Regression with race-modifiers: towards equity and interpretability

Daniel R. Kowal[a,1]

[a]Department of Statistics, Rice University, Houston, TX 77005

The pervasive effects of structural racism and racial discrimination are well-established and offer strong evidence that the effects of many important variables on health and life outcomes vary by race. Alarmingly, standard practices for statistical regression analysis introduce racial biases into the estimation and presentation of these race-modified effects. We introduce *abundance-based constraints* (ABCs) to eliminate these racial biases. ABCs offer a remarkable invariance property: estimates and inference for main effects are nearly unchanged by the inclusion of race-modifiers. Thus, quantitative researchers can estimate race-specific effects "for free"—without sacrificing parameter interpretability, equitability, or statistical efficiency. The benefits extend to prominent statistical learning techniques, especially regularization and selection. We leverage these tools to estimate the joint effects of environmental, social, and other factors on 4th end-of-grade readings scores for students in North Carolina ($n = 27,638$) and identify race-modified effects for racial (residential) isolation, $PM_{2.5}$ exposure, and mother's age at birth.

DRAFT

Health and life outcomes are inextricably linked to race (1, 2). Racial disparities exist in birth outcomes, mortality, disease onset and progression, socioeconomic status, and police-involved deaths, along with many other health and life outcomes (2–4). These disparities persist even after adjusting for socioeconomic status and occur through multiple pathways (1). Structural racism contributes to significant differences in the quality of education, housing, employment opportunities, accumulation of wealth, access to medical care, and treatment in the criminal justice system (1, 2, 5–7). Perceived racial discrimination impacts both mental and physical health through heightened stress responses, health behaviors, and traumatic experiences (8, 9). Thus, rigorous studies of health and life outcomes must carefully consider race as a primary factor.

That race permeates so many aspects of an individual's life course is a strong indicator that the effects of important factors (X) on health and life outcomes (Y) may be *race-specific* (10). Regression analysis—the primary statistical tool to quantify how these covariates X determine, predict, or associate with an outcome Y—must therefore consider *race-modifiers* for X. Indeed, there is abundant and growing evidence for race-specific modifiers, including the effects of red-lining, $PM_{2.5}$ exposure, and cigarette use on mortality risks (11–13); maternal age, poverty, education, and hypertension on infant birthweight, infant mortality, and maternal stroke risk (14–16); education level on multiple health outcomes (17); mood/anxiety disorder on chronic physical health conditions (18); perceived racism on mental health (9); age on allostatic load scores, known as "racial weathering" (19); and the timing of hypertension, insulin resistance, or diabetes onset (20, 21), among many others (22–26). The identification and quantification of race-modified effects are essential to understand and eliminate harmful race disparities in health and life outcomes (27).

[1]To whom correspondence should be addressed. E-mail: daniel.kowal@rice.edu

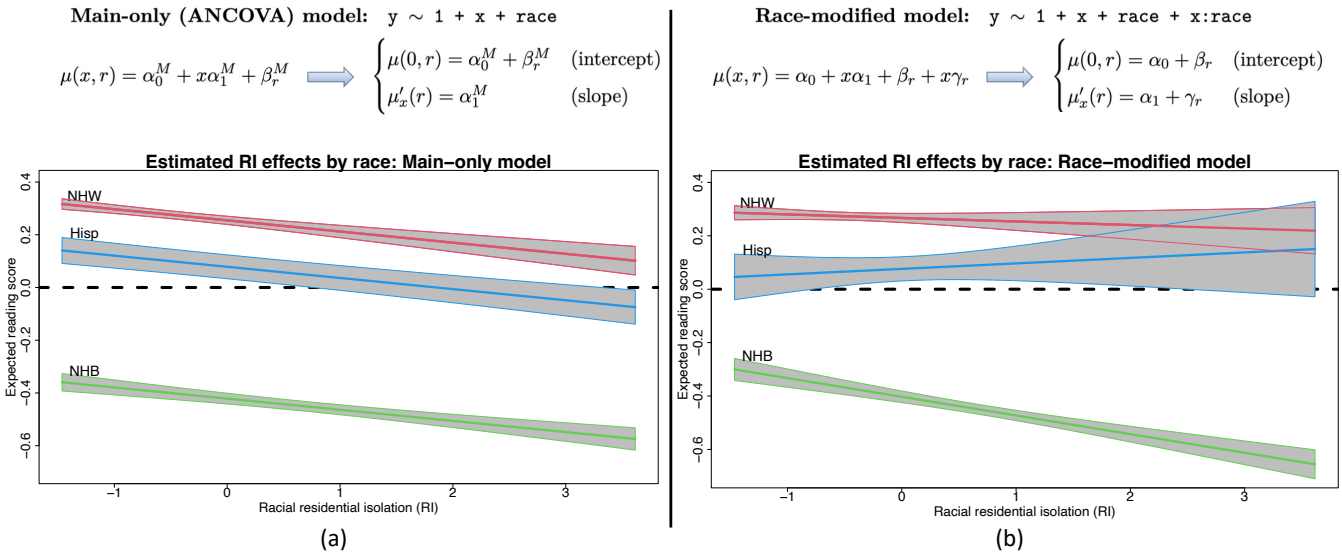

Fig. 1. Linear regression models for an outcome variable Y with a continuous covariate X and a categorical (or nominal) covariate `race`. The models parameterize the expected outcome, $\mathbb{E}(Y \mid X = x, \texttt{race} = r) = \mu(x,r)$, with corresponding x-effect (or slope) $\mu'_x(r) := \mu(x+1,r) - \mu(x,r)$. (a) The *main-only* model assumes a global (race-invariant) x-effect. (b) The *race-modified* model allows for race-specific x-effects. Fit to 4th end-of-grade reading scores in North Carolina, the main-only model obscures important race-specific differences in the effects of racial isolation (RI) that are uncovered by the race-modified model. The negative RI effect observed globally in (a) is driven by the negative RI effect *for non-Hispanic Black (NHB) students*, which does not persist for Hispanic (Hisp) or non-Hispanic White (NHW) students in (b).

DRAFT

To show the impact of race-modified effects, we present a regression of 4th end-of-grade reading scores on racial (residential) isolation (RI) and race (Figure 1). The dataset, detailed and reanalyzed subsequently, includes $n = 27,638$ students in North Carolina (58% non-Hispanic (NH) White, 36% NH Black, 6% Hispanic). RI measures the geographic separation of NH Black individuals and communities from other race groups, and thus is an important measure of structural racism (5, 6, 28–30). The *main-only* (or ANCOVA) model (Figure 1a) includes race only as an additive effect, which restricts the RI effect to be common across race groups. This widely-used model reports the same adverse effect of RI on reading scores for all students. However, the *race-modified* model (Figure 1b) provides the essential context: the RI effect is significantly negative for NH Black students, but not for other race groups. Thus, a race-modified model is necessary to uncover and quantify these racial discrepancies in the effects of structural racism on educational outcomes.

Despite these benefits, there are significant racial biases that occur in commonplace estimation, inference, and presentation of results for regression analysis with race as a covariate. Both the main-only and race-modified models (Figure 1) are overparametrized: neither $\{\alpha_0, \beta_r\}_r$ nor $\{\alpha_1, \gamma_r\}_r$ are identifiable without further constraints. Any constant could be added to α_0 and subtracted from each $\{\beta_r\}$, and similarly for α_1 and $\{\gamma_r\}$, which alters each parameter but leaves the model unchanged. Critically, neither the main nor the race-specific parameters can be estimated or interpreted without additional constraints. Undoubtedly, the most common approach is *reference group encoding* (RGE): a reference group is selected, typically NH White, and removed ($\beta_{\texttt{NHW}} = \gamma_{\texttt{NHW}} = 0$). This is the default

for all major statistical software implementations of (generalized) linear regression, including `R`, SAS, Python, MATLAB, and Stata. However, RGE output is racially biased (31), difficult to interpret, and obscures important main and race-modified effects. We categorize these significant limitations into *presentation bias* and *statistical bias*.

Presentation bias. Table 1 (left) displays standard output for a race-modified model. Under RGE, the RI effect (red) actually refers to the RI effect *only for NH White individuals*, $\alpha_1 = \alpha_1 + \gamma_{\texttt{NHW}} = \mu'_x(\texttt{NHW})$. Similarly, Intercept refers to the NH White intercept, $\alpha_0 = \alpha_0 + \beta_{\texttt{NHW}} = \mu(0, \texttt{NHW})$. We emphasize that the presentation format in Table 1 (left) is predominant in scientific journals. Among recent publications in social science journals, it was found that 92% of such tables used NH White as the reference group, while less than half explicitly stated the reference group (31).

First, this output is *inequitable*: it elevates a single race group above others. Further, all other race-specific effects are presented relative to the NH White group. For instance, `RI:NH Black` refers to the difference between the RI effects for NH Black students and NH White students: $\gamma_{\texttt{NHB}} = \mu'_x(\texttt{NHB}) - \alpha_1 = \mu'_x(\texttt{NHB}) - \mu'_x(\texttt{NHW})$. This framing presents NH White as "normal" and other race groups as "deviations from normal", which is known to bias interpretations of results (32). Second, this output is *unclear*: it is nowhere indicated that the intercept and RI effects are specific to NH White students. A cursory inspection of this output might result in a mistaken interpretation of the RI effect as a global effect, rather than a NH White effect. Finally, this output is *misleading*: the RI effect is reported to be small and insignificant, despite clear evidence to the contrary (Figure 1). Under RGE, the addition of the race-modifier substantially alters the estimates and reduces the statistical power for the RI main effect (α_1).

DRAFT

Table 1. Linear regression output under default reference group encoding (RGE; left) and *abundance-based constraints* (ABCs; right): race-modified effects of racial isolation (RI) on 4th end-of-grade reading scores for students in North Carolina (`y ~ 1 + RI + race + RI:race`).

Default: reference group encoding (RGE)

Variable	Estimate (SE)	p-value
Intercept	0.267 (0.009)	<0.001
Racial isolation (RI)	-0.013 (0.011)	0.220
Mother's race		
NH Black	-0.670 (0.014)	<0.001
Hispanic	-0.191 (0.025)	<0.001
RI × Mother's race		
RI:NH Black	-0.057 (0.014)	<0.001
RI:Hispanic	0.034 (0.027)	0.210

Proposed: abundance-based constraints (ABCs)

Variable	Estimate (SE)	p-value
Intercept	0.014 (0.007)	0.038
Racial isolation (RI)	-0.032 (0.007)	<0.001
Mother's race		
NH White	0.253 (0.006)	<0.001
NH Black	-0.417 (0.009)	<0.001
Hispanic	0.062 (0.023)	0.006
RI × Mother's race		
RI:NH White	0.018 (0.006)	0.001
RI:NH Black	-0.038 (0.009)	<0.001
RI:Hispanic	0.052 (0.024)	0.031

The default regression output (RGE, left) induces *presentation bias:* the estimated Intercept and RI effect (red) refer to the NH White group. This is inequitable, unclear, and misleading: it obfuscates the highly significant and detrimental effects of RI on reading scores (for the main-only model, $\hat{\alpha}_1^M = -0.042, \text{SE}(\hat{\alpha}_1^M) = 0.007, p < 0.001$). The regression output under ABCs (right) eliminates presentation bias, confirms the estimated RI effect and standard errors from the main-only model (blue), and clearly highlights and quantifies the critical result that the adverse RI effect is more than doubled for NH Black students ($\hat{\mu}'_{\texttt{RI}}(\texttt{NHB}) = \hat{\alpha}_{\texttt{RI}} + \hat{\gamma}_{\texttt{RI:NHB}} = -0.032 + -0.038 = -0.070$; Figure 1b).

Statistical bias. The racial inequity in RGE also permeates statistical estimation and inference. Modern statistical learning commonly features penalized regression, variable selection, and Bayesian inference (33). Broadly, these *regularization* strategies seek to stabilize (i.e., reduce the variance of) estimators, typically by "shrinking" coefficients toward zero. This approach is particularly useful in the presence of a moderate to large number of covariates that may be correlated. However, under RGE, shrinking or setting coefficients to zero introduces racial bias to the estimation. Critically, shrinkage or selection of the race-specific terms, $\gamma_r \to 0$, does not innocuously shrink toward a global slope; rather, it implies that the coefficient on x for race r is pulled toward that of the NH White group, $\mu'_x(r) = \alpha_1 + \gamma_r \to \alpha_1$, and $\alpha_1 = \mu'_x(\texttt{NHW})$. Not only is this estimator racially biased, but also it attenuates the estimated differences between the x-effects for each race and NH White individuals. Identification and quantification of such race-modified effects are precisely the goals of race-modified models. Furthermore, RGE cannot distinguish between shrinkage toward a global, race-invariant x-effect and shrinkage toward the NH White x-effect: both require $\gamma_r \to 0$ for all r. A fundamental goal of penalized estimation and selection in this context is to remove unnecessary race-modifiers. However, with RGE, the cost is racial bias in the shrinkage and selection. Thus, default RGE cannot fully and equitably leverage the state-of-the-art in statistical learning.

Although RGE is used in the overwhelming majority of regression analyses, there are several alternatives. *Data disaggregation* subsets the data by (race) groups and fits separate regression models (12, 25, 30, 34). This approach produces race-specific intercepts and slopes, and thus implicitly acknowledges the importance of race-modifiers. However, these separate models do not produce global (race-invariant) x-effect estimates or inference, cannot incorporate information-sharing or regularization across race groups (often leading to variance inflation and reduced power), and require separate model diagnostics. *Sum-to-zero* (STZ) constraints address the inequities in RGE, but the resulting model parameters are difficult to interpret and the estimators do not offer any of the appealing statistical properties provided by the proposed approach. Finally, *overparametrized estimation* omits any identifying constraints and relies on regularized regression to produce unique estimators. But the model parameters remain nonidentified, thus making the estimates extremely difficult to interpret. Again, the estimates fail to offer the useful statistical properties discussed subsequently.

The primary goal of this paper is to introduce, study, and validate alternative statistical methods that eliminate these racial biases. By carefully reframing the regression model, we ensure equitable and interpretable parameters with accompanying estimators that offer unique and appealing statistical properties. We apply these tools to identify and quantify the race-modified effects of multiple environmental, social, and other factors on 4th end-of-grade readings scores for students in North Carolina. Although we focus on race, the proposed methods remain applicable for other categorical covariates including sex, national origin, religion, and other protected groups.

