Measuring variance between groups in the presence of covariates

Clem Aeppli

February 27, 2025

Abstract

Social scientists often aim to measure the inequality or variation that occurs between groups, while taking into account characteristics that are not the focus of analysis. What exactly do we mean to estimate, in language that does not involve "fixed effects" or other model-bound terms? I argue that research questions typically reflect one of two concepts: first, the additional variation in the outcome that can be explained by incorporating group membership information; and second, the spread of average effects of group membership – or adjusted differences between groups, in a noncausal setting. I call the first the variance of group residuals (VGR) and the second the variance of group differences (VGD). The VGR measures the extra explanatory power of group information, while the VGD tells us about variability between groups. Unlike the widely used fixed and random effects methods, the VGD and the VGR are defined nonparametrically as population-level quantities and can be interpreted separately from linear functional forms. This means that both the VGR and the VGD can be estimated using various approaches from the causal-inference literature – including doubly-robust estimation and machine-learning methods – although they need not have a causal interpretation. Finally, I find that between-occupation inequality, as measured by either the VGD or VGR, is much smaller than the fixed-effects approach of Mouw & Kalleberg (2010) indicates.

1 Introduction

Social scientists often examine the amount of inequality in an outcome that occurs between groups such as occupations or neighborhoods (Xie, Killewald, and Near 2016; Tomaskovic-Devey et al. 2020; Leung-Gagné and Reardon 2023). The classic way to do so is to express the variance of an outcome Y as the sum of a between-group

component and a within-group component (see Liao 2022). This follows from the Law of Total Variance:

$$\mathbb{V}(Y) = \underbrace{\mathbb{V}(\mathbb{E}[Y \mid G])}_{\text{between-group}} + \underbrace{\mathbb{E}[\mathbb{V}(Y \mid G)]}_{\text{within-group}}. \tag{1}$$

where the G denote the groups of interest. When calculating this between-group variation, researchers frequently want to control or adjust for background characteristics that vary with both group membership and the outcome – for example, family background when studying between-classroom dispersion in test scores (Chetty et al. 2011); demographic characteristics when studying earnings variation by class strata (Wodtke 2016); or differences in worker's education when studying between-firm dispersion in earnings (Barth et al. 2016, 76; Haltiwanger, Hyatt, and Spletzer 2022). Across these topics appears a common question: How much of the variation in Y takes place between groups and cannot be attributed to covariate differences?

This paper argues that this question actually masks two different goals. First, we may be interested in the additional explanatory power of group membership *beyond* that which the covariates alone can explain. How much do we learn about the outcome by incorporating group membership into our analysis? Second, we may be interested in measuring the dispersion of coviarate-adjusted effects, or average differences, between groups. This paper formalizes these two quantities, naming the first the variance of group residuals (VGR) and the second the variance of group differences (VGD). It develops them in context of a known number of groups of possibly unequal size. The VGR allows us to measure the additional explanatory power gained by adding a new, categorical element to our study; the VGD, on the other hand, is better suited to summarizing the differences (or effects) between groups.

The VGR and VGD extend and formalize the intuition behind standard fixed- and random-effects approaches. In particular, the VGR echoes the intuition behind the random effects approach, while the VGD reflects the idea of the fixed effects approach. Yet both random effects and fixed effects have meaning only in terms of the particular functional form and parametric assumptions used to estimate them. It is typically left unstated what deeper substantive concept, whether descriptive or causal, they aim to capture. The group fixed effects approach, moreover, fails to even capture familiar average differences between groups (Goldsmith-Pinkham, Hull, and Kolesár 2022). This paper takes a step back to ask what we mean by inequality between groups, in terms that are not bound to particular modeling assumptions.

The VGR and VGD are of use to both causal and descriptive analyses. They are, moreover, entirely nonparametric estimands, and we can estimate them through a number of procedures. Additionally, they are still meaningful even in the common case that a single group does not have support over the whole range of covariates:

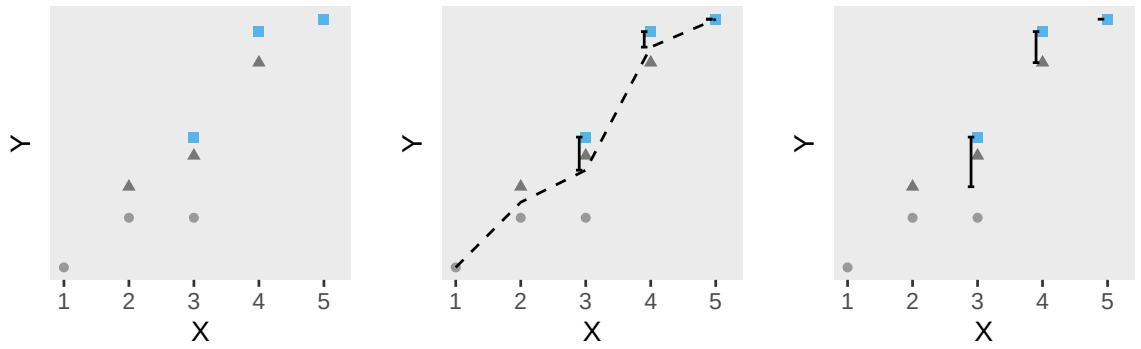
for example, with years of education and the occupation physicians. This paper proceeds in four steps: first, I formalize the VGR and VGD; second, I explain how to estimate them; third, I compare them to the variances of fixed and random effects; and fourth, I find that occupations explain less American income inequality than Mouw and Kalleberg (2010) find in their landmark paper. This exposition shows how even descriptive, noncausal research benefits from first considering what we aim to estimate *before* picking functional forms – a consideration that has typically been confined to causal research.

2 Two meanings of covariate adjustment

Consider some simple data in Figure 1a. There are three observations each from three groups: grey circles, grey triangles, and blue squares. There is a single covariate X, with respect to which the groups are not balanced – the blue square group is overrepresented at higher values of X (i.e., on the left side of the figure). In fact, the groups are not all represented at each level of X. We want to decompose the variation $\mathbb{V}(Y)$ of Y into a between-group component. If we were unconcerned about X, we could simply take the means of Y in each group and calculate the variance of these three means; this is the classic unadjusted decomposition (1). But we are concerned about X. What might we mean when we speak of between-group variance?

This section gives two answers. First, we may want to measure the additional explanatory power of group membership, beyond that which is explained by the covariate(s) alone. By how much does this information on groups – workplaces, occupations, classrooms, etc. – improve our explanation of a given outcome? This will be termed the variance of group residuals (VGR). Second, we may instead want to measure the variation in the effects or differences between groups, calculated conditionally on the covariates. The research question here is different: by how much do the adjusted differences between groups vary? I will refer to this as the variance of group differences (VGD). Before elaborating these two approaches, I first propose two principles which a meaningful measure of between-group variance ought to obey. I then develop two estimands consistent with these principles, and which capture the two different meanings of covariate adjustment.

Both expositions here are entirely nonparametric and focus on population-level quantities rather than estimation. As such, they clarify the questions behind group-variance analyses, and set variance decompositions on firmer grounding. Subsequent section will address the estimation of these quantities and their relations to commonly used approaches in the literature.



(a) Example data set.

(b) Compare blue observations to average of Y at each X, $\mathbb{E}[Y \mid X]$ (dashed line).

(c) Compare blue observations to their complement at each X.

Figure 1: Example with three groups (grey circles, grey triangles, blue squares) and a covariate X.

2.1 Principles

The prior section introduced two possible meanings of covariate-adjusted between-group variance: the first, the additional explanatory power of group identities *after* accounting for the covariates; the second, the spread in group average effects or, non-causally, average comparisons between groups across levels of X. Before formalizing these quantities, I identify two principles that a good measurement ought to obey.

First, consider a study of pay inequality between occupations, where a covariate X is individuals' highest level of education. Few if any physicians will have high school diplomas alone; yet we do not want to drop either this occupation or those without a higher level of formal education from our analysis. This gives rise to the first principle:

1. The quantity should exist and give a meaningful answer even if not every group is present at each level of X, including cases where no group has support over all of X.

This is consistent with applications of variance decomposition methods, where scholars generally aim to decompose inequality across the entire distribution of workers, students, or other units (e.g. Xie, Killewald, and Near 2016; Aeppli and Wilmers 2022), rather than focusing on identifying the effect(s) of a small number of treatments.

A second principle facilitates interpretation:

2. If the distribution of X is the same in each group, our quantity should yield the unadjusted between-group variance (1).

This principle ensures that estimands can be compared to the intuitive between-within variance decomposition in Equation 1. Other approaches to the study of variance, even if they have different goals, treat the basic between-within decomposition as an important interpretive stepping stone (Western and Bloome 2009; Western and Rosenfeld 2011).

