reversed.*

While the majority of our theoretical derivations assume the special case of $Z \perp G$, the empirical portion of this work explores the impact of feature under-reporting in the more general setting. To clarify which assumptions are sufficient for which finding in the paper, we supply a summary table in Appendix A, alongside descriptions in the main text.

4 DIFFERENTIAL FEATURE UNDER-REPORTING IN LINEAR REGRESSION

In this section, we examine the bias that differential feature under-reporting introduces into regression parameter estimates. We consider a setting in which true outcomes are a linear function of latent features, i.e. $Y = \alpha + \beta^T Z$, which implies that a linear model with access to the true Z recovers the true outcomes Y. Dropping subscripts, we write $Y = \hat{Y}_Z$ and $\hat{Y} = \hat{Y}_X$. Note that this section focuses on population-level regression.

Estimates and attenuation bias. Feature mismeasurement in the form of under-reporting leads to inconsistent parameter estimates in linear regression. When fitting a linear model on (X, Y), the least squares estimates become

$$\hat{\beta} = \Sigma_X^{-1} \Sigma_{XZ} \beta, \quad \hat{\alpha} = \alpha + \mathbb{E}[Z]^T \beta - \mathbb{E}[X]^T \hat{\beta}, \quad (1)$$

where Σ_{XZ} denotes the covariance matrix between X and Z and we write Σ_X for Σ_{XX}. At first glace, this solution resembles the regression estimates in the more commonly studied additive feature noise case. Assuming $X' = Z + U$ where U is independent zero-mean feature noise, we obtain $\hat{\beta} = \Sigma_{X'}^{-1} \Sigma_{X'Z} \beta = (\Sigma_Z + \Sigma_U)^{-1} \Sigma_Z \beta$. The factor $\lambda = (\Sigma_Z + \Sigma_U)^{-1} \Sigma_Z$ is commonly interpreted as a noise-to-signal ratio and, if Z is one-dimensional, we know that $|\hat{\beta}| = \lambda |\beta| < |\beta|$ which is generally referred to as attenuation bias [e.g. 31, 36]. In the under-reporting setting, Σ_X does not easily separate into terms depending on only the feature or only the mismeasurement. However, in the special case of one-dimensional Z and $Z \perp G$, we can still show that the parameter $\hat{\beta}$ is biased towards zero.

LEMMA 3 (ATTENUATION BIAS). *Assume the feature Z is one-dimensional and has the same distribution across groups. Then, the*

least squares regression of Y on the mismeasured feature X yields an estimated slope $\hat{\beta}$ with $|\hat{\beta}| \leq |\beta|$.

The d-dimensional case. Real-world prediction settings typically include multiple, often correlated, features. Assume the feature vector Z is d-dimensional, and under-reporting only occurs in the first feature, Z_1. This means that X coincides with Z in all but the first entry which is computed as $X_1 = Z_1 \xi_1$ where $\xi_1 = G \xi_1^1 + (1 - G) \xi_1^0$ and $\xi_1^1 \sim \text{Bern}(m_1^0)$, $\xi_1^0 \sim \text{Bern}(m_1^1)$. We further assume that features $Z_2, \ldots, Z_d$ are uncorrelated and the feature dependence structure is characterized entirely by the correlations between Z_1 and $Z_{[2:d]}$. This assumption is without loss of generality because we can always apply orthogonalization to the features. We explicitly exclude cases in which Z_1 is a perfect linear combination of other features to avoid problems of multicollinearity and assume $\mathbb{V}[Z_k] > 0$ for all $k \in [2 : d]$. Given this setting, Proposition B.1 provides a closed form representation for the parameter estimates $\hat{\beta}$. The bias in these estimates can be conceptualized as a generalization of omitted variable bias [5] which is discussed Appendix C. We now assume that $Z \perp G$ which implies that under-reporting is independent of the value that is under-reported. This resembles the assumptions made in previous work on the impact of additive feature noise on fairness [e.g. 3, 39, 48], and allows us to gain analytical insights that would otherwise remain intractable. First, we examine the behavior of the parameter estimate for the feature with under-reporting, i.e. $\hat{\beta}_1$.

PROPOSITION 4 (PROPERTIES OF $\hat{\beta}_1$). *If $Z \perp G$, the parameter estimate $\hat{\beta}_1$ has the following properties.*

(1) ***Sign invariance:*** *$\hat{\beta}_1$ has the same sign as β_1.*
(2) ***Attenuation bias:*** *$|\hat{\beta}_1| \leq |\beta_1|$.*
(3) ***Attenuation bias increasing with under-reporting:*** *If under-reporting $1 - m_1^g$ is increasing for one (or both) groups $g \in \{0, 1\}$, ceteris paribus, the magnitude of the parameter estimate, $|\hat{\beta}_1|$, is decreasing.*

This finding shows that feature under-reporting leads to attenuation bias in the respective parameter estimate even when other correlated and fully observed features are available. The attenuation bias gets more pronounced with more under-reporting. Next, we turn towards the estimates for the fully observed features $\hat{\beta}_k$ for $k \in [2 : k]$.

PROPOSITION 5 (PROPERTIES OF $\hat{\beta}_k$). *If $Z \perp G$, the parameter estimates $\hat{\beta}_k$ for $k \in [2 : d]$ have the following properties.*

(1) ***Correlation bias:*** *If $\hat{\beta}_k \neq \beta_k$, then $\rho(Z_1, Z_k) > 0$.*
(2) ***Shifting weight:*** *If under-reporting $1 - m_1^g$ is increasing for one (or both) groups $g \in \{0, 1\}$, ceteris paribus, $\hat{\beta}_k$ is increasing if $\text{sign}(\beta_1 \text{Cov}[Z_1, Z_k]) = +1$, and decreasing if $\text{sign}(\beta_1 \text{Cov}[Z_1, Z_k]) = -1$.*

In line with general intuition, under-reporting in Z_1 has no effect on the parameter estimate $\hat{\beta}_k$ if features Z_k and Z_1 are uncorrelated. If the features are correlated, the direction of the under-reporting effect on the parameter estimate depends on the signs of β_1 and $\text{Cov}[Z_1, Z_k]$. Note that this is independent of the value and sign of β_k.

Take-away. Proposition 4 and 5 tell a compelling story about the effect of under-reporting on parameter estimates in the studied setting. As more feature values default, the regression model places

less weight on the mismeasured feature and instead shifts weight to fully observed features with non-zero correlation. This can lead to increasing or decreasing parameter estimates. Analytically, there are cases in which 'shifting weight' means that the magnitude of a parameter estimate, $|\hat{\beta}_k|$, is decreasing, which may appear counterintuitive. For example, consider a setting in which both Z_1 and Z_k have positive true parameters but negative correlation, i.e. $\beta_1, \beta_k > 0$ and $Cov[Z_1, Z_k] < 0$. In practice, this could occur when there are several mutually exclusive paths to the same outcome. For example, consider prediction of general health risk scores with features including both the number of pediatrician visits in the last year and the number of internist visits. Presumably, these features are negatively correlated because they are relevant for two mutually exclusive parts of the population, i.e. children and adults, but in both columns larger values can be indicative of a high general health risk.

5 IMPACT ON SELECTION RATE DISPARITY

Selection rate disparity in Gaussian setting. We study the effect of differential feature under-reporting on selection rate disparities in linear regression. Similar to our previous discussion, we assume a d-dimensional feature setting in which only the first feature is subject to under-reporting and $Z \perp G$. We further assume that features are jointly Gaussian, i.e. $Z \sim \mathcal{N}(\mu, \Sigma)$ where $\mu \in \mathbb{R}^d$ and the covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ is positive definite. This has the benefit that predictions $\hat{Y} = \hat{Y}_X$ follow a Gaussian mixture distribution which allows us to directly analyze group selection rates. If under-reporting rates are the same across groups, there is no selection rate disparity as both groups have the same feature distributions. If the under-reporting rates vary between groups, we observe the following.

PROPOSITION 6. *For a sufficiently high threshold $\tilde{y}$, the group with more under-reporting is over-selected if*

$$\mathbb{V}\left[\hat{\alpha} + \hat{\beta}_{[2:d]} Z_{[2:d]}\right] > \mathbb{V}\left[\hat{\alpha} + \hat{\beta} Z\right]$$

(Case 1), or under-selected if the inequality is reversed (Case 2). For low thresholds, the cases are reversed.

We refer to Proposition B.2 for an expanded version of this finding including a discussion of sufficiently high thresholds. Proposition 6 shows that over-selection primarily depends on variance in predictions. When cutting off at a high threshold, the group with more feature under-reporting is over-selected if the variance in predictions for examples with defaulted feature exceeds the prediction variance for fully observed examples (Case 1). It is under-selected if the variance of predictions is larger for the examples with fully observed features (Case 2). Intuitively, at high thresholds, information deficiency in a group always leads to under-selection because it prompts the group's risk distribution to concentrate more closely around its mean moving more mass below the threshold. We find that, analytically, outcome disparities can go into either direction and sometimes groups with information deficiency are over-selected. While our findings suggest that this is mostly a question of variance in predictions, this is likely only part of the story in settings with group-dependent feature distributions. We study more general settings empirically in Section 7.

Combining parameter estimation and prediction steps. Feature under-reporting introduces bias both at estimation and prediction time. In the following, we combine our previous findings to examine the conditions under which under-reporting leads to over-selection and under-selection. As before, we assume a d-dimensional feature setting in which only the first feature Z_1 is impacted by under-reporting. Features are jointly Gaussian, and we further assume that $Z_2, \ldots, Z_d$ are uncorrelated.

COROLLARY 7. *Given the first and second moments of Z_1, the expected share of observed values $\mathbb{E}[\xi_1]$, and the fraction of variance in Z_1 that is explained by the remaining features $S^2 = \sum_{i=2}^d \rho(Z_1, Z_i)^2$, there exists a positive constant $c = c(\mathbb{E}[Z_1], \mathbb{V}[Z_1], \mathbb{E}[\xi_1], S^2)$ such that, at high thresholds, the group with more under-reporting is over-selected if*

$$\frac{1}{\beta_1} \sum_{j=2}^d \beta_j Cov[Z_1, Z_j] < -c,$$

(Case 1), and under-selected if the inequality is reversed (Case 2).

Thresholds are considered high if they exceed the turning point defined in Proposition B.2. The corollary shows that over-selection due to feature under-reporting depends on the signs and magnitudes of the true parameters β and the covariances between features. If $\text{sign}\left(\frac{1}{\beta_1} \sum_{j=2}^d \beta_j Cov[Z_1, Z_j]\right) = 1$, e.g. if all true parameters and covariances are non-negative, the group with more missingness will always be under-selected at high thresholds. If the sign is negative, the group with more under-reporting is over-selected if the covariance-weighted sum of true parameters is sufficiently large in absolute value. Otherwise, feature under-reporting still leads to under-selection.

6 SOLUTION APPROACHES

Sections 4 and 5 show how ignoring differential feature under-reporting can lead to disparities in selection rates across groups. In the following, we explore how conventional missing data methods can be adapted to the under-reporting setting. We then propose a new set of methods that is specifically tailored to this setting by separating the problem into two steps—estimation and prediction. For the estimation step, we provide a method that recovers the ground truth data generating model from observed data. For the prediction step, we derive optimal group-dependent imputation values. As before, we assume under-reporting occurs only in the first feature which is observed as $X_1 = Z_1 \xi_1$.

Standard missing data methods. Existing missing data methods typically assume that defaulted values are clearly marked which is not the case in the under-reporting setting. We explore adaptations of several methods.

(1) *Feature omission.* Discarding the mismeasured feature vector X_1 doesn't require missingness indicators and mitigates the bias introduced by under-reporting. However, this approach may decrease model performance significantly, and may itself introduce bias. When assuming a linear ground truth, feature omission leads to omitted variable bias in parameter estimates $\hat{\beta}_{[2:d]}$, which has been studied previously in the econometrics literature [5] (also see Appendix C).

(2) *Multiple imputation.* Multiple imputation draws plausible feature values while retaining variability. Since we do not observe indicators for missingness, we experiment with imputing all 0-entries in X_1 which includes correctly observed 0s. In each imputation run, we estimate the posterior $P(Z_1 \mid Z_2, \ldots, Z_d)$ on data rows with $X_1 \neq 0$, impute 0-entries, and train a prediction model for Y. At prediction time, we average the imputations and predictions over models to obtain a single prediction $\hat{Y}$. While this procedure successfully alleviates bias in some standard feature missingness settings, it is not a priori clear how well the method works with under-reporting.

(3) *Row omission.* Omission of rows with missing feature entries provides a convenient complete case analysis. Since true and false 0-entries are indistinguishable in our setting, we experiment with discarding all rows with $X_1 = 0$. If there is no model misspecification, e.g. in the linear case with $f(Z) = \alpha + \beta^T Z$ where we train a linear model on observed features, training on only complete rows is guaranteed to asymptotically retrieve the true parameters $\hat{\alpha} = \alpha$ and $\hat{\beta} = \beta$ if Z_1 is not binary. Even with access to the ground truth model, under-reporting introduces bias via the prediction step and $\hat{Y} = f(X)$ may not be the most accurate (or fairest) prediction.

Model estimation with augmented loss. Without model misspecification, row omission can recover the ground truth model $Y = f(Z)$. In practice, models are usually misspecified and discarding rows can significantly decrease performance. Instead, we propose an augmented loss function to recover the ground truth model. This proxy loss uses observed features X to provide an unbiased estimate of the loss of a model f on latent features Z. Similar approaches have previously been used in the label noise setting [46, 47]. Assume $Z \in \mathbb{R}^d$ has support $\mathcal{Z}$ and $y \in \mathbb{R}$ has support $\mathcal{Y}$. Let $\mathcal{F} : \mathcal{Z} \to \mathbb{R}$ be a class of real-valued functions and $l : \mathcal{F} \times \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$ be a bounded loss function. We assume $Z \perp G$ and denote the rate of observed values as $m_1 := \mathbb{E}[\xi_1] = r m_1^1 + (1 - r) m_1^0$.

LEMMA 8 (AUGMENTED LOSS). *Assume fixed $f \in \mathcal{F}, z \in \mathcal{Z}, y \in \mathcal{Y}$ and $X \in \mathbb{R}^d$ defined by $X_1 = Z_1 \xi_1$ and $X_{[2:d]} = z_{[2:d]}$. Define*
$$\tilde{l}(f, X, y) := \frac{1}{m_1} l(f, X, y) - \frac{1 - m_1^1}{m_1} l(f, [0, X_{[2:d]}]^T, y).$$

If $Z \perp G$, we have that $\mathbb{E}_{\xi_1}\left[\tilde{l}(f, X, y)\right] = l(f, z, y)$.

The fact that the augmented loss is unbiased with respect to under-reporting noise implies that a prediction model on observed data estimated with augmented loss, i.e. $\hat{f} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(X,Y)}\left[\tilde{l}(f, X, Y)\right]$, asymptotically recovers the Bayes optimal model on the true features Z. If $Y = \alpha + \beta^T Z$, $\mathcal{F}$ is the class of linear functions $f : \mathbb{R}^d \to \mathbb{R}$ and $l(f, z, y) = (f(z) - y)^2$ denotes squared error loss, the true parameters $\hat{\alpha} = \alpha$ and $\hat{\beta} = \beta$ are retrieved. Note that squared error loss is not bounded and estimating $\hat{f}$ requires the additional constraint $\tilde{l}(f, X, Y) \geq 0$. Lemma 8 operates in a group-agnostic setting with $Z \perp G$. Lemma B.3 provides a more general group-dependent version of the finding.

Optimal prediction imputation value. Assume we are in the linear case with $Y = \alpha + \beta^T Z$ and we have access to the true parameters

α and β, e.g. obtained via augmented loss. What is the best possible prediction for an example of the form $x = [0, z_2, \ldots, z_d]$? Since $x_1 = 0$ could mean $z_1 = 0$ or the entry is missing, it is intuitive that $\hat{y} = \alpha + \beta^T x$ does not minimize expected prediction error. Instead, we derive the optimal fixed prediction imputation value $x_1'^*$.

LEMMA 9 (OPTIMAL PREDICTION IMPUTATION VALUE). *Assume $Z \perp G$, $f(Z) = \alpha + \beta^T Z$ is the ground truth model and X the random vector of observed features. We set*
$$X' = \begin{cases} X \ if \ X_1 \neq 0, \\ [x_1', X_{[2:d]}] \ if \ X_1 = 0, \end{cases}$$
where x_1' is fixed. Then, $x_1'^* := \arg\min_{x_1'} \mathbb{E}_X[(f(X') - Y)^2] = \mathbb{E}[Z_1 \mid X_1 = 0]$ is the optimal prediction imputation value.