DRAFT

Results

Abundance-Based Constraints (ABCs) for Linear Regression. We update the race-modified model (Figure 1b) for multivariable regression with p covariates $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$, where the effect of each variable may be modified by race:

$$\begin{aligned} \mu(\boldsymbol{x}, r) &= \alpha_0 + \sum_{j=1}^{p} x_j \alpha_j + \beta_r + \sum_{j=1}^{p} x_j \gamma_{r,j} \\ &= \alpha_0 + \boldsymbol{x}^\top \boldsymbol{\alpha} + \beta_r + \boldsymbol{x}^\top \boldsymbol{\gamma}_r \end{aligned} \qquad [1]$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top$ are the main x-effects and $\boldsymbol{\gamma}_r = (\gamma_{r,1}, \ldots, \gamma_{r,p})^\top$ are the race-modifier effects. The main-only version omits all interactions ($\gamma_{r,j} = 0$). The intercepts are race-specific, $\mu(\mathbf{0}, r) = \alpha_0 + \beta_r$, while the race-modified model yields race-specific slopes for each variable $j = 1, \ldots, p$:

$$\begin{aligned} \mu'_{x_j}(r) &\coloneqq \mu(x_j + 1, \boldsymbol{x}_{-j}, r) - \mu(x_j, \boldsymbol{x}_{-j}, r) \\ &= \alpha_j + \gamma_{r,j}. \end{aligned}$$

The parameters $\{\alpha_0, \beta_r\}_r$ and $\{\alpha_j, \gamma_{r,j}\}_{r,j}$ must be further constrained to enable unique estimation and meaningful inference. Linear constraints of the form $\sum_r c_r \beta_r = 0$ and $\sum_r c_r \gamma_{r,j} = 0$ are most common: RGE sets $c_1 = 1$ and $c_r = 0$ for $r > 1$, while STZ uses $c_r = 1$ for all r. However, the equitability, interpretability, and statistical properties of the parameters and estimators depend critically on the choice of $\{c_r\}$.

We propose *abundance-based constraints (ABCs)* that use the race group abundances:

$$\begin{aligned} &\sum_r \hat{\pi}_r \beta_r = 0, \quad \sum_r \hat{\pi}_r \gamma_{r,j} = 0 \quad \text{for } j = 1, \ldots, p \\ &\hat{\pi}_r = \text{proportion in (race) group } r \end{aligned}$$

or equivalently, $\mathbb{E}_{\hat{\pi}}(\beta_R) = 0$ and $\mathbb{E}_{\hat{\pi}}(\gamma_{R,j}) = 0$ for all j, where the expectation is taken over a categorical random variable R with $\mathbb{P}(R = r) = \hat{\pi}_r$. If known, the population proportions may be used for $\{\hat{\pi}_r\}$; otherwise, we use the sample proportions. Historically, ABCs were considered for main-only models (35), but lacked sufficiently compelling advantages over alternative approaches (RGE, STZ, etc.) in that class of models to gain widespread adoption. Here, we advocate for ABCs in *race-modified* models based on equitability, interpretability, and special statistical properties.

To evaluate equitability and interpretability, we consider the meaning of each parameter in the race-modified model. Under ABCs, the race-modified model satisfies $\mathbb{E}_{\hat{\pi}}\{\mu(\boldsymbol{x}, R)\} = \alpha_0 + \boldsymbol{x}^\top \boldsymbol{\alpha}$, which produces a global (race-invariant) linear regression. As a consequence, each main x-effect may

DRAFT

be expressed as the *race-averaged slope* for the jth variable:

$$\alpha_j = \sum_r \hat{\pi}_r \mu'_{x_j}(r) = \mathbb{E}_{\hat{\pi}}\{\mu'_{x_j}(R)\}.$$

Unlike with RGE, where $\alpha_j = \mu'_{x_j}(\texttt{NHW})$, ABCs do *not* anchor each main x-effect to the NH White group and instead provide a global interpretation for these key parameters. The benefits cascade down to the other parameters:

$$\gamma_{r,j} = \mu'_{x_j}(r) - \alpha_j = \mu'_{x_j}(r) - \mathbb{E}_{\hat{\pi}}\{\mu'_{x_j}(R)\}$$

which is the difference between the *race-specific slope* and the *race-averaged slope* for variable j. The intercept also retains a convenient, more equitable interpretation. Suppose that each continuous covariate is centered, $\bar{x}_j = 0$. Then the intercept parameter is a marginal expectation:

$$\alpha_0 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\hat{\pi}}\{\mu(\boldsymbol{x}_i, R)\} = \mathbb{E}_{\hat{p}_{\boldsymbol{x}}}\mathbb{E}_{\hat{\pi}}\{\mu(\boldsymbol{X}, R)\}$$

where the expectation is taken (separately) over $\boldsymbol{X} \sim \hat{p}_{\boldsymbol{x}}$ for $\hat{p}_{\boldsymbol{x}}$ the empirical distribution of $\{\boldsymbol{x}_i\}_{i=1}^n$ and $R \sim \hat{\pi}$. The race-specific intercept coefficients proceed similarly:

$$\beta_r = \mu(\mathbf{0}, r) - \alpha_0 = \mathbb{E}_{\hat{p}_{\boldsymbol{x}}}\{\mu(\boldsymbol{X}, r)\} - \mathbb{E}_{\hat{p}_{\boldsymbol{x}}}\mathbb{E}_{\hat{\pi}}\{\mu(\boldsymbol{X}, R)\}.$$

Again, unlike for RGE, the parameters α_0 and β_r no longer elevate the NH White group. Instead, ABCs define all parameters as 1) global, race-averaged main effects or 2) race-specific deviations.

Estimation and Inference. Given data $\{y_i, \boldsymbol{x}_i, r_i\}_{i=1}^n$, the race-modified model with ABCs is estimated by applying linearly-constrained ordinary least squares (OLS) estimation. Standard errors, confidence intervals, and hypothesis testing are available as in traditional OLS estimation. Options for regularized (ridge, lasso, etc.) regression are readily available (see **Methods**). Because the estimators satisfy ABCs, they retain the same properties and interpretations as the parameters above.

Statistical Properties. A central obstacle with race-modified models is that, for default approaches (RGE, STZ, etc.), the inclusion of these interaction terms fundamentally alters the interpretations, estimates, and standard errors for the main x-effects. We observe this empirically (Table 1, left): compared to the main-only model, the race-modified model *under RGE* attenuates the RI main effect ($\hat{\alpha}_1^M = -0.042$ vs. $\hat{\alpha}_1 = -0.013$) and inflates the standard error ($\text{SE}(\hat{\alpha}_1^M) = 0.007$ vs. $\text{SE}(\hat{\alpha}_1) = 0.011$). These results are not contradictory: the RI effect is weaker for the NH White group (Figure 1b) than for the aggregate (Figure 1a), while NH White students represent a subset (58%) of the full sample. The broader implication is that analysts may be reluctant to include

DRAFT

race-modifiers. However, omitting race-modifiers can produce misleading results (Figure 1).

ABCs resolve these problems. The key result is *estimation invariance*: under ABCs, the OLS estimates of the main x-effects are nearly identical between the main-only model and the race-modified model, under appropriate conditions. For $p = 1$ (Figure 1), we establish the following:

$$\hat{\alpha}_1^M \approx \hat{\alpha}_1 \quad \text{whenever } \hat{\sigma}^2_{x[r]} \approx \hat{\sigma}^2_{x[1]} \text{ for all race groups } r \qquad [2]$$

where $\hat{\sigma}^2_{x[r]} = n_r^{-1} s_r^2 - \bar{x}_r^2$ is the (scaled) sample variance of $\{x_i\}_{i=1}^n$ within each race group r, with $n_r = n\hat{\pi}_r$, $s_r^2 = \sum_{r_i=r} x_i^2$ and $\bar{x}_r = n_r^{-1} \sum_{r_i=r} x_i$ (Theorem 1). Similar results are available for general $p > 1$ under suitable modifications of the equal-variance condition (Theorem 2).

The equal-variance condition in Eq. (2) requires that the scale of x is approximately the same for each race group. Otherwise, a one-unit change in x is not comparable across race groups. In that case, race-specific slopes are necessary, and the global slope from the main-only model (α_1^M) is not a suitable summary. However, the estimation invariance of ABCs is empirically robust to violations of the equal-variance condition. This condition is strongly violated for RI in Table 1 ($\hat{\sigma}_{\texttt{RI[NHW]}} = 0.688$, $\hat{\sigma}_{\texttt{RI[NHB]}} = 1.063$, $\hat{\sigma}_{\texttt{RI[Hisp]}} = 0.936$), yet the main effect estimates remain similar ($\hat{\alpha}_1^M = -0.042$ vs. $\hat{\alpha}_1 = -0.032$) and the standard errors (and p-values) are identical ($\text{SE}(\hat{\alpha}_1^M) = 0.007$ vs. $\text{SE}(\hat{\alpha}_1) = 0.007$). Similar results are observed for simulated data (Figure A.1), which further show that estimation invariance does *not* hold for RGE or STZ.

With ABCs, the analyst may include race-modifiers "for free": the estimated main x-effects are nearly unchanged by the addition of race-modifiers (`x:race`). This result is unique to ABCs and makes no assumptions about the true relationship between Y, X, and race. Notably, arbitrary dependencies are permitted between X and race—including varying means and distributions of X by race group—as long as the equal-variance condition holds. Thus, this result is distinct from classical estimation invariance results with OLS that require uncorrelatedness (36).

Sparsity. Sparsity is often prioritized to remove extraneous parameters, reduce estimation variability, and simplify interpretations. Regularized regression can produce sparse estimates, but depends critically on the parametrization. Importantly, sparsity of the race-modifiers, $\gamma_{r,j} = 0$, is meaningful under ABCs: it implies that the race-specific slope equals the race-averaged slope, $\mu'_{x_j}(r) = \mathbb{E}_{\hat{\pi}}\{\mu'_{x_j}(R)\} + \gamma_{r,j} = \mathbb{E}_{\hat{\pi}}\{\mu'_{x_j}(R)\}$. This eliminates the racial bias and inequity under RGE, where the same sparsity instead implies that the race-specific slope equals the NH White slope, $\mu'_{x_j}(r) = \mu'_{x_j}(\texttt{NHW}) + \gamma_{r,j} = \mu'_{x_j}(\texttt{NHW})$.

An especially concerning case arises when the race-modifier is nonzero ($\gamma_{r,j} \neq 0$), but the main x-effect is zero ($\alpha_j = 0$). Statistical approaches often eschew this scenario, and instead require that interactions are nonzero *only if* a main effect is nonzero (37, 38). Such restrictions are not necessary for ABCs: it is plausible that some race-specific x-effects are nonzero, $\mu'_{x_j}(r) = \gamma_{r,j} + \alpha_j = \gamma_{r,j} \neq 0$, while the race-averaged x-effect is zero, $\alpha_j = \mathbb{E}_{\hat{\pi}}\{\mu'_{x_j}(R)\} = 0$. Alarmingly, fitting a main-only model

DRAFT

would produce misleading results. Applying Eq. (2) (Theorems 1–2), the estimated x-effect would be near zero, $\hat{\alpha}_j^M \approx 0$, when in fact the x-effect is both *significant* and *race-specific*. Thus, it is possible that existing quantitative analyses based on main-only models (i.e., without race-modifiers) obscure both important and race-specific effects of certain variables (Figures 1 and 2).

NC Education Data Analysis. We apply the proposed methods to study the effects of multiple environmental, social, and other factors on educational outcomes—and assess whether, and how, these effects vary by race. Using ABCs, we fit equitable and interpretable race-modified models, empirically evaluate estimation and inference invariance properties, and study regularized (lasso) regression solution paths under competing parametrizations.

Data overview. We construct a cohort of $n = 27,638$ students in North Carolina (NC) by linking three administrative datasets:

NC Detailed Birth Records include maternal and infant characteristics for all documented live births in NC. We compute maternal covariates—mother's race, age (mAge), education level, marital status, and smoking status—and child covariates, sex and birthweight percentile for gestational age (BWTpct). RI is computed using residential addresses at birth.

NC Blood Lead Surveillance includes blood lead level (BLL) measurements for each child. Lead is an adverse environmental exposure with well-known effects on cognitive development and educational outcomes (39, 40).

NC Standardized Testing Data contains 4th end-of-grade standardized reading scores, economic disadvantage status (determined by participation in the National Lunch Program), and residential address at time-of-test. The reading scores, standardized by the year of test (2010, 2011, or 2012), serve as the outcome variable Y. The residential information is used to estimate the average $PM_{2.5}$ exposure ($PM_{2.5}$) over the year prior to the test, which is an adverse environmental exposure linked to educational outcomes (41).

Data characteristics are in Table A.1; additional details are provided elsewhere (30, 42, 43).

Race-modified regression with ABCs. We estimate a multivariable linear regression for 4th end-of-grade reading scores that includes these environmental, social, and other factors, as well as race-modifiers (Table 2). Each continuous covariate (BLL, $PM_{2.5}$, RI, mAge, and BWTpct) is centered and scaled and each categorical variable (mother's race, child's sex, mother's education level, mother's marital status, mother's smoking status, and economically disadvantaged) is identified using ABCs (see **Methods**). Race-modifiers are included for BLL, $PM_{2.5}$, RI, mAge, and BWTpct. Standard model diagnostics confirm linearity, homoskedasticity, and Gaussian error assumptions.

ABCs generate output for all main effects, all race-modifier effects, and each group in every categorical variable, which eliminates the presentation bias that would otherwise accompany *each* categorical variable under RGE. There are highly significant ($p < 0.01$) negative effects for BLL and RI, where the adverse RI effect *doubles* for NH Black students ($\hat{\mu}'_{\texttt{RI}}(\texttt{NHB}) = \hat{\alpha}_{\texttt{RI}} + \hat{\gamma}_{\texttt{RI:NHB}} =$

DRAFT

$-0.020 + -0.020 = -0.040$). This critical result for RI expands upon the previous model fit (Table 1): here, the model adjusts for many additional factors, yet the effect persists. Significantly lower test scores also occur for students who are NH Black, Male, or economically disadvantaged, and whose mothers who are less educated, unmarried, or smokers at time of birth. Significant positive effects are observed for the opposite categories—which is a byproduct of ABCs (e.g., the Male and Female proportions are identical, so the estimated effects must be equal and opposite)—as well as mAge and BWTpct. Finally, $PM_{2.5}$ is not identified as a significant main effect ($p = 0.403$), yet the race-specific effects *are* significant. Alarmingly, a fitted main-only model (*without* race-modifiers) conveys an insignificant $PM_{2.5}$ effect (Figure 2), which oversimplifies and misleads.