2.2 Approach 1: Variance of group residuals (VGR)

First, researchers may be interested in how much between-group variance there is *after* having accounted for covariates alone. Put differently, how much variation in the outcome is unexplained by the covariates and occurs between groups? This quantity, which I call the variance of group residuals (VGR), measures the additional explanatory power of incorporating group-membership information into our analysis – beyond any explanatory power of the covariates.

We can achieve this by, first, removing the average of Y at each level of X. The resulting purged quantity, $Y - \mathbb{E}[Y \mid X]$, has a conditional mean of 0 at each x. The group averages of this purged quantity therefore measure each group's mean level of Y after having excluded any variation in Y predicted by X. For example, the blue group's purged average is the average of the three vertical bars in Figure 1b weighted by the number of blue points at each level of X. Finally, we use these purged group averages to calculate the between-group decomposition from (1). This can be interpreted as the between-group spread after removing variation between Xs.

To formalize the notation, let X be a random column vector of length k with support $\mathcal{X} \subseteq \mathbb{R}^k$ and realizations x. Y is the real-valued outcome, and group membership G is a random variable taking values in $1, 2, ..., M$ with probabilities $\pi_1, \pi_2, ..., \pi_M$. Then the variance of group residuals is

$$\text{VGR} := \mathbb{V}\left(\mathbb{E}[Y - \mathbb{E}[Y \mid X] \mid G]\right). \tag{2}$$

This definition does *not* require that each group have observations at each value $x \in \mathcal{X}$ – a "common support"-type requirement familiar from causal inference (e.g., Iacus, King, and Porro 2012) – and thereby meets principle 1. We still want to decompose variances even when there is only a single group present at some x'. The purged Y will simply be 0 at $X = x'$, dragging a group's purged average towards 0. This attenuates the VGR. Attenuation may be desirable because we cannot distinguish variation between Xs from variation between group membership at $X = x'$; thus we

avoid attributing variation at this point to group membership. [1]

For brevity's sake, I will define

$$\delta_g := \mathbb{E}[Y \mid G = g] - \mathbb{E}[\mathbb{E}[Y \mid X] \mid G = g]. \tag{4}$$

The VGR can be expressed more compactly as the group-size-weighted variance $\mathbb{V}(\delta_g)$. Rewriting this way suggests another interpretation of the VGR. In the basic decomposition (1), the between-group component is the average square of the distance between g and the global mean, $\mathbb{E}[Y \mid G = g] - \mathbb{E}[Y]$. How much of this distance arises from the distribution of Xs, and how much is particular to g? The VGR makes use of one answer: δ_g is the average difference between group g and observations that look similar (in terms of X) to those in g. The remainder $\mathbb{E}[\mathbb{E}[Y \mid X] \mid G = g] - \mathbb{E}[Y]$ captures the differing covariate distribution. The VGR then measures the average square of the adjusted differences δ_g. In particular, if the X-distribution is same for each group, then principle 2.1 holds.

2.3 Approach 2: Variance of group differences (VGD)

Alternatively, we may be interested in the spread of average effects or differences – conditional on covariates – between groups. Say students in some classrooms score on higher on an exam than those in other classrooms, net of parents' socioeconomic status. How much variation is there between classrooms? If other causal assumptions are met, then this can be interpreted causally, as the variation in groups' causal effects. But absent those assumptions, it has a descriptive interpretation: how much variation is there in the covariate-adjusted differences between groups? This section walks through three considerations regarding such a quantity: the choice of comparisons or reference groups, the method of weighting across levels of X, and the use of rescaling to avoid double-counting.

One might begin by comparing each group to a single reference group g_0. This type of comparison is widespread (e.g., Haltiwanger, Hyatt, and Spletzer 2022; Western and Bloome 2009). However, it typically results in questionable comparisons and fails principle 1 (see also Hamjediers and Sprengholz 2023). For instance, to which occupation would it make sense to compare every other occupation at each level of education? As an alternative, we can instead compare each group g to its complement at each level of X, as in Figure 1c, and then construct an average of these local

1. The full decomposition of $\mathbb{V}(Y)$ is:

$$\mathbb{V}(Y) = \underbrace{\mathbb{V}(\mathbb{E}[Y \mid X])}_{\text{between } X} + \text{VGR} + \underbrace{\mathbb{E}(\mathbb{V}(Y - \mathbb{E}[Y \mid X] \mid G = g))}_{\text{residual: within } X\text{, within } G}. \tag{3}$$

comparisons. If causal assumptions are met, then this captures the average effect on outcome Y of switching into group g, after adjusting for X. Appendix B presents these assumptions and the causal interpretation in potential-outcomes notation. In the descriptive context, this can be worded: how does group g differ from g's complement when comparing within levels of X?

There are many strategies for averaging these comparisons over the covariate distribution. For example, the average treatment effect on the treated (ATT) is $\tau_g^{\text{ATT}} = \mathbb{E}\big[\mathbb{E}[Y \mid X, G = g] - \mathbb{E}[Y \mid X, G \neq g] \mid G = g\big]$. This is also known, in a non-causal setting, as the average gain of group g (Słoczyński 2020). A particularly useful scheme is the variance-weighted effect τ_g^V, where each level of x is weighted by the variation in group membership that occurs there:

$$\tau_g^V = \mathbb{E}\left[w_g(X)\left(\mathbb{E}[Y \mid X, G = g] - \mathbb{E}[Y \mid X, G \neq g]\right)\right], \text{ for } w_g(x) = \frac{p_g(x)\left(1 - p_g(x)\right)}{\mathbb{E}[p_g(X)(1 - p_g(X)]}. \tag{5}$$

This variance-weighting up-weights xs where there is more variation in g membership. Upweighting is useful methodologically and conceptually: more variation in "treatment" means the effect can be measured more precisely, and we may also be more interested in parts of the covariate distribution where treatment varies more (as in Angrist and Pischke 2009). Variance-weighting also ensures that principle 1 is met, as $w(x) = 0$ at levels of x where all observations or none are in g. The variance-weighted τ_g^V simply ignores these parts of $\mathcal{X}$.

In order to measure the dispersion in group effects/differences, we might then take the variance of these treatment effects τ. However, simply taking the variance of these τ terms will double-count each group's difference. Consider the simple example in Figure 2, where three groups all have the same distribution across the covariate X. The group effect of blue squares is shown in the blue arrow, which averages the difference between the squares and triangles and between the squares and circles; however, this latter difference is *also* included in the group effect of circles (green arrow). We effectively count each difference twice, so that the variance of the τ terms would over-exaggerate between-group inequality. One solution, shown on the right panel of Figure 2, is to rescale each τ_g by $(1 - \pi_g)$. I name the resulting quantity the variance of group differences (VGD):

$$\text{VGD} := \mathbb{V}\left((1 - \pi_G)\ \tau_G^V\right). \tag{6}$$

This scaled version meets principle 2: if the groups all have the same distribution of Xs, then the scaled-VGD yields the standard between-group decomposition. Regardless of the choice of weights – for example, if ATT-style comparisons are used instead

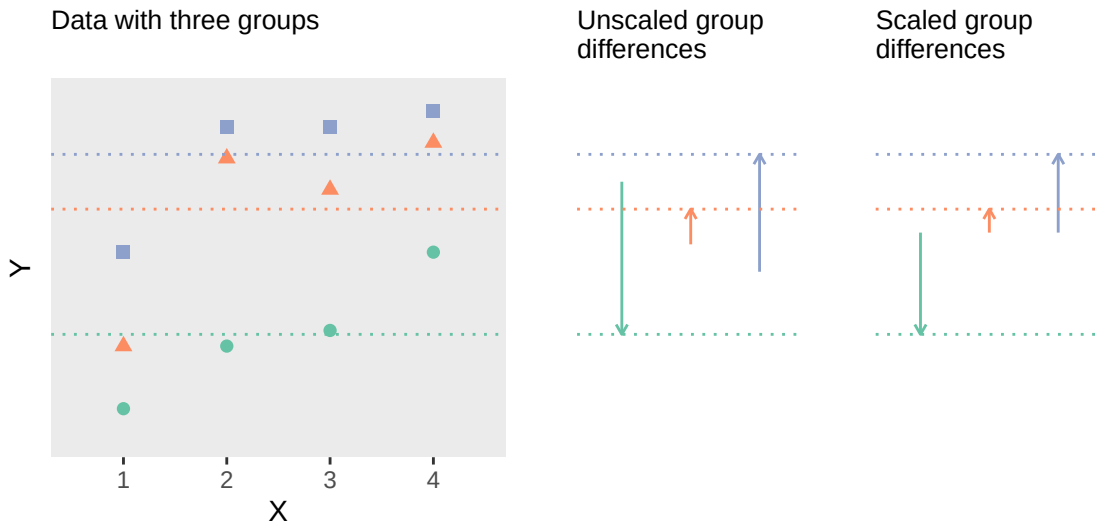


Figure 2: Example with three groups (green circles, orange triangles, blue squares) distributed equally across a covariate X. The "unscaled group differences" shows the τ_g terms for each group: how much does it differ from its complement? The "scaled group differences" shows the $(1-\pi_g)\tau_g$ terms, which no longer double-count differences between groups. Dashed lines show group-specific averages.