The Lemma shows that the loss-minimizing constant imputation value is the conditional mean Z_1 given the observed value is 0. This implies that, in alignment with earlier intuition, directly predicting with the observed $x = [0, z_2, \ldots, z_d]$ is sub-optimal in the under-reporting setting. The optimal value $x_1'^*$ in the setting of Lemma 9 can be written as
$$\mathbb{E}[Z_1 \mid X_1 = 0] = \frac{\frac{1}{m_1} \mathbb{E}[X_1] - P(X_1 \neq 0) \mathbb{E}[X_1 \mid X_1 \neq 0]}{P(X_1 = 0)},$$

which can be estimated directly from observed data if the under-reporting rate $1 - m$ is known. If feature distributions vary across groups, group-dependent optimal prediction imputation values can be derived as described in Lemma B.4.

Under-reporting rate estimation. Both augmented loss estimation and optimal prediction imputation require access to the reporting rate, m, which is typically unknown. In some cases, it may be possible to obtain supplementary data that can be used to estimate m. For example, in administrative data with under-reported health features for privately insured individuals, an external private insurance health claims dataset could be used to estimate the expected rate of true 0's. In most settings, estimation of under-reporting rates needs to rely directly on the observed data. Assume we have access to a dataset $V = \{(x, y)\}_{i=1}^n$. We split V into a training portion V_{train} and evaluation portion V_{eval}. Let P_{eval} denote the subset of examples from V_{eval} for which $x_1 \neq 0$. We draw on the literature on Positive and Unlabeled (PU) learning [24] and estimate under-reporting rates as follows. First, we fit a model h on V_{train} to estimate $P(X_1 \neq 0 \mid X_{[2:d]}, Y)$. Second, we evaluate h on P_{eval}. The estimator for the share of observed values m is given by $\hat{m} = \frac{1}{|P_{\text{eval}}|} \sum_{(x,y) \in P_{\text{eval}}} h(x_{[2:d]}, y)$. The estimation procedure assumes that under-reporting occurs completely at random. Our experimental setting assumes under-reporting completely at random within groups and thus m_0 and m_1 can be estimated with the described procedure by restricting V to examples from the respective group. For more details on the estimation procedure, we refer to Appendix F.

7 EXPERIMENTS
7.1 Publicly available datasets

Data. Both COMPAS data [6] and German credit data [52] are widely used across the algorithmic fairness literature. The American Community Survey (ACS) Income dataset is comprised of 2018

census data from California [21]. Datasets vary in size, number of features, and prediction tasks as shown in Table E.2. We conduct all experiments with both gender and race as group columns if available. All models are group-blind and race information is never included as predictive feature. Results for the German credit dataset are discussed in Appendix G.1.

Semi-synthetic outcomes. Since all of the prediction tasks are binary classification, we opt to generate semi-synthetic regression labels for our experiments. We first fit a logistic regression model to the entire dataset and extract the fitted probabilities. For the ACS Income data, the values are rescaled to center around the $50,000 income threshold. Next, we fit a linear regression model using the same features and the predicted probabilities as outputs. The fitted values from this linear model are chosen as the new "true" labels for our experiment. This outcome augmentation procedure allows us to generate artificial settings with a truly linear ground truth similar to the settings studied in Sections 4 and 5 while leaving realistic covariance structures intact. We further experiment with controlling the R^2 of the true linear model by adding additional noise to outcomes and report fairness implications in Appendix G.2.

Experiment stratification. Regression models are trained to predict semi-synthetic outcomes based on the features of the respective dataset. We select the top C share of the predictions as high risk and evaluate excess selection rates to assess the fairness impact of under-reporting. Artificial under-reporting is added to one feature column at a time and we repeat the experiments for each outcome column, group column, and under-reporting rate. Under-reporting rates range from 0-90% in 10 percentage point increments, and we add under-reporting to only one group at a time (e.g., we set 10% of a feature in the male group to 0 while leaving the features of the female group unchanged). Only numeric features are considered for under-reporting since, in administrative data, binary features are often categorical dummies or thresholded versions of continuous count features. All models are trained with 80% of the datasets while withholding 20% for testing. We experiment with various solution approaches as described in Section 6. This includes our proposed methods of group-dependent augmented loss estimation and group-dependent optimal prediction imputation.

7.2 County-level birth data

Data. We present an analysis of a private administrative dataset we obtained from a county in the US. The dataset contains information on newborn children and their families including demographics, child protective services history, birth record data, and mental and behavioral health information for those who used publicly funded services. We set up a prediction task that attempts to mimic the analysis described in the Hello Baby model methodology report from Allegheny County [29]. The Hello Baby model was developed to predict which families are at greatest risk of having their child removed by Child Protective Services (CPS) during their first three years of life, and is used to prioritize families with newborn children for opt-in, voluntary supportive services. Using our data, we train a similar model, and explore the effect of adding additional under-reporting to the behavioral and mental health data fields.

Experiment setup. We use the birth dataset with its original prediction outcomes to showcase a realistic example of the effect of under-reporting. As before, the data is separated into 80% for training and 20% for testing. We fit separate logistic regression models on three datasets. (1) Data as observed. (2) Data with behavioral health features set to 0 for privately insured individuals, i.e. mothers that are not insured through Medicaid. (3) Data without behavioral health features. For illustration, results are stratified by whether individuals are covered by Medicaid, and by whether the mother's race is recorded as Black. Medicaid coverage and race are not used as features in any of the models.

8 RESULTS

ACS Income data. Figure 2 summarizes the results for the experiments on the ACS Income data. We see that feature under-reporting in 'education attainment' and 'hours worked per week' consistently leads to under-selection of the group with under-reporting. This is true irrespective of whether feature under-reporting is injected into the female sub-group or the male sub-group, and we observe the same effect when under-reporting is added based on the individuals' racial group. The figure additionally suggests that more under-reporting generally leads to increasing under-selection. Intuitively, it makes sense that both education attainment and hours worked per week contribute positively to predicted income which is confirmed by the parameter estimates (Figure E.3). Exploration of the covariance matrix of the unbiased features further reveals that all numeric columns in the dataset are positively correlated which together creates a setting reminiscent of the Case 2 scenario studied in Section 5. At a high-level, our theoretical analysis predicts that the group with more under-reporting is under-selected in this setting which aligns with our observations. In addition to selection rate disparity, feature under-reporting in the data also leads to decreased model accuracy as displayed in Figure E.2, and the parameter estimates in Figure E.3 display an attenuation effect as predicted in Section 4.

COMPAS data. We focus on results for under-reporting in count features (Figure E.1) and point to Appendix G.3 for additional results. The feature 'priors count', i.e. the number of previous criminal offenses individuals have been convicted of, emerges as important feature with respect to under-reporting. Under-reporting in priors count leads to under-selection of the impacted group. This pattern repeats itself for any of the groups and both of the available prediction outcomes. The more feature under-reporting in a group, the larger the occurring outcome disparity. Similarly to the previous results, this suggests a setting of Case 2 as discussed in Section 5. As before, parameter estimates suggest an attenuation effect which is displayed in Figure 3. Under-reporting in priors count could be interpreted as an extreme case of crimes that do not result in arrest. Assuming that one demographic group is more likely to be convicted for committed crimes than the other group, the result implies that the already more frequently targeted group may additionally be flagged as high risk for recidivism at disproportionate rates. Racial disparities in arrest rates and police encounters are well-documented in the US [e.g. 4, 14, 28, 49] which highlights the importance of this finding.
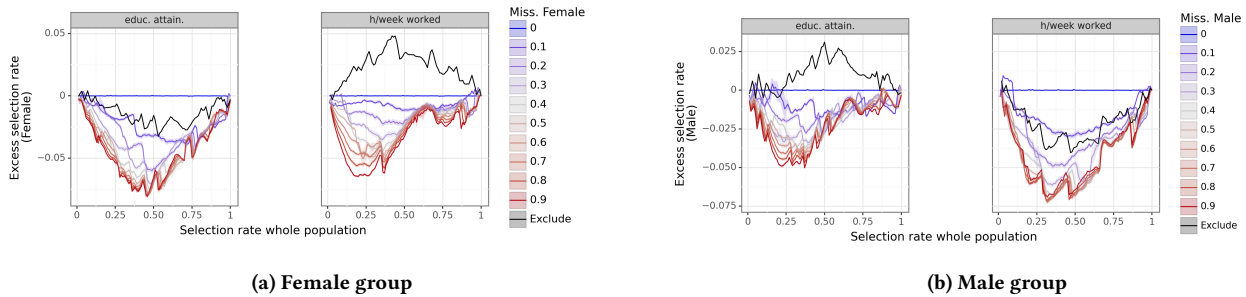
(a) Female group

(b) Male group

Figure 2: Group-wise excess selection rates using the ACS Income dataset. Each panel represents a feature that has been corrupted by under-reporting in independent runs of the experiment. Black curves show performance when omitting the entire feature column. Results are averaged over 50 runs on the test set. Shaded areas correspond to one standard deviation in each direction of the mean.

Standard missing data methods. Our experiments reveal that none of the standard missing data methods reliably mitigate the bias introduced through under-reporting and, instead, may themselves introduce disparities in selection rates. Omission of the feature 'hours worked per week' leads to over-selection of the female group and under-selection of the male group in the ACS Income data. This is because female individuals report to work on average less than male individuals (35.43h/week vs. 40.05h/week) while work hours contribute positively to income (Figure E.3). Omitting the feature blinds the model to these differences. Similar effects occur with the feature 'education attainment', and 'priors count' in the COMPAS data (Figures E.1 and 3). For multiple imputation on the COMPAS data, we see that the excess selection rate flips signs and the group with under-reported 'priors count' is over-selected (Figures 3 and E.4). This is because the feature has a lot of true 0-entries that are wrongly imputed as positive values. The cost incurred by these wrong imputations exceeds the benefit of imputation. In comparison to training on mismeasured features directly, the parameter estimation bias is considerable even for small amounts of under-reporting. For high under-reporting rates, the excess selection rate follows a similar pattern as the excess selection rate with feature omission since imputation is conducted using the features already present in the model adding little to no additional information. Since our models are well-specified, row omission recovers the true parameter estimates as displayed for the COMPAS data and feature 'priors count' in Figure 3. Despite access to the ground-truth, we observe that under-reporting bias introduced at prediction time increases the selection rate disparity. While the model without row omission is able to shift weight to correlated features as more and more entries for 'priors count' are under-reported, the row omission model cannot make use of the feature correlations ultimately leading to the increasing rather than decreasing disparities. With the same reasoning, the test set performance as measured by R^2 is decreased as displayed in the figure.

Augmented loss and optimal prediction imputation. We contrast the performance of our method and standard approaches for handling missing features at the example of the 'priors count' feature in the COMPAS dataset. Under-reporting rates are estimated with the

procedure described in Section 6. We refer to Appendix F for further details on the under-reporting rate estimation. The results in Figure 3 show that selection rate disparities decrease considerably when using group-dependent augmented loss and group-dependent optimal prediction imputation. In contrast multiple imputation, this fairness improvement comes at no visible cost in performance. In fact, the average test set R^2 of the corrected model is very similar to, and even slightly higher than, the test set R^2 of the model trained directly on under-reported data (see Figure E.6). Despite some variability, the average parameter estimates of the corrected model appear more stable across different amounts of under-reporting which suggests that the method successfully diminishes the bias introduced by under-reporting.

County-level birth data. For the birth data, Figure 4 suggest that under-reporting of all behavioral health data for the non-Medicaid population leads to over-selection of the Medicaid population. In the displayed overall selection rate range (<10%), the Medicaid population is selected about 10% more often than in the "true" data setting. Note that in reality this difference could be even larger because some of the "true" data features were likely already under-reported. As shown in the Figure, some of the resulting disadvantage is still observable when evaluating performance for Black families. This can be explained by the fact that the two group variables are positively correlated in the dataset ($\rho = 0.44$). Excluding behavioral health features altogether leads to a reversal of selection rate disparities in the Medicaid / non-Medicaid groups, and significantly increases selection of the Black sub-population for overall selection rates less than about 7%. This provides additional evidence highlighting that omitting features is an unreliable solution for addressing the disparities arising from differential under-reporting. It underscores the point that such an approach lacks precision, potentially leading to arbitrary and inequitable outcomes.

9 DISCUSSION

Differential feature under-reporting is a common phenomenon in administrative data. Data records are generally more complete for individuals who rely more consistently on public services (e.g. public health coverage). In many predictive risk assessment settings, the segment of the population with more complete observations
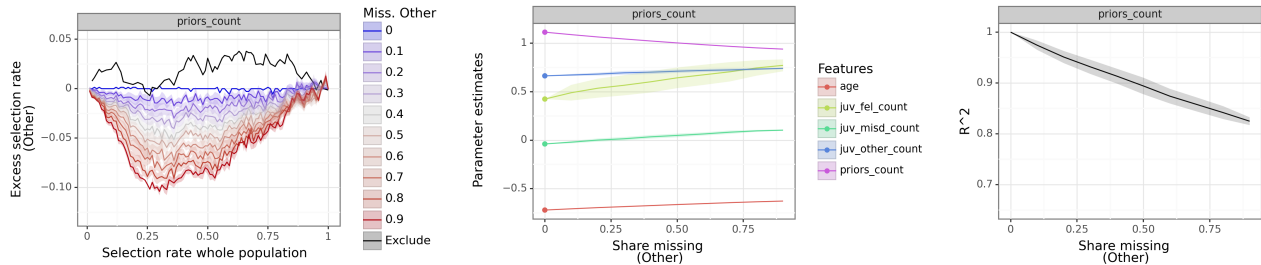
(a) Estimation and prediction with under-reported feature
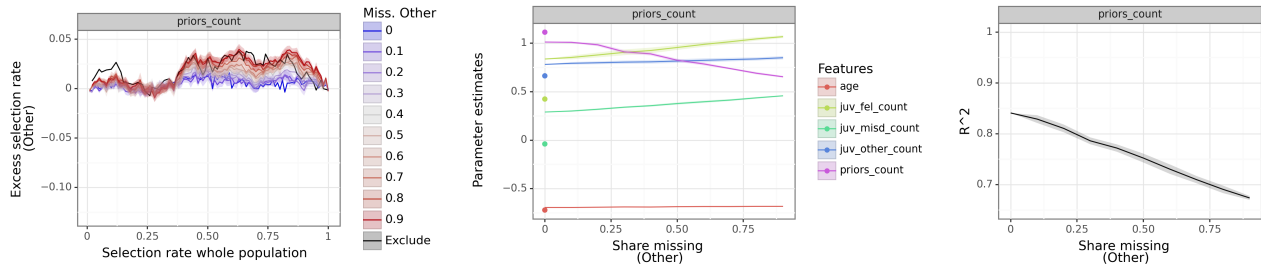
(b) Estimation and prediction with multiple imputation
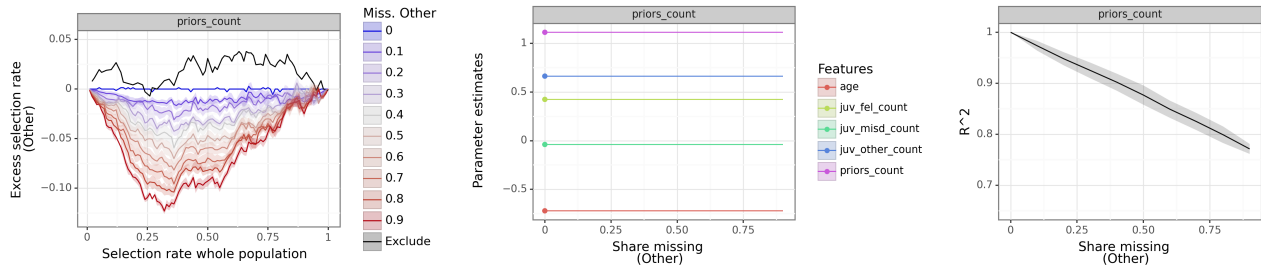
(c) Estimation on rows with non-zero entries and prediction with under-reported feature
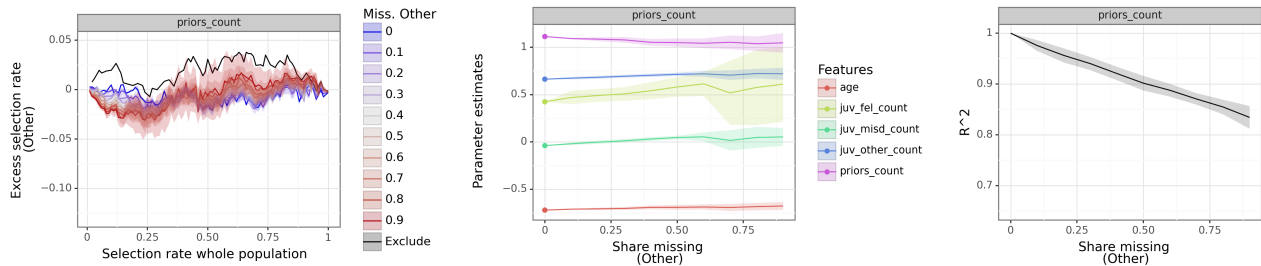
(d) Estimation with group-dependent augmented loss and prediction with optimal group-dependent imputation (our method)

Figure 3: Excess selection rates of group Other (i.e. not African-American) (left columns), parameter estimates (middle column), and test set R^2 (right columns) when under-reporting is injected into 'priors count' in group Other using the COMPAS dataset and synthetic two-year recidivism outcomes. In (a), the model is trained and evaluated using the under-reported feature. For (b), we first train a multiple imputation model and then train and evaluate the prediction model using probabilistic imputations. For (c), the model is trained on only rows without 0-entries in 'priors count' and evaluated on the under-reported data. In (d), we train with group-dependent augmented loss and use group-dependent optimal imputation values for prediction. Results are reported as averages over 30 runs. Shaded areas correspond to one standard deviation. The solid dots in the middle column correspond to true parameters. Note that in order to preserve readability, parameter estimates are only displayed for continuous features. Figure E.6 provides an overlay plot of the rightmost column for easy comparison.

