Regression with race-modifiers: towards equity and interpretability

Daniel R. Kowal[a,1]

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

This manuscript was compiled on January 4, 2024

The pervasive effects of structural racism and racial discrimination are alarming and well-established, and offer strong evidence that the effects of many important variables on health and life outcomes are *race-specific*. Regression analysis is the predominant statistical tool for estimating and reporting these race-modified effects. However, default numerical encodings of race implicitly elevate one (usually White) race group, leading to racial bias in the presentation of results and, in some cases, racial bias in the estimation and inference for model parameters. We provide an alternative statistical modeling approach using *abundance-based constraints* (ABCs), which eliminate these biases and deliver more interpretable model parameters. ABCs offer a remarkable invariance property: estimation and inference for main covariate effects are approximately unchanged by the inclusion of race-modifiers. Thus, quantitative researchers can include and estimate race-specific effects "for free"—both in terms of model interpretability and statistical efficiency. This framework is especially well-suited for penalized estimation and variable selection (e.g., lasso regression), with meaningful and equitable notions of parameter sparsity and efficient computing strategies. We leverage these statistical tools to estimate the joint effects of multiple environmental, social, and other key factors on 4th end-of-grade readings scores for a large cohort of students ($n = 27,638$) in North Carolina, and identify significant race-specific effects for racial (residential) isolation, $PM_{2.5}$ exposure, and mother's age at birth.

DRAFT

Introduction

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

Significance Statement

The far-reaching impacts of race suggest that the effects of many key variables on health and life outcomes are modified by race. Alarmingly, we show that standard practices for statistical regression analysis introduce racial biases into the estimation and presentation of these race-specific effects. We introduce an alternative approach that eliminates these racial biases, improves interpretability, and provides unique and appealing statistical properties for estimation and selection of race-specific effects. Using these tools, we identify and quantify race-modifiers for environmental, social, and other key factors on child educational outcomes in North Carolina. Our statistical approach offers high-impact potential to quantify, understand, and ultimately reduce the numerous and significant racial disparities in health and life outcomes.

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

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).

However, there are significant racial biases that occur in commonplace estimation, inference, and presentation of results for regression analysis with race as a covariate. We showcase these effects in the following example.

DRAFT

Motivating example. Suppose Y is a continuous outcome variable, X is a continuous covariate of interest, and R is a categorical variable (race). The most widely-used model to estimate both the *main* and *race-specific* effects of X on Y is the linear regression model for the expectation $\mu(x, r) := \mathbb{E}(Y \mid X = x, R = r)$, which we refer to as the *race-modified model*:

$$\mu(x, r) = \alpha_0 + x\alpha_1 + \beta_r + x\beta_r^* \qquad [1]$$

or equivalently, `y ~ 1 + x + race + x:race` in Wilkinson notation. By including X and race (the main effects) and their interaction (the race-modifier), Eq. (1) produces race-specific intercepts and slopes:

$$\mu(0, r) = \alpha_0 + \beta_r \qquad [2]$$

$$\mu_x'(r) := \mu(x + 1, r) - \mu(x, r) = \alpha_1 + \beta_r^* \qquad [3]$$

for each race r. By fitting Eq. (1) to observed data $\{x_i, r_i, y_i\}_{i=1}^n$, the analyst can quantify the effect of X on Y and determine whether this effect varies by race—which is essential information for constructing equitable policies and practices.

However, the race-modified model—which uses "one-hot" or "dummy variable" coding of the categorical variable R—is overparametrized: neither $\{\alpha_0, \beta_r\}_r$ nor $\{\alpha_1, \beta_r^*\}_r$ 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 $\{\beta_r^*\}$, which alters each parameter but leaves Eq. (1) unchanged. Thus, additional constraints are needed for estimation and inference with the race-modified model and its generalizations.

Undoubtedly, the most common identification strategy is *reference group encoding* (RGE): a reference group is selected, typically non-Hispanic (NH) White, and removed ($\beta_{\texttt{NHW}} = \beta^*_{\texttt{NHW}} = 0$, $r = \texttt{NHW}$). This is the default strategy for all major statistical software implementations of (generalized) linear regression, including `R`, SAS, Python, MATLAB, and Stata. However, we argue that RGE output is racially biased (28), difficult to interpret, and obscures important main and race-specific effects. We categorize these significant limitations into *presentation bias* and *statistical bias*.

Presentation bias. Table 1 (left) displays standard output for a regression of 4th end-of-grade reading scores on racial (residential) isolation (RI), race, and their interaction. 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, 29–31). The dataset, detailed subsequently, includes $n = 27,638$ students in North Carolina (58% NH White, 36% NH Black, 6% Hispanic). Using the race-modified model, the goal is to estimate the effect of RI on reading scores and determine if, and how, the RI effect varies by race.

Critically, under RGE, the RI effect (red) actually refers to the RI effect *only for NH White individuals*, $\alpha_1 = \alpha_1 + \beta^*_{\texttt{NHW}} = \mu'_x(\texttt{NHW})$. Similarly, the "Intercept" refers to the NH White intercept, $\alpha_0 = \alpha_0 + \beta_{\texttt{NHW}} = \mu(0, \texttt{NHW})$. First, this output is *inequitable*: it elevates a single race group (NH White) 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: $\beta^*_{\texttt{NHB}} = \mu'_x(\texttt{NHB}) - \alpha_1 = \mu'_x(\texttt{NHB}) - \mu'_x(\texttt{NHW})$. Implicitly, and problematically, this framing presents NH White effects 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. 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 (28). 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-specific effect. Finally, and related, this output is *misleading*: the RI effect is reported to be small and insignificant. This occurs because the RI effect for NH White individuals is estimated to be small and insignificant. However, this output—and RGE more broadly—does not provide a global, race-invariant x-effect (i.e., the coefficient on x). For comparison, consider the *main-only model* that omits the race-modifier,

$$\mu^M(x, r) = \alpha_0^M + x\alpha_1^M + \beta_r^M \qquad [4]$$

DRAFT

where the superscript M emphasizes that these parameters are distinct from those in the race-modified model. Here, the x-effect is global: $\alpha_1^M = \mu^M(x+1, r) - \mu^M(x, r)$ does not depend on the race group r. Fitting the main-only model, we identify a highly significant and detrimental *global* RI effect (Estimate $= -0.042, \text{SE} = 0.007, p < 0.001$; SI Table A1)—which is *not* identified in the default RGE output (Table 1, left). Of course, the main-only model does not estimate race-specific effects and thus is not a satisfactory solution.

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, which obfuscates the highly significant and detrimental effects of RI on reading scores. For reference, the main-only model `y ~ 1 + RI + race` estimates a significantly negative global RI effect (Estimate $= -0.042, \text{SE} = 0.007, p < 0.001$; SI Table A1). The regression output under ABCs (right) eliminates presentation bias, confirms the estimated global RI effect from the main-only model (blue), and clearly highlights the critical result that the adverse RI effect is more than doubled for NH Black students ($\hat{\mu}'_{\texttt{RI}}(\texttt{NHB}) = \hat{\alpha}_{\texttt{RI}} + \hat{\beta}^*_{\texttt{RI:NHB}} = -0.032 + -0.038 = -0.070$).

DRAFT

Statistical bias. The racial inequity in RGE is not limited to the presentation of results (Table 1), but 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, $\beta^*_r \to 0$, 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, $\mu'_x(r) = \alpha_1 + \beta^*_r \to \alpha_1 = \mu'_x(\texttt{NHW})$. First, such an estimator is racially biased. In particular, this bias attenuates the estimated differences between the x-effects for each race and NH White individuals. Identification and quantification of such race-specific effects are precisely the goal of race-modified models and are essential for equitable policy-making. Furthermore, RGE cannot distinguish between shrinkage toward a global, race-invariant x-effect and shrinkage toward the NH White x-effect: both require $\beta^*_r \to 0$ for all $r > 1$. Naturally, a fundamental goal of penalized estimation and selection in this context is to

remove unnecessary race-specific parameters. 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.

Overview. The primary goal of this paper is to introduce, study, and validate alternative statistical methods that eliminate these racial biases and deliver more interpretable results. By carefully reframing the regression model, we uniquely attain the best of both worlds from the race-modified model Eq. (1) and the main-only model Eq. (4): estimation and inference for *race-specific* x-effects as well as *global* x-effects (Table 1, right). This "global" effect interpretation—unavailable to RGE—is the key to removing the racial (presentation and statistical) biases, and is validated theoretically and empirically. We apply these tools to identify and quantify the race-specific effects of multiple environmental, social, and other key factors on 4th end-of-grade readings scores for students in North Carolina.

Alternative approaches. Although RGE is used in the overwhelming majority of regression analyses, there are several alternative identification strategies for Eq. (1) and its generalizations. The most common alternative is *data disaggregation*, which subsets the data by race groups and fits separate regression models (12, 25, 31, 34). This approach produces race-specific intercepts and slopes as in Eq. (2) and Eq. (3), and thus implicitly acknowledges the importance of including 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. Alternatively, *sum-to-zero* (STZ) constraints require that $\sum_r \beta_r = 0$ and $\sum_r \beta_r^* = 0$. While STZ constraints address the inequities in RGE, the resulting model parameters are difficult to interpret, and the estimators do not yield the useful properties offered by the proposed approach (SI Figure A9). Finally, *overparametrized estimation* considers Eq. (1) and its generalizations without any identifying constraints and relies on penalized (lasso, ridge, etc.) regression or Bayesian inference with informative priors to produce unique estimators. However, the model parameters remain nonidentified, which leaves the global x-effect α_1 and the race-specific deviations β_r^* (as well as α_0 and β_r) without a clear interpretation. These alternative strategies are subsequently evaluated using simulated data (Figure 3; SI Figures A1-A4).

We emphasize that the presentation bias and statistical bias of RGE are not resolved merely by changing the reference group (e.g., to NH Black or Hispanic). Furthermore, these challenges apply for any categorical covariates, including gender identity, national origin, religion, and other protected groups. Thus, there is urgent demand for more equitable and interpretable regression analysis with categorical covariates. Here, we focus on race, but the proposed methods apply more broadly.

DRAFT

Methods

We introduce statistical estimation, inference, and algorithms that resolve these racial biases in linear regression. The parameters in Eq. (1) can be identified by constraints of the form $\sum_r c_r \beta_r = 0$ and $\sum_r c_r \beta_r^* = 0$, which enables unique estimation and valid inference. RGE sets $c_1 = 1$ and $c_r = 0$ for $r > 1$, while STZ constraints use $c_r = 1$ for all r. To build intuition for the proposed approach, notice that the generic constraints imply $\alpha_1 = \alpha_1 + \sum_r c_r \beta_r^*$. Fixing $\sum_r c_r = 1$ without loss of generality, $\alpha_1 = \sum_r c_r(\alpha_1 + \beta_r^*) = \sum_r c_r \mu_x'(r)$ is the *average* of the race-specific slopes according to the "race probabilities" $\{c_r\}$ (assuming $c_r \geq 0$ for all r). Clearly, the interpretability and equitability of the parameters in Eq. (1)—as well as the statistical properties of the estimators (Theorem 1)—depend critically on the choice of $\{c_r\}$.