of variance-weighting – this scaling ensures the principle 2 is met. [2]

It is worth underscoring that the VGD in (6) need not have a causal interpretation. We are often interested in gaps in an outcome net of some covariates (e.g., Słoczyński 2020; Fortin, Lemieux, and Firpo 2011). For instance, we might be interested in gaps in test scores between classrooms, net of teacher's gender or parents' socioeconomic status; or we might compare pay between occupations adjusting for workers' age. We can make these comparisons without necessarily implying that they are causal: the covariate-adjusted gaps in scores between classrooms need not be causal, and classroom assignment need not be as good as randomly assigned conditional on parents'

2. In general, we can use a weighting function ψ of $p_g(x)$ to define a treatment effect

$$\tau_g^{\psi} = \frac{1}{\mathbb{E}[\psi(p_g(X))]}\mathbb{E}\left[\psi(p_g(X))\left\{\mathbb{E}[Y \mid G=g, X] - \mathbb{E}[Y \mid G \neq g, X]\right\}\right].$$

The ATT weight uses $\psi(t) = t$, while the variance weighting uses $\psi(t) = t(1-t)$. If all groups have the same distribution of X then $p_g(x) = \pi_g$ at all x. Then regardless of the weighting scheme ψ chosen, $\psi(p_g(x))$ is constant with respect to x and so τ_g^{ψ} becomes $\mathbb{E}[\mathbb{E}[Y \mid G=g, X] - \mathbb{E}[Y \mid G \neq g, X]]$. Since $\mathrm{d}F(X \mid G=g) = \mathrm{d}F(X \mid G \neq g) = \mathrm{d}F(X)$, τ_g^{ψ} further simplifies to $\mathbb{E}[Y \mid G=g] - \mathbb{E}[Y \mid G \neq g]$. So then $\mathbb{V}((1-\pi_g)\tau_g^{w})$ becomes $\mathbb{V}(\mathbb{E}[Y \mid G=g])$.

socioeconomic status. In this setting, the τ take on a purely descriptive meaning regardless of weighting scheme used. The VGD simply summarizes the spread of these descriptive adjusted gaps.

2.4 Relation between the VGR and the VGD

This section has introduced two estimands: the VGR measures how much more group membership tells us about an outcome, beyond a more parsimonious model with just the covariates, while the VGD measures the spread in effects or differences between groups. These two quantities relate in an intuitive way. The group-comparison term δ_g ($= \mathbb{E}[Y \mid g] - \mathbb{E}[\mathbb{E}[Y \mid X] \mid g]$ per Eq 4) of the VGR equals a re-scaled version of the variance-weighted treatment:

$$\delta_g = \frac{\mathbb{E}[\mathbb{V}(1_{G=g} \mid X)]}{\mathbb{V}(1_{G=g})}(1 - \pi_g)\tau_g^V. \tag{7}$$

(See A.1 for proof.) As noted in 2.2, the $(1-\pi_g)$ portion scales down g's contribution to the VGR as g grows in size: this way, we focus on decomposing the distance between $\mathbb{E}[Y \mid g]$ and $\mathbb{E}[Y]$, which will shrink as g grows. The quotient, meanwhile, captures the share of variation in group assignment that occurs *within* levels of x. If groups overlap greatly, then the quotient is close to 1. But if groups do not overlap much, then the fraction approaches zero and we scale down the treatment effect τ_g^V. This up- and down-scaling may be desirable in our descriptive decomposition: if there is a small amount of overlap on X, then variation in Y is occurring between values of X – which we may not want to attribute to groups (see Hamjediers and Sprengholz 2023).

In particular, observe that the quotient in (7) falls between 0 and 1 for all groups, since the numerator is a variance component of the denominator. A consequence of the relation is therefore that VGR $\leq$ VGD and $\mathbb{V}(\delta_g) \leq \mathbb{V}(\tau_g)$. (See A.3.) It is worth remarking that this ordering concerns the population-level VGR and VGD. Depending on the researcher's estimation strategies, the estimated VGR and VGD may reverse orders.

3 Estimation

Estimation of the VGD and VGR is straightforward and flexible. Take a sample (g_1, x_1, y_1), (g_2, x_2, y_2), ..., (g_n, x_n, y_n) where $g_i \in \{1, 2, ..., M\}$ denotes the group of observation i, $x_i \in \mathcal{X}$ is a column vector of i's covariates, and y_i is i's scalar outcome.

The VGD can be obtained by first estimating the group effects τ_g^w using some weighting scheme w (ATT, VWT, etc.); multiplying these by the complement of the

group proportions π_g; and then estimating the variance of the $(1-\pi_g)\hat{\tau}_g^w$. Similarly, the VGR involves first estimating the group comparisons δ_g, and then taking their group-size-weighted variance. Note that both of these "naive" plug-in procedures are biased in finite samples – I address this bias towards the end of this section.

Estimation of the group effects τ_g is familiar from the literature on causal inference, so I address them in less detail. The δ_g comparisons used for the VGR, on the other hand, are less familiar, so I begin this section by discussing their estimation.

3.1 Estimation of the VGR

In this section I sketch out three approaches to estimating the group comparisons, $\delta_g = \mathbb{E}[Y \mid G = g] - \mathbb{E}[\mathbb{E}[Y \mid X] \mid G = g]$, used in the VGR. These are, first, modeling the conditional expectation of Y at levels of X; second, modeling group membership conditional on X; and third, combining these two approaches to achieve double robustness.

3.1.1 Conditional expectation of Y given X

First estimate the conditional means $\mathbb{E}[Y \mid X = x]$ without using group information. Average the resulting estimates, which I will label $\hat{m}(x_j)$, within each group h and subtract from the average y_j in h:

$$\hat{\delta}_h^m = \frac{1}{n_h} \sum_{j:g_j=h} y_j - \hat{m}(x_j), \tag{8}$$

where n_h denotes the number of observations in group h. If $\hat{m}$ converges uniformly to the map $x \mapsto \mathbb{E}[Y \mid X = x]$, then $\hat{\delta}_h^m$ is consistent for δ_h (See A.6). Conditions for convergence are actually weaker than this: Appendix A.5 shows that $\hat{\delta}_m^h$ has a sort of incidental double robustness. If we are searching for the best-fitting $\hat{m}$ in some linear space $\mathcal{M}$ of maps of x, then $\hat{\delta}_m^h$ consistently estimates δ_m if either the map from x to $m(x)$ *or* the map from x to $p_g(x)$ are contained within $\mathcal{M}$.

The linear approach, which fits a pooled regression $y_i = x_i'\hat{\beta}^p + u_i$ and then sets $\hat{m}(x_j) = x_j'\hat{\beta}^p$, is the simplest. This pooled-regression VGR estimand is straightforward to implement and not computationally intensive. It can equally be interpreted as a reweighting version the VGR using a reweighting function that is linear in X (A.5).

However, the VGR is far from limited to this linear version. Since its definition in (3) is entirely nonparametric, the researcher can pick the method for estimating $\mathbb{E}[Y \mid X]$ that is most appropriate to her case. For example, if the X are discrete quantities such as, e.g., county of origin – or can be discretized as in coarsened

exact matching (Iacus, King, and Porro 2012) – then the researcher can use the sample means within each level of X to estimate $\mathbb{E}[Y \mid X]$. Equally viable are many nonparametric or supervised machine-learning estimators of $\mathbb{E}[Y \mid X]$, including recent advances such as super machine learners (Baćak and Kennedy 2019; see Molina and Garip 2019 for a review). Regardless of the method used, since the goal is to estimate $\mathbb{E}[Y \mid X]$ *without* separating by group, a more flexible function of X can be used with greater precision than in, for example, the fixed-effects method.

3.1.2 Group membership given X

An alternative estimation strategy is similar to inverse-propensity reweighting for treatment effects. First, estimate the relative probabilities $\frac{p_h(x)}{\pi_h}$ using a classification method. Methods can range from sample cell proportions (when X is discrete and has few levels) to random forest classification procedures. Or if we regress indicators for g membership on X, we end up with an OLS version of the VGR as discussed above. Next, average the product $y_i \hat{p}_h(x_i)/\hat{\pi}_h$ over *all* observations, regardless of group membership, and subtract from the mean y in h:

$$\hat{\delta}_h^p = \frac{1}{n_h} \sum_{j:g_j=h} y_j - \frac{1}{n} \sum_{j=1}^{n} y_j \frac{\hat{p}_h(x_j)}{\hat{\pi}_h}. \tag{9}$$

Repeat this process separately for each group. In the asymptotic setup where n and n_h approach infinity and n_h/n converges to a constant bounded away from zero, the uniform convergence of the map $x \mapsto \hat{p}_h(x)$ to the map $x \mapsto \mathbb{E}[\mathbf{1}_h \mid X = x]$ ensures that $\hat{\delta}_h^p$ converges to δ_h. (Appendix A.6)

The disadvantage of this second, membership-based approach is that it is likely more computationally intensive: whereas the first approach involves estimating a single model (of $\mathbb{E}[Y \mid X]$), the second requires a separate set of estimates of the relative probability of group membership *for each group*. However, the trade-off is that it makes no claims about the relation between X and Y. It is up to the researcher to determine the best method for her case; alternative approaches can also be compared to assess sensitivity. This flexibility distinguishes the VGR approach from fixed- or random-effect approaches, which cannot be separated from particular (linear) models of the outcome.