Figure 4: Selection rate fractions of different models. On the left, the results are displayed for the sub-population of Black individuals. On the right, the results are displayed for the sub-population that is insured through Medicaid. The selection rate of the whole population is considered to be 10% or lower which reflects a realistic range for predictive risk modeling.
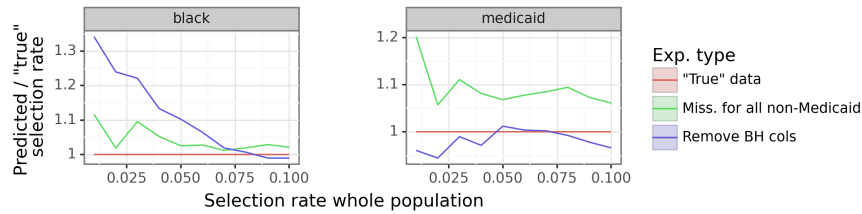
overlaps with sub-populations that are more commonly flagged as high-risk. Critics have argued that differential data availability is a key driver of the observed disparity in selection rates. When being classified as high risk subjects one to greater scrutiny of burden, this may disadvantage those with more complete data [26]. Overall, the results of our study lend further credence to the concern by demonstrating how feature under-reporting generally leads to under-selection of a group that is already less frequently identified as high-risk. While, as we demonstrate, groups with greater data availability can theoretically be under-selected, the feature dependence structure under which this occurs appears to be uncommon in practice.

We illustrate the increased selection rate that individuals who rely on public healthcare coverage may experience at a real world example. Following the idea of Allegheny County's Hello Baby program [29], we build a model that predicts the risk that a newborn child will be removed from their family by Child Protective Services (CPS) within three years based on county-level data. The dataset contains behavioral and mental health information on the parents which can be assumed to be more complete for families that rely on public insurance. We note that, for privately insured individuals, some of this information may still be observed, e.g. because the individual was publicly insured previously, or individual information has been collected explicitly, but a lot of the information can be assumed missing. Our experiments suggest that further under-reporting in behavioral health related information for the privately insured sub-population leads to an increase in high-risk predictions for the publicly insured group. We hypothesize that this effect would be even larger if the dataset was not already missing large portions of the feature observations for the privately insured group. This finding implies an unfair targeting of the publicly insured sub-population as high-risk. Since Black mothers in the data are publicly insured more frequently than mothers from other racial groups, these results also suggest that Black families are predicted to be at high-risk at unfairly inflated rates. Of course, depending on intervention type, a high risk classification may lead to an advantage or disadvantage for the families. In the Hello Baby setting, it is tied to eligibility for voluntary supportive services provided by county-funded service providers.

Our work proposes a technical remedy for the impact of under-reporting as a driver of disparities in selection rates. While standard missing data methods did not lead to more equitable outcomes in our experiments, these new methods reduced disparities considerably with little to no decrease in model accuracy. The applicability

and performance in real-world administrative data settings like the Hello baby program remains an interesting and important avenue for future work.

RESEARCH ETHICS AND SOCIAL IMPACT

Ethical considerations statement. The authors did not face ethical concerns that had to be mitigated while conducting this study. Experiments in this paper are based on the commonly used COMPAS and German credit datasets, 2018 US Census data from the American Community Survey, and a private county level data set. Result are aggregated over broad population groups and no identifiable information can be retrieved.

Researcher positionality statement. The authors recognize that their societal advantages give them certain benefits not shared by all individuals undergoing public sector risk assessment. Thus, great care was taken in reflecting on the question: Does this work benefit us or the community at large? Since inflated selection rates can lead to tangible advantages (e.g. qualification for further publicly funded services) and significant disadvantages (e.g. unfavorable bail decisions in criminal risk assessment) depending on the application area, we believe that addressing problems of feature under-reporting ultimately benefits the community at large.

Adverse impact statement. The authors believe that drawing attention to the problem of differential feature under-reporting has the potential to positively impact public sector risk assessment instruments for all individuals subjected to these systems. However, our work provides only a first step towards finding appropriate solutions to this problem. We propose a potential mitigation method and evaluate the method in a semi-synthetic setting, but assessment of potential adverse effects of the method in real-world applications are beyond the scope of this paper. We clearly state this at the end of the discussion and call for future work in this direction.

REFERENCES

[1] Roy Adams, Yuelong Ji, Xiaobin Wang, and Suchi Saria. 2019. Learning Models from Data with Measurement Error: Tackling Underreporting. In *International Conference on Machine Learning (IMCL '20)*.
[2] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. 2019. The Challenge of Imputation in Explainable Artificial Intelligence Models. *arXiv preprint, arXiv:1907.12669* (2019).
[3] Dennis J. Aigner and Glen G. Cain. 1977. Statistical Theories of Discrimination in Labor Markets. *Industrial and Labor Relations Review* 30, 2 (Jan. 1977), 175.
[4] Michelle Alexander. 2010. *The new Jim Crow: Mass Incarceration in the Age of Colorblindness*. New Press, New York, NY.

[5] J D Angrist and Jorn-Steffen Pischke. 2008. *Mostly harmless econometrics.* Princeton University Press, Princeton, NJ.

[6] Larson J. Mattu S. Angwin, J. and L. Kirchner. 2016. Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[7] Nicholas C Arpey, Anne H Gaglioti, and Marcy E Rosenbaum. 2017. How socioeconomic status affects patient perceptions of health care: A qualitative study. *J. Prim. Care Community Health* 8, 3 (July 2017), 169–175.

[8] T.C. Bailey, M.S. Carvalho, T.M. Lapa, W.V. Souza, and M.J. Brewer. 2005. Modeling of Under-detection of Cases in Disease Surveillance. *Annals of Epidemiology* 15, 5 (May 2005), 335–343.

[9] Geoffrey C Barnes and Jordan M Hyatt. 2012. Classifying adult probationers by forecasting future offending. *National Institute of Justice. Retrieved February* 4 (2012), 2020.

[10] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 50, 1 (July 2018), 3–44.

[11] Emily Black, Hadi Elzayn, Alexandra Chouldechova, Jacob Goldin, and Daniel Ho. 2022. Algorithmic fairness and vertical equity: Income fairness with IRS tax audit models. In *2022 ACM Conference on Fairness, Accountability, and Transparency.*

[12] Hermann Brenner and Dana Loomis. 1994. Varied Forms of Bias Due to Nondifferential Error in Measuring Exposure. *Epidemiology* 5, 5 (1994), 510–517.

[13] Irwin Bross. 1954. Misclassification in 2 X 2 Tables. *Biometrics* 10, 4 (Dec. 1954), 478.

[14] Bradley Butcher, Chris Robinson, Miri Zilka, Riccardo Fogliato, Carolyn Ashurst, and Adrian Weller. 2022. Racial Disparities in the Enforcement of Marijuana Violations in the US. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22).* 130–143.

[15] Danton S. Char, Nigam H. Shah, and David Magnus. 2018. Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *New England Journal of Medicine* 378, 11 (March 2018), 981–983.

[16] Irene Y. Chen, Fredrik D. Johansson, and David Sontag. 2018. Why is My Classifier Discriminatory?. In *32nd International Conference on Neural Information Processing Systems (NIPS'18).*

[17] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency (FAT*).*

[18] Haitao Chu, Zhaojie Wang, Stephen R. Cole, and Sander Greenland. 2006. Sensitivity Analysis of Misclassification: A Graphical and a Bayesian Approach. *Annals of Epidemiology* 16, 11 (2006), 834–841.

[19] Federico Cismondi, André S. Fialho, Susana M. Vieira, Shane R. Reti, João M.C. Sousa, and Stan N. Finkelstein. 2013. Missing data in medical databases: Impute, delete or classify? *Artificial Intelligence in Medicine* 58, 1 (May 2013), 63–72.

[20] Guilherme Lopes de Oliveira, Rosangela Helena Loschi, and Renato Martins Assunção. 2017. A random-censoring Poisson model for underreported data. *Statistics in Medicine* 36, 30 (Oct. 2017), 4873–4892.

[21] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems (Neurips '21).*

[22] Mustafa Dosemeci, Sholom Wacholder, and Jay H. Lubin. 1990. Does Nondifferential Misclassification of Exposure Always Bias a True Effect Toward the Null Value? *American Journal of Epidemiology* 132, 4 (1990), 746–748.

[23] Jessie K Edwards, Stephen R Cole, Melissa A Troester, and David B Richardson. 2013. Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. *Am. J. Epidemiol.* 177, 9 (May 2013), 904–912.

[24] Charles Elkan and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM.

[25] Zeynep Engin and Philip Treleaven. 2019. Algorithmic government: Automating public services and supporting civil servants in using data science technologies. *Comput. J.* 62, 3 (2019), 448–460.

[26] Virginia Eubanks. 2018. A Child Abuse Prediction Model Fails Poor Families. https://www.wired.com/story/excerpt-from-automating-inequality/

[27] Martínez-Plumed Fernando, Ferri Cèsar, Nieves David, and Hernández-Orallo José. 2021. Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems* 36, 7 (March 2021), 3217–3258.

[28] Riccardo Fogliato, Alice Xiang, Zachary Lipton, Daniel Nagin, and Alexandra Chouldechova. 2021. On the Validity of Arrest as a Proxy for Offense: Race and the Likelihood of Arrest for Violent Crimes. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21).* 100–111.

[29] Centre for Social Data Analytics at the Auckland University of Technology. 2020. Implementing the Hello Baby Prevention Program in Allegheny County. (2020). https://www.alleghenycountyanalytics.us/wp-content/uploads/2020/12/Hello-Baby-Methodology-v6.pdf

[30] Christian Fricke. 2020. Missing Fairness: The Discriminatory Effect of Missing Values inDatasets on Fairness in Machine Learning. *Master thesis* (2020).

[31] Wayne A Fuller. 1987. *Measurement Error Models.* John Wiley & Sons, Nashville, TN.

[32] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian Data Analysis* (3 ed.). Chapman & Hall/CRC, Philadelphia, PA.

[33] Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. 2018. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine* 178, 11 (Nov. 2018), 1544.

[34] Sander Greenland. 2014. Sensitivity Analysis and Bias Analysis. In *Handbook of Epidemiology.* Springer New York, 685–706.

[35] Rolf H H Groenwold and Olaf M Dekkers. 2020. Missing data: the impact of what is not there. *European Journal of Endocrinology* 183, 4 (Oct. 2020), E7–E9.

[36] Jerry Hausman. 2001. Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left. *Journal of Economic Perspectives* 15, 4 (2001), 57–67.

[37] Kimberly A Houser and Debra Sanders. 2016. The use of big data analytics by the IRS: Efficient solutions or the end of privacy as we know it. *Vand. J. Ent. & Tech. L.* 19 (2016), 817.

[38] Vincent Jeanselme, Maria De-Arteaga, Zhe Zhang, Jessica Barrett, and Brian Tom. 2022. Imputation Strategies Under Clinical Presence: Impact on Algorithmic Fairness. *Machine Learning for Health (ML4H '22).*

[39] Fereshte Khani and Percy Liang. 2020. Feature Noise Induces Loss Discrepancy Across Groups. In *International Conference on Machine Learning (ICML '20).*

[40] G. King, J. Honaker, A. Joseph, and K. Scheve. 2001. Analyzing incomplete political science data: an alternative algorithm for multipleimputation. *Am. Polit. Sci. Rev.* 95 (2001), 49–69.

[41] Chamari I Kithulgoda, Rhema Vaithianathan, and Dennis P Culhane. 2022. Predictive risk modeling to identify homeless clients at risk for prioritizing services using routinely collected data. *Journal of Technology in Human Services* 40, 2 (2022), 134–156.

[42] Karen Levy, Kyla E Chasalow, and Sarah Riley. 2021. Algorithms and decision-making in the public sector. *Annual Review of Law and Social Science* 17 (2021), 309–334.

[43] Sandra G Mayson. 2019. Bias in, bias out. *The Yale Law Journal* 128, 8 (2019), 2218–2300.

[44] John F McCarthy, Robert M Bossarte, Ira R Katz, Caitlin Thompson, Janet Kemp, Claire M Hannemann, Christopher Nielson, and Michael Schoenbaum. 2015. Predictive modeling and concentration of the risk of suicide: implications for preventive interventions in the US Department of Veterans Affairs. *American journal of public health* 105, 9 (2015), 1935–1942.

[45] P.E. McKnight, K.M. McKnight, S. Sidani, and A. J. Figueredo. 2007. *Missing Data: A Gentle Introduction.*

[46] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with Noisy Labels. In *Advances in Neural Information Processing Systems*, C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (Eds.), Vol. 26. Curran Associates, Inc.

[47] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. 2009. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization* 19, 4 (2009), 1574–1609.

[48] Edmund Phelps. 1972. The Statistical Theory of Racism and Sexism. *American Economic Review* 62, 4 (1972), 659–61.

[49] Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, and Sharad Goel. 2020. A large-scale analysis of racial disparities in police stops across the United States. *Nat. Hum. Behav.* 4, 7 (July 2020), 736–745.

[50] Dewi Rahardja and Dean M. Young. 2021. Confidence Intervals for the Risk Ratio Using Double Sampling with Misclassified Binomial Data. *Journal of Data Science* 9, 4 (2021), 529–548.

[51] Alvin Rajkomar, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. 2018. Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine* 169, 12 (Dec. 2018), 866.

[52] UCI Machine Learning Repository. 1994. German Credit data. (1994). https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[53] Rashida Richardson, Jason M Schultz, and Kate Crawford. 2019. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev. Online* 94 (2019), 15.

[54] Donald B. Rubin. 1976. Inference and missing data. *Biometrika* 63, 3 (1976), 581–592.

[55] Konstantinos Sechidis, Matthew Sperrin, Emily S. Petherick, Mikel Luján, and Gavin Brown. 2017. Dealing with under-reported variables: An information theoretic solution. *International Journal of Approximate Reasoning* 85 (2017), 159–177.

[56] Oliver Stoner, Theo Economou, and Gabriela Drummond Marques da Silva. 2019. A Hierarchical Framework for Correcting Under-Reporting in Count Data. *J. Amer. Statist. Assoc.* 114, 528 (apr 2019), 1481–1492.

[57] Rhema Vaithianathan, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney. 2017. Developing predictive models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation. *Center for Social data Analytics* (2017).

[58] Marvin Van Bekkum and Frederik Zuiderveen Borgesius. 2021. Digital welfare fraud detection and the Dutch SyRI judgment. *European Journal of Social Security* 23, 4 (2021), 323–340.

[59] Yanchen Wang and Lisa Singh. 2021. Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics* 12, 2 (May 2021), 101–119.