Table 2. Linear regression output (under ABCs) for the race-modified effects of environmental, social, and other factors on 4th end-of-grade reading scores for students in North Carolina.

Variable	Estimate (SE)	p-value
Intercept	0.000 (0.007)	0.976
Blood lead level (BLL)	-0.016 (0.006)	0.003
$PM_{2.5}$ exposure ($PM_{2.5}$)	-0.005 (0.005)	0.403
Racial isolation (RI)	-0.020 (0.007)	0.005
Mother's age (mAge)	0.038 (0.007)	<0.001
Birthweight percentile for gestational age (BWTpct)	0.026 (0.006)	<0.001
Mother's race		
NH White	0.173 (0.006)	<0.001
NH Black	-0.320 (0.010)	<0.001
Hispanic	0.258 (0.024)	<0.001
Sex		
Male	-0.064 (0.005)	<0.001
Female	0.064 (0.005)	<0.001
Mother's education level		
Did not complete high school	-0.179 (0.011)	<0.001
Completed high school	-0.074 (0.007)	<0.001
At least some postsecondary	0.181 (0.008)	<0.001
Mother's marital status		
Married at time of birth	0.018 (0.006)	0.003
Not married at time of birth	-0.024 (0.008)	0.003
Mother's smoking status		
Smoker	-0.039 (0.013)	0.003
Non-smoker	0.008 (0.003)	0.003
Economically disadvantaged		
Yes	-0.109 (0.005)	<0.001
No	0.171 (0.009)	<0.001

Variable (continued)	Estimate (SE)	p-value
BLL × Mother's race		
BLL:NH White	-0.005 (0.005)	0.240
BLL:NH Black	0.002 (0.007)	0.792
BLL:Hispanic	0.042 (0.022)	0.060
$PM_{2.5}$ × Mother's race		
$PM_{2.5}$:NH White	-0.011 (0.005)	0.018
$PM_{2.5}$:NH Black	0.021 (0.007)	0.005
$PM_{2.5}$:Hispanic	-0.017 (0.023)	0.472
RI × Mother's race		
RI:NH White	0.006 (0.006)	0.250
RI:NH Black	-0.020 (0.008)	0.016
RI:Hispanic	0.058 (0.023)	0.013
mAge × Mother's race		
mAge:NH White	0.028 (0.005)	<0.001
mAge:NH Black	-0.040 (0.007)	<0.001
mAge:Hispanic	-0.032 (0.024)	0.193
BWTpct × Mother's race		
BWTpct:NH White	0.000 (0.005)	0.929
BWTpct:NH Black	0.002 (0.007)	0.766
BWTpct:Hispanic	-0.018 (0.022)	0.427

Data restricted to individuals with 37-42 weeks gestation, mother's age 15-44 years old at birth, BLL $\leq 80\mu g/dL$ (and capped at $10\mu g/dL$), birth order ≤ 4, no current limited English proficiency, and residence in NC at the time of birth and time of 4th end-of-grade test. "Economically disadvantaged" is determined by participation in the National Lunch Program.

Estimation invariance with ABCs. Figure 2 presents the estimates and 95% confidence intervals for the main effects that are modified by race. We compare the main-only model (variables only in the left column of Table 2) to the race-modified model (all terms in Table 2), including both ABCs and RGE output for the race-modified models. Remarkably, under ABCs, the estimates and uncertainty quantification for the simpler, main-only model are nearly indistinguishable from those for the expanded, race-modified model. Effectively, ABCs allow estimation and inference for numerous

DRAFT

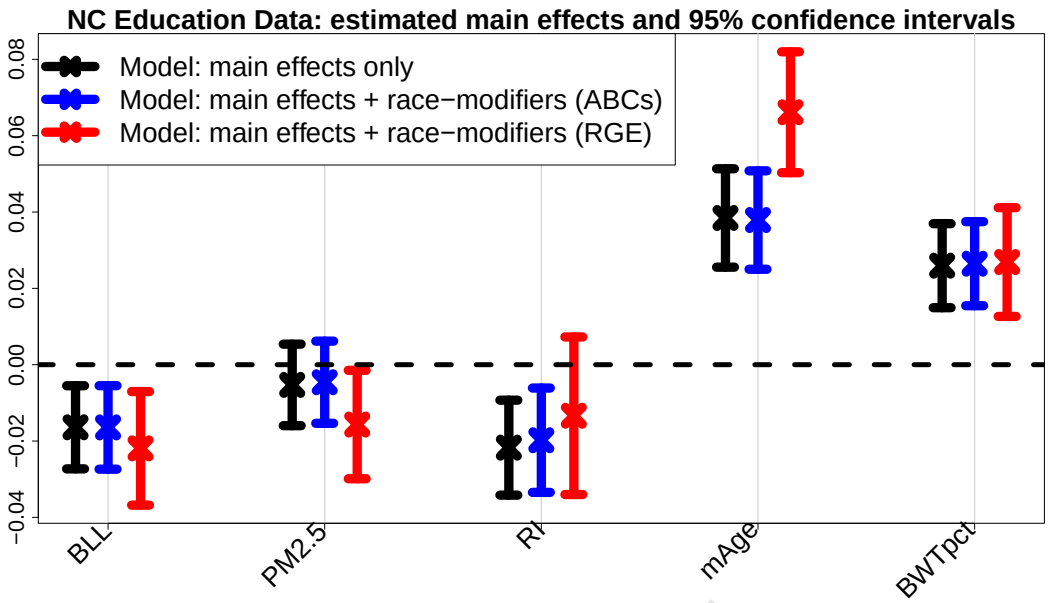

Fig. 2. Estimates and 95% confidence intervals for the main effects in the multivariable regression *without* race-modifiers (black) and the multivariable regression *with* race-modifiers under ABCs (blue) and RGE (red). Results are presented for blood lead level (BLL), $PM_{2.5}$ exposure ($PM_{2.5}$), racial isolation (RI), mother's age (mAge), and birthweight percentile for gestational age (BWTpct), each of which is interacted with race in the expanded model (blue, red); additional covariates include sex, mother's education level, mother's marital status, mother's smoking status, and economically disadvantaged (Table 2). ABCs exhibit invariance: despite the additional race-modifier parameters, the point and interval estimates for the main effects (blue) are nearly indistinguishable from those in the main effects-only model (black), thus effectively allowing the inclusion of race-modifiers "for free". In contrast, the RGE terms (red) correspond to the x-effects for the NH White group and deviate substantially for $PM_{2.5}$, RI, and mAge, including shifts in location and much wider intervals.

race-specific effects (Table 2, right column) "for free": the inferential summaries for the main effects are unchanged by the expansion of the model to include race-modifiers. This result empirically confirms the multivariable extension of Eq. (2) (Theorem 2), despite moderate violations of the equal-variance condition (Table A.2). Unsurprisingly, no such invariance holds for RGE (red): the point and interval estimates are substantially different, with uniformly wider intervals and conflicting conclusions about nonzero coefficients ($PM_{2.5}$, RI). These concerning discrepancies occur because the RGE "main effects" are exclusively for NH White students.

Regularized regression with ABCs. We assess regularized regression and variable selection with ABCs using lasso regression, including all variables from Table 2. We report estimates across tuning parameter values λ for the model coefficients $\{\hat{\alpha}_j, \hat{\gamma}_{r,j}\}_{r,j}$ and the race-specific slopes $\{\hat{\mu}'_{x_j}(r) = \hat{\alpha}_j + \hat{\gamma}_{r,j}\}_{r,j}$; $\lambda \to 0$ yields OLS estimates, while $\lambda \to \infty$ yields sparse estimates. Since the penalized estimates depend critically on the parameterization, we compare ABCs and RGE. The estimated λ-paths for RI are in Figure 3; results for the remaining race-modified effects (BLL, $PM_{2.5}$, mAge, and BWTpct) are in Figures A.2–A.5. RGE fixes $\hat{\gamma}_{r,j} = 0$ for all λ, which results in 1) racially-biased shrinkage of the race-specific effects toward the NH White-specific effect and 2) attenuation of the RI effect $\hat{\alpha}_j$ (Figure 3, top right). ABCs resolve these issues. First, the model parameters are separately and equitably pulled toward zero (Figure 3, top left). Second, the RI effect $\hat{\alpha}_{\texttt{RI}}$ is *not* attenuated, and preserves its magnitude until $\log \lambda \approx 5$ (Figure 3, top left). Finally, the race-specific

DRAFT

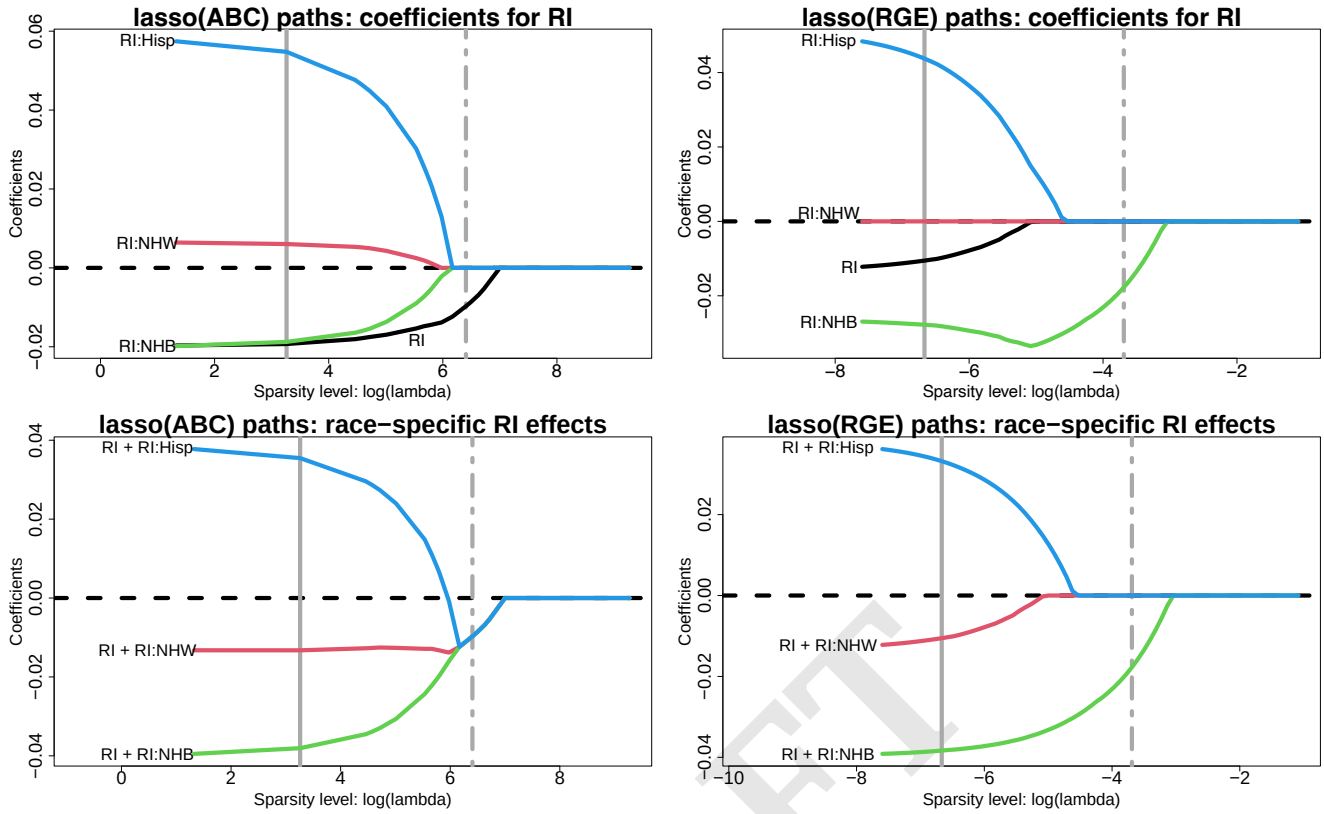

Fig. 3. Estimated lasso paths for RI across varying sparsity levels ($\log \lambda$) for the model coefficients $\hat{\alpha}_{\texttt{RI}}, \hat{\gamma}_{\texttt{RI:r}}$ (top) and the race-specific slopes $\hat{\mu}'_{\texttt{RI}}(r) = \hat{\alpha}_{\texttt{RI}} + \hat{\gamma}_{\texttt{RI:r}}$ (bottom) under ABCs (left) or RGE (right); vertical lines identify λ for the minimum cross-validated error (solid) and one-standard-error rule (dot-dashed). The outcome is 4th end-of-grade reading score and the covariates include all variables in Table 2. Small λ approximately corresponds to OLS, while increasing λ yields sparsity. Under RGE, the estimates are pulled toward the reference (NH White) estimate—inducing *statistical bias* by race—and the RI effect is attenuated. By comparison, ABCs offer more equitable shrinkage toward a global RI effect, which is nonzero and detrimental for 4th end-of-grade reading scores.

DRAFT

RI effects merge at a *global*, and negative, RI effect estimate, and this variable is selected by the one-standard-error rule (33) for choosing λ (Figure 3, bottom left).

These themes persist for the remaining race-modified effects (Figures A.2–A.5). We supplement the ABC and RGE lasso paths by including the lasso paths for *overparametrized estimation* (`Over`), which does not include any identifiability constraints. The parameters cannot be estimated uniquely by OLS, but can be estimated by lasso regression with $\lambda > 0$. In most cases, `Over` sets one of the coefficients $\{\alpha_j, \gamma_{r,j}\}_r$ to zero immediately (small λ) for each variable j. This effect reproduces RGE and thus `Over` inherits the same racial biases in estimation and selection. When this implicit selection sets $\hat{\gamma}_{j:\texttt{NHW}} = 0$, then the `Over` paths resemble those for RGE (RI, not shown; BWTpct, Figure A.5); when the selection corresponds to the smallest $|\hat{\gamma}_{r,j}|$ among race groups r from ABCs, then the `Over` and ABC paths are similar (BLL, Figure A.2; BWTpct, Figure A.5). However, when this selection sets the main effect to zero, $\hat{\alpha}_j = 0$ ($\text{PM}_{2.5}$, Figure A.3) or overshrinks multiple coefficients toward zero (mAge, Figure A.4), then the `Over` paths differ substantially from both the RGE and ABC paths and demonstrate erratic behavior (Figure A.4).