3.1.3 Double robustness

The above methods can further be combined into a "doubly robust" estimator. Say that $\hat{m}(x)$ is our estimate of $\mathbb{E}[Y \mid X = x]$ using any strategy (OLS, cell means, etc.)

and $\hat{p}_h(x)$ is our estimate of $p_h(x)$. Then a new estimator of δ_h is:

$$\hat{\delta}_h^{DR} = \frac{1}{n_h} \sum_{j:g_j=h} \{y_j - \hat{m}(x_j)\} + \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{p}_h(x_j)}{\hat{\pi}_h} \{\hat{m}(x_j) - y_j\}. \tag{10}$$

This estimator is a doubly-robust estimator of δ_g in that, if either the mean model $\hat{m}$ or the membership model $\hat{p}_g$ is uniformly consistent, then $\hat{\delta}_h^{DR}$ is consistent. A.6 formalizes this claim and presents additional assumptions.

3.2 Estimation of the VGD

This subsection considers estimation of the group-difference τ terms defined in (6) and used to calculate the VGD. In general, many of the methods from the causal-inference literature may be used to estimate the group effects τ: covariate adjustment, matching, reweighting, or more recent machine-learning methods can all be used to compare g to its complement across levels of X. In particular, the variance weighted effects defined in (5) can be recovered from the regression of Y on X plus an indicator for group g,

$$y_i = \lambda(g)' x_i + \tau_g^V 1_{g_i=g} + u_i, \tag{11}$$

per the classic result of Angrist and Pischke (2009). We would have to fit this regression once per group to estimate all the τs. The coefficients λ on X change in each of these models, hence the indexing by g, $\lambda(g)$.

An obvious concern is that this method would take a long time, especially if there are many groups or covariates. However, there is a shortcut that drastically reduces computation time: we can make use of the fact that the matrix $\mathbb{E}[XX']$ must be estimated and inverted in all these regressions. For simplicity, I will refer to the matrix of observed covariates as $\mathbf{X} = (x_1, x_2, ..., x_n)'$ and the observed outcomes as $\mathbf{Y} = (y_1, ..., y_n)'$. I will also make use of a vector formed by each observation's indicator for a given group g: $\mathbf{I}_g = (1_{g_1=g}, 1_{g_2=g}, ..., I_{g_n=g})'$. Appendix A.7 shows that one can then estimate τ_g as

$$\hat{\tau}_g^V = \frac{\mathbf{Y}' \left(\mathbf{I}_g - \mathbf{X} \left(\mathbf{X}'\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{I}_g \right)}{\mathbf{I}_g'\mathbf{I}_g - \mathbf{I}_g'\mathbf{X} \left(\mathbf{X}'\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{I}_g}, \quad \text{for all groups } g. \tag{12}$$

This means we can invert $\mathbf{X}'\mathbf{X}$ once and then reuse it to calculate all the τs. This dramatically speeds up computation. In tests (**APPENDIX**) and in the example in the next section, I am able to calculate the VGD in roughly the same amount of time as the VFE.

3.3 Bias of the plug-in estimators

A key issue with estimating the VGD and VGR – as well as other quadratic forms of regression coefficients such as the variance of fixed effects (Kline, Saggio, and Solvsten 2020) – is that unbiased estimators of the τ_g or δ_g do *not* imply an unbiased estimate of the VGD or VGR. Say we use any one of the procedures above to obtain unbiased estimates $\hat{\delta}_g$ of the group comparison terms δ_g. Then the naive plug-in estimator of the VGR is still biased:

$$\mathbb{E}[\widehat{\text{VGR}}] = \text{VGR} + \sum_g \pi_g \mathbb{V}(\hat{\delta}_g) + \sum_g \text{Cov}(\hat{\pi}_g, \hat{\delta}^2). \tag{13}$$

(See A.8.) In the presence of basic regularity conditions, then the covariance term is small. What remains is the middle term, which introduces a positive bias. That is, the sampling variance of the group terms $\hat{\delta}_g$ affects the bias of the VGR estimator, even if the group terms themselves are unbiased (if $\mathbb{E}[\hat{\delta}_g] = \delta_g$). Similarly, the naive plug-in estimator of the VGD is affected by the sampling variance of the $\hat{\tau}_g$ terms.

The bias in (13) applies regardless of the choice of estimators for δ or τ, though the specific nature of this bias depends on the choice of estimators. However, sample-splitting offers a simple method for mitigating the bias in (13). Randomly split the sample into three subsets A, B and C, and obtain unbiased estimates $\hat{\delta}_g$ separately on these two sets. The resulting estimates $\hat{\delta}_g^A$ and $\hat{\delta}_g^B$ are uncorrelated, and so $\sum_g \hat{\pi}_g^C \hat{\delta}_g^A \hat{\delta}_g^B$ is an unbiased estimate of the VGR. This extends the sample-splitting approach used by Godechot, Palladino, and Babet (2023) for estimating the variance of fixed effects. Similarly, uncorrelated and unbiased estimates $\hat{\tau}_g^A$ and $\hat{\tau}_g^B$ of the average treatment effects can be combined into an unbiased estimate of the VGD.

4 Comparisons to standard methods

The preceding sections presented two quantities and developed methods for their estimation. The VGR measures the additional explanatory power of group membership, beyond what can be explained by the covariates alone; and the VGD measures the spread in average group effects or controlled differences. These echo two common tools in the social scientist's kit: the variance of random effects (VRE) and the variance of fixed effects (VFE), respectively. This section considers these two approaches. Despite their differences, they suffer from a common fault: both are defined only with respect to the model used to estimate them, and it is generally left unstated what the target estimand is. Nonetheless, the VRE is typically interpreted similarly to the VGR, and the VFE is typically interpreted similarly to the VGD. The estimands introduced in the last section can therefore be seen as attempts to clarify the intuitions

behind the random- or fixed-effects approaches. However, once the researcher clarifies their goal – VGR and VGD – it likely turns out that random- and fixed-effects approaches are inconsistent estimators.

This section does not focus on better-known issues of estimation and small-sample bias here, but rather on what the quantities are taken to capture in the population.

4.1 Variance of random effects (VRE)

The first familiar approach is the variance of random effects. This approach is particularly common in education and demographic research, where groups often correspond to classrooms, schools, or families (Chetty et al. 2011; Van Winkle 2018). The researcher supposes that data has been generated through the following mixed-effects model:

$$y_i = \beta^{\text{ME}\prime} x_i + v_{g_i} + w_i. \tag{14}$$

Here, the group effect (or intercept), v, is taken to be a draw from a centered random variable. The distribution of group effects has variance σ_v^2, while the X have fixed returns β. This group variance σ_v^2 is then often reported as a measure of between-group variation (Raudenbush and Bryk 2002). The group effects v are assumed to be independent of or at least orthogonal to the individual-level errors w and the X. Thus, unlike the VFE discussed next, the VRE excludes any variation in Y that can be explained by a linear function of the covariates. This is also the key difference between the VRE and the VGD.

The VRE is close in spirit to the VGR described above. In fact, in the setting where the number of groups is unbounded and where group sizes are equal, the mixed effects approach will asymptotically measure the same population parameter as the VGR approach where the conditional mean of Y is linear in x.[3] In practice, however, the two approaches usually deliver different results – particularly if groups are unequal in size or the number of groups is small. As described earlier, the VGR is defined in terms of a known, finite number of groups such as occupations; the VRE, on the other hand, supposes that we observe a sample of groups from a larger population of group effects. The VRE has meaning only with respect to this particular data-generating process, in which the covariates relate linearly to the outcome and the group effect follows a specific (usually Normal) distribution. The VGR, meanwhile, is clearly defined without a data-generating model in mind. This means that it can be estimated using nonlinear functions of X, including by machine learning. The

3. To see why: at the population level, (14) coupled with the assumptions that $\mathbb{E}[Xv] = \mathbb{E}[Xw] = 0$ implies $\mathbb{E}[XY] = \mathbb{E}[XX'\beta^{\text{ME}}]$, so $\beta^{\text{ME}} = \mathbb{E}[XX']^{-1}\mathbb{E}[XY] = \beta^{\text{P}}$; β^{P} is the pooled slope obtained from regression of Y on X without any group information. See 3.1 for more information about estimating the VGR with a pooled slope.

concern that a naive sample will over-state variance between groups – one motivation for regularizing via random effects – can instead by addressed by the sample-splitting procedure outlined in Section 3.3. Nonetheless, if we are interested in a setting with an increasing number of small groups, such as families, imposing a random-effects assumption makes the problem tractable.