[60] Kevin Wu, Dominik Dahlem, Christopher Hane, Eran Halperin, and James Zou. 2023. Collecting data when missingness is unknown: a method for improving model performance given under-reporting in patient populations. In *Proceedings of the Conference on Health, Inference, and Learning*, Vol. 209. PMLR, 229–242.

[61] Yiliang Zhang and Qi Long. 2021. Assessing Fairness in the Presence of Missing Data. In *Advances in Neural Information Processing Systems (Neurips '21)*.

[62] Helen Zhou, Balakrishnan Sivaraman, and Zachary C. Lipton. 2022. Domain Adaptation Under Missingness Shift. *arXiv preprint, arXiv:2211.02093* (2022).

A SUMMARY OF ASSUMPTIONS

Result / Assumption	Missingness indicators	Default value $m = 0$	Linear ground truth	$G \perp Z$, i.e. MCAR	$Z_{[2:d]}$ uncorrelated	$Z \sim \mathcal{N}$ jointly Gaussian
Lemma 3		x	x	x		
Proposition 4 & 5		x	x	x	x	
Proposition 6 & B.2		x		x		x
Corollary 7		x	x	x	x	x
Lemma 8 & B.3		x		(x)		
Lemma 9 & B.4		x	x	(x)		
Under-reporting rate estimation		x		(x)		

Table A.1: Summary of assumptions. Rows represent paper segments, columns indicate sufficient assumptions for corresponding findings.

B ADDITIONAL THEOREMS

PROPOSITION B.1. *In the d-dimensional case setting described in Section 4, the parameter estimates from Equation 1 take the form*

$$\hat{\beta}_1 = \beta_1 \frac{1}{1 - R^2} \sqrt{\frac{\mathbb{V}[Z_1]}{\mathbb{V}[X_1]}} \left(\rho(X_1, Z_1) - \sum_{i=2}^{d} \rho(X_1, Z_i)\rho(Z_1, Z_i) \right),$$

$$\hat{\beta}_k = \beta_1 \sqrt{\frac{\mathbb{V}[Z_1]}{\mathbb{V}[Z_k]}} \left(\rho(Z_k, Z_1) - \frac{1}{1 - R^2}\rho(X_1, Z_k) \left(\rho(X_1, Z_1) - \sum_{i=2}^{d} \rho(X_1, Z_i)\rho(Z_1, Z_i) \right) \right) + \beta_k \tag{2}$$

for $k \in [2 : d]$. Here, $R^2 = \sum_{i=2}^{d} \rho(X_1, Z_i)^2 \in [0, 1)$.

Here, R^2 is the squared coefficient of multiple correlation between $Z_1\xi_1$ and $Z_{[2:d]} = [Z_2, \ldots, Z_d]$ which can be interpreted as the fraction of variance in $Z_1\xi_1$ that can be explained by the independent variables $Z_{[2:d]}$. If all features are observed, the factor $\rho(X_1, Z_1) - \sum_{i=2}^{d} \rho(X_1, Z_i)\rho(Z_1, Z_i)$ collapses to $(1 - R^2)$, and the estimates are unbiased. With under-reporting the bias introduced into the parameter estimates depends on the strength of correlations between features, as well as how this correlation changes with the mismeasurement of Z_1. The bias in Equation 2 can be conceptualized as a generalization of omitted variable bias [5] which is further explored in Appendix C.

PROPOSITION B.2. *Define the threshold turning point T as*

$$T = \hat{\alpha} + \hat{\beta}_{[2:d]}^T \mu_{[2:d]} + \frac{sd\left(\hat{\beta}_{[2:d]}^T Z_{[2:d]} \right)}{sd\left(\hat{\beta}_{[2:d]}^T Z_{[2:d]} \right) - sd\left(\hat{\beta}^T Z \right)} \hat{\beta}_1 \mu_1.$$

Then, for a high threshold $\tilde{y}$ with $\tilde{y} > T$, the group with more under-reporting is

- **Case 1:** *Over-selected if* $\mathbb{V}\left[\hat{\alpha} + \hat{\beta}_{[2:d]} Z_{[2:d]} \right] > \mathbb{V}\left[\hat{\alpha} + \hat{\beta}Z \right]$, *or*
- **Case 2:** *Under-selected if* $\mathbb{V}\left[\hat{\alpha} + \hat{\beta}_{[2:d]} Z_{[2:d]} \right] < \mathbb{V}\left[\hat{\alpha} + \hat{\beta}Z \right]$.

For low thresholds $\tilde{y} < T$, the cases are reversed.

In practical applications, thresholds are usually set such that only a small portion of predictions exceeds the threshold. For example, we can only decide to screen a small portion of calls in the child welfare setting. In particular, realistic thresholds are generally well above the average $\hat{Y}$. On a high level, the turning point T in Proposition B.2 represent an adjusted mean predicted value where the influence of the feature with under-reporting is weighed depending on a ratio determined by prediction variances with and without the feature.

LEMMA B.3 (GROUP-DEPENDENT AUGMENTED LOSS). *Assume fixed $f \in \mathcal{F}, z \in \mathcal{Z}, y \in \mathcal{Y}, g \in \{0, 1\}$ and $X \in \mathbb{R}^d$ defined by $X_1 = Z_1\xi_1^g$ and $X_{[2:d]} = z_{[2:d]}$. Define*

$$\tilde{l}(f, X, y, g) := \frac{1}{m_1^g} l(f, X, y) - \frac{1 - m_1^g}{m_1^g} l(f, [0, X_{[2:d]}]^T, y)$$

Then, we have that $\mathbb{E}_{\xi_1^g}\left[\tilde{l}(f, X, y, g) \right] = l(f, z, y)$.

LEMMA B.4 (GROUP-DEPENDENT OPTIMAL PREDICTION IMPUTATION VALUES). *Assume $f(Z) = \alpha + \beta^T Z$ is the ground truth model, X a random vector of observed features, and G the group membership. We set*

$$X' = \begin{cases} X \text{ if } X_1 \neq 0, \\ [x_1'^0, X_{[2:d]}] \text{ if } X_1 = 0 \text{ and } G = 0, \\ [x_1'^1, X_{[2:d]}] \text{ if } X_1 = 0 \text{ and } G = 1, \end{cases}$$

where $x_1'^0, x_1'^1$ are group-dependent fixed imputation values. Then,

$$\arg\min_{x_1'^g} \mathbb{E}_X \left[(f(X') - Y)^2 \right] = \mathbb{E}\left[Z_1 \mid X_1 = 0, G = g \right]$$

are the optimal group-dependent prediction imputation values for $g \in \{0, 1\}$.

Similar to before, the optimal imputation values can be written as

$$\mathbb{E}\left[Z_1 \mid X_1 = 0, G = g \right] = \frac{\frac{1}{m_1^g} \mathbb{E}\left[X_1, G = g \right] - P(X_1 \neq 0 \mid G = g) \mathbb{E}\left[X_1 \mid X_1 \neq 0, G = g \right]}{P(X_1 = 0 \mid G = g)},$$

which can be estimated directly from observed data.

C   CONNECTION BETWEEN PROPOSITION B.1 AND OMITTED VARIABLE BIAS

Econometrics literature uses the term omitted variable bias to refer to the model estimation bias that is introduced when omitting an independent variable that influences both other independent variables and the dependent outcome [5]. In the setting of Proposition B.1, omitting the first feature entirely corresponds to a setting in which all feature entries are under-reported, i.e. default to 0. The k-th parameter estimate in this case can be written as

$$\hat{\beta}_k = \beta_1 \frac{\text{Cov}\left[Z_k, Z_1 \right]}{\mathbb{V}\left[Z_k \right]} + \beta_k$$

which is known as omitted variable bias formula [5]. Here, $\gamma_{Z_1,Z_k} = \text{Cov}\left[Z_1, Z_k \right] / \mathbb{V}\left[Z_k \right]$ corresponds to the population regression coefficient of a linear regression of Z_1 on Z_k which can be written as

$$Z_1 = \alpha_{Z_1,Z_k} + \gamma_{Z_1,Z_k} Z_k,$$

where α_{Z_1,Z_k} is an intercept. Omitting Z_1 from the regression induces a confounding relationship where the effects of Z_1 on Z_k become intertwined. Instead of isolating the effect of Z_k on Y, $\hat{\beta}_k$ also includes a partial effect of Z_1 on Y. This effect is scaled by γ_{Z_1,Z_k} to account for the linear relationship between Z_1 and Z_k.

In the setting of this paper, we are interested in cases in which some but not necessarily all of the feature entries are missing. Maintaining the same notation as before, $\hat{\beta}_1$ from Equation 2 in this general case can be written as

$$\hat{\beta}_1 = \beta_1 \frac{1}{1 - R^2} \left(\frac{\text{Cov}\left[X_1, Z_1 \right]}{\mathbb{V}\left[X_1 \right]} - \sum_{i=2}^{d} \frac{\text{Cov}\left[X_1, Z_i \right] \text{Cov}\left[Z_1, Z_i \right]}{\mathbb{V}\left[X_1 \right] \mathbb{V}\left[Z_i \right]} \right)$$

$$= \beta_1 \left(\frac{\gamma_{Z_1,X_1} - \sum_{i=2}^{d} \gamma_{Z_i,X_1} \gamma_{Z_1,Z_i}}{1 - R^2} \right).$$

Here, the numerator of the biasing factor reflects how much information about Z_1 remains encoded in X_1 without drawing on associations through the other features $Z_2, \ldots, Z_d$ (i.e, arrows of the form $X_1 \rightarrow Z_i \rightarrow Z_1$). The denominator measures how much of the variance in X_1 is explained by $Z_2, \ldots, Z_d$. For $k \in [2 : d]$, we receive

$$\hat{\beta}_k = \beta_k + \beta_1 \frac{\text{Cov}\left[Z_k, Z_1 \right]}{\mathbb{V}\left[Z_k \right]} - \beta_1 \frac{1}{1 - R^2} \frac{\text{Cov}\left[X_1, Z_k \right]}{\mathbb{V}\left[Z_k \right]} \left(\frac{\text{Cov}\left[X_1, Z_1 \right]}{\mathbb{V}\left[X_1 \right]} - \sum_{i=2}^{d} \frac{\text{Cov}\left[X_1, Z_i \right] \text{Cov}\left[Z_1, Z_i \right]}{\mathbb{V}\left[Z_i \right] \mathbb{V}\left[X_1 \right]} \right)$$

$$= \beta_k + \beta_1 \gamma_{Z_1,Z_k} - \beta_1 \frac{1}{1 - R^2} \gamma_{X_1,Z_k} \left(\gamma_{Z_1,X_1} - \sum_{i=2}^{d} \gamma_{Z_i,X_1} \gamma_{Z_1,Z_i} \right)$$

$$= \beta_k + \underbrace{\beta_1 \gamma_{Z_1,Z_k}}_{\text{omitted variable bias}} - \underbrace{\hat{\beta}_1 \gamma_{X_1,Z_k}}_{\text{Correction since partially observed}}.$$

Instead of just encoding the effect of Z_k on Y and partial effect of Z_1 on Y like before, the estimate $\hat{\beta}_k$ now also corrects for the fact that Z_1 is partially observed. The magnitude of the correction depends on the parameter estimate for the partially observed variable as well as the linear relationship between X_1 and Z_k.

D PROOFS

In this section, we provide the full proofs for the results in the main text.

Lemma 3. We have $\xi \perp Z$ and $\mathbb{E}\left[\xi^2\right] = \mathbb{E}\left[\xi\right]$. Since $\mathbb{E}\left[\xi\right] \in [0, 1]$, we have

$$\mid \hat{\beta} \mid = \mid \frac{\text{Cov}\left[X, Z\right]}{\text{Cov}\left[X, X\right]} \beta \mid = \mid \frac{\mathbb{E}\left[\xi Z^2\right] - \mathbb{E}\left[\xi Z\right]\mathbb{E}\left[Z\right]}{\mathbb{E}\left[\xi^2 Z^2\right] - \mathbb{E}\left[\xi Z\right]^2} \beta \mid = \frac{\mathbb{E}\left[\xi\right]\mathbb{V}\left[Z\right]}{\mathbb{E}\left[\xi\right]\left(\mathbb{E}\left[Z^2\right] - \mathbb{E}\left[\xi\right]\mathbb{E}\left[Z\right]^2\right)} \mid \beta \mid \leq \mid \beta \mid .$$

Proposition B.1. In order to derive the equations for $\hat{\beta}_i$ for $i \in [1 : d]$, we start by inverting the covariance matrix

$$\Sigma_X = \begin{pmatrix} \mathbb{V}\left[Z_1\xi_1\right] & \text{Cov}\left[Z_1\xi_1, Z_2\right] & \text{Cov}\left[Z_1\xi_1, Z_3\right] & \cdots & \text{Cov}\left[Z_1\xi_1, Z_d\right] \\ \text{Cov}\left[Z_2, Z_1\xi_1\right] & \mathbb{V}\left[Z_2\right] & \text{Cov}\left[Z_2, Z_3\right] & \cdots & \text{Cov}\left[Z_2, Z_d\right] \\ \text{Cov}\left[Z_3, Z_1\xi_1\right] & \text{Cov}\left[Z_3, Z_2\right] & \mathbb{V}\left[Z_3\right] & \cdots & \text{Cov}\left[Z_3, Z_d\right] \\ \vdots & & & \ddots & \vdots \\ \text{Cov}\left[Z_d, Z_1\xi_1\right] & \text{Cov}\left[Z_d, Z_2\right] & & \cdots & \mathbb{V}\left[Z_d\right] \end{pmatrix}.$$

For this, we separate the matrix into the blocks $A = (\Sigma_X)_{11}$, $B = ((\Sigma_X)_{1j})_{j \in [2:d]}$, $C = ((\Sigma_X)_{i1})_{i \in [2:d]}$, and $D = ((\Sigma_X)_{ij})_{i,j \in [2:d]}$. Note that $A \in \mathbb{R}^{1 \times 1}$, $B = C^T \in \mathbb{R}^{1 \times (d-1)}$, and $D \in \mathbb{R}^{(d-1) \times (d-1)}$. Using matrix inversion theorem, the inverse of Σ_X can be written as

$$(\Sigma_X)^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}, \tag{3}$$

where $D - CA^{-1}B = D - CA^{-1}C^T$ is the Schur complement of A in Σ_X. Recall that, by assumption, $\text{Cov}\left[Z_i, Z_j\right] = 0$ for $i, j > 1$ with $i \neq j$ which means that D is a diagonal matrix. We also note that $\text{rank}(CC^T) = 1$. Denoting $g = \text{trace}(-A^{-1}CC^TD^{-1})$, the inverse of the Schur complement can be written as

$$(D - CA^{-1}C^T)^{-1} = (D - A^{-1}CC^T)^{-1} = D^{-1} + \frac{1}{1 + g}D^{-1}A^{-1}CC^TD^{-1}.$$

Here, D^{-1} is a diagonal matrix with values $1/\mathbb{V}\left[Z_i\right]$ for $i \in [2 : d]$ on the diagonal, $A^{-1} = 1/\mathbb{V}\left[Z_1\xi_1\right]$, and the diagonal of CC^T is $\text{Cov}\left[Z_1\xi_1, Z_i\right]^2$ for $i \in [2 : d]$. It follows that

$$g = -\sum_{i=2}^{d} \rho(Z_1\xi_1, Z_i)^2$$

which corresponds to the negative of the R^2 between $Z_1\xi$ and $Z_2, \ldots, Z_d$. We hence write R^2 for $-g$ in the following.