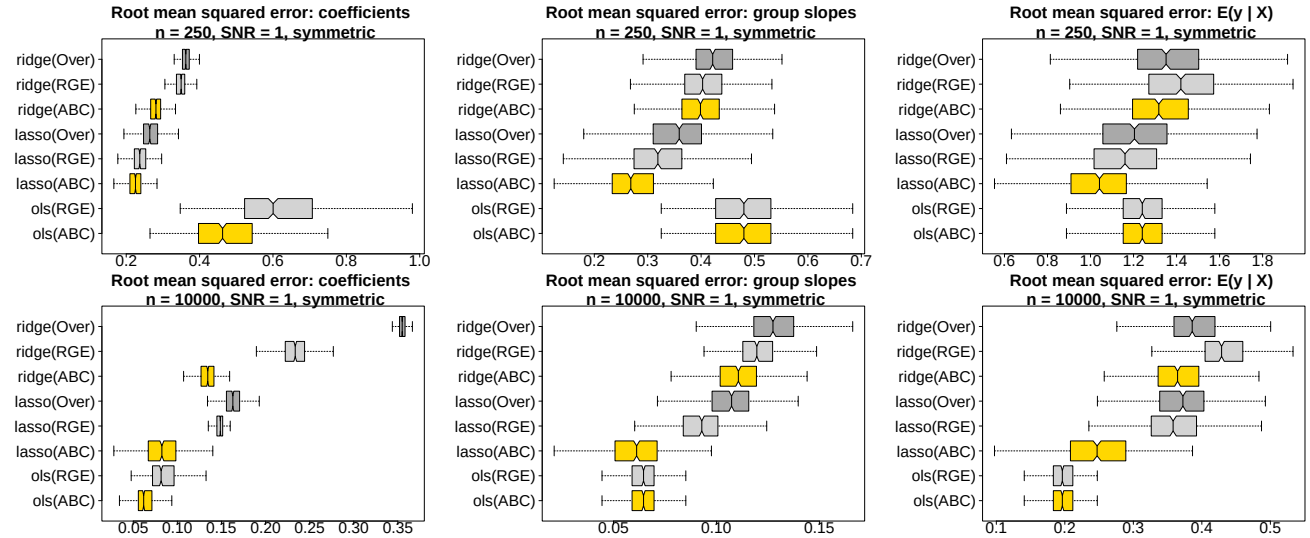

Fig. 4. Estimation and prediction accuracy for the regression coefficients (left), the race-specific slopes (center), and the fitted values (right) for $n = 250$ (top) and $n = 10{,}000$ (bottom) across 500 simulated datasets; nonoverlapping notches indicate significant differences between medians. Data are generated from a Gaussian main-only model with $p = 10$ covariates and a categorical variable with symmetric proportions $\boldsymbol{\pi} = (0.15, 0.35, 0.15, 0.35)^\top$; both RGE and ABCs are satisfied in the true data-generating process. All fitted models use the race-modified model Eq. (1) and thus contain extraneous race-modifiers. ABCs (gold) outperform both RGE (light gray) and `Over` (dark gray) within each estimation method (ridge, lasso, OLS). By definition, the OLS race-specific slopes and fitted values are invariant to the constraints (ABCs or RGE), and `Over` cannot be computed for OLS.

DRAFT

Estimation and predictive accuracy for simulated data. We evaluate estimation and prediction for ABCs, RGE, and `Over` across several estimation methods: OLS, ridge, and lasso regression (Figure 4). Data are simulated from a Gaussian main-only model with $p = 10$ covariates and a categorical variable with four levels. For fair comparisons, the data-generating process satisfies both RGE and ABCs. To mimic the challenges of real data analysis, the fitted models are misspecified as Eq. (1), and thus contain extraneous race-modifiers. Root mean squared errors are computed for the regression coefficients $\{\alpha_0, \alpha_j, \beta_r, \gamma_{r,j}\}_{r,j}$, the race-specific slopes $\{\alpha_j + \gamma_{r,j}\}_{r,j}$, and the model expectations $\mu(\boldsymbol{x}, r)$ across 500 simulated datasets. In each case, ABCs are substantially more accurate within each estimation method (OLS, ridge, lasso). The estimation invariance of ABCs offers a plausible explanation: whereas each fitted model includes extraneous variables (the race-modifiers), only ABCs reproduce the main effect estimates from the main-only model, which here is the ground truth. Thus, this unique statistical property of ABCs is not only convenient for interpreting race-modified models, but also provides more accurate estimates and predictions under both OLS and regularized regression.

Discussion

The path to more equitable decision-making and policy requires a precise and comprehensive understanding of the links between race and health and life outcomes. Alarmingly, the primary statistical tool for this task—regression analysis with race as a covariate and a modifier—in its current form propagates racial bias in both the *presentation* of results and the *estimation* of model

parameters. We introduced an alternative approach, abundance-based constraints (ABCs), with several unique benefits. First, ABCs eliminate these racial biases in both presentation and statistical estimation of linear regression models. Second, ABCs produce more interpretable parameters for race-modified models. Third, estimation with ABCs features an appealing invariance property: the estimated main effects are approximately unchanged by the inclusion of race-modifiers. Thus, analysts can include and estimate race-specific effects "for free"—without sacrificing parameter interpretability, equitability, or statistical efficiency. Finally, ABCs are especially convenient for regularized regression and variable selection, with meaningful and equitable notions of parameter sparsity and efficient computational algorithms.

Using this new approach, we estimated the effects of multiple environmental, social, and other factors on 4th end-of-grade readings scores for a large cohort of students ($n = 27,638$) in North Carolina. In aggregate, this analysis 1) identified significant race-specific effects for racial (residential) isolation, $PM_{2.5}$ exposure, and mother's age at birth; 2) showcased the racial biases and potentially misleading results obtained under default approaches; and 3) provided more equitable and interpretable estimates, uncertainty quantification, and selection, both for main effects and race-modified effects. Simulation studies demonstrated substantially more accurate estimates and predictions with OLS, ridge, and lasso regression compared to alternative approaches.

We acknowledge that the interpretation of any "race" effect requires great care (44). Race encompasses a vast array of social and cultural factors and life experiences, with effects that vary across time and geography (27, 45). In some settings, race data are unreliable or partially missing (46, 47). These overarching challenges are not addressed in this paper.

DRAFT

1. DR Williams, JA Lawrence, BA Davis, Racism and health: evidence and needed research. *Annu. Rev. Public Heal.* **40**, 105–125 (2019).
2. K Churchwell, et al., Call to action: structural racism as a fundamental driver of health disparities: a presidential advisory from the american heart association. *Circulation* **142**, e454–e468 (2020).
3. DR Williams, Miles to go before we sleep: Racial inequities in health. *J. health social behavior* **53**, 279–295 (2012).
4. F Edwards, H Lee, M Esposito, Risk of being killed by police use of force in the united states by age, race–ethnicity, and sex. *Proc. national academy sciences* **116**, 16793–16798 (2019).
5. K White, LN Borrell, Racial/ethnic residential segregation: framing the context of health risk and health disparities. *Heal. Place* **17**, 438–448 (2011).
6. DR Williams, C Collins, Racial residential segregation: a fundamental cause of racial disparities in health. *Public Heal. Reports* (2016).
7. ZD Bailey, et al., Structural racism and health inequities in the usa: evidence and interventions. *The Lancet* **389**, 1453–1463 (2017).
8. EA Pascoe, LS Richman, Perceived discrimination and health: a meta-analytic review. *Psychol. Bull.* **135**, 531 (2009).
9. Y Paradies, et al., Racism as a determinant of health: a systematic review and meta-analysis. *PloS one* **10**, e0138511 (2015).
10. DR Williams, SA Mohammed, Racism and health i: Pathways and scientific evidence. *Am. behavioral scientist* **57**, 1152–1173 (2013).
11. CA Haiman, et al., Ethnic and racial differences in the smoking-related risk of lung cancer. *New Engl. J. Medicine* **354**, 333–342 (2006).
12. MS Mujahid, X Gao, LP Tabb, C Morris, TT Lewis, Historical redlining and cardiovascular health: the multi-ethnic study of atherosclerosis. *Proc. Natl. Acad. Sci.* **118**, e2110986118 (2021).
13. KP Josey, et al., Air pollution and mortality at the intersection of race and social class. *New Engl. J. Medicine* **388**, 1396–1404 (2023).

14. KC Schoendorf, CJR Hogue, JC Kleinman, D Rowley, Mortality among infants of black as compared with white college-educated parents. *New Engl. journal medicine* **326**, 1522–1526 (1992).
15. VA Rauh, HF Andrews, RS Garfinkel, The contribution of maternal age to racial disparities in birthweight: a multilevel perspective. *Am. journal public health* **91**, 1815–1824 (2001).
16. EC Miller, et al., Maternal race/ethnicity, hypertension, and risk for stroke during delivery admission. *J. Am. Hear. Assoc.* **9**, e014775 (2020).
17. RT Kimbro, S Bzostek, N Goldman, G Rodríguez, Race, ethnicity, and the education gradient in health. *Heal. Aff.* **27**, 361–372 (2008).
18. V Johnson-Lawrence, DM Griffith, DC Watkins, The effects of race, ethnicity, and mood/anxiety disorders on the chronic physical health conditions of men from a national sample. *Am. journal men's health* **7**, 58S–67S (2013).
19. AT Geronimus, M Hicken, D Keene, J Bound, "weathering" and age patterns of allostatic load scores among blacks and whites in the united states. *Am. journal public health* **96**, 826–833 (2006).
20. A Reeves, et al., Study selection bias and racial or ethnic disparities in estimated age at onset of cardiometabolic disease among midlife women in the us. *JAMA Netw. Open* **5**, e2240665–e2240665 (2022).
21. X Huang, K Lee, MC Wang, NS Shah, SS Khan, Age at diagnosis of hypertension by race and ethnicity in the us from 2011 to 2020. *JAMA cardiology* **7**, 986–987 (2022).
22. JH Tamayo-Sarver, SW Hinze, RK Cydulka, DW Baker, Racial and ethnic disparities in emergency department analgesic prescription. *Am. journal public health* **93**, 2067–2073 (2003).
23. LA Schmidt, Y Ye, TK Greenfield, J Bond, Ethnic disparities in clinical severity and services for alcohol problems: results from the national alcohol survey. *Alcohol. Clin. Exp. Res.* **31**, 48–56 (2007).
24. J Liang, X Xu, AR Quiñones, JM Bennett, W Ye, Multiple trajectories of depressive symptoms in middle and late life: racial/ethnic variations. *Psychol. aging* **26**, 761 (2011).
25. M Agénor, N Krieger, SB Austin, S Haneuse, BR Gottlieb, At the intersection of sexual orientation, race/ethnicity, and cervical cancer screening: assessing pap test use disparities by sex of sexual partners among black, latina, and white us women. *Soc. science medicine* **116**, 110–118 (2014).
26. E Ramirez, et al., Insurance status as a modifier of the association between race and stage of prostate cancer diagnosis in florida during 1995 and 2013. *Cancer Epidemiol.* **59**, 104–108 (2019).
27. TA LaVeist, Beyond dummy variables and sample selection: what health services researchers ought to know about race as a variable. *Heal. Serv. Res.* **29**, 1 (1994).
28. SF Reardon, D O'Sullivan, Measures of spatial segregation. *Sociol. methodology* **34**, 121–162 (2004).
29. R Anthopolos, SA James, AE Gelfand, ML Miranda, A spatial measure of neighborhood level racial isolation applied to low birthweight, preterm birth, and birthweight in north carolina. *Spatial spatio-temporal epidemiology* **2**, 235–246 (2011).
30. M Bravo, D Zephyr, DR Kowal, KB Ensor, ML Miranda, Racial residential segregation shapes relationships between early childhood lead exposure and 4th grade standardized test scores. *Proc. Natl. Acad. Sci.* **119**, e2117868119 (2022).
31. SS Johfre, J Freese, Reconsidering the reference category. *Sociol. Methodol.* **51**, 253–269 (2021).
32. EK Chestnut, EM Markman, "girls are as good as boys at math" implies that boys are probably better: A study of expressions of gender equality. *Cogn. science* **42**, 2229–2249 (2018).
33. T Hastie, R Tibshirani, J Friedman, *The Elements of Statistical Learning*. (Springer) Vol. 2, (2009).
34. LN Borrell, BA Burt, HW Neighbors, GW Taylor, Social factors and periodontitis in an older population. *Am. journal public health* **98**, S95–S101 (2008).
35. RE Sweeney, EF Ulveling, A transformation for simplifying the interpretation of coefficients of binary variables in regression analysis. *The Am. Stat.* **26**, 30–32 (1972).
36. R Frisch, FV Waugh, Partial time regressions as compared with individual trends. *Econom. J. Econom. Soc.* pp. 387–401 (1933).
37. J Nelder, A reformulation of linear models. *J. Royal Stat. Soc. Ser. A: Stat. Soc.* **140**, 48–63 (1977).
38. J Bien, J Taylor, R Tibshirani, A lasso for hierarchical interactions. *Annals Stat.* **41**, 1111 (2013).
39. RL Canfield, et al., Intellectual impairment in children with blood lead concentrations below 10 g per deciliter. *New Engl. journal medicine* **348**, 1517–1526 (2003).
40. ML Miranda, et al., The relationship between early childhood blood lead levels and performance on end-of-grade tests. *Environ. health perspectives* **115**, 1242–1247 (2007).
41. Y Gao, DR Kowal, Bayesian adaptive and interpretable functional regression for exposure profiles. *Annals Appl. Stat.* **18**, 642–663 (2024).
42. CEH Initiative, Linked births, lead surveillance, grade 4 end-of-grade (eog) scores [data set] (2020).
43. DR Kowal, et al., Bayesian variable selection for understanding mixtures in environmental exposures. *Stat. Medicine* **40**, 4850–4871 (2021).