We propose **abundance-based constraints (ABCs)** that elicit these "race probabilities" directly from the data:

$$\sum_r \hat{\pi}_r \beta_r = 0, \quad \sum_r \hat{\pi}_r \beta_r^* = 0, \quad \hat{\pi}_r = \text{proportion in (race) group } r \qquad [5]$$

or equivalently, $\mathbb{E}_{\hat{\pi}}(\beta_R) = 0$ and $\mathbb{E}_{\hat{\pi}}(\beta_R^*) = 0$, where the expectation is taken over the categorical random variable R with $\mathbb{P}(R = r) = \hat{\pi}_r$. If known, the population (race) proportions may be used for $\{\hat{\pi}_r\}$; otherwise, we use the sample proportions based on the data $\{r_i\}_{i=1}^n$. ABCs uniquely identify the model parameters and will be enforced during estimation.

ABCs immediately deliver equity and interpretability for the race-modified model parameters. First, the linear model satisfies $\mathbb{E}_{\hat{\pi}}\{\mu(x, R)\} = \alpha_0 + \alpha_1 x$, which averages Eq. (1) over race to produce a global linear regression. Thus, while the race-modified model produces race-specific intercepts and slopes, the parameters α_0 and α_1 retain a global, race-averaged correspondence. Most important, the main x-effect satisfies

$$\alpha_1 = \sum_r \hat{\pi}_r \mu_x'(r) = \mathbb{E}_{\hat{\pi}}\{\mu_x'(R)\} \qquad [6]$$

which is the *race-averaged slope* for the x-variable. Unlike with RGE, for which $\alpha_1 = \mu_x'(\texttt{NHW})$, ABCs do *not* anchor α_1 to the NH White group and instead provide a global interpretation for this key parameter. The benefits of this interpretation cascade down to the other parameters:

$$\beta_r^* = \mu_x'(r) - \alpha_1 = \mu_x'(r) - \mathbb{E}_{\hat{\pi}}\{\mu_x'(R)\}, \qquad [7]$$

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

$$\alpha_0 = \frac{1}{n}\sum_{i=1}^n (\alpha_0 + \alpha_1 x_i) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\hat{\pi}}\{\mu(x_i, R)\} = \mathbb{E}_{\hat{p}_x}\mathbb{E}_{\hat{\pi}}\{\mu(X, R)\}$$

DRAFT

where the expectation is taken (separately) over $X \sim \hat{p}_x$ for $\hat{p}_x$ the empirical distribution of $\{x_i\}_{i=1}^n$ and $R \sim \hat{\pi}$. In this sense, α_0 is a *marginal* expectation of Y (SI Theorem S1) and is not anchored to the NH White group as in RGE. The race-specific intercepts proceed similarly:

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

Equivalently, the intercept is $\alpha_0 = \mathbb{E}_{\hat{\pi}}\{\mu(\bar{x}, R)\}$ and the race-specific intercepts are $\beta_r = \mu(\bar{x}, r) - \mathbb{E}_{\hat{\pi}}\{\mu(\bar{x}, R)\}$. Again, unlike RGE, the parameter β_r is no longer relative to the NH White group.

In conjunction, these results highlight the interpretability of model parameters under ABCs, including both global and race-specific effects—while critically avoiding the elevation of any single race group (Table 1). These advantages are accentuated by several important, and unique, invariance properties for statistical estimation described subsequently (Theorems 1–2).

Multivariable regression. More commonly, regression analysis features multiple covariates $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$, where the effect of each variable may be modified by race. We generalize the race-modified model Eq. (1) accordingly:

$$\mu(\boldsymbol{x}, r) = \alpha_0 + \boldsymbol{x}^\top \boldsymbol{\alpha} + \beta_r + \boldsymbol{x}^\top \boldsymbol{\beta}_r^* = \alpha_0 + \sum_{j=1}^p x_j \alpha_j + \beta_r + \sum_{j=1}^p x_j \beta_{r,j}^* \qquad [8]$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top$ and $\boldsymbol{\beta}_r^* = (\beta_{r,1}^*, \ldots, \beta_{r,p}^*)^\top$. For model identification, we apply ABCs as in Eq. (5), but now additionally for each covariate, $\sum_r \hat{\pi}_r \beta_{r,j}^* = 0$ for $j = 1, \ldots, p$. ABCs imply that $\mathbb{E}_{\hat{\pi}}\{\mu(\boldsymbol{x}, R)\} = \alpha_0 + \boldsymbol{x}^\top \boldsymbol{\alpha}$, so that averaging Eq. (8) over race yields a traditional multivariate regression model with covariates $\boldsymbol{x}$. Individually, each x_j-effect satisfies

$$\alpha_j = \mathbb{E}_{\hat{\pi}}\{\mu'_{x_j}(R)\} \qquad [9]$$

where $\mu'_{x_j}(r) \coloneqq \mu(x_j + 1, \boldsymbol{x}_{-j}, r) - \mu(x_j, \boldsymbol{x}_{-j}, r)$ is the race-specific slope for the jth covariate. Despite the additional complexity of the multivariable model Eq. (8) relative to Eq. (1), ABCs identify race-averaged, global parameters that do not elevate any single race group. Similarly, $\beta_{r,j}^* = \mu'_{x_j}(r) - \alpha_j = \mu'_{x_j}(r) - \mathbb{E}_{\hat{\pi}}\{\mu'_{x_j}(R)\}$ is the difference between the x_j-effect for race r and the race-averaged x_j-effect. The interpretations of α_0 and $\{\beta_r\}$ proceed as in the previous case.

In contrast, the inequities from RGE are compounded in the multivariable setting. RGE sets $\beta_1 = 0$ and $\beta_{1,j}^* = 0$ for *each* covariate j. Thus, each x_j-effect is presented for or relative to the NH White group, $\alpha_j = \mu'_{x_j}(\texttt{NHW})$ and $\beta_{r,j}^* = \mu'_{x_j}(r) - \mu'_{x_j}(\texttt{NHW})$. Not only are these parametrizations inequitable, but also they fail to provide a global interpretation for each x_j-effect.

ABCs, and the ensuing interpretations, are similarly generalizable for multiple categorical covariates and multiple continuous-categorical interactions (SI Appendix).

DRAFT

Estimation invariance with ABCs. An additional advantage of ABCs is that the global interpretation of the coefficients on x also extends to *estimation*. Specifically, we compare ordinary least squares (OLS) estimates for the main-only model and the race-modified model. These models have distinct parameters with distinct advantages and disadvantages. The primary benefit of the main-only model is that it produces a global x-effect, $\alpha_1^M = \mu^M(x+1, r) - \mu^M(x, r)$; however, it does not admit race-specific effects. By comparison, the race-modified model produces race-specific effects; however, under RGE, there is no global x-effect: $\alpha_1 = \mu(x+1, \texttt{NHW}) - \mu(x, \texttt{NHW})$ is the x-effect only for the NH White group.

Crucially, OLS estimation of the race-modified model *under ABCs* offers the best of both worlds. In addition to estimating race-specific effects, this approach delivers global x-effect estimates that are nearly identical between the main-only and race-modified models, under mild conditions. Specifically, let $\hat{\sigma}^2_{x[r]} = n_r^{-1} s_r^2 - \bar{x}_r^2$ be 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 $\bar{x}_r = n_r^{-1} \sum_{r_i=r} x_i$. When the variation in x is roughly the same within each race group, $\hat{\sigma}^2_{x[r]} \approx \hat{\sigma}^2_{x[1]}$ for all r, then the OLS estimates of the global x-effect are nearly invariant to whether or not the race-modifier (`x:race`) is included.

Theorem 1. *Under ABCs, the OLS estimates for the race-modified model* Eq. (1) *and the main-only model* Eq. (4) *satisfy* $\hat{\alpha}_1 \approx \hat{\alpha}_1^M$ *whenever the (scaled) sample variances of* $\{x_i\}_{i=1}^n$ *are approximately equal across race groups,* $\hat{\sigma}^2_{x[r]} \approx \hat{\sigma}^2_{x[1]}$ *for all* r.

This approximation is empirically robust to moderate violations of the equal-variance condition (SI Figure A9). Theorem 1 does *not* require independence or uncorrelatedness between X and R: the covariate may have dramatically different means and distributions across race groups, as long as the scales are about the same. Thus, the result is distinct from classical estimation invariance results with OLS (35). Further, Theorem 1 makes no assumptions about the true relationship between Y and the covariates. These results persist for multivariable regression:

Theorem 2. *Consider the multivariable race-modified model* Eq. (8) *and the multivariable main-only model*

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

Under ABCs, the OLS estimates satisfy $\hat{\boldsymbol{\alpha}} \approx \hat{\boldsymbol{\alpha}}^M$ *whenever the (scaled) sample covariances of* $\{\boldsymbol{x}_i\}_{i=1}^n$ *are approximately equal across race groups,* $\hat{\sigma}_{x[r]}(j,h) \approx \hat{\sigma}_{x[1]}(j,h)$ *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 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 among the p covariates is approximately the same for each race group. Importantly, this result allows for arbitrary dependencies among the p variables $\boldsymbol{X}$, includes a variety of dependencies between each

DRAFT

x_j and R, and does not impose any assumptions about the true relationship between Y and $\boldsymbol{X}$. Theorems 1 and 2 are presented using the sample (race) proportions for $\{\hat{\pi}_r\}$, but can be modified for population proportions.

We emphasize that Theorems 1 and 2 (and the equal-variance conditions) are *not* required for estimation and inference with ABCs. Rather, these results show that under ABCs, we may include race-modifiers "for free": despite the increase in complexity from main-only models to race-modified models, the estimated x-effects are nearly unchanged. Clearly, these results do not hold for RGE (Table 1, Figure 1) or STZ (SI Figure A9). Yet for ABCs, Theorems 1 and 2 remain accurate approximations (Figure 1), even under moderate violations of the conditions (SI Table A2).

Estimation. Statistical estimation with ABCs requires solving a linearly-constrained least squares problem for models of the form Eq. (8) given data $\{\boldsymbol{x}_i, r_i, y_i\}_{i=1}^n$. Define $\boldsymbol{\theta}$ to be the model parameters $\{\alpha_0, \boldsymbol{\alpha}, \beta_r, \boldsymbol{\beta}_r^*\}_r$ and $\tilde{\boldsymbol{x}}_i$ to include the intercept, covariates, race variable indicators (i.e., "dummy variables"), and covariate-race interactions such that Eq. (8) may be written $\mu(\boldsymbol{x}_i, r_i) = \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta}$. Let $\boldsymbol{C}$ encode ABCs such that $\boldsymbol{C\theta} = \boldsymbol{0}$ enforces $\sum_r \hat{\pi}_r \beta_r = 0$ and $\sum_r \hat{\pi}_r \beta_{r,j}^* = 0$ for $j = 1, \ldots, p$, so $\boldsymbol{C}$ has $m = p + 1$ rows corresponding to the number of constraints. The ABC OLS estimator is

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^n (y_i - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})^2 \quad \text{subject to } \boldsymbol{C\theta} = \boldsymbol{0}. \qquad [11]$$

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 $\boldsymbol{C}^\top = \boldsymbol{QR}$ 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 to verify that $\boldsymbol{\theta} = \boldsymbol{Q}_{-(1:m)}\boldsymbol{\zeta}$ satisfies $\boldsymbol{C\theta} = \boldsymbol{0}$ 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. (11) is equivalently solved using unconstrained OLS:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{Q}_{-(1:m)}\hat{\boldsymbol{\zeta}}, \quad \hat{\boldsymbol{\zeta}} = \arg\min_{\boldsymbol{\zeta}} \sum_{i=1}^n (y_i - \tilde{\boldsymbol{z}}_i^\top \boldsymbol{\zeta})^2. \qquad [12]$$

The QR-decomposition has minimal cost due to the efficiency of Householder rotations and the low dimensionality of $\boldsymbol{C}$ (36).