Now we can calculate the top left block of the inverse matrix in Equation 3 as

$$A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} = \frac{1}{\mathbb{V}\left[Z_1\xi_1\right]} \frac{1}{1 - R^2}.$$

The top right bock corresponds to the row vector

$$-AB(D - CA^{-1}B)^{-1} = \left(\left(-\frac{\text{Cov}\left[Z_1\xi_1, Z_i\right]}{\mathbb{V}\left[Z_i\right]\mathbb{V}\left[Z_1\xi_1\right]} \frac{1}{1 - R^2}\right)_i\right)_{i \in [2:d]},$$

while the bottom left block is the same transposed. Lastly, the bottom left block of Equation 3 can be computed as

$$(D - CA^{-1}B)^{-1} = \left(\text{diag}(1/\mathbb{V}\left[Z_i\right]) + \frac{1}{1 - R^2}\left(\frac{\text{Cov}\left[Z_1\xi_1, Z_i\right]\text{Cov}\left[Z_1\xi_1, Z_j\right]}{\mathbb{V}\left[Z_1\xi_1\right]\mathbb{V}\left[Z_i\right]\mathbb{V}\left[Z_j\right]}\right)_{ij}\right)_{i,j \in [2:d]}.$$

Inserting these values into Equation 1 yields the desired parameter estimates

$$\hat{\beta}_1 = \beta_1 \frac{1}{1 - R^2} \sqrt{\frac{\mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[X_1\right]}} \left(\rho(X_1, Z_1) - \sum_{i=2}^{d} \rho(X_1, Z_i)\rho(Z_1, Z_i)\right),$$

and

$$\hat{\beta}_k = \beta_1 \sqrt{\frac{\mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_k\right]}} \left(\rho(Z_k, Z_1) - \frac{1}{1 - R^2}\rho(X_1, Z_k)\left(\rho(X_1, Z_1) - \sum_{i=2}^{d} \rho(X_1, Z_i)\rho(Z_1, Z_i)\right)\right) + \beta_k.$$

Proposition 4. The estimates from Proposition B.1 simplify to

$$\hat{\beta}_1 = \frac{1}{1 - R^2} \left(\frac{\mathbb{E}\left[\xi_1 \right] \mathbb{V}\left[Z_1 \right]}{\mathbb{V}\left[Z_1 \xi_1 \right]} - \frac{R^2}{\mathbb{E}\left[\xi_1 \right]} \right) \beta_1$$

and

$$\hat{\beta}_k = \beta_1 \frac{1}{1 - R^2} \left(\frac{\text{Cov}\left[Z_1, Z_k \right] \left(\mathbb{V}\left[Z_1 \xi_1 \right] - \mathbb{V}\left[Z_1 \right] \mathbb{E}\left[\xi_1 \right]^2 \right)}{\mathbb{V}\left[Z_k \right] \mathbb{V}\left[Z_1 \xi_1 \right]} \right) + \beta_k,$$

for $k \in [2 : d]$.

For the first claim, recall that $\mathbb{V}\left[\xi \right] = \mathbb{E}\left[\xi_1 \right] - \mathbb{E}\left[\xi_1 \right]^2$ and $\mathbb{V}\left[Z_1 \xi_1 \right] = \mathbb{V}\left[Z_1 \right] \mathbb{E}\left[\xi_1 \right]^2 + \mathbb{V}\left[\xi_1 \right] \mathbb{E}\left[Z_1^2 \right]$. Note that $Z \perp \xi$ allows us to rewrite

$$
\begin{aligned}
R^2 &= \sum_{i=2}^{d} \rho(Z_1 \xi_1, Z_i)^2 \\
&= \sum_{i=2}^{d} \frac{\mathbb{E}\left[\xi_1 \right]^2 \text{Cov}\left[Z_1, Z_i \right]^2 \mathbb{V}\left[Z_1 \right]}{\mathbb{V}\left[Z_1 \xi_1 \right] \mathbb{V}\left[Z_i \right] \mathbb{V}\left[Z_1 \right]} \\
&= \frac{\mathbb{E}\left[\xi_1 \right]^2 \mathbb{V}\left[Z_1 \right]}{\mathbb{V}\left[Z_1 \xi_1 \right]} \sum_{i=2}^{d} \rho(Z_1, Z_i)^2
\end{aligned}
$$

and thus

$$
\begin{aligned}
& \frac{1}{1 - R^2} \left(\frac{\mathbb{E}\left[\xi_1 \right] \mathbb{V}\left[Z_1 \right]}{\mathbb{V}\left[Z_1 \xi_1 \right]} - \frac{R^2}{\mathbb{E}\left[\xi_1 \right]} \right) \\
&= \frac{\mathbb{E}\left[\xi_1 \right] \mathbb{V}\left[Z_1 \right] \mathbb{E}\left[\xi_1 \right] - R^2 \mathbb{V}\left[Z_1 \xi_1 \right]}{(1 - R^2) \mathbb{V}\left[Z_1 \xi_1 \right] \mathbb{E}\left[\xi_1 \right]} \\
&= \frac{\mathbb{E}\left[\xi_1 \right] \mathbb{V}\left[Z_1 \right] \mathbb{E}\left[\xi_1 \right] - \frac{\mathbb{E}\left[\xi_1 \right]^2 \mathbb{V}\left[Z_1 \right]}{\mathbb{V}\left[Z_1 \xi_1 \right]} \sum_{i=2}^{d} \rho(Z_1, Z_i)^2 \mathbb{V}\left[Z_1 \xi_1 \right]}{(1 - \frac{\mathbb{E}\left[\xi_1 \right]^2 \mathbb{V}\left[Z_1 \right]}{\mathbb{V}\left[Z_1 \xi_1 \right]} \sum_{i=2}^{d} \rho(Z_1, Z_i)^2) \mathbb{V}\left[Z_1 \xi_1 \right] \mathbb{E}\left[\xi_1 \right]} \\
&= \frac{\mathbb{E}\left[\xi_1 \right] \mathbb{V}\left[Z_1 \right] \mathbb{E}\left[\xi_1 \right] - \mathbb{E}\left[\xi_1 \right]^2 \mathbb{V}\left[Z_1 \right] \sum_{i=2}^{d} \rho(Z_1, Z_i)^2}{(\mathbb{V}\left[Z_1 \right] \mathbb{E}\left[\xi_1 \right]^2 + (\mathbb{E}\left[\xi_1 \right] - \mathbb{E}\left[\xi_1 \right]^2) \mathbb{E}\left[Z_1^2 \right]) \mathbb{E}\left[\xi_1 \right] - \mathbb{E}\left[\xi_1 \right]^2 \mathbb{V}\left[Z_1 \right] \sum_{i=2}^{d} \rho(Z_1, Z_i)^2 \mathbb{E}\left[\xi_1 \right]} \\
&= \frac{\mathbb{V}\left[Z_1 \right] (1 - \sum_{i=2}^{d} \rho(Z_1, Z_i)^2)}{\mathbb{V}\left[Z_1 \right] \mathbb{E}\left[\xi_1 \right] (1 - \sum_{i=2}^{d} \rho(Z_1, Z_i)^2) + (1 - \mathbb{E}\left[\xi_1 \right]) \mathbb{E}\left[Z_1^2 \right]},
\end{aligned}
$$

which is positive as long as Z_1 is not a linear combination of other features which was explicitly excluded from consideration. The claim follows.

For the second claim, we show that

$$
\begin{aligned}
& \frac{\mathbb{V}\left[Z_1 \right] (1 - \sum_{i=2}^{d} \rho(Z_1, Z_i)^2)}{\mathbb{V}\left[Z_1 \right] \mathbb{E}\left[\xi_1 \right] (1 - \sum_{i=2}^{d} \rho(Z_1, Z_i)^2) + (1 - \mathbb{E}\left[\xi_1 \right]) \mathbb{E}\left[Z_1^2 \right]} < 1 \\
&\Leftrightarrow (1 - \sum_{i=2}^{d} \rho(Z_1, Z_i)^2) \mathbb{V}\left[Z_1 \right] (1 - \mathbb{E}\left[\xi_1 \right]) < (1 - \mathbb{E}\left[\xi_1 \right]) \mathbb{E}\left[Z_1^2 \right] \\
&\Leftrightarrow (1 - \sum_{i=2}^{d} \rho(Z_1, Z_i)^2) (\mathbb{E}\left[Z_1^2 \right] - \mathbb{E}\left[Z_1 \right]^2) < \mathbb{E}\left[Z_1^2 \right] \\
&\Leftrightarrow (1 - \sum_{i=2}^{d} \rho(Z_1, Z_i)^2) \left(1 - \frac{\mathbb{E}\left[Z_1 \right]^2}{\mathbb{E}\left[Z_1^2 \right]} \right) < 1.
\end{aligned}
$$

Since Z_1 is not a linear combination of other features, we know that $1 - \sum_{i=2}^{d} \rho(Z_1, Z_i)^2 \in (0, 1]$ and this inequatily is always true. The claim follows with the first part of the proposition.

For the third claim, recall that $\mathbb{E}\left[\xi_1\right] = rm_1^1 + (1-r)m_1^0$. Assume we have two sets of parameters (m_1^0, m_1^1) and $(m_1^{0'}, m_1^{1'})$. If $m_1^0 < m_1^{0'}$ and $m_1^1 \leq m_1^{1'}$, we have

$$\mathbb{E}\left[\xi\right] = rm_1^1 + (1-r)m_1^0 < rm_1^{1'} + (1-r)m_1^{0'} = \mathbb{E}\left[\xi'\right].$$

The same holds true if $m_1^0 \leq m_1^{0'}$ and $m_1^1 < m_1^{1'}$ which shows that the expected share of observed features $\mathbb{E}\left[\xi_1\right]$ is decreasing if and only if we are increasing under-reporting in either (or both) of the groups while leaving everything else fixed. Thus, instead of changes in m_1^g, we argue directly about changes in $\mathbb{E}\left[\xi_1\right]$ in the following.

Denote $S^2 = \sum_{i=2}^d \rho(Z_1, Z_i)^2$ and consider the function

$$f : (0, 1] \to \mathbb{R}$$

$$\mathbb{E}\left[\xi_1\right] = x \mapsto \frac{\mathbb{V}\left[Z_1\right](1 - S^2)}{\mathbb{V}\left[Z_1\right] x(1 - S^2) + (1-x)\mathbb{E}\left[Z_1^2\right]}.$$

We show that f is monotonically increasing from which the claim follows directly. It holds that

$$\frac{d}{dx} f(x) = \frac{-\mathbb{V}\left[Z_1\right](1 - S^2)(\mathbb{V}\left[Z_1\right](1 - S^2) - \mathbb{E}\left[Z_1^2\right])}{\left(\mathbb{V}\left[Z_1\right] x(1 - S^2) + (1-x)\mathbb{E}\left[Z_1^2\right]\right)^2}.$$

Since $1 - S^2 \in [0, 1)$ and $\mathbb{V}\left[Z_1\right] > 0$, the numerator is positive iff

$$-\mathbb{V}\left[Z_1\right](1 - S^2)(\mathbb{V}\left[Z_1\right](1 - S^2) - \mathbb{E}\left[Z_1^2\right]) > 0$$

$$\Leftrightarrow \mathbb{V}\left[Z_1\right](1 - S^2) < \mathbb{E}\left[Z_1^2\right]$$

$$\Leftrightarrow \left(1 - \frac{\mathbb{E}\left[Z_1\right]^2}{\mathbb{E}\left[Z_1^2\right]}\right)(1 - S^2) < 1,$$

which is a true statement. We conclude that f is monotonically increasing in x and the claim follows.

Proposition 5. Given a $k \in [2 : d]$, we know that

$$\hat{\beta}_k = \beta_1 \frac{1}{1 - R^2}\left(\frac{\text{Cov}\left[Z_1, Z_k\right](\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right]^2)}{\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\xi_1\right]}\right) + \beta_k.$$

The first claim is obvious from this expression.

For the second claim, we follow similar steps as for the third claim in the proof of Proposition 4. We know from the previous proof that

$$R^2 = \frac{\mathbb{E}\left[\xi_1\right]^2 \mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1\xi_1\right]} \sum_{i=2}^d \rho(Z_1, Z_i)^2$$

and thus, denoting $S^2 = \sum_{i=2}^d \rho(Z_1, Z_i)^2$,

$$\frac{1}{1 - R^2}\left(\frac{\text{Cov}\left[Z_1, Z_k\right](\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{V}\left[Z_2\right]\mathbb{E}\left[\xi\right]^2)}{\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\xi_1\right]}\right)$$

$$= \frac{\text{Cov}\left[Z_1, Z_k\right]\mathbb{V}\left[\xi_1\right]\mathbb{E}\left[Z_1^2\right]}{(1 - \frac{\mathbb{E}\left[\xi_1\right]^2 \mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1\xi_1\right]} S^2)\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\xi_1\right]}$$

$$= \frac{\text{Cov}\left[Z_1, Z_k\right]\mathbb{V}\left[\xi_1\right]\mathbb{E}\left[Z_1^2\right]}{\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{E}\left[\xi_1\right]^2 \mathbb{V}\left[Z_1\right]S^2\mathbb{V}\left[Z_k\right]}$$

$$= \frac{\text{Cov}\left[Z_1, Z_k\right](\mathbb{E}\left[\xi_1\right] - \mathbb{E}\left[\xi_1\right]^2)\mathbb{E}\left[Z_1^2\right]}{\mathbb{V}\left[Z_k\right](\mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right]^2 + (\mathbb{E}\left[\xi_1\right] - \mathbb{E}\left[\xi_1\right]^2)\mathbb{E}\left[Z_1^2\right]) - \mathbb{E}\left[\xi_1\right]^2 \mathbb{V}\left[Z_1\right]S^2\mathbb{V}\left[Z_k\right]}$$

$$= \frac{\text{Cov}\left[Z_1, Z_k\right](1 - \mathbb{E}\left[\xi_1\right])\mathbb{E}\left[Z_1^2\right]}{(1 - S^2)\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right] + \mathbb{V}\left[Z_k\right](1 - \mathbb{E}\left[\xi_1\right])\mathbb{E}\left[Z_1^2\right]},$$

since $\mathbb{V}\left[Z_1\xi_1\right] = \mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right]^2 + \mathbb{V}\left[\xi_1\right]\mathbb{E}\left[Z_1^2\right]$ and $\mathbb{V}\left[\xi_1\right] = \mathbb{E}\left[\xi_1\right] - \mathbb{E}\left[\xi_1\right]^2$.

Now, consider the function

$$g : (0, 1] \to \mathbb{R}$$

$$\mathbb{E}\left[\xi_1\right] = x \mapsto \frac{\text{Cov}\left[Z_1, Z_k\right](1 - x)\mathbb{E}\left[Z_1^2\right]}{(1 - S^2)\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\right] x + \mathbb{V}\left[Z_k\right](1 - x)\mathbb{E}\left[Z_1^2\right]}.$$

We compute

$$
\begin{aligned}
\frac{d}{dx}g(x) = & \frac{-\mathrm{Cov}\,[Z_1, Z_k]\,\mathbb{E}\left[Z_1^2\right]\left((1-S^2)\mathbb{V}\,[Z_k]\,\mathbb{V}\,[Z_1]\,x + \mathbb{V}\,[Z_k]\,(1-x)\mathbb{E}\left[Z_1^2\right]\right)}{\left((1-S^2)\mathbb{V}\,[Z_k]\,\mathbb{V}\,[Z_1]\,x + \mathbb{V}\,[Z_k]\,(1-x)\mathbb{E}\left[Z_1^2\right]\right)^2} \\
& - \frac{(1-x)\mathrm{Cov}\,[Z_1, Z_k]\,\mathbb{E}\left[Z_1^2\right]\left((1-S^2)\mathbb{V}\,[Z_k]\,\mathbb{V}\,[Z_1] - \mathbb{V}\,[Z_k]\,\mathbb{E}\left[Z_1^2\right]\right)}{\left((1-S^2)\mathbb{V}\,[Z_k]\,\mathbb{V}\,[Z_1]\,x + \mathbb{V}\,[Z_k]\,(1-x)\mathbb{E}\left[Z_1^2\right]\right)^2}.
\end{aligned}
$$

Further,

$$
\begin{aligned}
& \frac{d}{dx}g(x) > 0 \\
\Leftrightarrow & -\mathrm{Cov}\,[Z_1, Z_k]\,\mathbb{E}\left[Z_1^2\right]\left((1-S^2)\mathbb{V}\,[Z_k]\,\mathbb{V}\,[Z_1]\,x + \mathbb{V}\,[Z_k]\,(1-x)\mathbb{E}\left[Z_1^2\right]\right) \\
& > (1-x)\mathrm{Cov}\,[Z_1, Z_k]\,\mathbb{E}\left[Z_1^2\right]\left((1-S^2)\mathbb{V}\,[Z_k]\,\mathbb{V}\,[Z_1] - \mathbb{V}\,[Z_k]\,\mathbb{E}\left[Z_1^2\right]\right) \\
\Leftrightarrow & -\mathrm{Cov}\,[Z_1, Z_k]\,(1-S^2)\mathbb{V}\,[Z_k]\,\mathbb{V}\,[Z_1]\,x > (1-x)\mathrm{Cov}\,[Z_1, Z_k]\,(1-S^2)\mathbb{V}\,[Z_k]\,\mathbb{V}\,[Z_1] \\
\Leftrightarrow & 0 > \mathrm{Cov}\,[Z_1, Z_k]\,.
\end{aligned}
$$

This shows that factor determining the influence of β_1 on $\hat{\beta}_k$ is increasing with decreasing under-reporting if $\mathrm{Cov}\,[Z_1, Z_k] < 0$ and decreasing with decreasing under-reporting otherwise. The claim follows.