DRAFT

44. TJ VanderWeele, WR Robinson, On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology* **25**, 473 (2014).
45. DR Williams, Race and health: basic questions, emerging directions. *Annals Epidemiol.* **7**, 322–333 (1997).
46. EV Klinger, et al., Accuracy of race, ethnicity, and language preference in an electronic health record. *J. general internal medicine* **30**, 719–723 (2015).
47. FCG Polubriaginof, et al., Challenges with quality of race and ethnicity data in observational databases. *J. Am. Med. Informatics Assoc.* **26**, 730–736 (2019).
48. S Wood, *Generalized additive models: an introduction with R.* (CRC press), (2006).
49. TB Arnold, RJ Tibshirani, *genlasso: Path Algorithm for Generalized Lasso Problems*, (2022) R package version 1.6.1.
50. MT Grotenhuis, et al., When size matters: advantages of weighted effect coding in observational studies. *Int. J. Public Heal.* **62**, 163–167 (2017).
51. L Brehm, PM Alday, Contrast coding choices in a decade of mixed models. *J. Mem. Lang.* **125**, 104334 (2022).
52. R Nieuwenhuis, HF te Grotenhuis, BJ Pelzer, Weighted effect coding for observational data with wec. (2017).
53. MC Lovell, Seasonal adjustment of economic time series and multiple regression analysis. *J. Am. Stat. Assoc.* **58**, 993–1010 (1963).

Methods

Abundance-Based Constraints (ABCs) with Multiple Categorical Covariates. Regression analysis often features multiple continuous covariates $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ *and* multiple categorical covariates $\boldsymbol{R} = (R_1, \ldots, R_L)^\top$ such as race, sex, education level, etc.:

$$\mu(\boldsymbol{x}, \boldsymbol{r}) = \alpha_0 + \boldsymbol{x}^\top \boldsymbol{\alpha} + \sum_{\ell=1}^{L} \beta_{\ell, r_\ell} + \sum_{\ell=1}^{L} \boldsymbol{x}^\top \boldsymbol{\gamma}_{\ell, r_\ell} \qquad [3]$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top$ and $\boldsymbol{\gamma}_{\ell, r_\ell} = (\gamma_{\ell, r_\ell, 1}, \ldots, \gamma_{\ell, r_\ell, p})^\top$ are p-dimensional and r_ℓ denotes the level of the ℓth categorical variable, $\ell = 1, \ldots, L$. Eq. (3) includes all continuous-categorical interactions and requires $L(1+p)$ constraints for identification; RGE sets $\beta_{\ell,1} = 0$ for all ℓ and $\gamma_{\ell,1,j} = 0$ for all ℓ, j.

We extend ABCs based on the *joint* distribution of the categorical variables $\boldsymbol{R}$. Specifically, let $\hat{\boldsymbol{\pi}} = \hat{\pi}_{r_1, \ldots r_L} = \mathbb{P}(R_1 = r_1, \ldots, R_L = r_L)$. If known, the population proportions may be used for $\hat{\boldsymbol{\pi}}$; otherwise, we use the sample proportions based on the observed data $\{\boldsymbol{r}_i\}_{i=1}^n$, i.e., $\hat{\pi}_{r_1, \ldots, r_L} = n^{-1} \sum_{i=1}^n \mathbb{I}\{r_{i,1} = r_1, \ldots, r_{i,L} = r_L\}$. Concisely, the generalized ABCs are

$$\mathbb{E}_{\hat{\boldsymbol{\pi}}}(\boldsymbol{\beta}_{\boldsymbol{R}}) = \boldsymbol{0}_L, \quad \mathbb{E}_{\hat{\boldsymbol{\pi}}}(\boldsymbol{\gamma}_{\boldsymbol{R},j}) = \boldsymbol{0}_L, \quad j = 1, \ldots, p \qquad [4]$$

where $\boldsymbol{\beta}_{\boldsymbol{R}} = (\beta_{1,R_1}, \ldots, \beta_{L,R_L})^\top$, $\boldsymbol{\gamma}_{\boldsymbol{R},j} = (\gamma_{1,R_1,j}, \ldots, \gamma_{L,R_L,j})^\top$, and $\boldsymbol{0}_L$ is an L-dimensional vector of zeros. Eq. (4) may be equivalently represented via separate marginal expectations for the L sets of categorical covariate parameters: for instance, $\mathbb{E}_{\hat{\boldsymbol{\pi}}}(\beta_{\ell,R_\ell}) = \mathbb{E}_{\hat{\pi}_\ell}(\beta_{\ell,R_\ell}) = \sum_{r_\ell} \hat{\pi}_{\ell,r_\ell} \beta_{\ell,r_\ell} = 0$ for each categorical covariate ℓ.

ABCs in Eq. (4) provide interpretable parameter identifications with equitable presentation and estimation. These interpretations are unchanged if some or all interaction terms are omitted from Eq. (3), which may occur if multiple categorical variables (e.g., sex, education level) are included as

DRAFT

covariates, but only race is included as a modifier. ABCs imply that $\mathbb{E}_{\hat{\boldsymbol{\pi}}}\{\mu(\boldsymbol{x}, \boldsymbol{R})\} = \alpha_0 + \boldsymbol{x}^\top \boldsymbol{\alpha}$, so that averaging the regression Eq. (3) over all categorical variables (jointly) yields a multivariate regression with only continuous variables. Individually, each x_j-effect satisfies

$$\alpha_j = \mathbb{E}_{\hat{\boldsymbol{\pi}}}\{\mu'_{x_j}(\boldsymbol{R})\} \qquad [5]$$

where $\mu'_{x_j}(\boldsymbol{r}) = \mu(x_j+1, \boldsymbol{x}_{-j}, \boldsymbol{r}) - \mu(x_j, \boldsymbol{x}_{-j}, \boldsymbol{r})$ is the slope in the jth direction. To further simplify the interpretation, the expectation under $\hat{\boldsymbol{\pi}}$ in Eq. (5) need only be taken with respect to the categorical variables that are *interacted* with x_j (e.g., race). By comparison, the RGE parametrization yields $\alpha_j = \mu'_{x_j}(r_1 = 1, \ldots, r_L = 1)$, which is the group-specific slope for x_j with each group set to its reference category (e.g., NH White, Male, etc.). Clearly, this representation compounds inequity across each categorical variable and fails to deliver a global interpretation of the x_j-effect.

Interpretation of group-specific slopes and the parameters $\gamma_{\ell,r_\ell,j}$ proceeds by considering *partial* expectations $\hat{\boldsymbol{\pi}}_{-\ell}$, which is analogous to the joint distribution $\hat{\boldsymbol{\pi}}$ but omits the ℓth categorical variable. Here, as with Eq. (5), this expectation need only consider the categorical variables that are interacted with x_j; if the ℓth categorical variable is the only interaction term, then no expectation is needed at all. Then the x_j-effect when the ℓth categorical variable has level r_ℓ, averaged over the remaining categorical variables, is

$$\mathbb{E}_{\hat{\boldsymbol{\pi}}_{-\ell}}\{\mu'_{x_j}(r_\ell, \boldsymbol{R}_{-\ell})\} = \alpha_j + \gamma_{\ell,r_\ell,j} \qquad [6]$$

or equivalently, $\gamma_{\ell,r_\ell,j} = \mathbb{E}_{\hat{\boldsymbol{\pi}}_{-\ell}}\{\mu'_{x_j}(r_\ell, \boldsymbol{R}_{-\ell})\} - \mathbb{E}_{\hat{\boldsymbol{\pi}}}\{\mu'_{x_j}(\boldsymbol{R})\}$. The interpretation is simpler than the notation: Eq. (6) directly extends the usual notion of race-specific slopes to average over any other categorical variables that modify x_j.

Estimation. Statistical estimation with ABCs requires solving a linearly-constrained least squares problem given data $\{\boldsymbol{x}_i, \boldsymbol{r}_i, y_i\}_{i=1}^n$. Define $\boldsymbol{\theta}$ to be the model parameters $\{\alpha_0, \boldsymbol{\alpha}, \beta_{\ell,r_\ell}, \boldsymbol{\gamma}_{\ell,r_\ell}\}_{r_\ell,\ell}$ and $\tilde{\boldsymbol{x}}_i$ to include the intercept, covariates, race variable indicators (i.e., "dummy variables"), and covariate-race interactions such that Eq. (3) may be written $\mu(\boldsymbol{x}_i, \boldsymbol{r}_i) = \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta}$. Let $\boldsymbol{C}$ encode ABCs such that $\boldsymbol{C}\boldsymbol{\theta} = \boldsymbol{0}$ enforces Eq. (4), so $\boldsymbol{C}$ has $m = L(1+p)$ rows corresponding to the number of constraints. The OLS estimator under ABCs is

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^n (y_i - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})^2 \quad \text{subject to } \boldsymbol{C}\boldsymbol{\theta} = \boldsymbol{0}. \qquad [7]$$

To compute $\hat{\boldsymbol{\theta}}$—and subsequently provide inference and penalized estimation—we reparametrize the problem into an unconstrained space with m fewer parameters. Let $\boldsymbol{C}^\top = \boldsymbol{Q}\boldsymbol{R}$ be the QR-decomposition of the transposed constraint matrix with columnwise partitioning of the orthogonal matrix $\boldsymbol{Q} = (\boldsymbol{Q}_{1:m} : \boldsymbol{Q}_{-(1:m)})$ with $\boldsymbol{R}^\top = (\boldsymbol{R}_{1:m,1:m} : \boldsymbol{0})$, since $\boldsymbol{C}^\top$ has rank m. It is straightforward

DRAFT

to verify that $\boldsymbol{\theta} = \boldsymbol{Q}_{-(1:m)}\boldsymbol{\zeta}$ satisfies $\boldsymbol{C}\boldsymbol{\theta} = \boldsymbol{0}$ for *any* $\boldsymbol{\zeta}$. Then, using the adjusted covariate matrix $\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{X}}\boldsymbol{Q}_{-(1:m)}$ with $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n)^\top$, the solution to Eq. (7) is equivalently solved using unconstrained OLS:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{Q}_{-(1:m)}\hat{\boldsymbol{\zeta}}, \quad \hat{\boldsymbol{\zeta}} = \arg\min_{\boldsymbol{\zeta}} \sum_{i=1}^{n} (y_i - \tilde{\boldsymbol{z}}_i^\top \boldsymbol{\zeta})^2. \qquad [8]$$

The QR-decomposition has minimal cost due to the efficiency of Householder rotations and the low dimensionality of $\boldsymbol{C}$ (48).

Although alternative computing strategies are available, the reparametrization in Eq. (8) is especially convenient for generalizations to regularized (lasso, ridge, etc.) estimation. Let $\mathcal{P}(\boldsymbol{\theta})$ denote a complexity penalty on the regression coefficients. The penalized least squares estimator under ABCs is

$$\hat{\boldsymbol{\theta}}(\lambda) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y_i - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})^2 + \lambda \mathcal{P}(\boldsymbol{\theta}) \quad \text{subject to } \boldsymbol{C}\boldsymbol{\theta} = \boldsymbol{0} \qquad [9]$$

where $\lambda \geq 0$ controls the tradeoff between goodness-of-fit and complexity (measured via $\mathcal{P}$). Following Eq. (8), we instead compute

$$\hat{\boldsymbol{\theta}}(\lambda) = \boldsymbol{Q}_{-(1:m)}\hat{\boldsymbol{\zeta}}(\lambda), \quad \hat{\boldsymbol{\zeta}}(\lambda) = \arg\min_{\boldsymbol{\zeta}} \sum_{i=1}^{n} (y_i - \tilde{\boldsymbol{z}}_i^\top \boldsymbol{\zeta})^2 + \lambda \mathcal{P}(\boldsymbol{Q}_{-(1:m)}\boldsymbol{\zeta})$$

which requires the solution to an *unconstrained* penalized least squares problem.

We focus on complexity penalties of the form

$$\mathcal{P}(\boldsymbol{\theta}) = \sum_j \omega_j |\theta_j|^\delta$$

where $\omega_j > 0$ are known weights, $\delta = 1$ produces sparse coefficients (adaptive lasso regression) and $\delta = 2$ guards against collinearity (adaptive ridge regression). Under ridge regression ($\delta = 2$), the solution is

$$\hat{\boldsymbol{\theta}}(\lambda) = \boldsymbol{Q}_{-(1:m)}(\tilde{\boldsymbol{Z}}^\top \tilde{\boldsymbol{Z}} + \lambda \boldsymbol{D}^\top \boldsymbol{D})^{-1} \tilde{\boldsymbol{Z}}^\top \boldsymbol{y}$$

where $\boldsymbol{D} = \text{diag}\{\omega_j^{1/2}\}_{j=1}^{p} \boldsymbol{Q}_{-(1:m)}$. The lasso version ($\delta = 1$) can be solved efficiently using the `genlasso` package in `R` (49).

For practical use, we set ω_j to be the sample standard deviation of the jth column of $\tilde{\boldsymbol{X}}$ (with $\omega_1 = 1$ for the intercept). This strategy applies a standardized penalty to each covariate, which is especially important for ABCs. In particular, the magnitudes of the race-specific coefficients vary according to the abundance of the group: by construction, low abundances in group r will correspond to larger group r-specific coefficients. The standardized penalty adjusts for this effect to avoid overpenalization of group-specific coefficients for groups with low abundance.

Inference. The reparametrization strategy in Eq. (8) allows direct application of classical inference theory to the ABC OLS estimator: $\hat{\boldsymbol{\theta}}$ is a known, linear function of the (unconstrained) OLS

DRAFT

estimator $\hat{\boldsymbol{\zeta}}$. Thus, it is straightforward to derive the (Gaussian) sampling distribution of the ABC OLS estimator, which can be used to compute standard errors, hypothesis tests, and confidence intervals, and to establish unbiasedness and efficiency of the estimator. Under minimal assumptions, the (unconstrained) OLS estimator satisfies $\sqrt{n}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) \xrightarrow{d} N(0, \mathcal{I}(\boldsymbol{\zeta})^{-1})$ and thus the ABC OLS estimator satisfies

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(0, \boldsymbol{Q}_{-(1:m)}\mathcal{I}(\boldsymbol{\zeta})^{-1}\boldsymbol{Q}_{-(1:m)}^{\top})$$

where $\mathcal{I}$ is the Fisher information and $\boldsymbol{\zeta}, \boldsymbol{\theta}$ are the true parameter values. When the regression model is paired with independent and identically distributed Gaussian errors $\epsilon_i := y_i - \mu(\boldsymbol{x}_i, \boldsymbol{r}_i)$ with variance σ^2, the unconstrained OLS estimator satisfies $\hat{\boldsymbol{\zeta}} \sim N\{\boldsymbol{\zeta}, \sigma^2(\tilde{\boldsymbol{Z}}^{\top}\tilde{\boldsymbol{Z}})^{-1}\}$ and thus

$$\hat{\boldsymbol{\theta}} \sim N\{\boldsymbol{\theta}, \sigma^2\boldsymbol{Q}_{-(1:m)}(\tilde{\boldsymbol{Z}}^{\top}\tilde{\boldsymbol{Z}})^{-1}\boldsymbol{Q}_{-(1:m)}^{\top}\}$$

even in finite samples. This sampling distribution for the OLS estimator under ABCs ensures unbiasedness and efficiency, and provides the means to compute standard errors, hypothesis tests, and confidence intervals.