Although alternative computing strategies are available, the reparametrization in Eq. (12) is especially convenient for generalizations to penalized (lasso, ridge, etc.) estimation. Let $\mathcal{P}(\boldsymbol{\theta})$ denote a complexity penalty on the regression coefficients, typically of the form $\mathcal{P}(\boldsymbol{\theta}) = \sum_j |\theta_j|^\gamma$; $\gamma = 1$ produces sparse coefficients (lasso regression) and $\gamma = 2$ guards against collinearity (ridge regression). The ABC penalized least squares estimator is

$$\hat{\boldsymbol{\theta}}(\lambda) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^n (y_i - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})^2 + \lambda \mathcal{P}(\boldsymbol{\theta}) \quad \text{subject to } \boldsymbol{C\theta} = \boldsymbol{0} \qquad [13]$$

DRAFT

where $\lambda \geq 0$ controls the tradeoff between goodness-of-fit and complexity (measured via $\mathcal{P}$). Following Eq. (12), we instead solve the (unconstrained) penalized least squares problem $\hat{\boldsymbol{\zeta}}(\lambda) = \arg\min_{\boldsymbol{\zeta}} \sum_{i=1}^{n}(y_i - \tilde{\boldsymbol{z}}_i^\top \boldsymbol{\zeta})^2 + \lambda \mathcal{P}(\boldsymbol{Q}_{-(1:m)}\boldsymbol{\zeta})$ and return $\hat{\boldsymbol{\theta}}(\lambda) = \boldsymbol{Q}_{-(1:m)}\hat{\boldsymbol{\zeta}}(\lambda)$. Details are provided in SI Appendix.

Inference. The reparametrization strategy in Eq. (12) allows direct application of classical inference theory to the ABC OLS estimator: $\hat{\boldsymbol{\theta}}$ is a known, linear function of the (unconstrained) OLS 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 (SI Appendix).

Sparsity. In the presence of many covariates, sparsity—enforced by variable selection or penalized (lasso) estimation—becomes increasingly important to eliminate unnecessary parameters, reduce estimation variability, and simplify interpretations. However, the meaning of a zero coefficient depends on the identification strategy. Consider sparsity of the race-modifier, $\beta_r^* = 0$. Under RGE, $\beta_r^* = 0$ implies that the race-specific slope is equal to the NH White group slope, $\mu_x'(r) = \mu_x'(\texttt{NHW}) + \beta_r^* = \mu_x'(\texttt{NHW})$, which elevates the NH White group. Under ABCs, this same sparsity $\beta_r^* = 0$ implies that the race-specific slope is equal to the race-averaged slope, $\mu_x'(r) = \mathbb{E}_{\hat{\pi}}\{\mu_x'(R)\} + \beta_r^* = \mathbb{E}_{\hat{\pi}}\{\mu_x'(R)\}$, thus eliminating the bias and asymmetry from RGE.

An especially concerning case arises when the global x-effect is zero ($\alpha_1 = 0$) but the race-modifier is nonzero ($\beta_r^* \neq 0$). Under ABCs, this occurs when the race-averaged x-effect is zero, $\alpha_1 = \mathbb{E}_{\hat{\pi}}\{\mu_x'(R)\} = 0$, but there exists a nonzero race-specific x-effect, $\mu_x'(r) = \beta_r^* + \alpha_1 = \beta_r^* \neq 0$. However, if the main-only model *without* race-modifiers Eq. (4) were estimated in place of the race-modified model Eq. (1), then the estimated x-effect would be (nearly) zero, $\hat{\alpha}_1^M \approx 0$ (Theorem 2), when in fact the x-effect is both *significant* and *race-specific*. Alarmingly, it is possible that quantitative analyses based on regression models that exclude race-modifiers obscure both important and race-specific effects of certain covariates (Figure 1).

Although it is generally recommended to include interactions ($\beta_r^* \neq 0$) only if main effects are included ($\alpha_1 \neq 0$ or $\beta_r \neq 0$), we argue that ABCs are exempt from this sparsity restriction (37, 38). Specifically, $\beta_r^* \neq 0$ remains meaningful even when $\alpha_1 = 0$ or $\beta_r = 0$: the case of $\alpha_1 = 0$ is considered above, while the interpretation of β_r^* (Eq. (7)) does not depend on β_r at all. Thus, ABCs allow variable selection and penalized estimation without modification.

Results

NC Education Data Analysis. The proposed methods are applied to study the effects of multiple environmental, social, and other key factors on educational outcomes—and assess whether, and how, these effects vary by race. The dataset features a cohort of $n = 27,638$ students in North Carolina (NC) formed by linking three administrative datasets, summarized below and described

DRAFT

in detail elsewhere (31, 39, 40). First, *NC Detailed Birth Records* include maternal and infant characteristics for all documented live births in NC. We construct 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). From residential addresses, we compute racial isolation (RI) at birth, which measures the geographic separation of NH Black individuals and communities from other race groups (29, 30). Second, *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 (41, 42). Third, the *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 another adverse environmental exposure that has been linked to educational outcomes (43). Data characteristics are in SI Table A3.

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, specifically using the sample proportions for each group (SI Table A3). 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 each group in every categorical variable, which eliminates the presentation bias that would otherwise accompany *each* categorical effect (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 ($\hat{\mu}'_{\texttt{RI}}(\texttt{NHB}) = \hat{\alpha}_{\texttt{RI}} + \hat{\beta}^*_{\texttt{RI:NHB}} = -0.020 + -0.020 = -0.040$). 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, and interestingly, $PM_{2.5}$ is not identified as a significant main effect ($p = 0.403$), yet the race-specific effects *are* significant. Alarmingly, a fitted model *without* the race-modifiers conveys an insignificant $PM_{2.5}$ effect (Figure 1), which oversimplifies and misleads.

To showcase the utility and empirical validity of the ABCs invariance property (Theorem 2), Figure 1 presents the estimates and 95% confidence intervals for the main effects that are modified

DRAFT

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 $\times$ 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
$PM_{2.5}$ $\times$ 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 $\times$ 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 $\times$ 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 $\times$ 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.

DRAFT

by race. Specifically, we compare the model output for regressions *with* race-modifiers (i.e., all variables in Table 2) and *without* race-modifiers (i.e., variables only in the left column of Table 2), including both ABC and RGE output for the race-modified models. Remarkably, the estimates and uncertainty quantification for the simpler, main-only model are nearly indistinguishable from those for the expanded, race-modified model under ABCs. Effectively, ABCs allow estimation and inference for numerous 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 Theorem 2, despite moderate violations of the equal-variance condition (SI Table A2). 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.

Finally, we assess penalized estimation and variable selection with ABCs using lasso regression, including all the covariates from Table 2. We report estimates across tuning parameter values λ for the model coefficients $\hat{\alpha}_j, \{\hat{\beta}^*_{r,j}\}_r$ and the race-specific slopes $\{\hat{\mu}'_{x_j}(r) = \hat{\alpha}_j + \hat{\beta}^*_{r,j}\}_r$; $\lambda \to 0$ yields OLS estimates, while $\lambda \to \infty$ yields sparse estimates. Since the penalized estimates depend critically on

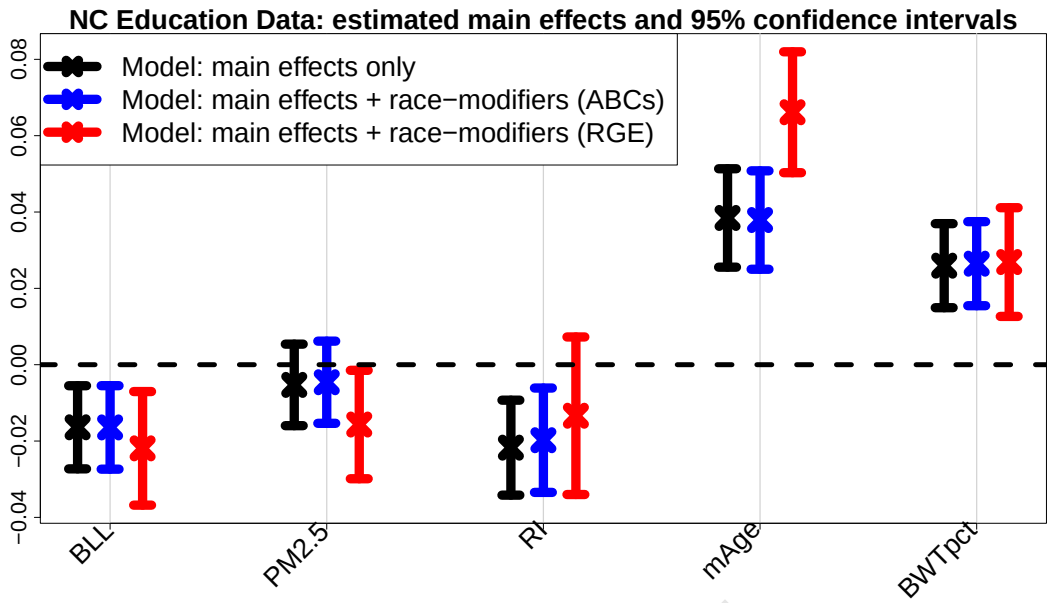


Fig. 1. Estimates and 95% confidence intervals for the main effects in the multivariable regression *without* race-modifiers (black) and the multivariable regression *with* race-modifiers under ABCs (blue) and RGE (red). Results are presented for blood lead level (BLL), $PM_{2.5}$ exposure ($PM_{2.5}$), 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 (Theorem 2): 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_j-effects for the NH White group and deviate substantially for $PM_{2.5}$, RI, and mAge, including shifts in location and much wider intervals.

DRAFT

the parameterization, we compare ABCs and RGE. The estimated λ-paths for RI are in Figure 2; results for the remaining race-modified effects (BLL, $PM_{2.5}$, mAge, and BWTpct) are in SI Appendix (Figures A1-A4). Notably, RGE fixes $\hat{\beta}^*_{r,j} = 0$ 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 2, top right). ABCs resolve these issues. First, the model parameters are separately and equitably pulled toward zero (Figure 2, top left). Second, the RI effect $\hat{\alpha}_j$ is *not* attenuated, and preserves its magnitude until $\log \lambda \approx 5$ (Figure 2, top left). Finally, the race-specific RI effects merge at a *global*, and negative, RI effect estimate, which is selected by the one-standard-error rule (33) for choosing λ (Figure 2, bottom left).