4.2 Variance of fixed effects (VFE)

The variance of fixed effects is commonly used in labor-market research (e.g. Card et al. 2018; Wilmers and Aeppli 2021). It has us (a) regress Y on X and a vector of group indicators in order to obtain group fixed effects:

$$y_i = \beta^{\mathrm{w}'} x_i + \sum_{g>1} \alpha_g 1_{g_i=g} + u_i, \tag{15}$$

where $G = 1$ is the reference group, and u is mean independent of g. Then (b) we calculate the variance of the fixed effects, $\mathbb{V}(\alpha_g)$, weighting by group size. The fixed effects are often interpreted as measurement of a groups' effect or premium controlling for X (McNeish and Kelley 2019), and their variance as the between-group variation adjusting for X. It is well known that naive estimation will tend to over-estimate the variance of fixed effects (Kline, Saggio, and Solvsten 2020), though such finite-sample performance is not the focus of this paper.

There are two substantial differences between the VFE and the VGD advanced in this paper. First, the FE approach assumes group effect homogeneity: group g differs from the reference group by the same value α_g at all levels of X. If this assumption is not met, then even the population-level FEs do not give any familiar average of the differences between g and the reference group at levels of X. Instead, the group fixed effects α_g are 'contaminated' by the comparisons between other groups (Goldsmith-Pinkham, Hull, and Kolesár 2022). The VFE can therefore be seen as an inconsistent estimator of the variance of group effects relative to some reference group. We may wish to estimate this target more accurately in the presence of group-effect heterogeneity; in that case, it would be more accurate to fit separate models comparing each group to the reference group, and then take the variance of the group indicator coefficients.[4]

Second, the VGD spelled out earlier focuses not on the comparison between g and a specific reference group, but between g and all other observations. Section 2.3

4. Letting 1 be the reference group, we fit $y_i = \beta^{g\prime} x_i + \lambda_g 1_{g_i=g} + u_i$ for $g = 2, 3, \ldots$. When estimating this model, we include observations only from groups 1 and g. Per Angrist and Pischke (2009), λ_g then captures the variance-weighted treatment effect of g as compared to the reference group.

argues that this latter comparison is more meaningful, since in many real cases it is hard to find a single group to serve as a reasonable reference category across all levels of X. Instead, we can compare each group to its complement, as long as we rescale before calculating variance in order to avoid double-counting differences (See Fig 2). In practice, this change in comparisons means that the VGD is often less than the VGR; however, this is not necessarily the case and will depend on the data at hand.

The VFE thus differs from the VGD in two respects: the presence of contamination bias and the choice of comparisons. To show how these play out in practice, I present the results of a simple simulation. I first assign observations to one of three groups $g = 1, 2, 3$ chosen equally at random. I generate a single Bernoulli covariate x for which the probability of success is governed by a parameter a: at $a = 0$, x is evenly split between 0 and 1 for all groups; at $a = 1$, x is always 0 in group 1 and always 1 in group 3.[5] By using this simple, dichotomous x, we can focus on the comparisons made by each method, rather than issues of functional form. I then generate observations according to a simple linear model,

$$y = g + (1 + b \times g)x + u, \tag{17}$$

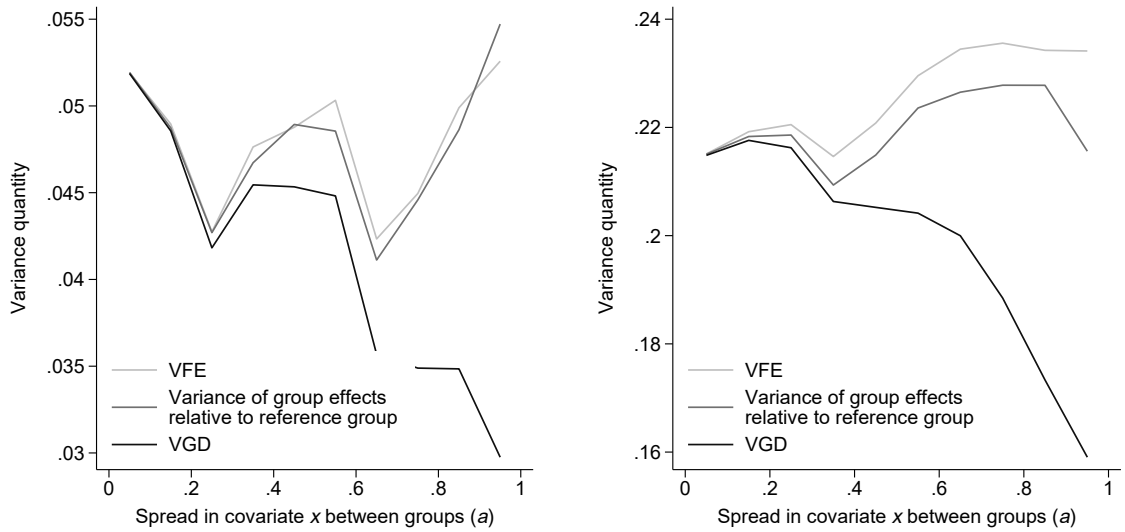
with noise $u \sim \mathcal{N}(0, 1)$. Here, b governs the extent of group-effect heterogeneity across levels of x: as b increases, the slope on x differs more between groups.

I reproduce this simulation at various levels of a, and for low and high levels of effect heterogeneity ($b = 0.2, b = 0.8$ respectively). For each simulation I draw 10,000 observations. I calculate three versions of the between-group variation: the VFE; the variance of group coefficients from the procedure described above; and the VGD.

The results of this exercise are plotted in Figure 3. The VGD is generally lower than the VFE, and this difference expands as groups grow further apart with respect to X. As noted above, the VFE and VGD differ due to the choice of comparison groups and due to the presence of contamination bias (Goldsmith-Pinkham, Hull, and Kolesár 2022). We can assess the degree of contamination bias by comparing the VFE to the variance of group effects relative to the reference group; these are very similar when group effects are homogeneous (Fig 3a), but unsurprisingly grow apart when group effects are heterogeneous (Fig 3b). However, the distance between the VFE and the variance of group effects relative to a reference group is generally quite small relative to the difference between the VFE and VGD. This means that, in this simulation, the bulk of the difference between VFE and VGD is due to the choice of comparisons rather than contamination bias.

5. Specifically, x equals 1 with probability

$$p = a\left(\frac{g-1}{2}\right) + (1-a)\frac{1}{2}. \tag{16}$$



(a) With a low level of group effect heterogeneity ($b = 0.2$).

(b) With a high level of group effect heterogeneity ($b = 0.8$).

Figure 3: Between-group variance quantities calculated using (17). Horizontal axis is the level of between-group dispersion in the covariate, specified by the parameter a in (16).

5 Example: income inequality between occupations.

Mouw and Kalleberg (2010) examine the role of occupations in structuring income inequality. Their investigation speaks to a rich and long-running sociology of occupations (e.g., Blau and Duncan 1967), as well as groups and boundaries more generally (Weber 1978). Here occupations are social groups saddled with meaning and boundaries, providing the basis for claims to resources (Weeden and Grusky 2012). Occupational memberships are not necessarily "treatments" in the causal sense, and the goal is not to estimate the causal effect of belonging to particular occupations. Rather, the authors aim to measure the amount of overall economic stratification that occurs between occupations in order to assess the extent to which occupational dynamics structure inequality. Occupational disparities can grow as the relative earnings in different occupations changes, or as occupations' sizes fluctuate – most prominently in the theory of occupational polarization, wherein the ranks of high- and low-paying occupations have swollen (for ex., Dwyer and Wright 2019). Crucially, Mouw and Kalleberg (2010) want to adjust for the role of other factors, such as age and occupations, which covary with occupational membership and with compensation. They do so using the fixed effects approach. In this replication, I show how the quantities developed in this paper can inform our understanding of earnings inequality between occupations.

5.1 Data

I reproduce the original analysis with the Outgoing Rotation Group of the Current Population Survey (ORG) and a vector of covariates X that include: gender, race, union status, education, and age. In preparing the ORG for analysis, I hew closely to the authors' specifications (see p.411 of the original). I include only employed individuals between ages 18 and 65 with earnings between \$1 and \$100 per hour in real 1975 USD (\$4.95 to \$494.65 in 2021). I replace all top-coded earnings with 1.4 times the topcode value, and omit observations with allocated earnings. Because records from 1994 and 1995 do not include trustworthy allocation flags, I drop these years from my analysis. The one difference is that I will use the harmonized `OCC1990` occupation codes, which IPUMS has made available since the original 2010 paper.