Proposition B.2. Predictions are obtained from the model $\hat{Y} = \hat{\alpha} + \hat{\beta}^T X$. Since $Z \sim \mathcal{N}(\mu, \Sigma)$ is jointly Gaussian, we know that

$$
\hat{\beta}^T Z \sim \mathcal{N}\left(\hat{\beta}^T \mu, \hat{\beta}^T \Sigma \hat{\beta}\right) = \mathcal{N}\left(\sum_{i=1}^d \hat{\beta}_i \mu_i, \sum_{i=1}^d \hat{\beta}_i^2 \sigma_i^2 + \sum_{i=1}^d \sum_{j=i+1}^d 2\hat{\beta}_i \hat{\beta}_j \mathrm{Cov}\,[Z_i, Z_j]\right)
$$

and

$$
\hat{\beta}_{[2:d]}^T Z_{[2:d]} \sim \mathcal{N}\left(\hat{\beta}_{[2:d]}^T \mu_{[2:d]}, \hat{\beta}_{[2:d]}^T \Sigma_{[2:d,2:d]} \hat{\beta}_{[2:d]}\right) = \mathcal{N}\left(\sum_{i=2}^d \hat{\beta}_i \mu_i, \sum_{i=2}^d \hat{\beta}_i^2 \sigma_i^2 + \sum_{i=2}^d \sum_{j=i+1}^d 2\hat{\beta}_i \hat{\beta}_j \mathrm{Cov}\,[Z_i, Z_j]\right)
$$

where $\sigma_i^2 = \mathbb{V}\,[Z_i]$ for $i \in [1:d]$.

The cdf of predictions $\hat{Y}$ in group $g \in \{0, 1\}$ can be written as

$$
\begin{aligned}
F_{\hat{Y}|G=g}(x) = & P\left(\hat{\beta}_1 Z_1 \xi_1^g + \hat{\beta}_{[2:d]}^T Z_{[2:d]} \leq x - \hat{\alpha}\right) \\
= & (1 - m_1^g)P\left(\hat{\beta}_{[2:d]}^T Z_{[2:d]} \leq x - \hat{\alpha}\right) + m_1^g P\left(\hat{\beta}^T Z \leq x - \hat{\alpha}\right).
\end{aligned}
$$

Let $C \in [0, 1]$ and denote $\tilde{y} = F_{\hat{Y}}^{-1}(1 - C)$. Without loss of generality, assume that $m_1^0 < m_1^1$. If $m_1^0 = m_1^1$ the selection rate disparity is 0, if $m_1^0 > m_1^1$ the following calculation can easily be adjusted. The inequality $m_1^0 < m_1^1$ means that group 0 has the same or more expected under-reporting in feature Z_1 than group 1. Group 0 is over-selected at threshold C according to Definition 2 if and only if

$$
\begin{aligned}
& 1 - F_{\hat{Y}|G=1}(\tilde{y}) < 1 - F_{\hat{Y}|G=0}(\tilde{y}) \\
\Leftrightarrow & (1 - m_1^1)P\left(\hat{\beta}_{[2:d]}^T Z_{[2:d]} \leq \tilde{y} - \hat{\alpha}\right) + m_1^1 P\left(\hat{\beta}^T Z \leq \tilde{y} - \hat{\alpha}\right) > (1 - m_1^0)P\left(\hat{\beta}_{[2:d]}^T Z_{[2:d]} \leq \tilde{y} - \hat{\alpha}\right) + m_1^0 P\left(\hat{\beta}^T Z \leq \tilde{y} - \hat{\alpha}\right) \\
\Leftrightarrow & (m_1^1 - m_1^0)P\left(\hat{\beta}^T Z \leq \tilde{y} - \hat{\alpha}\right) > (m_1^1 - m_1^0)P\left(\hat{\beta}_{[2:d]}^T Z_{[2:d]} \leq \tilde{y} - \hat{\alpha}\right) \\
\Leftrightarrow & P\left(\hat{\beta}^T Z \leq \tilde{y} - \hat{\alpha}\right) > P\left(\hat{\beta}_{[2:d]}^T Z_{[2:d]} \leq \tilde{y} - \hat{\alpha}\right).
\end{aligned}
$$

Expanding on this in the jointly Gaussian case, we can see that

$$
\begin{aligned}
& P\left(\hat{\beta}^T Z \leq \tilde{y} - \hat{\alpha}\right) > P\left(\hat{\beta}_{[2:d]}^T Z_{[2:d]} \leq \tilde{y} - \hat{\alpha}\right) \\
\Leftrightarrow & \Phi\left(\frac{\tilde{y} - \hat{\alpha} - \hat{\beta}^T \mu}{\sqrt{\hat{\beta}^T \Sigma \hat{\beta}}}\right) > \Phi\left(\frac{\tilde{y} - \hat{\alpha} - \hat{\beta}_{[2:d]}^T \mu_{[2:d]}}{\sqrt{\hat{\beta}_{[2:d]}^T \Sigma_{[2:d,2:d]} \hat{\beta}_{[2:d]}}}\right) \\
\Leftrightarrow & \left(\tilde{y} - \hat{\alpha} - \hat{\beta}^T \mu\right)\sqrt{\hat{\beta}_{[2:d]}^T \Sigma_{[2:d,2:d]} \hat{\beta}_{[2:d]}} > \left(\tilde{y} - \hat{\alpha} - \hat{\beta}_{[2:d]}^T \mu_{[2:d]}\right)\sqrt{\hat{\beta}^T \Sigma \hat{\beta}} \\
\Leftrightarrow & \tilde{y}\left(\sqrt{\hat{\beta}_{[2:d]}^T \Sigma_{[2:d,2:d]} \hat{\beta}_{[2:d]}} - \sqrt{\hat{\beta}^T \Sigma \hat{\beta}}\right) > \left(\hat{\alpha} + \hat{\beta}_{[2:d]}^T \mu_{[2:d]}\right)\left(\sqrt{\hat{\beta}_{[2:d]}^T \Sigma_{[2:d,2:d]} \hat{\beta}_{[2:d]}} - \sqrt{\hat{\beta}^T \Sigma \hat{\beta}}\right) + \left(\hat{\beta}_1 \mu_1\right)\sqrt{\hat{\beta}_{[2:d]}^T \Sigma_{[2:d,2:d]} \hat{\beta}_{[2:d]}}.
\end{aligned}
$$

Here, Φ is the standard normal cdf. If

$$\sqrt{\hat{\beta}_{[2:d]}^T \Sigma_{[2:d,2:d]} \hat{\beta}_{[2:d]}} - \sqrt{\hat{\beta}^T \Sigma \hat{\beta}} > 0,$$

group 0 is over-selected if and only if C implies a threshold $\tilde{y}$ with

$$\tilde{y} > \left(\hat{\alpha} + \hat{\beta}_{[2:d]}^T \mu_{[2:d]}\right) + \left(\hat{\beta}_1 \mu_1\right) \frac{\sqrt{\hat{\beta}_{[2:d]}^T \Sigma_{[2:d,2:d]} \hat{\beta}_{[2:d]}}}{\sqrt{\hat{\beta}_{[2:d]}^T \Sigma_{[2:d,2:d]} \hat{\beta}_{[2:d]}} - \sqrt{\hat{\beta}^T \Sigma \hat{\beta}}}.$$

If

$$\sqrt{\hat{\beta}_{[2:d]}^T \Sigma_{[2:d,2:d]} \hat{\beta}_{[2:d]}} - \sqrt{\hat{\beta}^T \Sigma \hat{\beta}} < 0,$$

group 0 is over-selected if and only if C implies a threshold $\tilde{y}$ with

$$\tilde{y} < \left(\hat{\alpha} + \hat{\beta}_{[2:d]}^T \mu_{[2:d]}\right) + \left(\hat{\beta}_1 \mu_1\right) \frac{\sqrt{\hat{\beta}_{[2:d]}^T \Sigma_{[2:d,2:d]} \hat{\beta}_{[2:d]}}}{\sqrt{\hat{\beta}_{[2:d]}^T \Sigma_{[2:d,2:d]} \hat{\beta}_{[2:d]}} - \sqrt{\hat{\beta}^T \Sigma \hat{\beta}}}.$$

The proposition follows.

Corollary 7. We combine Proposition B.2 with the parameter estimates given in the proof of Proposition 4. For a high threshold $\tilde{y}$, the group with more under-reporting is over-selected if

$$\mathbb{V}\left[\hat{\beta}_{[2:d]}^T Z_{[2:d]}\right] > \mathbb{V}\left[\hat{\beta} Z\right]$$

$$\Leftrightarrow \sum_{i=2}^{d} \hat{\beta}_i^2 \sigma_i^2 + \sum_{i=2}^{d} \sum_{j=i+1}^{d} 2\hat{\beta}_i \hat{\beta}_j \mathrm{Cov}\left[Z_i, Z_j\right] > \sum_{i=1}^{d} \hat{\beta}_i^2 \sigma_i^2 + \sum_{i=1}^{d} \sum_{j=i+1}^{d} 2\hat{\beta}_i \hat{\beta}_j \mathrm{Cov}\left[Z_i, Z_j\right]$$

$$\Leftrightarrow \hat{\beta}_1^2 \sigma_1^2 + \sum_{j=2}^{d} 2\hat{\beta}_1 \hat{\beta}_j \mathrm{Cov}\left[Z_1, Z_j\right] < 0.$$

Recall that

$$R^2 = \frac{\mathbb{E}\left[\xi_1\right]^2 \mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1 \xi_1\right]} \sum_{i=2}^{d} \rho(Z_1, Z_i)^2$$

and denote $S^2 = \sum_{i=2}^{d} \rho(Z_1, Z_i)^2$. Then

$$\frac{1}{1 - R^2} = \frac{\mathbb{V}\left[Z_1 \xi_1\right]}{\mathbb{V}\left[Z_1 \xi_1\right] - \mathbb{E}\left[\xi_1\right]^2 \mathbb{V}\left[Z_1\right] S^2}.$$

Using Proposition 4 and inserting the parameter estimates gives

$$\hat{\beta}_1^2 \sigma_1^2 + \sum_{j=2}^{d} 2\hat{\beta}_1 \hat{\beta}_j \text{Cov}\left[Z_1, Z_j\right] < 0$$

$$\Leftrightarrow \left(\frac{1}{1-R^2}\right)^2 \left(\frac{\mathbb{E}\left[\xi_1\right] \mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1\xi_1\right]} - \frac{R^2}{\mathbb{E}\left[\xi_1\right]}\right)^2 \beta_1^2 \mathbb{V}\left[Z_1\right]$$

$$+ \frac{1}{1-R^2}\left(\frac{\mathbb{E}\left[\xi_1\right] \mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1\xi_1\right]} - \frac{R^2}{\mathbb{E}\left[\xi_1\right]}\right)\beta_1 \sum_{j=2}^{d} 2\left(\beta_1 \frac{1}{1-R^2}\left(\frac{\text{Cov}\left[Z_1, Z_j\right]\left(\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right]^2\right)}{\mathbb{V}\left[Z_j\right]\mathbb{V}\left[Z_1\xi_1\right]}\right) + \beta_j\right)\text{Cov}\left[Z_1, Z_j\right] < 0$$

$$\Leftrightarrow \left(\frac{1}{1-R^2}\right)\left(\frac{\mathbb{E}\left[\xi_1\right] \mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1\xi_1\right]} - \frac{R^2}{\mathbb{E}\left[\xi_1\right]}\right)\beta_1^2 \mathbb{V}\left[Z_1\right]$$

$$+ 2\frac{1}{1-R^2}\beta_1^2 \sum_{j=2}^{d} \left(\frac{\text{Cov}\left[Z_1, Z_j\right]\left(\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right]^2\right)}{\mathbb{V}\left[Z_j\right]\mathbb{V}\left[Z_1\xi_1\right]}\right)\text{Cov}\left[Z_1, Z_j\right] + 2\beta_1 \sum_{j=2}^{d}\beta_j \text{Cov}\left[Z_1, Z_j\right] < 0$$

$$\Leftrightarrow \frac{1}{1-R^2}\left(\frac{\mathbb{E}\left[\xi_1\right] \mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1\xi_1\right]} - \frac{R^2}{\mathbb{E}\left[\xi_1\right]}\right)\beta_1^2 \mathbb{V}\left[Z_1\right]$$

$$+ 2\beta_1^2 \frac{1}{1-R^2}\frac{\left(\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right]^2\right)\mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1\xi_1\right]} \sum_{j=2}^{d}\left(\frac{\text{Cov}\left[Z_1, Z_j\right]^2}{\mathbb{V}\left[Z_j\right]\mathbb{V}\left[Z_1\right]}\right) + 2\beta_1 \sum_{j=2}^{d}\beta_j \text{Cov}\left[Z_1, Z_j\right] < 0$$

$$\Leftrightarrow \frac{\mathbb{E}\left[\xi_1\right] \mathbb{V}\left[Z_1\right]^2}{\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{E}\left[\xi_1\right]^2 \mathbb{V}\left[Z_1\right] S^2}(1-S^2)\beta_1^2 + 2\beta_1^2 \frac{\left(\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right]^2\right)S^2 \mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{E}\left[\xi_1\right]^2 \mathbb{V}\left[Z_1\right] S^2} + 2\beta_1 \sum_{j=2}^{d}\beta_j \text{Cov}\left[Z_1, Z_j\right] < 0$$

$$\Leftrightarrow \beta_1^2 \mathbb{V}\left[Z_1\right] \frac{\mathbb{V}\left[Z_1\right](1-S^2) + 2(1-\mathbb{E}\left[\xi\right])\mathbb{E}\left[Z_1^2\right]S^2}{\mathbb{V}\left[Z_1\right](1-S^2)\mathbb{E}\left[\xi_1\right] + (1-\mathbb{E}\left[\xi_1\right])\mathbb{E}\left[Z_1^2\right]} + 2\beta_1 \sum_{j=2}^{d}\beta_j \text{Cov}\left[Z_1, Z_j\right] < 0$$

where we used that $\mathbb{V}\left[Z_1\xi_1\right] = \mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right]^2 + \mathbb{V}\left[\xi_1\right]\mathbb{E}\left[Z_1^2\right]$ and $\mathbb{V}\left[\xi_1\right] = \mathbb{E}\left[\xi_1\right] - \mathbb{E}\left[\xi_1\right]^2$. Note that the first term on the left side is always positive. Thus the inequality is fulfilled if and only if

$$\text{sign}\left(\beta_1 \sum_{j=2}^{d}\beta_j \text{Cov}\left[Z_1, Z_j\right]\right) = -1$$

and

$$2\beta_1 \sum_{j=2}^{d}\beta_j \text{Cov}\left[Z_1, Z_j\right] < -\beta_1^2 \mathbb{V}\left[Z_1\right] \frac{\mathbb{V}\left[Z_1\right](1-S^2) + 2(1-\mathbb{E}\left[\xi_1\right])\mathbb{E}\left[Z_1^2\right]S^2}{\mathbb{V}\left[Z_1\right](1-S^2)\mathbb{E}\left[\xi_1\right] + (1-\mathbb{E}\left[\xi_1\right])\mathbb{E}\left[Z_1^2\right]}$$

$$\Leftrightarrow \frac{1}{\beta_1} \sum_{j=2}^{d}\beta_j \text{Cov}\left[Z_1, Z_j\right] < -\frac{\mathbb{V}\left[Z_1\right]^2(1-S^2) + 2(1-\mathbb{E}\left[\xi_1\right])\mathbb{E}\left[Z_1^2\right]\mathbb{V}\left[Z_1\right]S^2}{2\mathbb{V}\left[Z_1\right](1-S^2)\mathbb{E}\left[\xi_1\right] + 2(1-\mathbb{E}\left[\xi_1\right])\mathbb{E}\left[Z_1^2\right]}$$

$$\Leftrightarrow \frac{1}{\beta_1} \sum_{j=2}^{d}\beta_j \text{Cov}\left[Z_1, Z_j\right] < -\frac{\mathbb{V}\left[Z_1\right]^2(1-S^2) + 2(1-\mathbb{E}\left[\xi_1\right])(\mathbb{V}\left[Z_1\right] + \mathbb{E}\left[Z_1\right]^2)\mathbb{V}\left[Z_1\right]S^2}{2\mathbb{V}\left[Z_1\right](1-S^2)\mathbb{E}\left[\xi_1\right] + 2(1-\mathbb{E}\left[\xi_1\right])(\mathbb{V}\left[Z_1\right] + \mathbb{E}\left[Z_1\right]^2)}.$$

Since we know that the fraction on the right side is always positive, this can be rewritten as presented in the corollary. If the inequality is not fulfilled, the group with more under-reporting is under-selected at a high threshold.