These themes persist for the remaining race-modified effects (SI Figures A1-A4). We supplement the ABC and RGE lasso paths by including the lasso paths for *overparametrized estimation* (`Over`), which also uses dummy variables to encode all categorical variables and race-modifiers, but does not include any identifiability constraints. These parameters cannot be estimated by OLS, but can be estimated by lasso regression with $\lambda > 0$. Perhaps unsurprisingly, `Over` lasso estimation typically results in one of the coefficients $\{\alpha_j, \beta^*_{r,j}\}_r$ being set to zero immediately (i.e., for small λ) for each covariate j. This induces an identification similar to RGE, but without selecting the reference group in advance. Thus, it suffers from the same racial biases in estimation and selection that plague RGE. When this implicit selection sets $\hat{\beta}^*_{j:\texttt{NHW}} = 0$, then the `Over` paths resemble those for RGE (RI, not

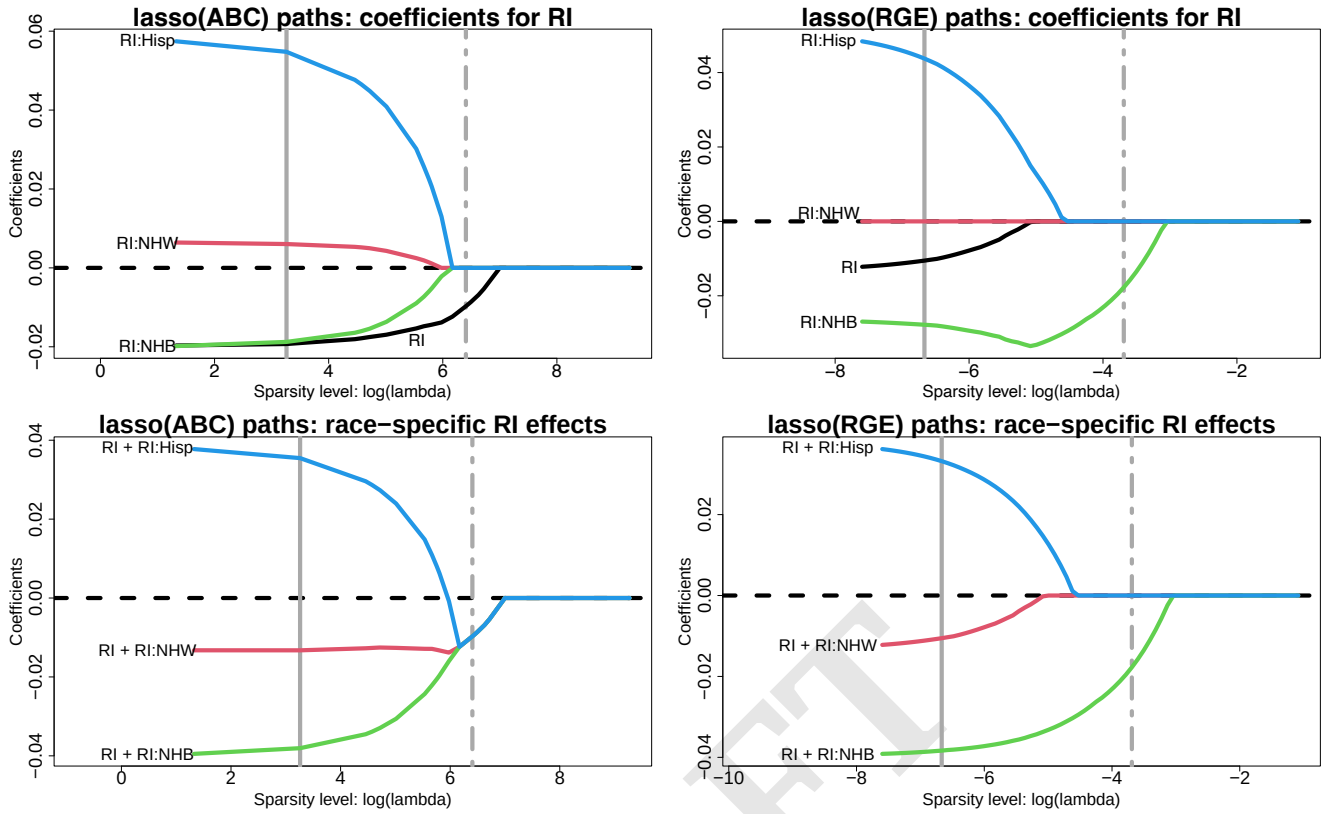


Fig. 2. Estimated lasso paths for RI across varying sparsity levels ($\log \lambda$) for the model coefficients $\hat{\alpha}_{\texttt{RI}}, \hat{\beta}^*_{\texttt{RI:r}}$ (top) and the race-specific slopes $\hat{\mu}'_{\texttt{RI}}(r) = \hat{\alpha}_{\texttt{RI}} + \hat{\beta}^*_{\texttt{RI:r}}$ (bottom) 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

shown; BWTpct, SI Figure A4); when the selection corresponds to the smallest $|\hat{\beta}^*_{r,j}|$ among race groups r from ABCs, then the `Over` and ABC paths are similar (BLL, SI Figure A1; BWTpct, SI Figure A4). However, when this selection sets the main effect to zero, $\hat{\alpha}_j = 0$ ($\text{PM}_{2.5}$, SI Figure A2) or overshrinks multiple coefficients toward zero (mAge, SI Figure A3), then the `Over` paths differ substantially from both the RGE and ABC paths and demonstrate erratic behavior (SI Figure A3).

Simulation results. We evaluate the performance of ABCs for estimation, prediction, and inference using simulated data. Each simulated dataset is constructed from a Gaussian linear regression model Eq. (10) with $p = 10$ continuous covariates and one categorical (race) covariate with 4 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 from small ($n = 50$), moderate ($n = 250$), and large ($n = 10{,}000$)

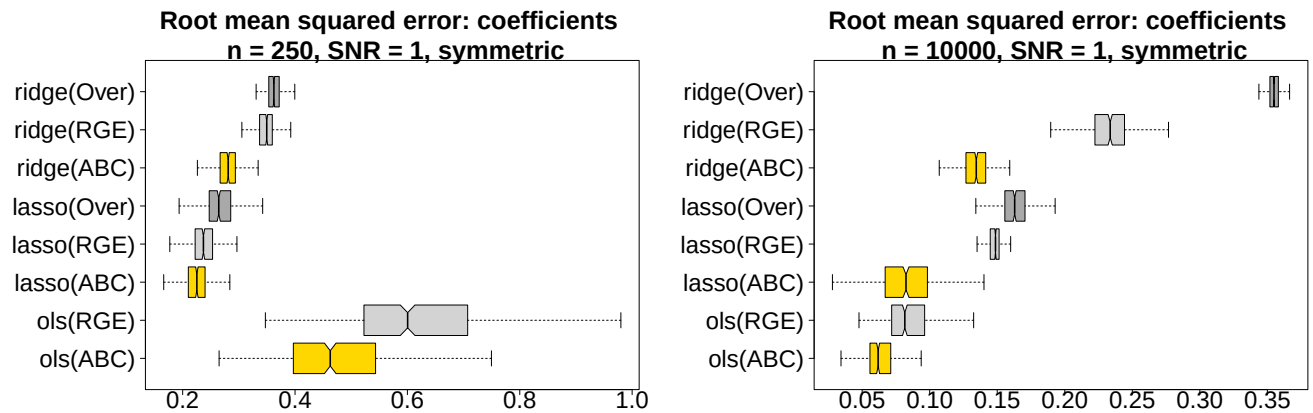


Fig. 3. Evaluating point estimates for the regression coefficients for $n = 250$ (left) and $n = 10{,}000$ (right) across 500 simulated datasets; nonoverlapping notches indicate significant differences between medians. These estimands are *not* invariant to the categorical encoding (or constraints), but ABCs and RGE are both satisfied in the data-generating process. ABCs (gold) outperform the other encodings (gray) within each estimation method.

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 construct a data-generating model that satisfies both RGE and ABCs. We consider two designs for the categorical proportions: symmetric weights $\boldsymbol{\pi} = (0.15, 0.35, 0.15, 0.35)^\top$ and uniform weights $\boldsymbol{\pi} = (0.25, 0.25, 0.25, 0.25)^\top$ (SI Appendix). 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. (8), 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 $\beta^*_{r,j}$: we consider ABCs, RGE, and `Over`. This latter option is not identified for OLS, and thus is presented only for ridge and lasso regression.

Estimation of the model parameters $\boldsymbol{\theta}$ is evaluated using root mean squared error (Figure 3). Within each estimation method (OLS, ridge, lasso), ABCs offer substantial improvements over both RGE and `Over`. Similar gains occur for prediction (SI Figure A5), estimation of the race-specific slopes $\alpha_j + \beta^*_{r,j}$ (SI Figure A6), and confidence intervals (SI Figure A8), and specifically, ABCs produce narrower confidence intervals—and thus more powerful inference—while maintaining nominal coverage.

DRAFT

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", i.e., without sacrificing the interpretability of the global (race-invariant) main effects. Finally, ABCs are especially convenient for penalized estimation and variable selection (e.g., lasso regression), 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 key 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 previous approaches; and 3) provided more equitable and interpretable estimates, uncertainty quantification, and selection, both for main effects and race-modified effects.

We compared ABCs against several alternative strategies for linear regression with categorical variables, with evaluations on real and simulated data for OLS and penalized (lasso and ridge) regression. For OLS estimation, a related approach to encode identifiability constraints is *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. 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 simple ANOVA or main-only models and are difficult to combine with penalized estimation, variable selection, and Bayesian regression. 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" (44) or "certain types of unbalanced data that are missing not at random" (45), with regression output that suffers from the same presentation bias that afflicts RGE (46). We argue more forcefully that ABCs provide a necessary resolution to certain racial biases that occur under default (RGE) approaches. Similar challenges arise for other categorical covariates, including gender identity, national origin, religion, and other protected

DRAFT

groups, and thus the potential impact of ABCs extends well beyond race.

We acknowledge that the interpretation of any "race" effect requires great care (47). Race encompasses a vast array of social and cultural factors and life experiences, with effects that vary across time and geography (27, 48). In some settings, race data are unreliable or partially missing (49, 50). These overarching challenges are not addressed in this paper.

ACKNOWLEDGMENTS. We thank Amy Willis for helpful feedback and discussions that greatly improved this work. Research was sponsored by the National Institute of Environmental Health Sciences (R01ES028819) and the National Science Foundation (SES-2214726). The content is solely the responsibility of the author(s) and does not necessarily represent the official views of the NIH or the U.S. government. The findings and conclusions in this publication are those of the author(s) and do not necessarily represent the views of the North Carolina Department of Health and Human Services, Division of Public Health.

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).

DRAFT

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. SS Johfre, J Freese, Reconsidering the reference category. *Sociol. Methodol.* **51**, 253–269 (2021).
29. SF Reardon, D O'Sullivan, Measures of spatial segregation. *Sociol. methodology* **34**, 121–162 (2004).
30. 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).
31. 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).
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. R Frisch, FV Waugh, Partial time regressions as compared with individual trends. *Econom. J. Econom. Soc.* pp. 387–401 (1933).
36. S Wood, *Generalized additive models: an introduction with R*. (CRC press), (2006).
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. CEH Initiative, Linked births, lead surveillance, grade 4 end-of-grade (eog) scores [data set] (2020).
40. DR Kowal, et al., Bayesian variable selection for understanding mixtures in environmental exposures. *Stat. Medicine* **40**, 4850–4871 (2021).
41. 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).
42. 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).
43. Y Gao, DR Kowal, Bayesian adaptive and interpretable functional regression for exposure profiles. *Annals Appl. Stat.* (2023).
44. MT Grotenhuis, et al., When size matters: advantages of weighted effect coding in observational studies. *Int. J. Public Heal.* **62**, 163–167 (2017).
45. L Brehm, PM Alday, Contrast coding choices in a decade of mixed models. *J. Mem. Lang.* **125**, 104334 (2022).
46. R Nieuwenhuis, HF te Grotenhuis, BJ Pelzer, Weighted effect coding for observational data with wec. (2017).
47. TJ VanderWeele, WR Robinson, On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology* **25**, 473 (2014).
48. DR Williams, Race and health: basic questions, emerging directions. *Annals Epidemiol.* **7**, 322–333 (1997).
49. EV Klinger, et al., Accuracy of race, ethnicity, and language preference in an electronic health record. *J. general internal medicine* **30**, 719–723 (2015).
50. FCG Polubriaginof, et al., Challenges with quality of race and ethnicity data in observational databases. *J. Am. Med. Informatics Assoc.* **26**, 730–736 (2019).
51. 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).
52. MC Lovell, Seasonal adjustment of economic time series and multiple regression analysis. *J. Am. Stat. Assoc.* **58**, 993–1010 (1963).