Theory.

Theorem 1. *Under ABCs, the OLS estimates for the main–only model (Figure 1a) and the race-modified model (Figure 1b) satisfy*

$$\hat{\alpha}_1^M \approx \hat{\alpha}_1 \quad \textit{whenever } \hat{\sigma}^2_{x[r]} \approx \hat{\sigma}^2_{x[1]} \textit{ for all } r$$

where $\hat{\sigma}^2_{x[r]} = n_r^{-1}s_r^2 - \bar{x}_r^2$ is the (scaled) sample variance of $\{x_i\}_{i=1}^n$ within each (race) group r, with $n_r = n\hat{\pi}_r$, $s_r^2 = \sum_{r_i=r} x_i^2$ and $\bar{x}_r = n_r^{-1}\sum_{r_i=r} x_i$.

Theorem 2. *Consider the multivariable race-modified model* Eq. (1) *and the multivariable main-only model*

$$\mu(\boldsymbol{x}, r) = \alpha_0^M + \boldsymbol{x}^{\top}\boldsymbol{\alpha}^M + \beta_r^M. \qquad [10]$$

Under ABCs, the OLS estimates satisfy

$$\hat{\boldsymbol{\alpha}} \approx \hat{\boldsymbol{\alpha}}^M \textit{ whenever } \hat{\sigma}_{x[r]}(j,h) \approx \hat{\sigma}_{x[1]}(j,h) \textit{ for all } r$$

where $\hat{\sigma}_{x[r]}(j,h) = n_r^{-1}s_r(j,h) - \bar{x}_r(j)\bar{x}_r(h)$ is the (scaled) sample covariance between x_j and x_h in (race) group r, with $s_r(j,h) = \sum_{r_i=r} x_{ij}x_{ih}$, and $\bar{x}_r(j) = n_r^{-1}\sum_{r_i=r} x_{ij}$.

Theorem 2 shows that *all* p estimated coefficients on $x_1, \ldots, x_p$ are approximately invariant to the inclusion of race-modifiers for *all* p variables whenever the (scaled) sample covariance is approximately the same for each race group. Importantly, this result allows for arbitrary dependencies among the

DRAFT

p variables $\boldsymbol{X}$, includes a variety of dependencies between each x_j and R, and does not impose any assumptions about the true relationship between Y and $\boldsymbol{X}$.

Simulation design. Each simulated dataset is constructed from a Gaussian main-only model Eq. (10) with $p = 10$ continuous covariates and one categorical (race) covariate with four levels. The $p = 10$ continuous covariates include six independent covariates, $X_j \sim N(0, 1)$ for $j = 1, 2, 3, 6, 7, 8$, and four covariates that depend on the categorical variable, $[X_j \mid R = r] \sim N(r, 1)$, i.e., mean one for group one, mean two for group two, etc., for $j = 4, 5, 9, 10$. The categorical variable is generated based on population proportions $\boldsymbol{\pi}$, which we describe below. In addition to the intercept $\alpha_0 = 1$, the true coefficients for the continuous covariates are $\alpha_1 = \cdots = \alpha_5 = 1$ (signals) and $\alpha_6 = \cdots = \alpha_{10} = 0$ (noise). Only main effects (i.e., no race-modifiers or interactions) are included in the data-generating process. We vary the sample size, $n \in \{250, 10{,}000\}$, and use a signal-to-noise ratio of one.

The categorical covariate and accompanying coefficients are constructed carefully to ensure fair comparisons. Different identifications correspond to different parameterizations, which may lead to unfair evaluations. Thus, we use a data-generating model that satisfies both RGE and ABCs. We consider two designs for the categorical proportions: symmetric weights $\boldsymbol{\pi} = (0.15, 0.35, 0.15, 0.35)^\top$ (Figure 4) and uniform weights $\boldsymbol{\pi} = (0.25, 0.25, 0.25, 0.25)^\top$ (Figure A.6). The true categorical variable coefficients are $\boldsymbol{\beta} = (0, 1, 0, -1)^\top$, so RGE is satisfied, $\beta_1 = 0$. ABCs are satisfied for the true population proportions, $\sum_r \pi_r \beta_r = 0$, but we use the sample proportions $\hat{\boldsymbol{\pi}}$ for estimation. The omission of race-modifiers from the data-generating process satisfies both RGE and ABCs. Finally, we require at least $p+1$ observations for each categorical level, which is necessary for OLS estimation of the interaction effects. We simulate 500 such datasets for each $(n, \boldsymbol{\pi})$ design.

All competing methods follow Eq. (1), which includes all continuous covariates, race, and all race-modifiers. Thus, all competing models are overparametrized relative to the ground truth, with 55 columns of the unconstrained designed matrix $\tilde{\boldsymbol{X}}$ and 44 identifiable model parameters to estimate. The estimation approaches are OLS, ridge regression, and lasso regression. The tuning parameter for ridge and lasso regression is selected using the one-standard-error rule (33). The parametrizations determine the identification constraints on β_r and $\gamma_{r,j}$: we consider ABCs, RGE, and `Over`. `Over` is not identified for OLS, and thus is presented only for ridge and lasso regression.

Empirical verification of estimation invariance with ABCs. To empirically verify Theorem 1, we generate 500 synthetic datasets that mildly violate the equal-variance condition. Iteratively, we sample a categorical variable R with groups $\{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}\}$ and respective probabilities $\boldsymbol{\pi} = (0.55, 0.20, 0.10, 0.15)^\top$

DRAFT

and then sample a continuous variable with the distribution determined by the group:

$$[X \mid R = r] \sim \begin{cases} 5 + N(0,1) & r = \texttt{A} \\ \sqrt{12}\ \text{Uniform}(0,1) & r = \texttt{B} \\ 5 + t_4(0,1) & r = \texttt{C} \\ \text{Gamma}(1,1) & r = \texttt{D} \end{cases} \tag{11}$$

By design, X depends on R in both mean and distribution. The R-specific *population* variances of X are each one, but Theorem 1 requires that the R-specific *sample* variances are identical, which will not be satisfied for any simulated dataset. Thus, Eq. (11) includes a mild deviation from the equal-variance condition of Theorem 1.

The response variable Y is simulated with expectation

$$\mu(x, r) = 1 + x + \gamma x \mathbb{I}\{r = \texttt{A}\} - \gamma x \mathbb{I}\{r = \texttt{B}\}$$

plus $t_4(0,1)$-distributed errors, i.e., standard t-distributions with 4 degrees of freedom. The coefficient γ determines the strength of the race-modifier effect: we consider $\gamma = 0$ (no race-modifier effect), $\gamma = 0.5$ (50% of the main x-effect), and $\gamma = 1.5$ (150% of the main x-effect). This data-generating process includes a R-modifier (when $\gamma \neq 0$) but does not satisfy ABCs or traditional Gaussian error assumptions. Repeating this process 500 times, each simulated dataset contains $\{x_i, r_i, y_i\}_{i=1}^n$. We consider $n \in \{100, 500\}$.

We fit the main-only and race-modified models and record the estimated x-effects $\hat{\alpha}_1^M$ and $\hat{\alpha}_1$, respectively, for each simulated dataset. These estimated coefficients depend on the constraints: we compare ABCs, RGE (with reference group A), and STZ constraints for $\{\beta_r, \gamma_r\}$ (Figure A.1). Although the conditions of Theorem 1 are not satisfied, the ABC estimates lie along the 45 degree line with $\hat{\alpha}_1 = \hat{\alpha}^M$; the estimated x-effect is nearly unchanged by the addition of the race-modifier. This invariance is *not* satisfied for RGE or STZ. The estimated x-effects under RGE or STZ vary considerably between the main-only and race-modified models, with greater discrepancies as the magnitude of the race-modifier effect increases. By comparison, the estimation invariance of ABCs is robust to the magnitude of the race-modifier effect. The mild deviations from the equal-variance condition of Theorem 1 are most impactful when $\gamma = 1.5$, which represents the unusual setting in which the interaction effect is much larger than the main effect. Even in this challenging case, the ABC estimates remain nearly invariant between the main-only and race-modified models, especially when compared to the RGE and STZ counterparts.

Contrast coding. For OLS estimation, identifiability constraints may be imposed using *contrasts*. In this approach, the linear model is fit under any minimally sufficient identification (RGE, STZ, ABCs, etc.) and the categorical variable coefficients are post-processed using linear contrast matrices.

DRAFT

Examples include dummy coding (akin to RGE), effects coding (akin to STZ), weighted effects coding (WEC; akin to ABCs), and Helmert coding (for ordered categories). However, contrasts are typically reserved for main-only models and are difficult to combine with regularized regression and variable selection. Further, these previous approaches do not consider or resolve the inequities of reporting or estimating race-specific effects. In particular, WEC has been advocated only in cases when "a categorical variable has categories of different sizes, and if these differences are considered relevant" (50) or "certain types of unbalanced data that are missing not at random" (51), with regression output that suffers from the same presentation bias that afflicts RGE (52). We do not agree with such restrictions for ABCs, and instead argue that this approach offers an equitable and interpretable parametrization with unique and appealing statistical properties, including both estimation invariance and regularized regression. These estimation invariance results and regularized regression analyses are notably absent from previous contrast coding approaches.

DRAFT

A. Supporting Information

To prove Theorems 1 and 2, we leverage a classical property of OLS (36, 53).

Theorem S2 (Frisch-Waugh-Lovell Theorem). For a partition of the $n \times p$ covariate matrix $\boldsymbol{X} = (\boldsymbol{X}_0 : \boldsymbol{X}_1)$ into p_0 and p_1 columns, the partition of the ordinary least squares estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_0^\top, \hat{\boldsymbol{\beta}}_1^\top)^\top$ satisfies $\hat{\boldsymbol{\beta}}_0 = (\boldsymbol{X}_0^\top \boldsymbol{E}_{01})^{-1} \boldsymbol{E}_{01}^\top \boldsymbol{y} = (\boldsymbol{E}_{01}^\top \boldsymbol{E}_{01})^{-1} \boldsymbol{E}_{01}^\top \boldsymbol{y}$, where $\boldsymbol{E}_{01} = (\boldsymbol{I}_n - \boldsymbol{H}_{X_1})\boldsymbol{X}_0$ is the $n \times p_0$ matrix of residuals from regressing each column of $\boldsymbol{X}_0$ on $\boldsymbol{X}_1$, $\boldsymbol{H}_{X_1} = \boldsymbol{X}_1(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top$ is the corresponding hat matrix for $\boldsymbol{X}_1$, and $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ is the vector of outcomes.

Although Theorems 1 and 2 and stated in terms of approximate equivalence under approximate equal-variance conditions, here we prove exact equivalence under exact equal-variance conditions.

Proof (Theorem 1). Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$, and $\boldsymbol{Z}$ be the matrix of categorical dummy variables with entries $[\boldsymbol{Z}]_{ir} = 1$ if $r_i = r$ and zero otherwise. Using Theorem S2, the estimated coefficients under the main-only model satisfy $\hat{\alpha}_1^M = (\boldsymbol{x}^\top \hat{\boldsymbol{e}}_{x\sim r})^{-1} \hat{\boldsymbol{e}}_{x\sim r}^\top \boldsymbol{y}$, where $\hat{\boldsymbol{e}}_{x\sim r}$ is the vector of residuals from regressing the continuous variable $\{x\}_{i=1}^n$ on the categorical variable $\{r_i\}_{i=1}^n$ (i.e., $\boldsymbol{Z}$). Similarly, the estimated coefficients under the race-modified model satisfy $\hat{\alpha}_1 = (\boldsymbol{x}^\top \hat{\boldsymbol{e}}_{x\sim r+Z_{XQ}})^{-1} \hat{\boldsymbol{e}}_{x\sim r+Z_{XQ}}^\top \boldsymbol{y}$, where $\hat{\boldsymbol{e}}_{x\sim r+Z_{XQ}}$ are the residuals from regressing the continuous variable $\{x\}_{i=1}^n$ on the categorical variable $\{r_i\}_{i=1}^n$ (i.e., $\boldsymbol{Z}$) *and* the reparametrized interaction term that enforces ABCs, $\boldsymbol{Z}_{XQ} = \boldsymbol{Z}_X \boldsymbol{Q}_{-(1:m)}$, where $\boldsymbol{Z}_X = \boldsymbol{D}_X \boldsymbol{Z}$ and $\boldsymbol{D}_X = \text{diag}(\boldsymbol{x})$. Thus, it suffices to show that $\hat{\boldsymbol{e}}_{x\sim r} = \hat{\boldsymbol{e}}_{x\sim r+Z_{XQ}}$, which occurs when the additional (interaction) coefficients from the latter model, say $\hat{\boldsymbol{b}}_{Z_{XQ}}$ (corresponding to $\boldsymbol{Z}_{XQ}$), are identically zero. Again using Theorem S2, these estimated coefficients are $\hat{\boldsymbol{b}}_{Z_{XQ}} = \boldsymbol{Q}_{-(1:m)}(\boldsymbol{Z}_{XQ}^\top \boldsymbol{E}_{Z_{XQ}})^{-1} \boldsymbol{E}_{Z_{XQ}}^\top \boldsymbol{x}$, where $\boldsymbol{E}_{Z_{XQ}}$ is the matrix of residuals from regressing $\boldsymbol{Z}_{XQ}$ on $\boldsymbol{Z}$, i.e., $\boldsymbol{E}_{Z_{XQ}} = \boldsymbol{Z}_{XQ} - \boldsymbol{H}_Z \boldsymbol{Z}_{XQ}$ where $\boldsymbol{H}_Z = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$ is the hat matrix for the categorical covariate. Thus, showing $\boldsymbol{x}^\top \boldsymbol{E}_{Z_{XQ}} = \boldsymbol{0}$ is sufficient. Factoring $\boldsymbol{x}^\top \boldsymbol{E}_{Z_{XQ}} = (\boldsymbol{x}^\top \boldsymbol{Z}_X - \boldsymbol{x}^\top \boldsymbol{H}_Z \boldsymbol{Z}_X)\boldsymbol{Q}_{-(1:m)}$, first observe that $\boldsymbol{x}^\top \boldsymbol{Z}_X = \boldsymbol{x}^\top \boldsymbol{D}_X \boldsymbol{Z} = (\ldots, s_r^2, \ldots)$ is the vector of $s_r^2 = \sum_{r_i=r} x_i^2$ across groups. Next, observe that $(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{Z}_X = \text{diag}(\{\bar{x}_r\}_r)$ contains the sample means of $\{x\}_{i=1}^n$ by each group r, and therefore $\boldsymbol{x}^\top \boldsymbol{H}_Z \boldsymbol{Z}_X = \boldsymbol{x}^\top \boldsymbol{Z}\text{diag}(\{\bar{x}_r\}_r) = (\ldots, n_r \bar{x}_r^2, \ldots)$ with $n_r = n\hat{\pi}_r$. Combining these results, we have $\boldsymbol{x}^\top \boldsymbol{E}_{Z_{XQ}} = \boldsymbol{v}^\top \boldsymbol{Q}_{-(1:m)}$, where $\boldsymbol{v}^\top = (\ldots, s_r^2 - n_r \bar{x}_r^2, \ldots) = n(\ldots, \hat{\pi}_r \hat{\sigma}_{x[r]}^2, \ldots) = n\hat{\sigma}_{x[1]}^2 \hat{\boldsymbol{\pi}}^\top$ under the assumption that $\hat{\sigma}_{x[r]}^2 = \hat{\sigma}_{x[1]}^2$ is common for all r. Finally, the definition of $\boldsymbol{Q}_{-(1:m)}$ via ABCs implies that $\hat{\boldsymbol{\pi}}^\top \boldsymbol{Q}_{-(1:m)} = \boldsymbol{0}$, which proves the result. □