5.2 Methods

I begin with the fixed-effects variance used in Mouw and Kalleberg (2010). For each year t I fit the regression

$$y_i = x_i' \beta_t^{\text{w}} + \sum_{\text{occ}} \alpha_{\text{occ},t} 1_{\text{occ}_i=\text{occ}} + u_i,$$

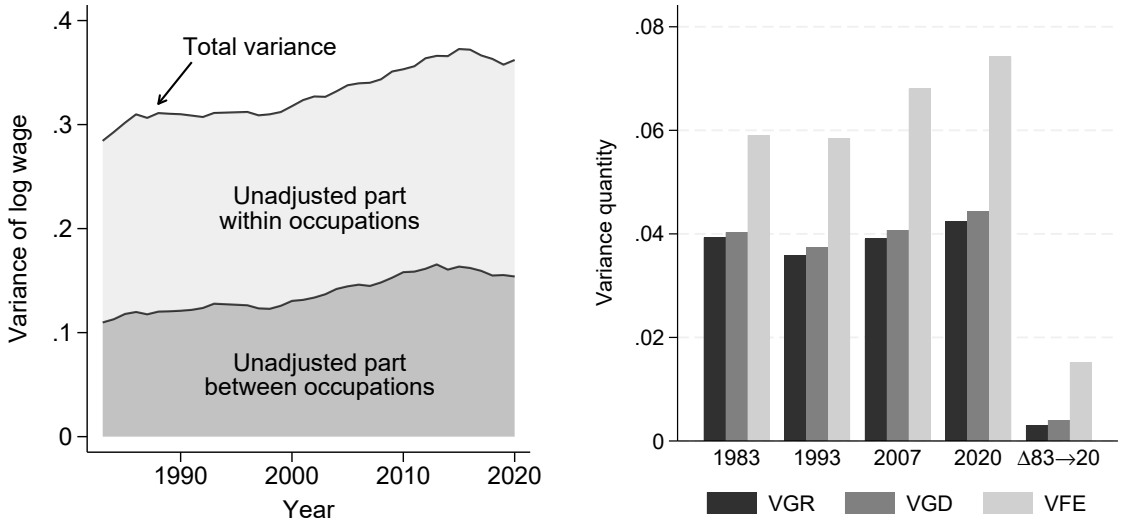
where the β_t^{w} are the returns to the covariates x_i in year t. Here $\alpha_{\text{occ}_i,t}$ are the occupation fixed effects in year t. The superscript w of β^{w} indicates that this is the *within*-group coefficient, obtained after removing occupational averages. I obtain a set of occupation fixed effects for each year, which are intended to capture occupational pay premiums net of the covariates. The employment-weighted variance of the estimated fixed effects is used to measure the between-occupation inequality adjusting for X.

I compare the VFE to estimates of the VGD and VGR quantities that I have proposed. First, I calculate a simple pooled-regression VGR quantity, as described in Sections 3 and 4. I (a) regress y on x each year; (b) calculate the residuals from these regressions; (c) average these residuals for each occupation; (d) subtract this from the overall average of y in each group, thereby estimating the differences δ_{occ}; and (e) take the employment-weighted variance of these these differences. Second, I calculate the linear version of the variance-weighted VGD. This is equivalent to regressing y_i on x_i and an indicator of a given occupation, separately for each occupation to obtain each effect τ_g; however, I use the faster procedure described in section 3.2. I also estimate several additional variance terms, presented in the appendix: the VRE fit via maximum likelihood estimation; the variance terms calculated with sample-splitting as described in Section 3.3; and more flexible approaches using Lasso and logistic regression instead of OLS to estimate the VGR and VGD.

5.3 Comparison

Figure 4b shows the resulting between-occupation components under the three adjustment strategies. I present results for three of the years – 1983, 1993, and 2007 – examined in Table 5 of the original paper (420). Figure 4b also includes results for 2020, in order to follow the evolution of between-occupation inequality in more recent years. The VFE results can be found in the fully-specified Model 3, "between-occupation" column of the original paper's Table 5. The VFE calculated here is 0.002 to 0.004 log-points smaller than in the original. This is likely due to differences in earnings imputation for observations with missing hourly or weekly pay. However, the patterns observed over time are close to those in the original paper.

The fixed-effects method finds much more inequality between occupations than do either of the methods proposed here: it accounts for almost a quarter of total



(a) Total variance of log hourly earnings (top line), with unadjusted decomposition (1).

(b) Between-occupation components using three strategies to adjust for covariates.

Figure 4: Analysis of the between-occupation component of income inequality in the United States. Data: CPS-ORG from IPUMS.

earnings variance in the last decade (which averaged around 0.35, top line of Figure 4a). By comparison, the VGR and VGD hardly break one tenth of total log wage variance. This divergence grew in recent years. A split-sample approach to correct for upward bias among the VFE, VGD, and VGR yields very similar results (Appendix C, Figure 6). Approaches based on logistic regression and Lasso – instead of OLS – result in similar, though slightly lower, estimates of the VGD and VGR (Figure 8).

What explains the gap between the VGD and VFE? Section 4.2 noted two key sources of difference: the fixed effects underlying the VFE are typically contaminated by other occupations' effects; and, even when adjusting for this contamination, the VFE approach compares each group to a single reference group. The role of these two differences depends on the nature of the data. As shown in the simulation (Fig 3), we can tease apart the role of these two differences. To do so, I use the largest occupation (managers and administrators) as a reference group in all years. I fit models of the form

$$y_i = \beta'_{\text{occ,t}} x_i + \lambda_{\text{occ},t} 1_{\text{occ}_i=\text{occ}} + u_i, \tag{18}$$

separately for each occupation, on the subsample of individuals whose occupation is either 'occ' or the reference occupation. The λ_{occ} then measures the variance-weighted effect of occupation 'occ' relative to reference occupation, without contamination bias

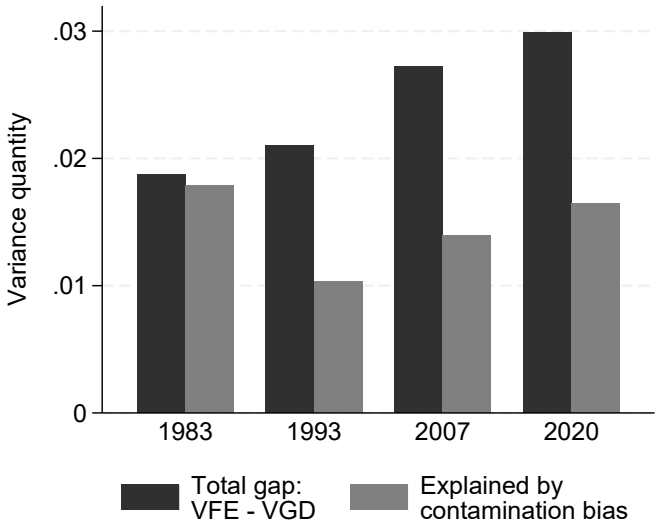


Figure 5: Decomposition of difference between two measures of between-occupation inequality. Black bars show the full gap between VFE and VGD; gray bars show the amount of this gap that can be explained by contamination bias using Equation 18.

(Goldsmith-Pinkham, Hull, and Kolesár 2022). The total gap between VFE and VGD can be decomposed into a part explained by contamination bias – the VFE minus the employment-weighted variance of the λs – and a part due to the difference in comparisons.

Figure 5 presents the full difference between VFE and VGD, as well as the part due to contamination bias. At the beginning of the period, the gap between the two is almost entirely due to the bias of the fixed effects approach. Over subsequent years, however, this contamination bias makes up only half of the total difference. The remaining half is explained not by the bias of the fixed-effects approach, but by the fact that it compares all occupations to a single reference occupation. As argued earlier, this choice results in unrealistic comparisons: for any given occupation, we compare it to the single reference occupation at each level of each covariate. The VGD compares each occupation to *all* other occupations, resulting in a more interpretable quantity.

However, there are related research questions for which the methods developed in this paper are poorly suited. Later in their paper, Mouw and Kalleberg (2010) explore how changing occupational composition has affected earnings inequality over time. Other examples include the changing effect of ownership on between-class inequality, and the role of class composition in widening between-class income stratification Wodtke (2016) and Zhou and Wodtke (2019). These topics typically involve pro-

ducing counterfactual estimates, where some features of the distribution of earnings, occupations, and other characteristics are held constant while others are modified. To calculate these counterfactuals, researchers usually need to make stronger modeling assumptions. Stronger modeling assumptions – such as constant returns to occupations – allows recent research to reveal, for instance, that rising occupational premiums have helped to reduce earnings inequality (Autor, Dube, and McGrew 2023). The VGR and VGD will be less useful for this type of longitudinal counterfactual question.