DRAFT

A. Supporting Information

Regression 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.:

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

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top$ and $\boldsymbol{\beta}^*_{\ell, r_\ell} = (\beta^*_{\ell, r_\ell, 1}, \ldots, \beta^*_{\ell, r_\ell, p})^\top$ are p-dimensional and r_ℓ denotes the level of the ℓth categorical variable, $\ell = 1, \ldots, L$. Eq. (A.1) includes all continuous-categorical interactions, but the subsequent interpretations remain applicable when only a subset of these interactions are included. Eq. (A.1) requires $L(1+p)$ constraints for identification; reference group encoding (RGE) sets $\beta_{\ell,1} = 0$ for all ℓ and $\beta^*_{\ell,1,j} = 0$ 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

$$\mathbb{E}_{\hat{\boldsymbol{\pi}}}(\boldsymbol{\beta}_{\boldsymbol{R}}) = \boldsymbol{0}_L, \quad \mathbb{E}_{\hat{\boldsymbol{\pi}}}(\boldsymbol{\beta}_{\boldsymbol{R},j}) = \boldsymbol{0}_L, \quad j = 1, \ldots, p \qquad [A.2]$$

where $\boldsymbol{\beta}_{\boldsymbol{R}} = (\beta_{1,R_1}, \ldots, \beta_{L,R_L})^\top$, $\boldsymbol{\beta}_{\boldsymbol{R},j} = (\beta^*_{1,R_1,j}, \ldots, \beta^*_{L,R_L,j})^\top$, and $\boldsymbol{0}_L$ is a vector of zeros. The joint constraints in Eq. (A.2) may be equivalently represented via separate marginal expectations for the L sets of categorical covariate parameters: for instance, $\mathbb{E}_{\hat{\boldsymbol{\pi}}}(\beta_{\ell,R_\ell}) = \mathbb{E}_{\hat{\pi}_\ell}(\beta_{\ell,R_\ell}) = \sum_{r_\ell} \hat{\pi}_{\ell,r_\ell} \beta_{\ell,r_\ell} = 0$.

ABCs in Eq. (A.2) provide interpretable parameter identifications with equitable presentation and estimation. These interpretations are unchanged if some or all interaction terms are omitted from Eq. (A.1), which may occur if multiple categorical variables (e.g., sex, education level) are included as 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. (A.1) over all categorical variables (jointly) yields a multivariate regression with only continuous variables. Individually, each x_j-effect satisfies

$$\alpha_j = \mathbb{E}_{\hat{\boldsymbol{\pi}}}\{\mu'_{x_j}(\boldsymbol{R})\} \qquad [A.3]$$

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. (A.3) 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 $\beta^*_{\ell,r_\ell,j}$ 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. (A.3), 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

DRAFT

x_j-effect when the ℓth categorical variable has level r_ℓ, averaged over the remaining categorical variables, is

$$\mathbb{E}_{\hat{\boldsymbol{\pi}}_{-\ell}}\{\mu'_{x_j}(r_\ell, \boldsymbol{R}_{-\ell})\} = \alpha_j + \beta^*_{\ell,r_\ell,j} \qquad [A.4]$$

or equivalently, $\beta^*_{\ell,r_\ell,j} = \mathbb{E}_{\hat{\boldsymbol{\pi}}_{-\ell}}\{\mu'_{x_j}(r_\ell, \boldsymbol{R}_{-\ell})\} - \mathbb{E}_{\hat{\boldsymbol{\pi}}}\{\mu'_{x_j}(\boldsymbol{R})\}$. The interpretation is simpler than the notation: Eq. (A.4) directly extends the usual notion of race-specific slopes to average over any other categorical variables that modify x_j.

Estimation invariance of ABCs: proofs and additional results. Estimation with ABCs delivers multiple invariance properties. First, we supplement the results from the main paper with a result for ordinary least squares (OLS) estimation of the intercept.

Theorem S1. For any main-only regression model of the form Eq. (A.1) *without* interactions ($\beta^*_{\ell,r_1,j} = 0$), the OLS-estimated intercept equals the sample mean, $\hat{\alpha}_0 = \bar{y}$, whenever 1) the categorical variable coefficients satisfy ABCs Eq. (A.2) and 2) the continuous variables are centered ($\bar{x}_j = 0$).

Proof (Theorem S1). Under OLS, $\bar{y}$ equals the sample mean of the fitted values $\{\hat{y}_i\}_{i=1}^n$; this is true for ABCs, RGE, STZ, etc. Then we simplify:

$$\begin{aligned}
\bar{y} &= n^{-1}\sum_{i=1}^n \hat{y}_i = n^{-1}\sum_{i=1}^n (\hat{\alpha}_0 + \boldsymbol{x}_i^\top \hat{\boldsymbol{\alpha}} + \sum_{\ell=1}^L \hat{\beta}_{\ell,r_\ell}) \\
&= \hat{\alpha}_0 + \bar{\boldsymbol{x}}^\top \hat{\boldsymbol{\alpha}} + \sum_{\ell=1}^L \sum_{r_\ell} \hat{\pi}_{\ell,r_\ell}\hat{\beta}_{\ell,r_\ell} \\
&= \hat{\alpha}_0
\end{aligned}$$

since the continuous covariates are centered and each set of categorical variable coefficients satisfies ABCs. □

Under ABCs, the estimated intercept is invariant across *all* main-only models, regardless of the true association between Y and the covariates. This result is unique to ABCs and provides motivation to use ABCs even without a continuous-categorical interaction, such as for traditional ANOVA models. A similar result was noted by (51) in a simplified setting.

Next, we proceed to prove Theorems 1 and 2. To do so, we leverage a classical property of OLS (35, 52).

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.

DRAFT

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 $\boldsymbol{D}_X = \mathrm{diag}(\boldsymbol{x})$. 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 $(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{Z}_X = \mathrm{diag}(\{\bar{x}_r\}_r)$ contains the sample means of $\{x\}_{i=1}^n$ by each group r, and therefore $\boldsymbol{x}^\top \boldsymbol{H}_Z \boldsymbol{Z}_X = \boldsymbol{x}^\top \boldsymbol{Z} \mathrm{diag}(\{\bar{x}_r\}_r) = (\ldots, n_r \bar{x}_r^2, \ldots)$ 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 $\boldsymbol{D}_{X_j} = \mathrm{diag}(\boldsymbol{x}_j)$. 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 latter 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. □

Penalized estimation with ABCs. Recall that $\boldsymbol{\theta}$ is the model parameters $\{\alpha_0, \boldsymbol{\alpha}, \beta_r, \boldsymbol{\beta}_r^*\}_r$, $\tilde{\boldsymbol{x}}_i$ includes the intercept, covariates, race variable indicators, and covariate-race interactions such that $\mu(\boldsymbol{x}_i, r_i) = \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta}$, and $\boldsymbol{C}$ encodes ABCs such that $\boldsymbol{C\theta} = \boldsymbol{0}$ enforces $\sum_r \hat{\pi}_r \beta_r = 0$ and $\sum_r \hat{\pi}_r \beta_{r,j}^* = 0$ for $j = 1, \ldots, p$. The ABC penalized least squares estimator is

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

DRAFT

where $\mathcal{P}(\boldsymbol{\theta})$ is a complexity penalty on the regression coefficients and $\lambda \geq 0$ controls the tradeoff between goodness-of-fit and complexity (measured via $\mathcal{P}$). Here, we consider complexity penalties of the form

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

where $\omega_j > 0$ are known weights and typically $\gamma = 1$ for (adaptive) lasso regression or $\gamma = 2$ for (adaptive) ridge regression.

Adopting the same reparametrization strategy as in the unpenalized case, 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. 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 ABC penalized least squares estimator is

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

which requires the solution to an *unconstrained* penalized least squares problem. For instance, under ridge regression ($\gamma = 2$), 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 ($\gamma = 1$) may be solved efficiently using the `genlasso` package in `R`.

For practical use, we set ω_j to be the sample standard deviation of the $j$the 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 race group r: by design, small values of $\hat{\pi}_r$ inflate the corresponding coefficients β_r and $\beta^*_{r,j}$. The standardized penalty adjusts for this effect to avoid overpenalization of race-specific coefficients for groups with low abundance.

Inference with ABCs. 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 OLC 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 $\epsilon_i := y_i - \mu(\boldsymbol{x}_i, r_i)$ 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

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

DRAFT

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.

Simulation results: approximate 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 $\{\texttt{A}, \texttt{B}, \texttt{C}, \texttt{D}\}$ and respective probabilities $\boldsymbol{\pi} = (0.55, 0.20, 0.10, 0.15)^\top$ and then sample a continuous variable with the distribution determined by the group:

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

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. (A.5) includes a mild deviation from the equal-variance condition of Theorem 1.

The response variable Y is simulated with expectation

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

plus $t_4(0,1)$-distributed errors, i.e., standard t-distributions with 4 degrees of freedom. The coefficient η determines the strength of the race-modifier effect: we consider $\eta = 0$ (no race-modifier effect), $\eta = 0.5$ (50% of the main x-effect), and $\eta = 1.5$ (150% of the main x-effect). This data-generating process includes a R-modifier (when $\eta \neq 0$) 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 $\{\beta_r, \beta_r^*\}$ in Figure A.9. 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 preserves the interpretations from the simpler model, and is *not* satisfied for RGE or STZ constraints. The estimated x-effects under RGE or STZ constraints vary considerably between the main-only and race-modified models, with greater discrepancies as the magnitude of the race-modifier effect η increases. By comparison, the estimation invariance of ABCs is quite robust to the magnitude of the race-modifier effect. The mild deviations from the equal-variance condition of Theorem 1 are most impactful when $\eta = 1.5$, which represents the unusual setting in which the interaction effect is much larger than the main effect. Yet 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 estimates.

DRAFT

Table A.1. Linear regression output for racial isolation (RI) and mother's race on 4th end-of-grade reading scores for students in North Carolina (`y ~ 1 + RI + race`).

Variable	Estimate (SE)	p-value
Intercept	0.000 (0.006)	1.000
Racial isolation (RI)	-0.042 (0.007)	<0.001
Mother's race		
NH White	0.255 (0.006)	<0.001
NH Black	-0.421 (0.009)	<0.001
Hispanic	0.078 (0.022)	<0.001

In contrast with Table 1, this model uses only main effects (RI and mother's race) without any race-modifiers, thus providing a global estimate of the RI effect. ABCs are used for mother's race; however, the RI estimates, SEs, and p-values are the same under RGE. The RI output (blue) is nearly identical to the race-modified model output under ABCs (Table 1, right), which confirms the estimation invariance of ABCs (Theorem 1). By comparison, the race-modified model output under RGE (Table 1, left) deviates substantially, and instead refers to the RI effect *only* for NH White students.