Proof (Theorem 2). Let $\boldsymbol{X}$ denote the $n \times p$ matrix of continuous covariates, $\boldsymbol{Z}$ the matrix of categorical dummy variables with entries $[\boldsymbol{Z}]_{ir} = 1$ if $r_i = r$ and zero otherwise, and $\boldsymbol{Z}_{XQ} = (\boldsymbol{Z}_{X_1Q}, \ldots, \boldsymbol{Z}_{X_pQ})$ with $\boldsymbol{Z}_{X_jQ} = \boldsymbol{Z}_{X_j}\boldsymbol{Q}_{-(1:m)}$, $\boldsymbol{Z}_{X_j} = \boldsymbol{D}_{X_j}\boldsymbol{Z}$, and $\boldsymbol{D}_{X_j} = \text{diag}(\boldsymbol{x}_j)$. By Theorem S2, it suffices to show that $\boldsymbol{E}_M = \boldsymbol{E}$, where $\boldsymbol{E}_M = (\boldsymbol{I}_n - \boldsymbol{H}_Z)\boldsymbol{X}$ are the residuals from regressing each column of $\boldsymbol{X}$ on $\boldsymbol{Z}$ and $\boldsymbol{E}$ are similarly the residuals from regressing each column of $\boldsymbol{X}$ on $\boldsymbol{Z}$ *and* $\boldsymbol{Z}_{XQ}$. Thus, it is sufficient to show that the coefficients associated with $\boldsymbol{Z}_{XQ}$ in the lat-

DRAFT

ter regression are identically zero. Again using Theorem S2, we see that this occurs whenever $\boldsymbol{X}^\top \boldsymbol{E}_{Z_{XQ}} = \boldsymbol{0}$, where $\boldsymbol{E}_{Z_{XQ}} = \boldsymbol{Z}_{XQ} - \boldsymbol{H}_Z \boldsymbol{Z}_{XQ} = (\boldsymbol{Z}_{X_1Q} - \boldsymbol{H}_Z \boldsymbol{Z}_{X_1Q}, \ldots, \boldsymbol{Z}_{X_pQ} - \boldsymbol{H}_Z \boldsymbol{Z}_{X_pQ})$. Noticing that $\boldsymbol{X}^\top \boldsymbol{E}_{Z_{XQ}} = (\boldsymbol{X}^\top(\boldsymbol{Z}_{X_1Q} - \boldsymbol{H}_Z \boldsymbol{Z}_{X_1Q}), \ldots, \boldsymbol{X}^\top(\boldsymbol{Z}_{X_pQ} - \boldsymbol{H}_Z \boldsymbol{Z}_{X_pQ}))$, we consider the individual components $\boldsymbol{x}_h^\top(\boldsymbol{Z}_{X_jQ} - \boldsymbol{H}_Z \boldsymbol{Z}_{X_jQ}) = (\boldsymbol{x}_h^\top \boldsymbol{D}_{x_j} \boldsymbol{Z} - \boldsymbol{x}_h^\top \boldsymbol{H}_Z \boldsymbol{D}_{x_j} \boldsymbol{Z}) \boldsymbol{Q}_{-(1:m)}$, each of which must equal the zero vector with dimension equal to the number of categories. Noting that $\boldsymbol{x}_h^\top \boldsymbol{D}_{x_j} \boldsymbol{Z} = (\ldots, s_r(j,h), \ldots)$ with $s_r(j,h) = \sum_{r_i = r} x_{ij} x_{ih}$ and $\boldsymbol{x}_h^\top \boldsymbol{H}_Z \boldsymbol{D}_{x_j} \boldsymbol{Z} = (\ldots, n_r \bar{x}_r(j) \bar{x}_r(h), \ldots)$ with $\bar{x}_r(j) = n_r^{-1} \sum_{r_i = r} x_{ij}$, we apply the same arguments as in Theorem 1. $\square$

DRAFT

Table A.1. Characteristics of the North Carolina data ($n = 27,638$).

Variable	Proportion
Mother's race	
NH White	57.9%
NH Black	36.1%
Hispanic	6.0%
Sex	
Male	50.0%
Female	50.0%
Mother's education level	
Did not complete high school	24.2%
Completed high school	36.8%
At least some postsecondary	39.0%
Mother's marital status	
Married at time of birth	56.2%
Not married at time of birth	43.8%
Mother's smoking status	
Smoker	16.9%
Non-smoker	83.1%
Economically disadvantaged	
Yes	61.0%
No	39.0%

Sample proportions by group for each categorical variable. These sample proportions are used for ABCs with each categorical variable. "Economically disadvantaged" is determined by participation in the National Lunch Program.

DRAFT

Table A.2. The (scaled) sample standard deviations $\hat{\sigma}_{x[r]}(j)$ by race r for each covariate $j = 1, \ldots, p$.

Variable j	$\hat{\sigma}_{x[\texttt{NHW}]}(j)$	$\hat{\sigma}_{x[\texttt{NHB}]}(j)$	$\hat{\sigma}_{x[\texttt{Hisp}]}(j)$
Blood lead level	0.949	1.043	0.976
$PM_{2.5}$ exposure	0.999	1.004	0.924
Racial isolation (RI)	0.688	1.063	0.936
Mother's age	0.998	0.974	0.886
Birthweight percentile for gestational age	0.998	0.955	0.984

The invariance result for estimators with and without race-modifiers (Theorem 2) requires $\hat{\sigma}_{x[\texttt{NHW}]}(j) = \hat{\sigma}_{x[\texttt{NHB}]}(j) = \hat{\sigma}_{x[\texttt{Hisp}]}(j)$ for each covariate j (and similarly for the cross-covariances). Although this condition is clearly violated, the estimates and SEs maintain invariance (Figure 2), which suggests strong empirical robustness for the desirable invariance property of ABCs.

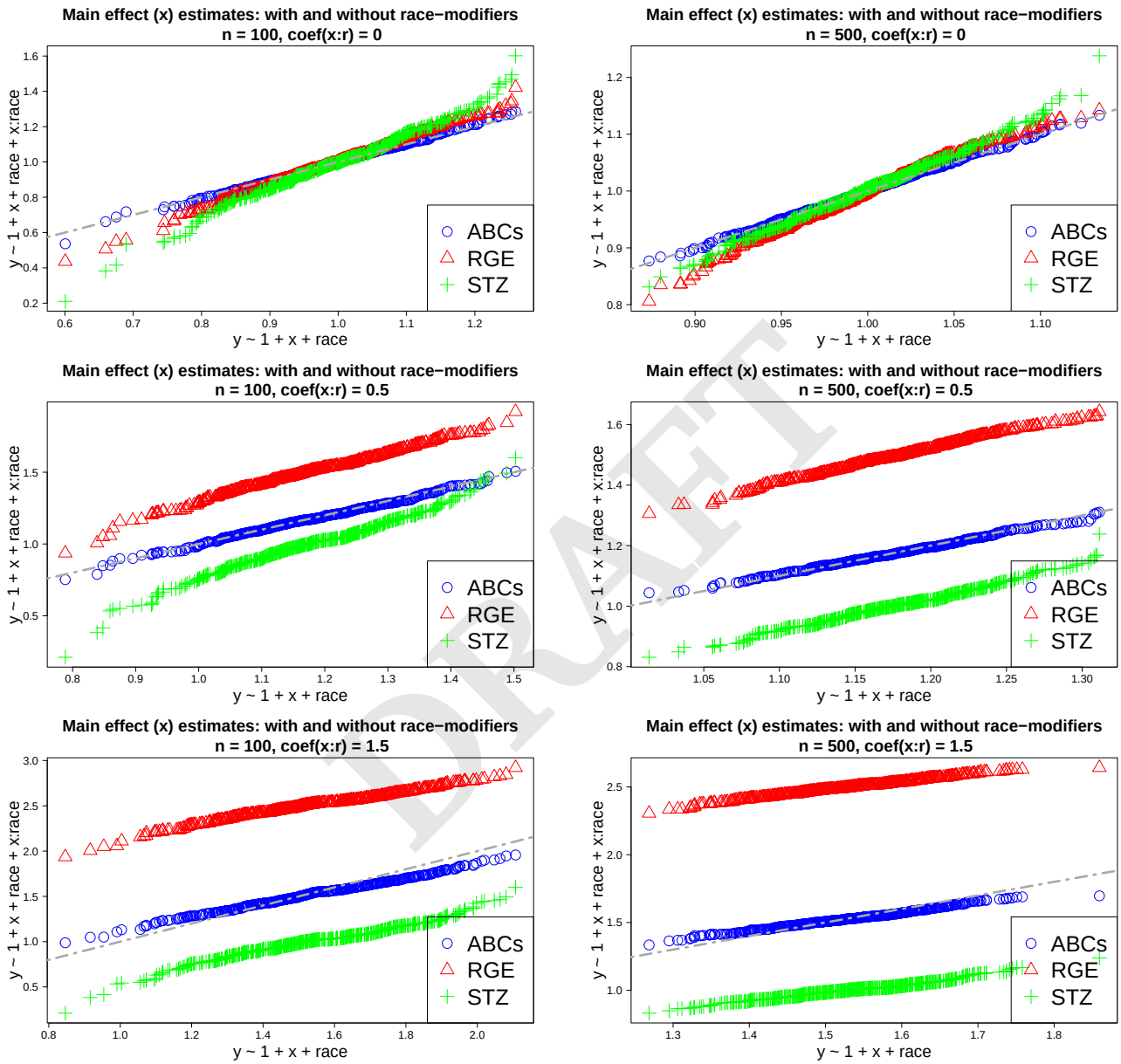

Fig. A.1. Estimated x-effects $(\hat{\alpha}_1^M, \hat{\alpha}_1)$ under different categorical encodings across 500 simulated datasets for $n = 100$ (left) and $n = 500$ (right) and varying race-modifier effects $\gamma \in \{0, 0.5, 1.5\}$ (top to bottom). Uniquely, ABCs produce nearly identical x-effect estimates with and without the race-modifier (45° line), which preserves the interpretations from the simpler (main-only) model.

DRAFT

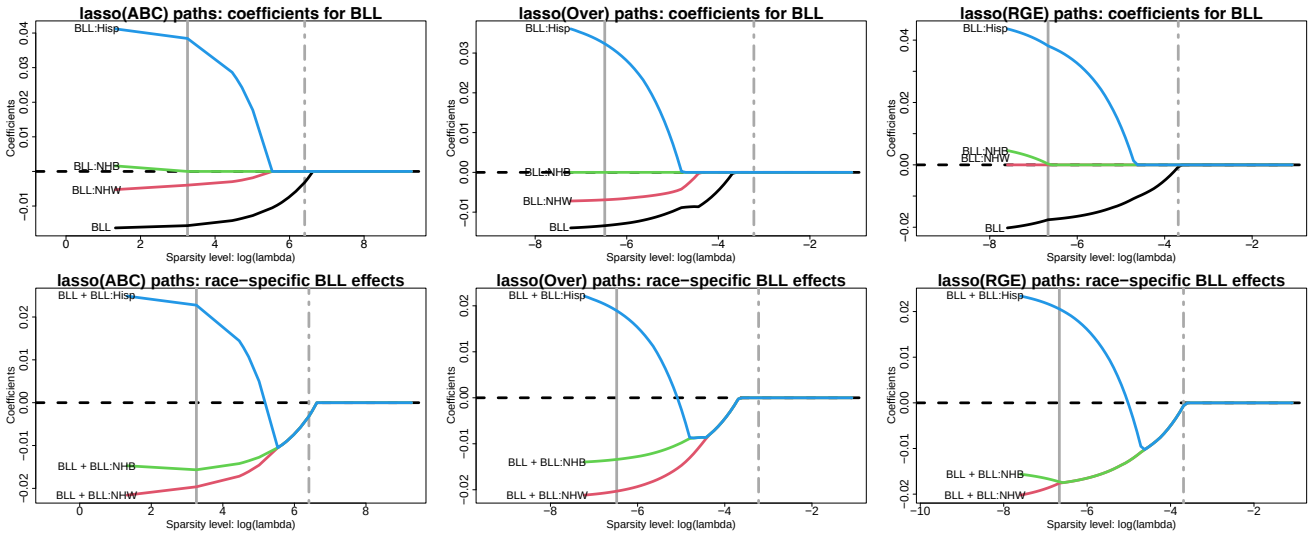

Fig. A.2. Estimated lasso paths for blood lead level (BLL) across varying sparsity levels ($\log \lambda$) for the model coefficients $\hat{\alpha}_{\texttt{BLL}}, \hat{\gamma}_{\texttt{BLL:r}}$ (top) and the race-specific slopes $\hat{\mu}'_{\texttt{BLL}}(r) = \hat{\alpha}_{\texttt{BLL}} + \hat{\gamma}_{\texttt{BLL:r}}$ (bottom) under ABCs (left), overparametrized estimation (center), and RGE (right); vertical lines identify λ for the minimum CV error (solid) and one-standard-error rule (dot-dashed). The outcome is 4th end-of-grade reading score and the covariates include all variables in Table 2. Small λ approximately corresponds to OLS, while increasing λ yields sparsity. Under RGE, the race-specific effects are pulled toward the NH White estimate (bottom right). For overparametrized estimation, the paths are similar to the ABC versions (center and left), but estimate $\hat{\gamma}_{\texttt{BLL:NHB}} = 0$ for all λ and thus implicitly selects NH Black as the reference group. This explains the differences from RGE, which uses a NH White reference group ($\hat{\gamma}_{\texttt{BLL:NHW}} = 0$). Under ABCs, the race-specific effects are pulled toward a global BLL effect (bottom left), which is is nonzero and detrimental for 4th end-of-grade reading scores.