6 Conclusion

Sociologists and economists often examine the degree of inequality in an outcome occurring between groups. Typical groups are occupations, workplaces, schools, or social classes. In these analyses, a common goal is to adjust for variation that could be attributed to differing levels of covariates. However, it is typically unclear just what it means to calculate the degree of inequality between groups adjusting for this additional information.

In this paper, I explore two goals that a researcher may have in mind. First, she may intend to measure the additional explanatory power of incorporating group membership into her model, beyond the explanatory power of the covariates alone. Second, she may want to measure the dispersion in group effects calculated net of the covariates – or, in descriptive terms, the dispersion in comparisons between groups, where the comparisons are made conditional on covariates. I formulate two estimands reflecting these two objectives: the variance of group residuals (VGR) and the variance of group differences (VGD). Both are well defined even when groups do not overlap perfectly in their covariate distributions, and both simplify to the classic unconditional between-group variance term (1) when all groups have identical covariate distributions.

Both estimands are defined in nonparametric terms given a known, finite number of groups. As such, they can be estimated in any number of ways – including flexible machine-learning techniques, which can further be combined into a doubly-robust estimator. The simplest estimators, which use linear models of the outcome, are transparent and easy to implement. This sets them apart from standard methods that employ group fixed effects or random effects, since these result in quantities that only makes sense with respect to the particular (linear) model used. While the VGD reflects the intuition behind the variance of fixed effects, it differs from the latter by comparing each group to all other groups, rather than to a single reference group.

In a follow-up to Mouw and Kalleberg (2010), I calculate the between-occupation component of American earnings inequality using several estimators of my quantities

as well as the authors' fixed-effects method. Both VGR and VGD are about half the size of the fixed-effect quantity. Around half of this difference is due to contamination bias in the estimation of fixed effects (Goldsmith-Pinkham, Hull, and Kolesár 2022). The remaining half is due to the fact that the fixed-effects approach compares all occupations to a single reference occupation, which results in surprising comparisons among very low- or highly-education workers. For these reasons, the fixed-effect variance overstates the role of occupations in structuring income inequality.

This application does not hinge on occupations having a *causal* effect: it still gives a meaningful summary of inequality between occupations net of covariates, even absent the assumptions needed for causal identification. In general, neither the VGR nor the VGD must be causal quantities; their purpose is to precisely describe dispersion between groups, whether or not such dispersion counts as an "effect" or just an adjusted comparison. Of course, the VGD can give a useful summary of the spread in multiple groups' effects. But the idea of comparing groups at similar levels of a covariate is both useful and common in many descriptive projects (Słoczyński 2020), and this paper offers a natural way to summarize the spread of these comparisons across several groups. This exposition shows that Lundberg et al.'s (2021) call – to define our estimands prior to developing our methods – is not limited to causal questions. Descriptive social statistics, too, can be fruitfully improved by thinking through the comparisons we want to make.

References

Aeppli, Clem, and Nathan Wilmers. 2022. "Rapid wage growth at the bottom has offset rising US inequality." *Proceedings of the National Academy of Sciences* 119, no. 42 (2022): e2204305119.

Angrist, Joshua, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.

Autor, David, Arindrajit Dube, and Annie McGrew. 2023. *The Unexpected Compression: Competition at Work in the Low Wage Labor Market.* w31010. Cambridge, MA: National Bureau of Economic Research.

Baćak, Valerio, and Edward H. Kennedy. 2019. "Principled Machine Learning Using the Super Learner: An Application to Predicting Prison Violence." Publisher: SAGE Publications Inc, *Sociological Methods & Research* 48 (3): 698–721.

Barth, Erling, Alex Bryson, James C. Davis, and Richard Freeman. 2016. "It's Where You Work: Increases in the Dispersion of Earnings across Establishments and Individuals in the United States." *Journal of Labor Economics* 34 (S2): S67–S97.

Blau, Peter M., and Otis Dudley Duncan. 1967. *The American Occupational Structure.*

Card, David, Ana Rute Cardoso, Joerg Heining, and Patrick Kline. 2018. "Firms and Labor Market Inequality: Evidence and Some Theory." *Journal of Labor Economics* 36 (1): 13–70.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star." *The Quarterly Journal of Economics* 126 (4): 1593–1660.

Dwyer, Rachel E., and Erik Olin Wright. 2019. "Low-Wage Job Growth, Polarization, and the Limits and Opportunities of the Service Economy." *RSF: The Russell Sage Foundation Journal of the Social Sciences* 5 (4): 56–76.

Fortin, Nicole, Thomas Lemieux, and Sergio Firpo. 2011. "Decomposition Methods in Economics." In *Handbook of Labor Economics,* 4:1–102. Elsevier.

Godechot, Olivier, Marco Palladino, and Damien Babet. 2023. "In the Land of AKM: Explaining the Dynamics of Wage Inequality in France." *Working paper.*

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár. 2022. "Contamination Bias in Linear Regressions." *Working paper,* 46.

Haltiwanger, John C, Henry R Hyatt, and James Spletzer. 2022. "Industries, Mega Firms, and Increasing Inequality." NBER Working Paper Series.

Hamjediers, Maik, and Maximilian Sprengholz. 2023. "Comparing the Incomparable? Issues of Lacking Common Support, Functional-Form Misspecification, and Insufficient Sample Size in Decompositions." *Sociological Methodology* 53 (2): 344–365.

Iacus, Stefano M., Gary King, and Giuseppe Porro. 2012. "Causal inference without balance checking: Coarsened exact matching." *Political analysis* 20 (1): 1–24.

Kline, Patrick. 2011. "Oaxaca-Blinder as a Reweighting Estimator." *American Economic Review* 101 (3): 532–537.

Kline, Patrick, Raffaele Saggio, and Mikkel Solvsten. 2020. "Leave-out estimation of variance components." *Econometrica* 88 (5): 1859–1898.

Leung-Gagné, Josh, and Sean F Reardon. 2023. "It Is Surprisingly Difficult to Measure Income Segregation." *Demography* 60 (5): 1387–1413.

Liao, Tim Futing. 2022. "Individual Components of Three Inequality Measures for Analyzing Shapes of Inequality." *Sociological Methods & Research* 51 (3): 1325–1356.

Lundberg, Ian, Rebecca Johnson, and Brandon M Stewart. 2021. "What is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory." *American Sociological Review* 86 (3): 532–565.

McNeish, Daniel, and Ken Kelley. 2019. "Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations." *Psychological Methods* 24 (1): 20–35.

Molina, M., and F. Garip. 2019. "Machine Learning for Sociology." *Annual Review Of Sociology* 45 (1): 27–45.

Mouw, Ted, and Arne Kalleberg. 2010. "Occupations and the Structure of Wage Inequality in the United States, 1980s to 2000s." *American Sociological Review* 73 (3): 402–431.

Raudenbush, Stephen, and Anthony Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Chicago: University of Chicago Press.

Słoczyński, Tymon. 2020. "Average Gaps and Oaxaca–Blinder Decompositions: A Cautionary Tale about Regression Estimates of Racial Differences in Labor Market Outcomes." *ILR Review* 73 (3): 705–729.

Tomaskovic-Devey, Donald, Anthony Rainey, Dustin Avent-Holt, Nina Bandelj, István Boza, David Cort, Olivier Godechot, et al. 2020. "Rising between-workplace inequalities in high-income countries." *Proceedings of the National Academy of Sciences* 117 (17): 9277–9283.

Van Winkle, Zachary. 2018. "Family Trajectories Across Time and Space: Increasing Complexity in Family Life Courses in Europe?" *Demography* 55, no. 1 (2018): 135–164.

Weber, Max. 1978. *Economy and Society.* Edited by Claus Wittich and Guenther Roth.

Weeden, Kim A., and David B. Grusky. 2012. "The Three Worlds of Inequality." *American Journal of Sociology* 117 (6): 1723–1785.

Western, Bruce, and Deirdre Bloome. 2009. "Variance Function Regressions for Studying Inequality." *Sociological Methodology* 39 (1): 293–326.

Western, Bruce, and Jake Rosenfeld. 2011. "Unions, Norms, and the Rise in U.S. Wage Inequality." *American Sociological Review,* 25.

Wilmers, Nathan, and Clem Aeppli. 2021. "Consolidated Advantage: New Organizational Dynamics of Wage Inequality." *American Sociological Review* 86 (6): 1100–1130.

Wodtke, Geoffrey T. 2016. "Social Class and Income Inequality in the United States: Ownership, Authority, and Personal Income Distribution from 1980 to 2010." *American Journal of Sociology* 121 (5): 1375–1415.