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 demonstrate near-perfect invariance (Figure 1), which suggests strong empirical robustness for the desirable invariance property of ABCs.

Table A.3. 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

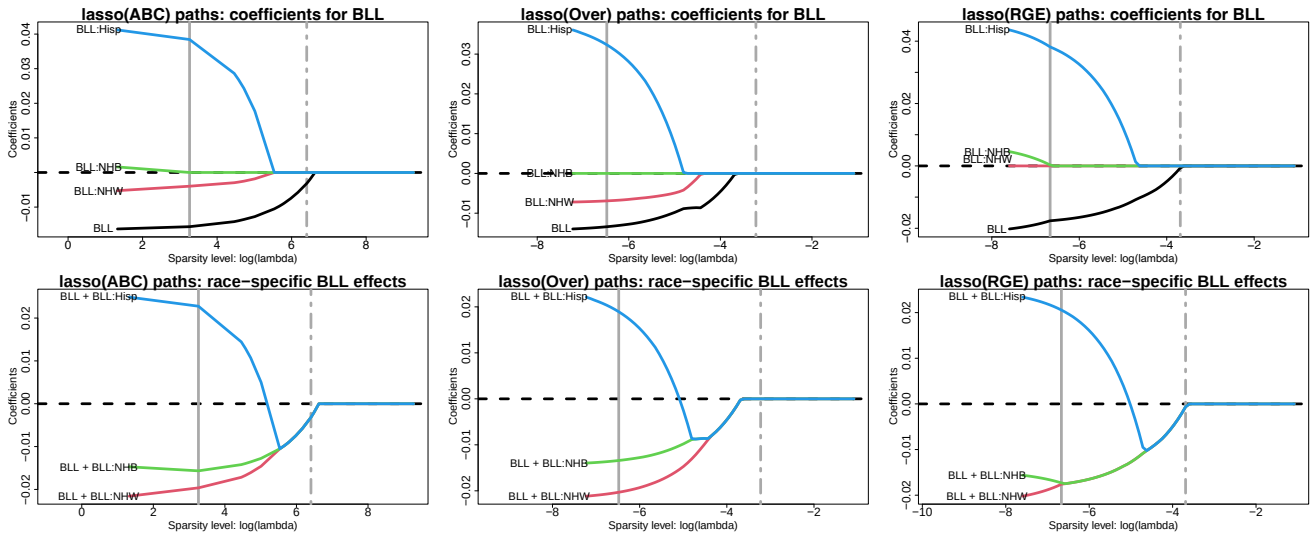


Fig. A.1. Estimated lasso paths for blood lead level (BLL) across varying sparsity levels ($\log \lambda$) for the model coefficients $\hat{\alpha}_{\texttt{BLL}}, \hat{\beta}^*_{\texttt{BLL:r}}$ (top) and the race-specific slopes $\hat{\mu}'_{\texttt{BLL}}(r) = \hat{\alpha}_{\texttt{BLL}} + \hat{\beta}^*_{\texttt{BLL:r}}$ (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 $\hat{\beta}^*_{\texttt{BLL:NHB}} = 0$ 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 ($\hat{\beta}^*_{\texttt{BLL:NHW}} = 0$). 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

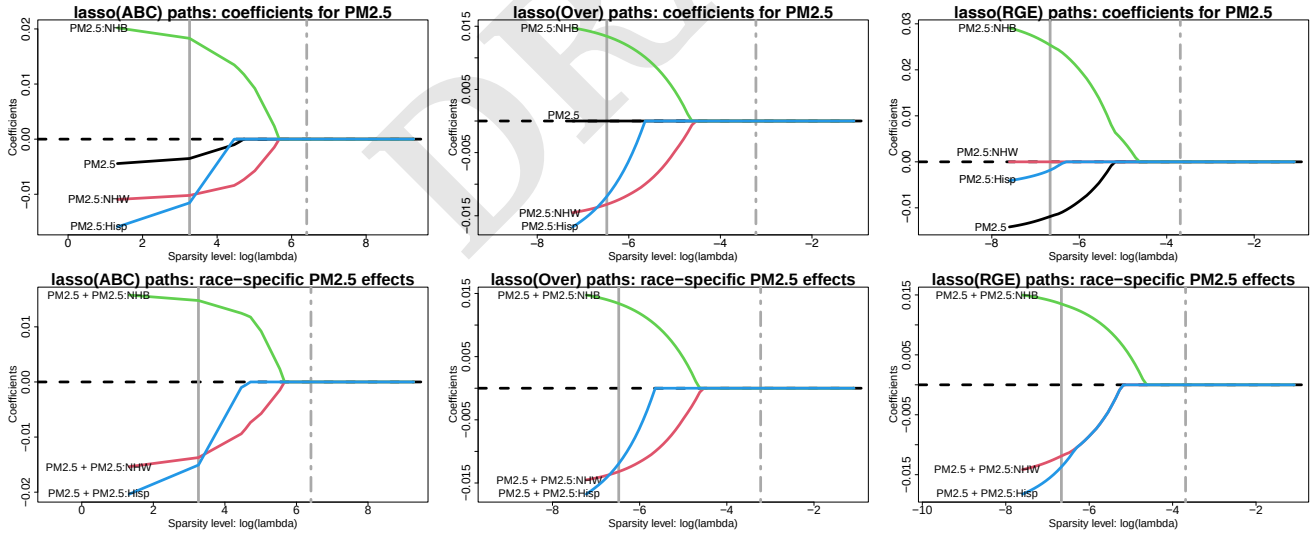


Fig. A.2. Estimated lasso paths for $\text{PM}_{2.5}$ exposure ($\text{PM}_{2.5}$) across varying sparsity levels ($\log \lambda$) for the model coefficients $\hat{\alpha}_{\texttt{PM2.5}}, \hat{\beta}^*_{\texttt{PM2.5:r}}$ (top) and the race-specific slopes $\hat{\mu}'_{\texttt{PM}}(r) = \hat{\alpha}_{\texttt{PM2.5}} + \hat{\beta}^*_{\texttt{PM2.5:r}}$ (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).

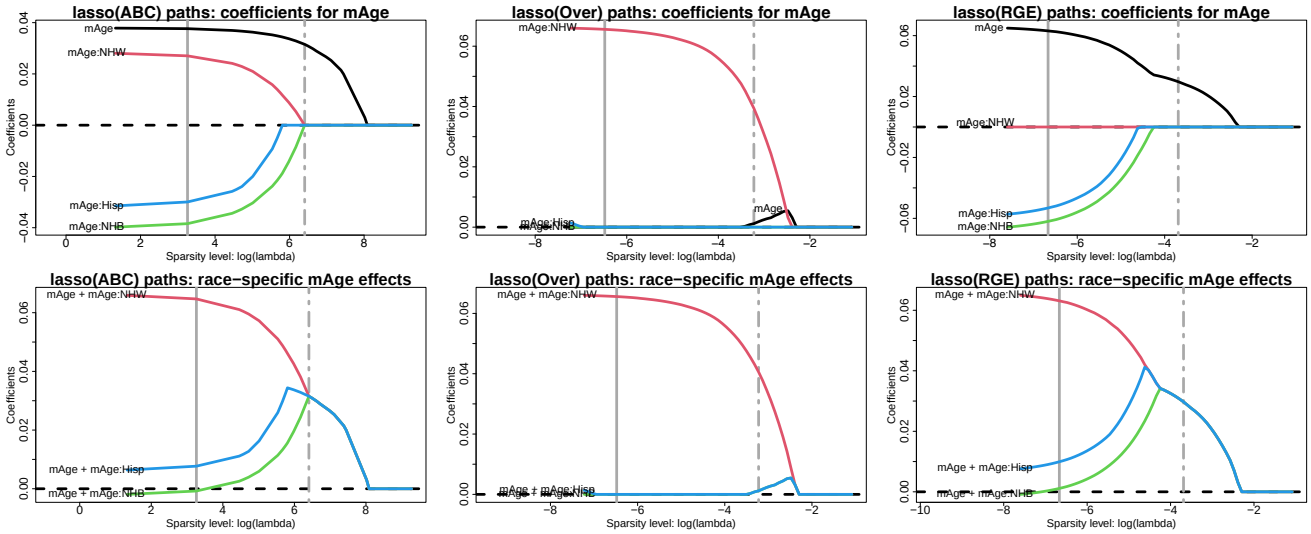


Fig. A.3. Estimated lasso paths for mother's age (mAge) across varying sparsity levels ($\log \lambda$) for the model coefficients $\hat{\alpha}_{\texttt{mAge}}, \hat{\beta}^*_{\texttt{mAge:r}}$ (top) and the race-specific slopes $\hat{\mu}'_{\texttt{PM}}(r) = \hat{\alpha}_{\texttt{mAge}} + \hat{\beta}^*_{\texttt{mAge:r}}$ (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

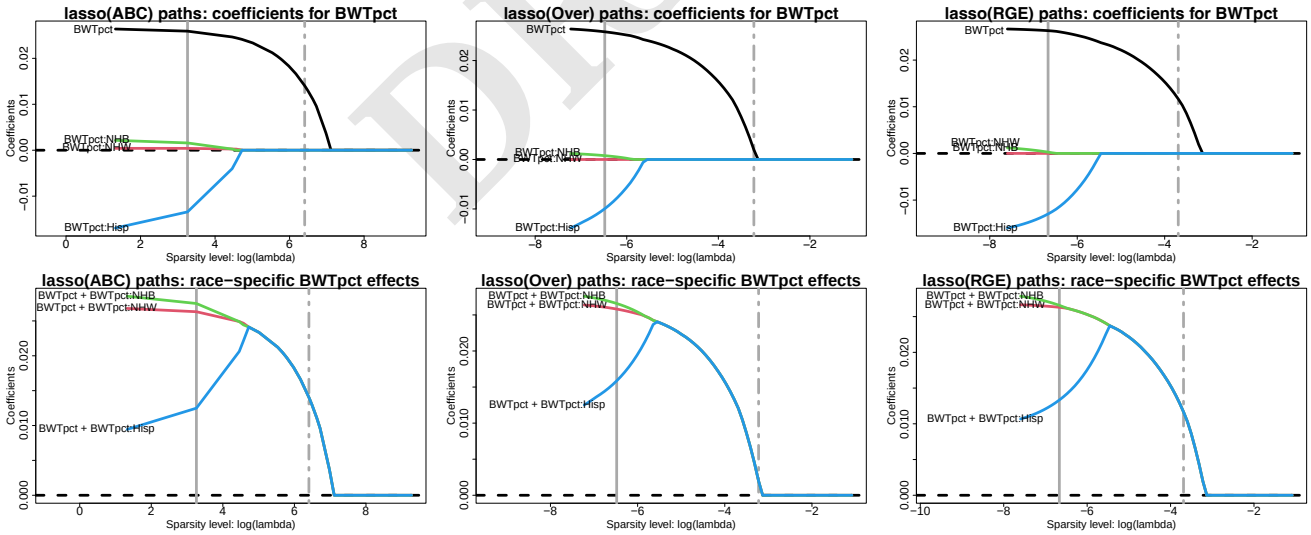


Fig. A.4. Estimated lasso paths for birthweight percentile for gestational age (BWTpct) across varying sparsity levels ($\log \lambda$) for the model coefficients $\hat{\alpha}_{\texttt{BWTpct}}, \hat{\beta}^*_{\texttt{BWTpct:r}}$ (top) and the race-specific slopes $\hat{\mu}'_{\texttt{PM}}(r) = \hat{\alpha}_{\texttt{BWTpct}} + \hat{\beta}^*_{\texttt{BWTpct:r}}$ (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.