Lemma 8. In the setting of the Lemma, we can write

$$\mathbb{E}_\xi \left[l(f, X, y)\right] = \mathbb{E}_\xi \left[\frac{1}{m}l(f, X, y) - \frac{1-m}{m}l(f, [0, X_{[2:d]}]^T, y)\right]$$

$$= \frac{1}{m}\mathbb{E}_\xi \left[l(f, X, y)\right] - \frac{1-m}{m}l(f, [0, z_{[2:d]}]^T, y)$$

$$= \frac{1}{m}\left(P(\xi_1 = 1)l(f, z, y) + P(\xi_1 = 0)l(f, [0, z_{[2:d]}]^T, y)\right) - \frac{1-m}{m}l(f, [0, z_{[2:d]}]^T, y)$$

$$= l(f, z, y) + \frac{1-m}{m}l(f, [0, z_{[2:d]}]^T, y) - \frac{1-m}{m}l(f, [0, z_{[2:d]}]^T, y)$$

$$= l(f, z, y).$$

Here, the first equality holds since only the first feature has under-reporting and the second equality holds because $Z \perp G$ which implies $Z \perp \xi$.

Lemma B.3. Follows the same as Lemma 8. Instead of under-reporting completely at random, the under-reporting is completely at random within group g.

Lemma 9. Since f is linear and under-repoting only occurs in the first feature, the expected prediction error for an imputation value x_1' can be written as

$$R(f) = \mathbb{E}_X \left[(f(X') - Y)^2\right]$$

$$= \beta_1^2 \mathbb{E}_X \left[(X_1' - Z_1)^2\right]$$

$$= \beta_1^2 P(X_1 = 0)\mathbb{E}_X \left[(X_1' - Z_1)^2 \mid X_1 = 0\right] + \beta_1^2 P(X_1 \neq 0)\mathbb{E}_X \left[(X_1' - Z_1)^2 \mid X_1 \neq 0\right]$$

$$= \beta_1^2 P(X_1 = 0)\mathbb{E}_Z \left[(x_1' - Z_1)^2 \mid X_1 = 0\right] + \beta_1^2 P(X_1 \neq 0)\mathbb{E}_Z \left[(Z_1 - Z_1)^2 \mid X_1 \neq 0\right]$$

$$= \beta_1^2 P(X_1 = 0)\left(x_1'^2 - 2x_1'\mathbb{E}\left[Z_1 \mid X_1 = 0\right] + \mathbb{E}\left[Z_1^2 \mid X_1 = 0\right]\right).$$

Then

$$\frac{dR(f)}{dx_1'} = \beta_1^2 P(X_1 = 0)(2x_1' - 2\mathbb{E}\left[Z_1 \mid X_1 = 0\right]) \stackrel{!}{=} 0$$

$$\Leftrightarrow x_1' = \mathbb{E}\left[Z_1 \mid X_1 = 0\right].$$

We implicitly assume that $\beta_1 \neq 0$ and the probability of 0-entries is positive.

Lemma B.4. Recall that $G \sim \text{Bern}(r)$. Similar to the proof of Lemma 9, the expected prediction error can be written as

$$R(f) = \mathbb{E}_X \left[(f(X') - Y)^2\right]$$

$$= \beta_1^2 P(X_1 = 0)\mathbb{E}_Z \left[(X_1' - Z_1)^2 \mid X_1 = 0\right]$$

$$= \beta_1^2 P(X_1 = 0)\left(r\mathbb{E}_Z \left[(X_1' - Z_1)^2 \mid X_1 = 0, G = 1\right] + (1-r)\mathbb{E}_Z \left[(X_1' - Z_1)^2 \mid X_1 = 0, G = 0\right]\right)$$

$$= \beta_1^2 P(X_1 = 0)\left(r\mathbb{E}_Z \left[(x_1'^1 - Z_1)^2 \mid X_1 = 0, G = 1\right] + (1-r)\mathbb{E}_Z \left[(x_1'^0 - Z_1)^2 \mid X_1 = 0, G = 0\right]\right).$$

This prediction error is minimal when

$$\frac{dR(f)}{dx_1'^g} = \beta_1^2 P(X_1 = 0)P(G = g)(2x_1'^g - 2\mathbb{E}\left[Z_1 \mid X_1 = 0, G = g\right]) \stackrel{!}{=} 0$$

$$\Leftrightarrow x_1'^g = \mathbb{E}\left[Z_1 \mid X_1 = 0, G = g\right]$$

for $g \in \{0, 1\}$.

E SUPPLEMENTARY FIGURES

Name	#Obs.	#Feat.	Groups	Binary outcomes
COMPAS [6]	7,214	6	Race (51% African-American, 49% other), Gender (81% male, 19% female)	Two-year recidivism, violent recidivism
German credit [52]	1,000	19	Gender (69% male, 31% female)	Good credit
ACS Income (CA, 2018) [21]	195,665	6	Race (62% White, 38% other), Gender (53% male, 47% female)	Yearly income over $50,000
Birth data	39,365	51	Medicaid (no 72%, yes 28%), Race (African-American 21%, other 79%)	Child placed in foster care within 3 years

Table E.2: Statistics of the datasets used in experiments. Data is split randomly into 80% for training and 20% for testing. For the first three datasets, we iterate over all outcome types, groups, and numerical features for missingness injection.
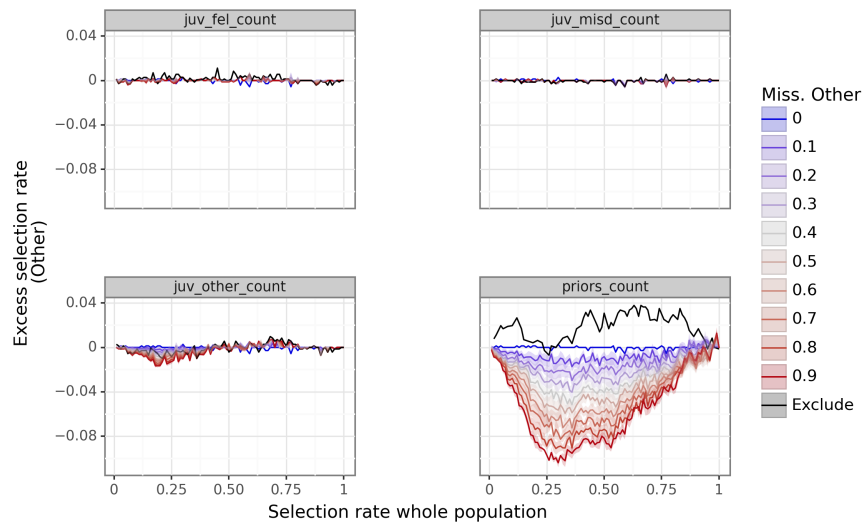
Figure E.1: Excess selection rate of racial group Other (i.e. not African-American) at different selection rates of the whole population with synthetic two-year recidivism outcomes using the COMPAS dataset. Each panel represents a feature that has been corrupted by under-reporting in independent runs of the experiment. Feature under-reporting is added to the Other group with 0-90% missing in 10 percentage point increments. The black curves show performance when excluding the whole feature column from modeling. Results are reported as averages over 30 runs on a test set. Shaded areas correspond to one standard deviation in each direction of the mean.
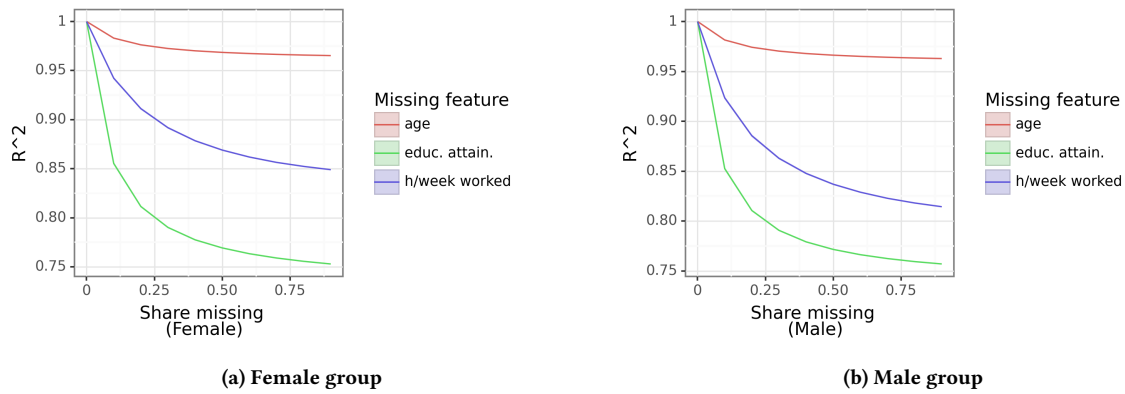
(a) Female group

(b) Male group

Figure E.2: Test set R^2 over varying levels of feature under-reporting in the ACS Income dataset. Results are reported as averages over 50 runs on the test set. Variation in results was minimal.
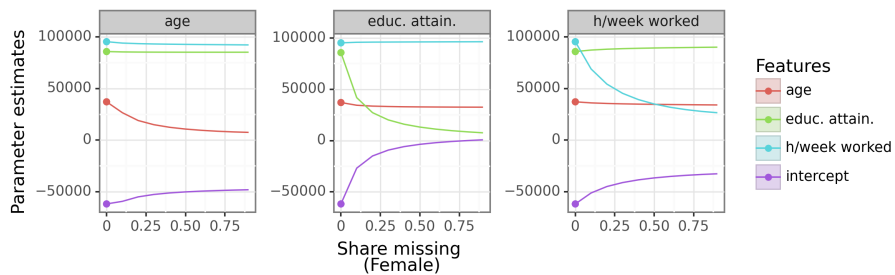

Figure E.3: Parameter estimates over varying levels of feature under-reporting in the ACS Income dataset. Each panel indicates a different feature selected for under-reporting injection. Points indicate the true parameters from the semi-synthetic ground truth model. Results are reported as averages over 50 runs. Variation in estimates over runs was minimal. Note that only estimates for continuous features are displayed and the estimated parameters for the levels of the categorical variables worker class, marital status, and relationship to reference person are omitted for readability.
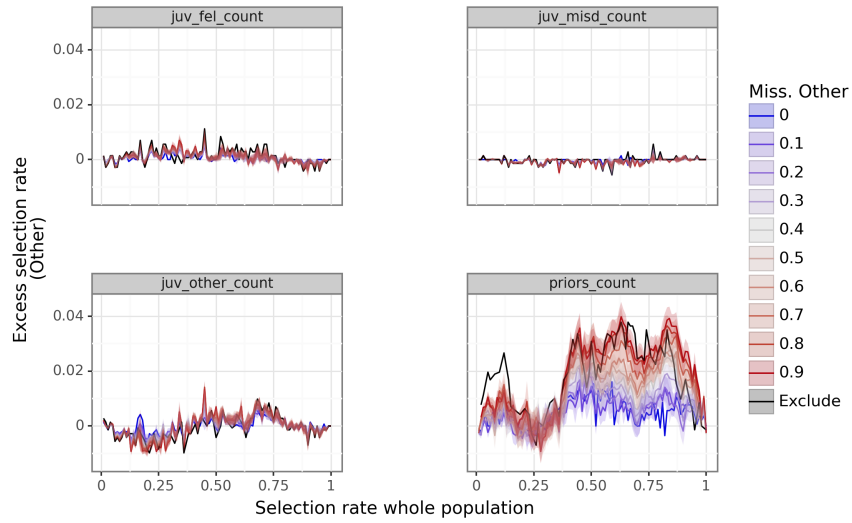
Figure E.4: Multiple imputation excess selection rate of racial group Other (i.e. not African-American) at different selection rates of the whole population with synthetic two-year recidivism outcomes using the COMPAS dataset. Results are reported as averages over 30 runs on a test set. Shaded areas correspond to one standard deviation in each direction of the mean.
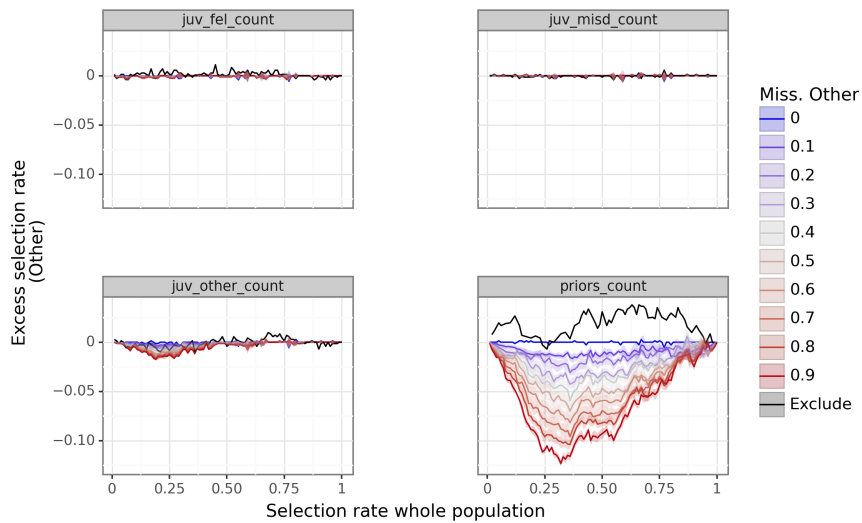

Figure E.5: Excess selection rate of racial group Other (i.e. not African-American) when training on rows without 0-entries at different selection rates of the whole population with synthetic two-year recidivism outcomes using the COMPAS dataset. Results are reported as averages over 30 runs on a test set. Shaded areas correspond to one standard deviation in each direction of the mean.
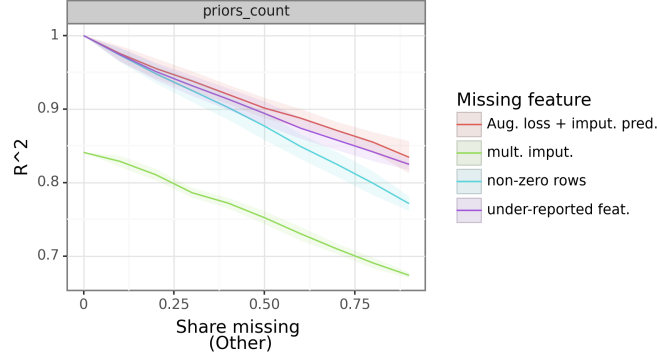
Figure E.6: Test set R^2 of different solution approaches using the COMPAS dataset with synthetic two-year recidivism outcomes. Under-reporting is injected into the feature 'priors count' of group Other (i.e. not African-American). Results are reported as averages over 30 runs. Shaded areas correspond to one standard deviation.