DRAFT

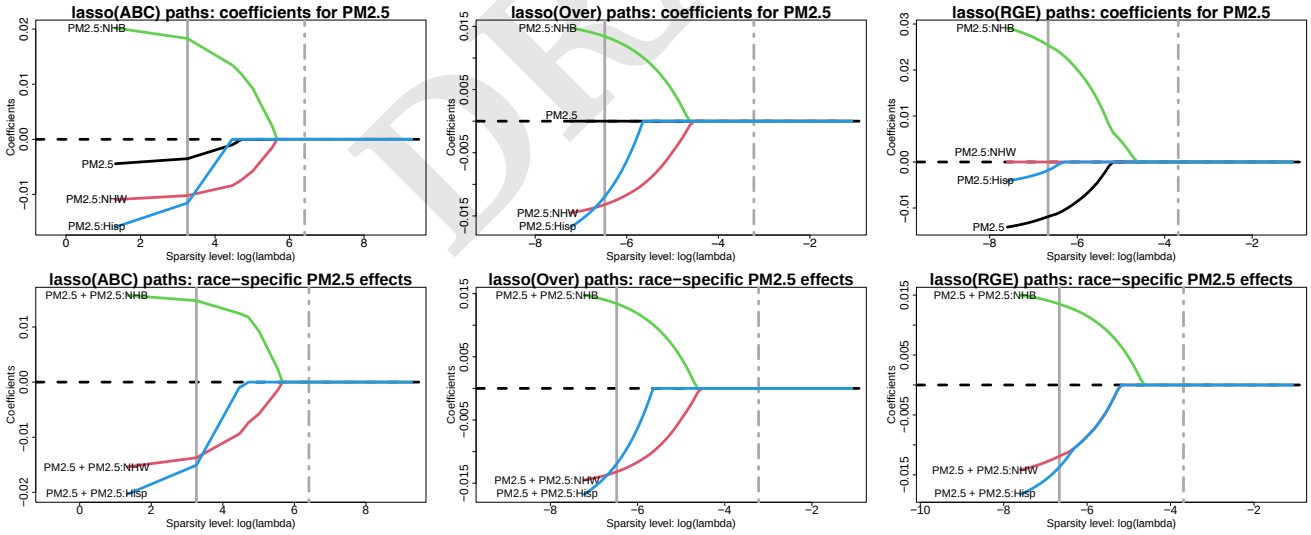

Fig. A.3. Estimated lasso paths for $\text{PM}_{2.5}$ exposure ($\text{PM}_{2.5}$) across varying sparsity levels ($\log \lambda$) for the model coefficients $\hat{\alpha}_{\texttt{PM2.5}}, \hat{\gamma}_{\texttt{PM2.5:r}}$ (top) and the race-specific slopes $\hat{\mu}'_{\texttt{PM}}(r) = \hat{\alpha}_{\texttt{PM2.5}} + \hat{\gamma}_{\texttt{PM2.5:r}}$ (bottom) under ABCs (left) overparametrized estimation (center), and RGE (right); vertical lines identify λ for the minimum CV error (solid) and one-standard-error rule (dot-dashed). The outcome is 4th end-of-grade reading score and the covariates include all variables in Table 2. Small λ approximately corresponds to OLS, while increasing λ yields sparsity. The ABC paths confirm the OLS output: the global $\text{PM}_{2.5}$ effect is pulled toward zero in advance of the race-specific deviations (top left), so the race-specific slopes merge at a global estimate of zero (bottom left). The RGE estimates demonstrate the shrinkage of race-specific effects toward the NH White estimate. The overparameterized paths for the race-specific effects resemble those for ABCs (bottom center and bottom left), but the overparametrized version sets the main effect to zero, $\hat{\alpha}_{\texttt{PM2.5}} = 0$ and thus results in different coefficients compared to either ABCs or RGE (top).

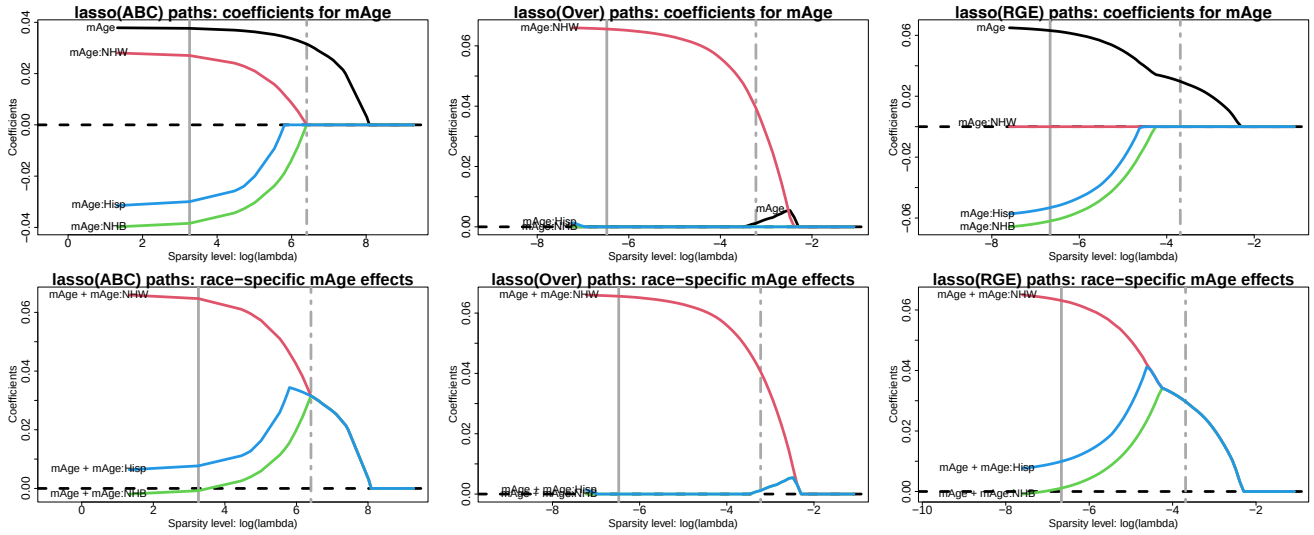

Fig. A.4. Estimated lasso paths for mother's age (mAge) across varying sparsity levels ($\log \lambda$) for the model coefficients $\hat{\alpha}_{\texttt{mAge}}, \hat{\gamma}_{\texttt{mAge:r}}$ (top) and the race-specific slopes $\hat{\mu}'_{\texttt{PM}}(r) = \hat{\alpha}_{\texttt{mAge}} + \hat{\gamma}_{\texttt{mAge:r}}$ (bottom) under ABCs (left) overparametrized estimation (center), and RGE (right); vertical lines identify λ for the minimum CV error (solid) and one-standard-error rule (dot-dashed). The outcome is 4th end-of-grade reading score and the covariates include all variables in Table 2. Small λ approximately corresponds to OLS, while increasing λ yields sparsity. The racial bias of RGE is clear (bottom right): the race-specific effects are each pulled toward the NH White estimate. By comparison, under ABCs, the race-specific effects are pulled toward a global mAge effect (bottom left), which is nonzero and positive for 4th end-of-grade reading scores. The overparametrized estimation cannot determine a reference group, and exhibits erratic behavior that does not resemble either alternative.

DRAFT

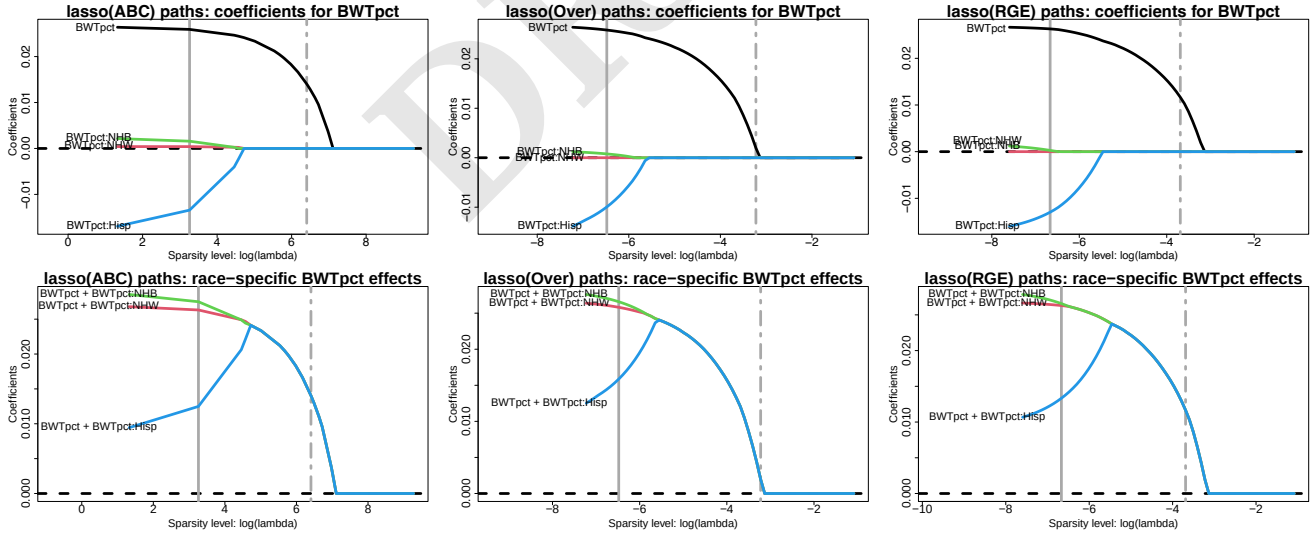

Fig. A.5. Estimated lasso paths for birthweight percentile for gestational age (BWTpct) across varying sparsity levels ($\log \lambda$) for the model coefficients $\hat{\alpha}_{\texttt{BWTpct}}, \hat{\gamma}_{\texttt{BWTpct:r}}$ (top) and the race-specific slopes $\hat{\mu}'_{\texttt{PM}}(r) = \hat{\alpha}_{\texttt{BWTpct}} + \hat{\gamma}_{\texttt{BWTpct:r}}$ (bottom) under ABCs (left) overparametrized estimation (center), and RGE (right); vertical lines identify λ for the minimum CV error (solid) and one-standard-error rule (dot-dashed). The outcome is 4th end-of-grade reading score and the covariates include all variables in Table 2. Small λ approximately corresponds to OLS, while increasing λ yields sparsity. Under ABCs, the race-specific effects are all positive, and merge at a positive global effect of BWTpct. Because the race-specific deviations for NH White individuals are near zero (top left), the RGE paths—which fix these coefficients at zero by design—are very similar to the ABC paths. This effect is similar for overparametrized estimation.

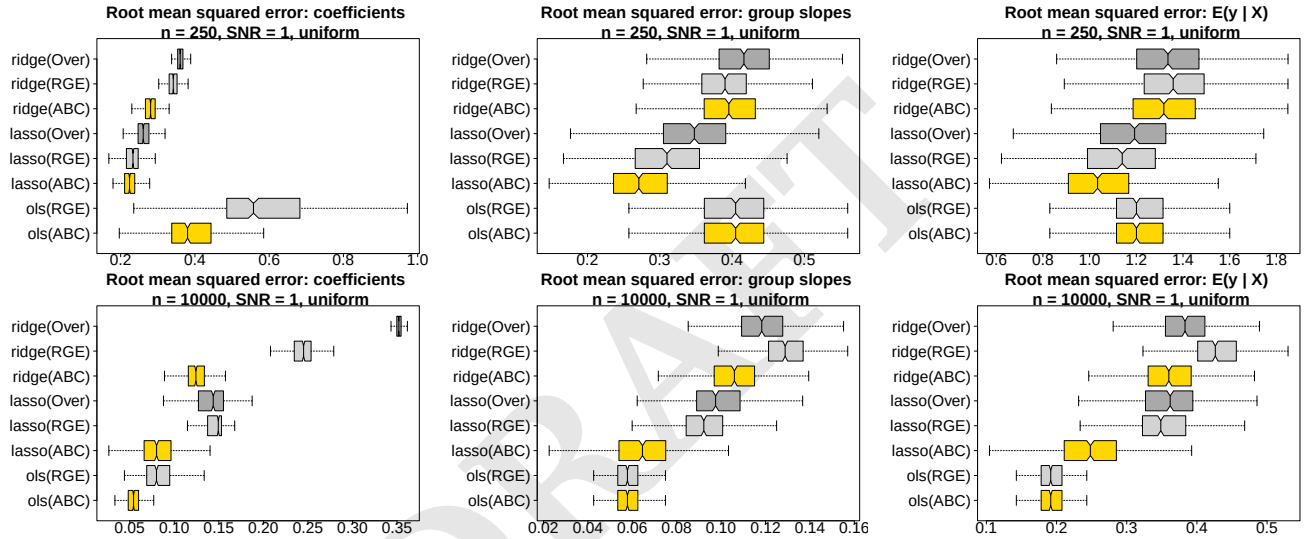

Fig. A.6. Estimation and prediction accuracy for the regression coefficients (left), the race-specific slopes (center), and the fitted values (right) for $n = 250$ (top) and $n = 10{,}000$ (bottom) across 500 simulated datasets; nonoverlapping notches indicate significant differences between medians. Data are generated from a Gaussian main-only model with $p = 10$ covariates and a categorical variable with uniform proportions $\boldsymbol{\pi} = (0.25, 0.25, 0.25, 0.25)^\top$; both RGE and ABCs are satisfied in the true data-generating process. All fitted models use the race-modified model Eq. (1). ABCs (gold) outperform both RGE (light gray) and `Over` (dark gray) within each estimation method (ridge, lasso, OLS). By definition, the OLS race-specific slopes and fitted values are invariant to the constraints (ABCs or RGE), and `Over` cannot be computed for OLS.

DRAFT