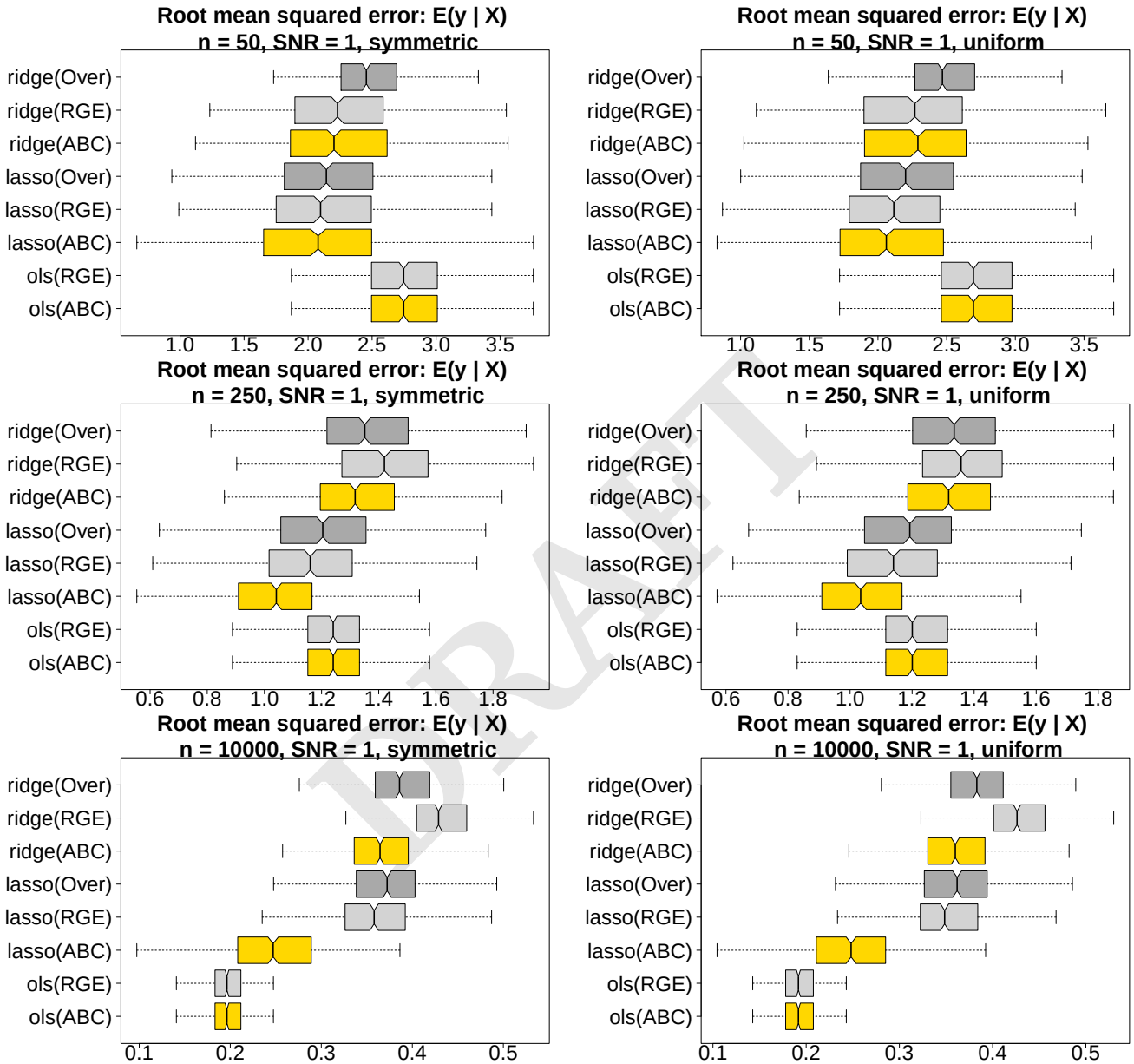
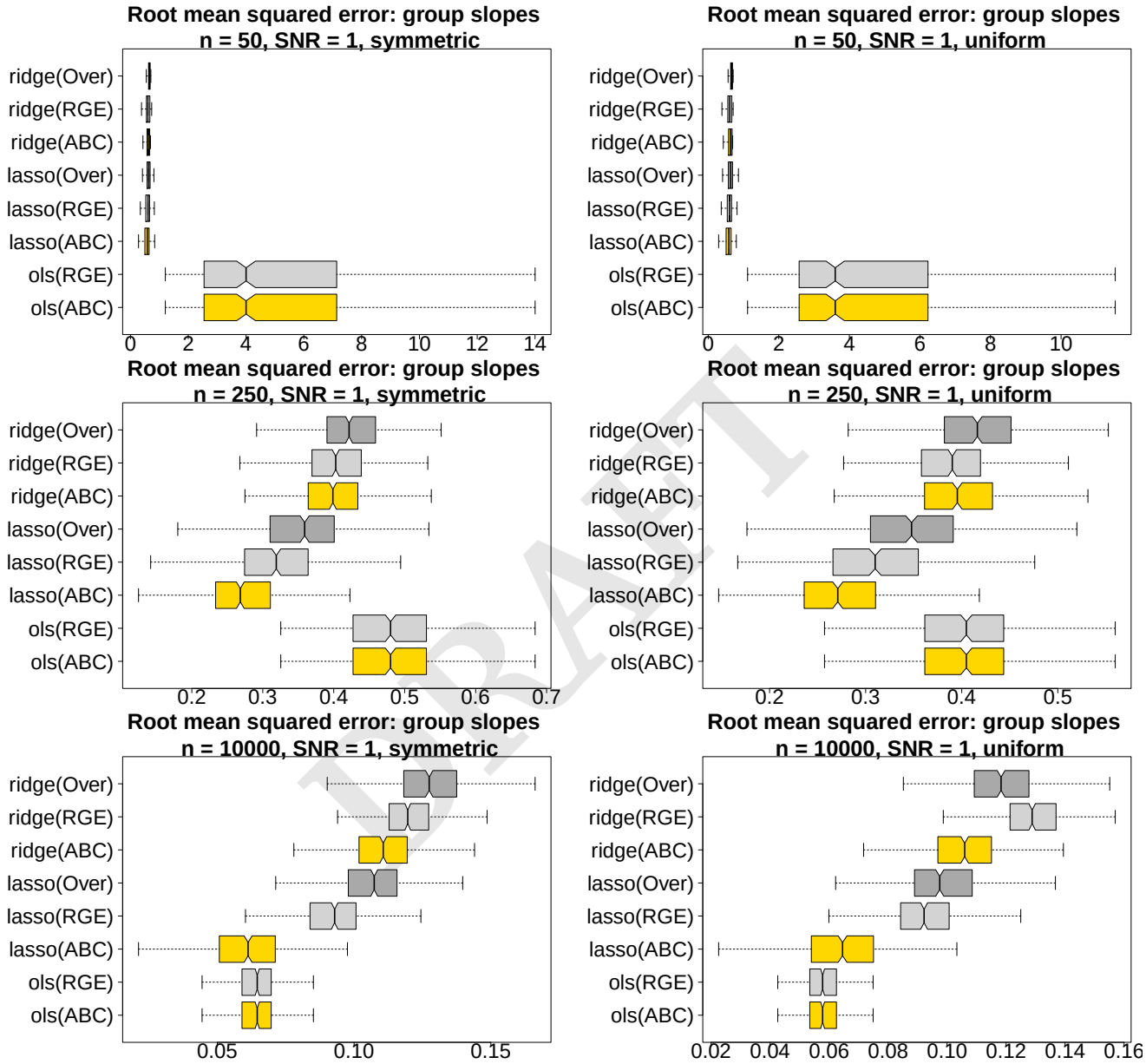


Fig. A.5. Evaluating point predictions for the conditional expectation $\mu(\boldsymbol{x}, r) := \mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}, R = r)$ for symmetric category proportions $\boldsymbol{\pi} = (0.15, 0.35, 0.15, 0.35)^\top$ (left) and uniform category proportions $\boldsymbol{\pi} = (0.25, 0.25, 0.25, 0.25)^\top$ (right) and sample sizes $n \in \{50, 250, 10{,}000\}$ across 500 simulations. These conditional expectations are invariant to the categorical variable encoding, and by definition, the OLS fitted values are identical for ABCs and RGE. For ridge and lasso regression, ABCs (gold) outperform both RGE (light gray) and the overparametrized model (dark gray) in predictive accuracy.

DRAFT



DRAFT

Fig. A.6. Evaluating point estimation for the race-specific slopes for symmetric category proportions $\boldsymbol{\pi} = (0.15, 0.35, 0.15, 0.35)^\top$ (left) and uniform category proportions $\boldsymbol{\pi} = (0.25, 0.25, 0.25, 0.25)^\top$ (right) and sample sizes $n \in \{50, 250, 10{,}000\}$ across 500 simulations. These estimands are invariant to the categorical variable encoding, and by definition, the OLS estimates are identical for ABCs and RGE. For ridge and lasso regression, ABCs (gold) outperform both RGE (light gray) and the overparametrized model (dark gray) in estimation accuracy.