F UNDER-REPORTING RATE ESTIMATION

F.1 Estimation procedure

We draw on PU-learning literature to estimate under-reporting rates. Elkan and Noto [24] assume a classification setting with a latent indicator $s \in \{0, 1\}$ that encodes whether an example is labeled or not. Only positive examples are labeled, i.e. $s = 1$ implies that $y = 1$, but when the example is unlabeled $s = 0$, we don't know if $y = 0$ or $y = 1$. In our case, the outcome y translates to the indicator $\mathbb{1}(Z_1 \neq 0)$ while the labeling indicator s translates to $\mathbb{1}(X_1 \neq 0)$. If $X_1 \neq 0$, we know that the example is not under-reported and $Z_1 \neq 0$. But, if $X_1 = 0$, we don't have any information about the value of $\mathbb{1}(Z_1 \neq 0)$ because the example may not be 'labeled'. Similar to the assumption in [24], our setting fulfills an under-reported completely at random assumption which can be expressed as

$$P(X_1 \neq 0 \mid X_{[2:d]}, Y, Z_1 \neq 0) = P(X_1 \neq 0 \mid Z_1 \neq 0).$$

In our analysis, this assumption is fulfilled either over the whole population or within the group considered for estimation of m_g. Given this notation, the rate of correctly observed feature entries m can be written as

$$m = P(\xi_1 \neq 0) = P(\xi_1 Z_1 \neq 0 \mid Z_1 \neq 0) = P(X_1 \neq 0 \mid Z_1 \neq 0).$$

This implies that we can estimate m without having to consider correctly recorded 0-entries in the feature vector. As described in the main text, we assume access to a data set $V = \{(x, y)_{i=1}^{n}\}$ that is split into a training portion V_{train} and evaluation portion V_{eval}. Let P_{eval} denote the subset of examples from V_{eval} for which $x_1 \neq 0$. We now fit a model h on V_{train} to estimate $P(X_1 \neq 0 \mid X_{[2:d]}, Y)$ and evaluate h on P_{eval}. The estimator for the share of observed values m is given by

$$\hat{m} = \frac{1}{|P_{\text{eval}}|} \sum_{(x,y) \in P_{\text{eval}}} h(x_{[2:d]}, y).$$

Assuming $h(x_{[2:d]}, y) = P(X_1 \neq 0 \mid X_{[2:d]} = x_{[2:d]}, Y = y)$, i.e. no error is introduced through the estimation of h, this provides an unbiased estimate of m. To see this, we show that $h(x_{[2:d]}, y) = m$ for all $(x, y) \in P_{\text{eval}}$. We can write

$$
\begin{aligned}
h(x_{[2:d]}, y) =& P(X_1 \neq 0 \mid X_{[2:d]} = x_{[2:d]}, Y = y) \\
=& P(X_1 \neq 0 \mid X_{[2:d]} = x_{[2:d]}, Y = y, Z_1 \neq 0) P(Z_1 \neq 0 \mid X_{[2:d]} = x_{[2:d]}, Y = y) \\
& + P(X_1 \neq 0 \mid X_{[2:d]} = x_{[2:d]}, Y = y, Z_1 = 0) P(Z_1 = 0 \mid X_{[2:d]} = x_{[2:d]}, Y = y) \\
=& P(X_1 \neq 0 \mid X_{[2:d]} = x_{[2:d]}, Y = y, Z_1 \neq 0) \\
=& P(X_1 \neq 0 \mid Z_1 \neq 0) = m.
\end{aligned}
$$

Here, the third equality follows because $(x, y) \in P_{\text{eval}}$.

F.2 Estimation results

We estimate under-reporting rates using the procedure described above where V is taken to be the 80% training data fold conditioned on the group with under-reporting. Half of the data is used for training of h while the other half is used as evaluation data to compute $\hat{m}_g$. As model class for h, we use XGBoost classifiers with 100 trees of maximum depth 3, and learning rate 0.1. Figure F.7 depicts the results of the estimation procedure for under-reporting in the feature priors count for the racial group Other and synthetic two-year recidivism outcomes

using the COMPAS dataset. We see that estimation of m_g works particularly well when under-reporting in the feature is high. When the feature is fully observed, i.e. $m_g = 1$, the estimator returns $\hat{m}_g = 0.811$ on average. We hypothesize that this is due to the high share of true 0-entries occurring in the priors count feature which impacts the estimator more when fewer values are missing due to under-reporting.
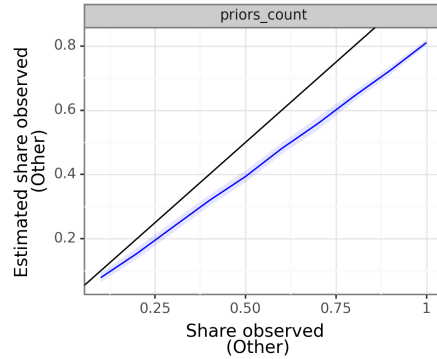

Figure F.7: True vs. estimated observed rate (i.e. 1 - under-reporting rate) of feature priors count in racial group Other (i.e. not African-American) with sythetic two-year recidivism outcomes using the COMPAS dataset. Results are reported as averages over 30 runs. Shaded areas correspond to one standard deviation in each direction of the mean. The black line shows $y = x$ for comparison.

G ADDITIONAL EXPERIMENTS AND RESULTS
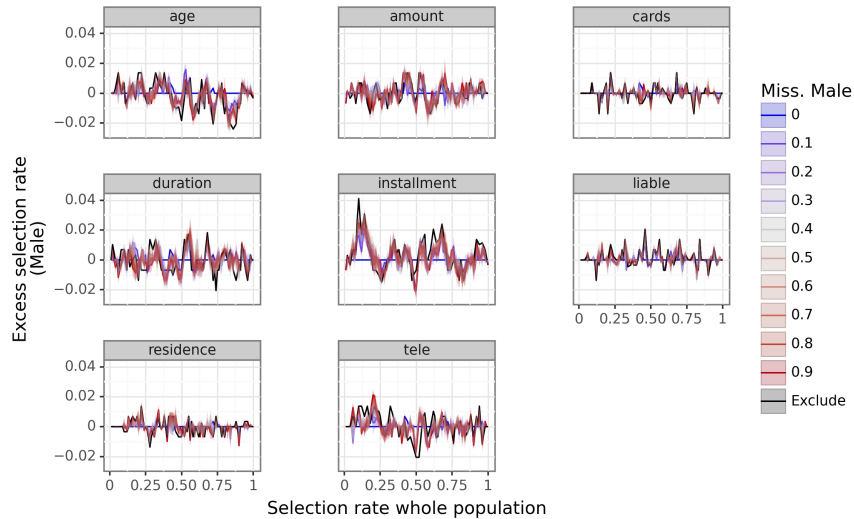
G.1 German credit data

Figure G.8: Excess selection rate of male group at different selection rates of the whole population with synthetic outcomes using the German credit dataset. Each panel represents a feature that has been corrupted by under-reporting in independent runs of the experiment. Feature under-reporting is added to the male group with 0-90% missing in 10 percentage point increments. The black curves show performance when excluding the whole feature column from modeling. Results are reported as averages over 50 runs on a test set. Shaded areas correspond to one standard deviation in each direction of the mean. Note that feature under-reporting is only injected into continuous features and models are estimated using the displayed features as well as the available categorical features checking account status, credit history, purpose, savings, employment, marital status, type of owned property, other installment plans, housing type, and job type.

Our experiments suggest that addition of feature under-reporting to one of the two gender groups in the German credit dataset has only marginal fairness implications. Figure G.8 depicts the results for synthetic outcomes and addition of different amounts of under-reporting

to the features of the male group. We can see that, for any of the considered features, the amount of under-reporting injected has little to no effect on the excess selection rate of the male group. However, when selecting rates of around 10-15% from the whole population *any* amount of under-reporting in the installment feature appears to results in a slight over-selection of the male group. The installment feature in the German credit dataset is discretized into four values with lower values indicating a higher installment rate. Incorrectly observed 0-values may thus suggest a high installment rate which is indicative of good credit.

G.2 Beyond the noise-free setting

Motivation. The experiments on the publicly available datasets discussed in Sections 7 and 8 rely on semi-synthetic outcomes that are computed as deterministic linear functions of correctly measured features. The implicit simplifying assumptions are that, without feature under-reporting, a linear model on the data can retrieve the true data generating model $f(X) = \alpha + \beta^T Z$ and the exact outcomes Y as recorded in the data. This modeling choice facilitates isolation of the effect of feature under-reporting by explicitly excluding potential effects of model misspecification and regression noise. In real-life applications, we can generally not predict outcomes exactly even if correctly measured features are available. In the following additional set of experiments, we loosen the assumption of a noise-free regression setting to allow for more general settings.

Experimental setup. We follow a similar experimental setup as described in Section 7.1. Instead of relying on noise-free regression labels $Y = \alpha + \beta^T Z$, we add some noise back into the system by setting

$$Y = \alpha + \beta^T Z + \varepsilon.$$

Here, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is i.i.d. and assumed to have mean zero. Like before, fitting a linear regression of Y on features Z with sufficient data yields the correct parameter estimates $\hat{\beta} = \beta$ and $\hat{\alpha} = \alpha$. However, in contrast to the noise-free setting, the prediction model $\hat{Y}_Z = \alpha + \beta^T Z$ can only retrieve labels Y up to random noise. The R^2 of this prediction model can be controlled via the variance σ^2 by setting

$$\sigma^2 = \frac{1 - R^2}{R^2} \mathbb{E}\left[\left((\alpha + \beta^T Z) - \mathbb{E}\left[\alpha + \beta^T Z\right] \right)^2 \right].$$

We experiment with R^2 values between 0.1 and 1.0 in increments of 0.1. Instead of comparing predictions $\hat{Y}_Z$ to predictions under feature under-reporting $\hat{Y}_X$, we compare thresholded versions of outcomes Y and $\hat{Y}_X$ directly to measure both the impact of under-reporting and regression noise.

Results. Figure G.9 depicts a subset of the results for the COMPAS dataset. Comparing against the results of the noise-free setting summarized in Figure E.1, we observe that the group Other is under-selected to a greater extend with additional noise. Under-selection occurs even if no feature under-reporting is added (dark blue curves) and increases with increasing noise, i.e. decreasing R^2 of the model on Z. On a high level, this occurs because the predictions $\hat{Y}_Z$ concentrate more closely around their group-level means as compared to the true values Y. The mean of $\hat{Y}_Z$ is smaller for the group Other than the group African-American which leads to under-selection of the group Other as compared to the true Y at many thresholds. We note that the group-level variances in outcomes Y and predictions $\hat{Y}_Z$ play a role in this dynamic as well. The isolated effect of feature under-reporting in the studied setting appears to be similar to the effect in the noise-free setting. As under-reporting is introduced into the group Other via the feature 'priors count', the group Other is further under-selected. The more under-reporting is injected, the more the group is under-selected. The magnitude of under-selection due to under-reporting is comparable across different levels of regression noise. Overall, the results give us some insight into what to expect in more realistic settings of feature under-reporting. Instead of selection rate disparities that are exclusively due to differential feature under-reporting, disparities in the studied setting also depend on regression noise which, together, leads to increased disparities overall.

G.3 Possibility of decreasing disparities

We conduct our main experiments on three publicly available datasets, i.e. COMPAS data [6], German credit data [52], and ACS Income data [21], where each numerical feature is considered for the effect of under-reporting. As discussed in Section 8, the results suggest that, if an effect is present, feature under-reporting generally leads to under-selection of the group with under-reporting which aligns with Case 2 from the theoretical derivations in Section 5. If the group with under-reporting aligns with the group that is less frequently selected in the ground truth model, this implies that differential feature under-reporting leads to increased selection rate disparities.

All three datasets have a numerical age feature which was considered for under-reporting but omitted for the discussion of results in the main text. In contrast to most other features (e.g. the counts in the COMPAS data), the default value of 0 is somewhat unintuitive for age and lies outside of the feature's support in each of the datasets. Studying the effect of fitting a model on differentially available data directly is less compelling in this setting since we essentially have indicators for under-reporting and could hope to use missing data methods like imputation directly. Nevertheless, we discuss the results for feature under-reporting in age for the COMPAS dataset in the following as it presents the only empirical example for decreasing disparities we encounter in our experiments.

Figure G.10 depicts the parameter estimates and excess selection rate of group Other (i.e. not African-American) when fitting a model on data with feature under-reporting in the feature 'age' for group Other. We see that under-reporting in this setting leads to over-selection of the group with under-reporting. This over-selection is increasing with increasing levels of under-reporting. As the figure shows, the regression
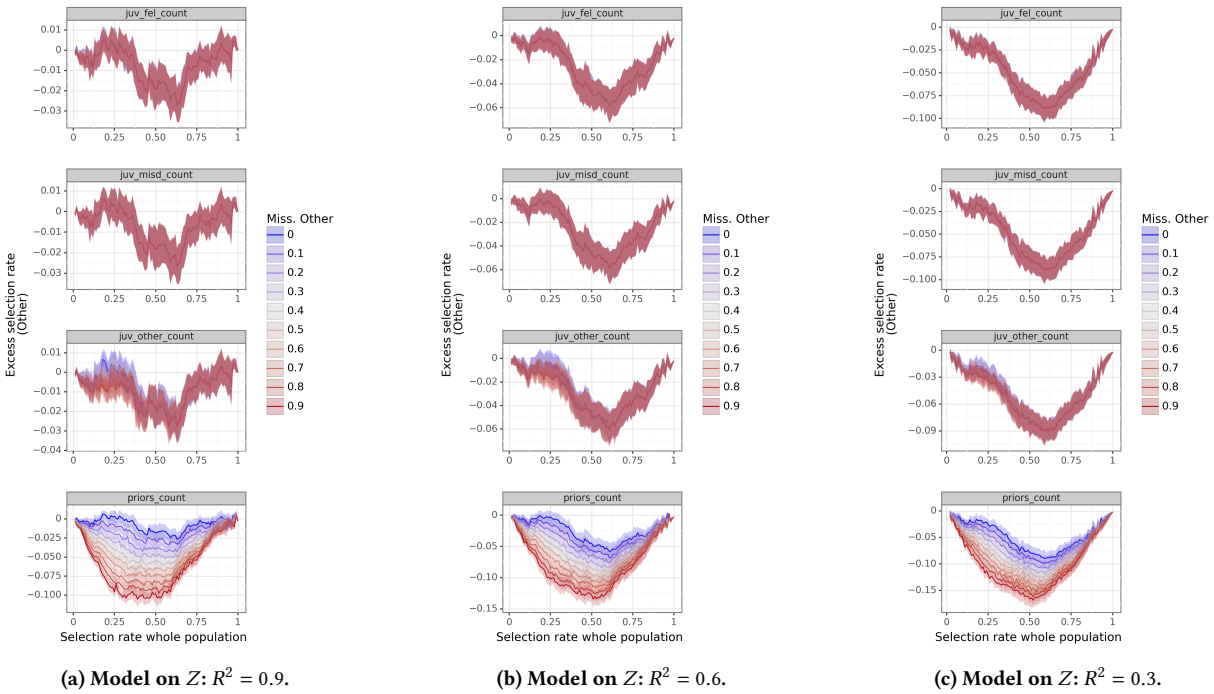
(a) Model on Z: $R^2 = 0.9$.

(b) Model on Z: $R^2 = 0.6$.

(c) Model on Z: $R^2 = 0.3$.

Figure G.9: Excess selection rate under under-reporting over true labels Y with different levels of R^2 for the model on correctly measured features Z. Low R^2 indicates a high level of noise and vice versa. Under-reporting injected into the features of group Other (i.e. not African-American) in the COMPAS dataset. Results are reported as averages over 30 simulation runs with shaded areas representing one standard deviation in each direction.
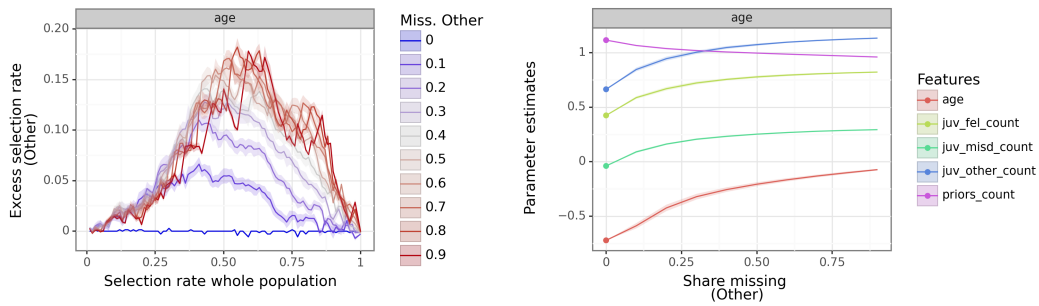

Figure G.10: Excess selection rate of group Other (i.e. not African-American) at different population selection rates with synthetic outcomes using the COMPAS dataset, and the respective parameter estimates. Under-reporting is added to the feature 'age' in group Other. Results are reported as averages over 30 simulation runs with shaded areas representing one standard deviation in each direction. Note that parameter estimates are only displayed for continuous count features and age to preserve readability. The models additionally take sex and the categorical feature charge degree into consideration.

parameter for age is negative with an attenuation effect when under-reporting is injected. This means that in the semi-synthetic ground truth model and in the prediction models under under-reporting younger defendants are more likely to reoffend than older defendants. The feature correlations between age and juvenile crime counts (felony, misdemeanor, and other) are negative in the data while the correlation between age and the feature 'priors count' is positive. This leads to parameter estimates that are increasing for increasing under-reporting in age for juvenile crime counts and decreasing for increasing under-reporting for priors count exactly as predicted by the theoretical analyses in Proposition 5. Ultimately, this example shows how, in some settings, disparities may decrease as a function of under-reporting which aligns with Case 1 from the theoretical discussions in Section 5. However, the example presented here is somewhat artificial and we find that typically disparities are increasing with differential feature under-reporting.