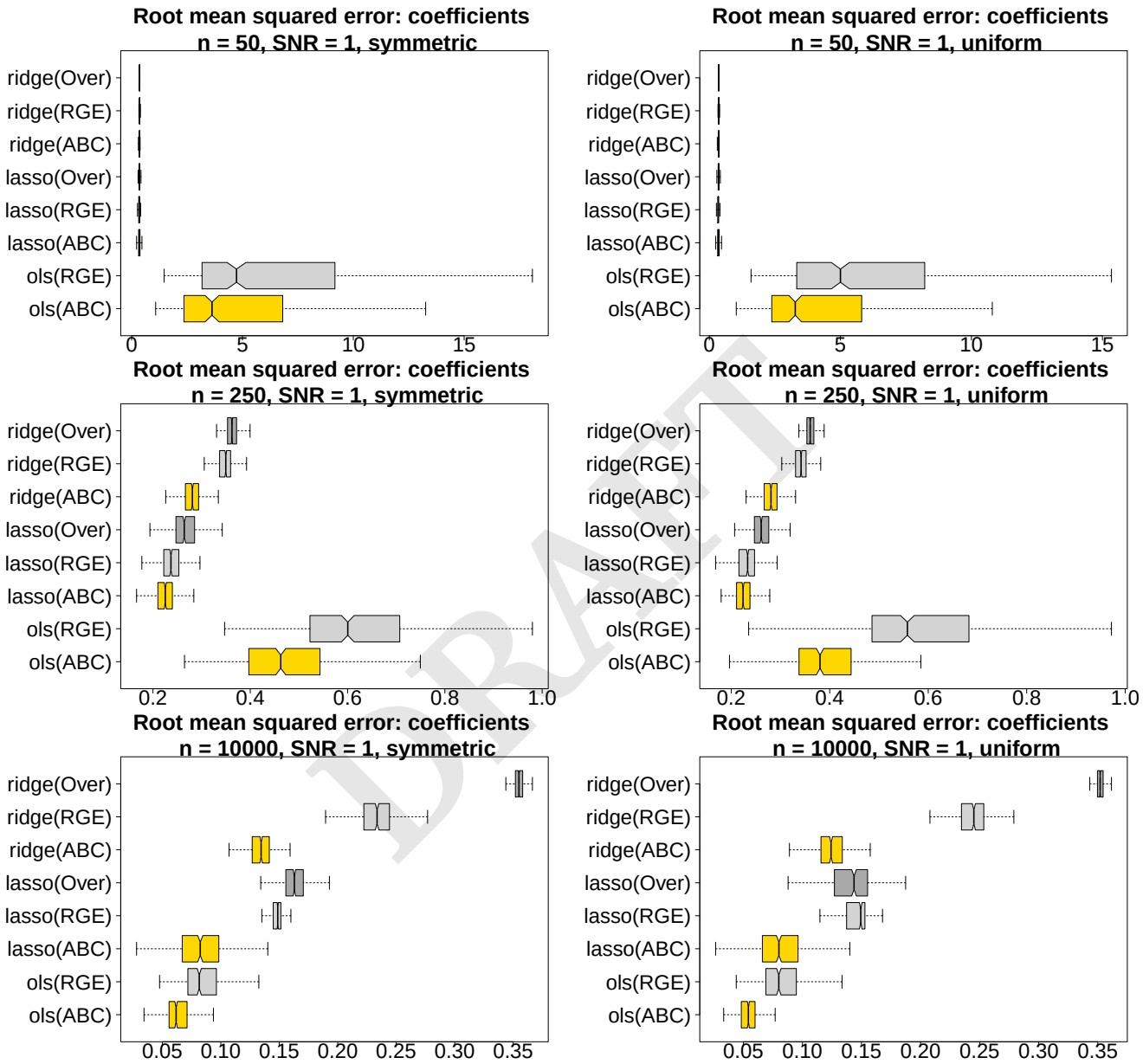
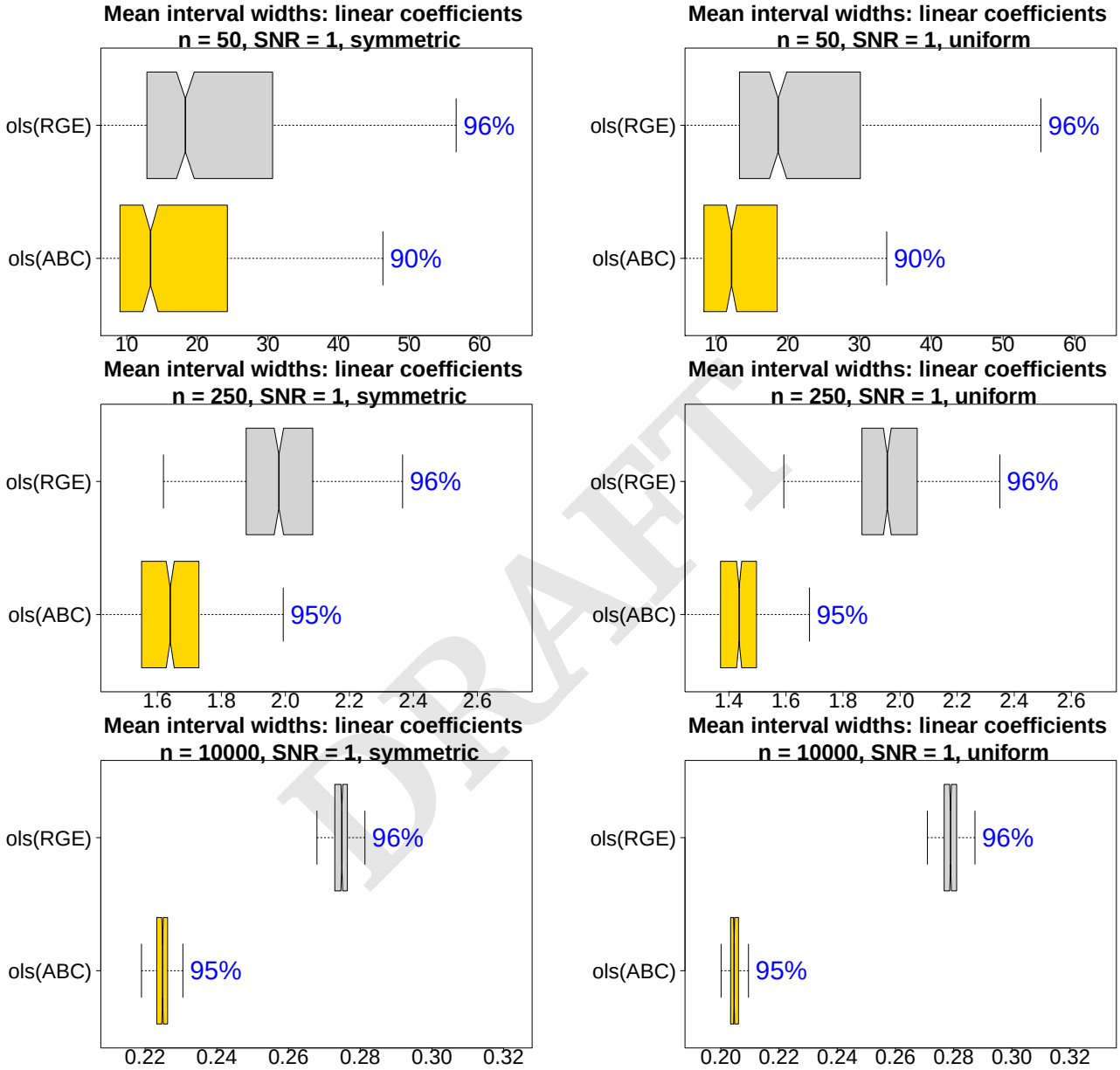


Fig. A.7. Evaluating point estimation for the regression coefficients for symmetric category proportions $\boldsymbol{\pi} = (0.15, 0.35, 0.15, 0.35)^\top$ (left) and uniform category proportions $\boldsymbol{\pi} = (0.25, 0.25, 0.25, 0.25)^\top$ (right) and sample sizes $n \in \{50, 250, 10{,}000\}$ across 500 simulations. These estimands are *not* invariant to the categorical encoding, but ABCs and RGE are satisfied in the data-generating process. ABCs (gold) outperform the other encodings (gray) within each estimation method.



DRAFT

Fig. A.8. Evaluating interval estimation for the regression coefficients for symmetric category proportions $\boldsymbol{\pi} = (0.15, 0.35, 0.15, 0.35)^\top$ (left) and uniform category proportions $\boldsymbol{\pi} = (0.25, 0.25, 0.25, 0.25)^\top$ (right) and sample sizes $n \in \{50, 250, 10{,}000\}$ across 500 simulations. 95% confidence intervals are computed for each regression coefficient, and the average widths are computed for each simulated dataset; nonoverlapping notches indicate significant differences between medians and annotations denote empirical coverage. These estimands are *not* invariant to the categorical encoding, but ABCs and RGE are satisfied in the data-generating process. ABCs (gold) provide more precise intervals that maintain nominal coverage.

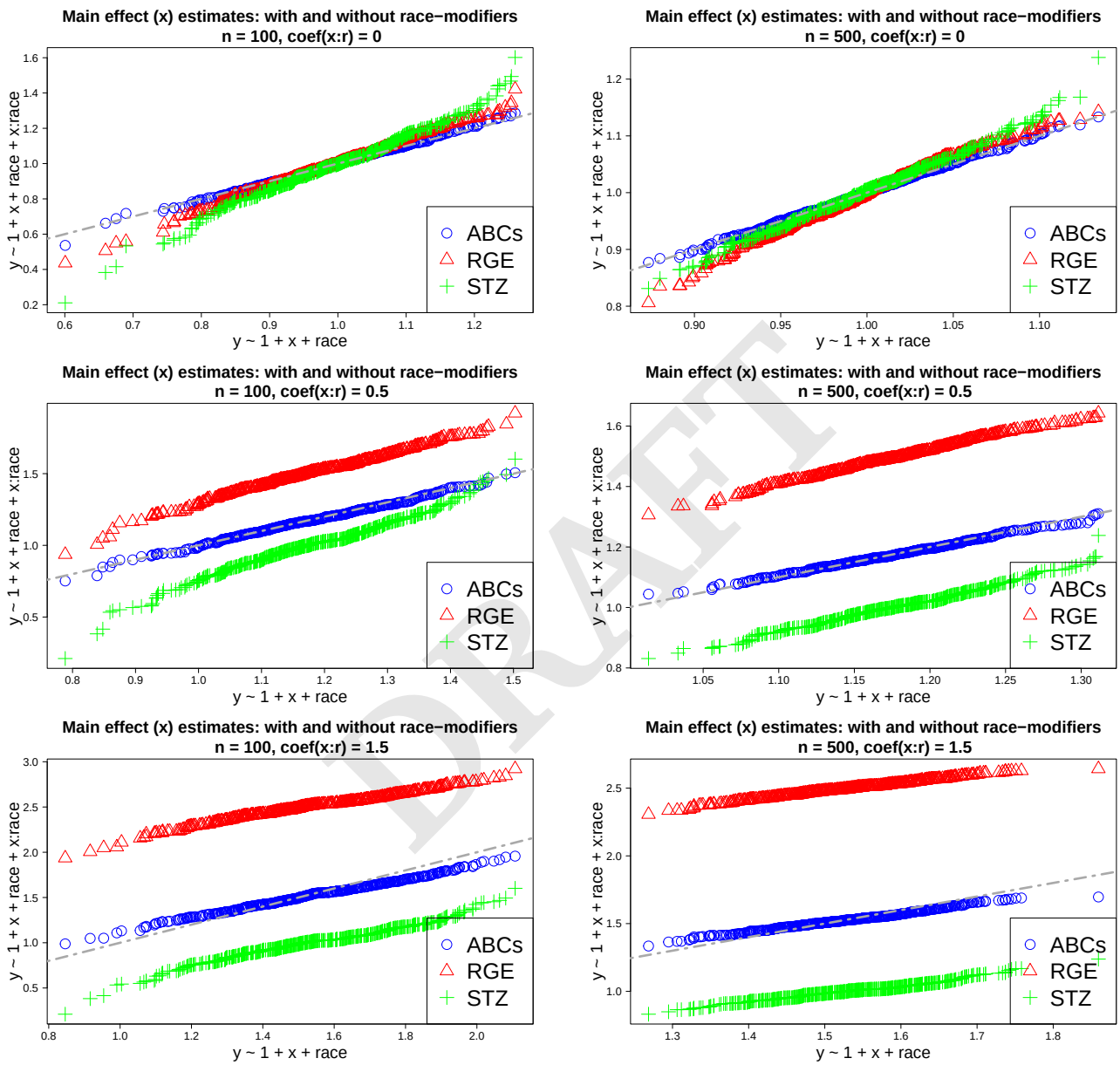


Fig. A.9. Estimated x-effects $(\hat{\alpha}_1^M, \hat{\alpha}_1)$ under different categorical encodings across 500 simulated datasets for $n = 100$ (left) and $n = 500$ (right) and varying race-modifier effects $\eta \in \{0, 0.5, 1.5\}$ (top to bottom). Uniquely, ABCs produce nearly identical x-effect estimates with and without the race-modifier (45° line), which preserves the interpretations from the simpler (main-only) model.

DRAFT