Xie, Yu, Alexandra Killewald, and Christopher Near. 2016. "Between- and Within-Occupation Inequality: The Case of High-Status Professions." *The ANNALS of the American Academy of Political and Social Science* 663 (1): 53–79.

Zhou, Xiang, and Geoffrey T Wodtke. 2019. "Income Stratification among Occupational Classes in the United States." *Social Forces* 97, no. 3 (2019): 945–972.

A Proofs of claims

The purpose of this appendix is to formalize and prove the key claims made in the paper. The first few subsections work through my claims about population-level quantities and their relations (A.1, A.2, A.3, A.4, A.5); later I take up the claims about estimation and asymptotics (A.6, A.8).

Notation. Outcome Y has support in $\mathbb{R}$ and measurable cumulative distribution function F_Y. X is a centered length-k random column vector with support $\mathcal{X} \subseteq \mathbb{R}^k$, with CDF F_X. The lowercase x refers to a specific realization of X; x_i to the column of covariates for observation i from a sample of size n; and $\mathbf{X} = (x_1, x_2, ..., x_n)^T$ to the design matrix.

Group membership is given by a random variable G that takes values $1, 2, ..., M$, where M is the (finite) number of distinct groups. The lowercase g refers to a specific group, or realization of G, and observation i is in group g_i. $1_{G=g}$ is an indicator taking 1 in group g; $\pi_g = \mathbb{E}[1_{G=g}]$ is the overall prevalence of g while $p_g(x) = \mathbb{E}[1_{G=g} \mid X = x]$ is g's proportion at $X = x$.

Throughout, I will make the following assumptions. The global proportions are (strictly) bounded $0 < \pi_g < 1$ for all groups g. However, I do allow the local proportions $p_g(x)$ to be 0 or 1 for certain values of (x, g). The expectations $\mathbb{E}[Y]$, $\mathbb{E}[Y \mid X = x]$ (for all $x \in \mathcal{X}$), and $\mathbb{E}[X]$ exist and are finite. Additionally, $\mathbb{E}[Y^2]$ and $\mathbb{E}[Y^2 \mid X = x]$ are finite and nonzero at all $x \in \mathcal{X}$.

Define the $k \times k$ covariance matrix $\Sigma = \mathbb{E}[XX']$ of X. For brevity, write the group averages as $\bar{X}_g = \mathbb{E}[X \mid G = g]$ and $\bar{Y}_g = \mathbb{E}[Y \mid G = g]$; and write the group residuals as $\tilde{X} = X - \bar{X}_g$ and $\tilde{Y} = Y - \bar{Y}_g$. Then we can define the group-level covariance matrix $\overline{\Sigma} = \mathbb{E}\left[\bar{X}\bar{X}'\right]$ and the residual covariance matrix $\tilde{\Sigma} = \mathbb{E}\left[\tilde{X}\tilde{X}'\right]$. I assume Σ, $\overline{\Sigma}$, and $\tilde{\Sigma}$ are finite and invertible. With these definitions, we can construct the pooled slope $\beta^{\text{p}} := \Sigma^{-1}\mathbb{E}[XY]$; the within-group slope $\beta^{\text{w}} := \tilde{\Sigma}^{-1}\mathbb{E}\left[\tilde{X}\tilde{Y}\right]$; and the between-group slope $\beta^{\text{b}} := \overline{\Sigma}^{-1}\mathbb{E}\left[\bar{X}\bar{Y}\right]$. Finally, let $\mathbf{I}_k$ be the $k \times k$ identity matrix.

A.1 Relation between δ_g and τ_g^{VWT}

I show that $\delta_g = (1 - \pi_g)\frac{\mathbb{E}[\mathbb{V}(1_{G=g}|X)]}{\mathbb{V}(1_{G=g})}\tau_g^{\text{VWT}}$, where $\delta_g := \mathbb{E}[Y - \mathbb{E}[Y \mid X] \mid G = g]$ and τ_g^{VWT} is the variance-weighted treatment effect,

$$\tau_g^{\text{VWT}} = \mathbb{E}\left[\frac{\mathbb{V}(1_{G=g} \mid X)}{\mathbb{E}[\mathbb{V}(1_{G=g} \mid X)]}\Big\{\mathbb{E}[Y \mid G = g, X] - \mathbb{E}[Y \mid G \neq g, X]\Big\}\right],$$

where $\mathbb{V}(1_{G=g} \mid x) = p_g(x)(1 - p_g(x))$ is the variation in group g membership at $X = x$ and $\mathbb{V}(1_{G=g}) = \pi_g(1 - \pi_g)$ is the overall variation in group membership.

The relation follows from Bayes rule:

$$\begin{aligned}\delta_g = \mathbb{E}[Y - \mathbb{E}[Y \mid X] \mid G = g] &= \int_{\mathcal{X}} \Big\{\mathbb{E}[Y \mid G = g, X] - \mathbb{E}[Y \mid X]\Big\} \mathrm{d}F(x \mid G = g) \\ &= \int_{\mathcal{X}} (1 - p_g(x)) \Big\{\mathbb{E}[Y \mid G = g, x] - \mathbb{E}[Y \mid G \neq g, x]\Big\} \mathrm{d}F(x \mid G = g) \\ &= \int_{\mathcal{X}} (1 - p_g(x)) \frac{p_g(x)}{\pi_g} \Big\{\mathbb{E}[Y \mid G = g, x] - \mathbb{E}[Y \mid G \neq g, x]\Big\} \mathrm{d}F(x) \\ &= \frac{1}{\pi_g} \mathbb{E}\left[\mathbb{V}(1_{G=g} \mid X) \Big\{\mathbb{E}[Y \mid G = g, x] - \mathbb{E}[Y \mid G \neq g, x]\Big\}\right].\end{aligned}$$

Noting that $\frac{1}{\pi_g} = \frac{1-\pi_g}{\pi_g(1-\pi_g)} = \frac{1-\pi_g}{\mathbb{V}(1_{G=g})}$ yields the result.

A.2 Linear version of A.1

Observe first that the linear VGR term for group g can be written

$$\begin{aligned}\delta_g^L &= \mathbb{E}[Y \mid G = g] - \mathbb{E}[X \mid G = g]' \beta^{\mathrm{p}} \\ &= \frac{1}{\pi_g} \mathbb{E}\left((Y - X\beta^{\mathrm{p}})\, \mathbf{1}_g\right) \\ &= \frac{1}{\pi_g} \mathrm{Cov}(Y - X\beta^{\mathrm{p}}, \mathbf{1}_g), \end{aligned} \tag{19}$$

since $\mathbb{E}[Y - X\beta^p] = 0$ by construction of β. Now, returning to (**??**), notice that by Frisch-Waugh-Lovell we can obtain τ_g^L in $Y = 1_{G=g}\tau_g^L + X'\beta + e'$ from the residualized regression

$$Y^{\perp X} = \mathbf{1}_g^{\perp X} \tau_g^L + e' \tag{20}$$

where $Y^{\perp X}$ is the residual from a regression of Y on X alone, and $\mathbf{1}_g^{\perp X}$ is the residual from a regression of the group-g indicator $\mathbf{1}_g$ on X alone. (20) implies

$$\begin{aligned}\tau_g^L &= \mathbb{V}(\mathbf{1}_g^{\perp X})^{-1} \mathrm{Cov}(Y^{\perp X}, \mathbf{1}_g^{\perp X}) \\ &= \mathbb{V}(\mathbf{1}_g^{\perp X})^{-1} \mathrm{Cov}(Y^{\perp X}, \mathbf{1}_g), \end{aligned} \tag{21}$$

the second line following since $\left(\mathbf{1}_g - \mathbf{1}_g^{\perp X}\right)$ is orthogonal to $Y^{\perp X}$. Now, notice that $Y^{\perp X}$ is simply $Y - X\beta^{\mathrm{p}}$ from the pooled regression of Y on X! So $\mathrm{Cov}(Y - X'\beta^{\mathrm{p}}, \mathbf{1}_g) = \mathbb{V}(\mathbf{1}_g^{\perp X})\tau_g^L$. Substituting this into (19) yields

$$\begin{aligned}\delta_g^L &= \frac{1}{\pi_g} \mathbb{V}(\mathbf{1}_g^{\perp X}) \tau_g^L \\ &= (1 - \pi_g) \frac{\mathbb{V}(\mathbf{1}_g^{\perp X})}{\mathbb{V}(g)} \tau_g^L,\end{aligned}$$

which is the claim.

A.3 Ordering of $\mathbb{V}(\delta_g)$ and $\mathbb{V}((1-\pi_g)\tau^{\mathbf{VWT}})$

The fact that $\mathbb{V}(\delta_g) \leq \mathbb{V}((1-\pi_g)\tau^{\text{VWT}})$ is a corollary of the relation in A.1, which lets us write

$$(1-\pi_g)\tau_g^{\text{VWT}} = a_g\delta_g \text{ where } a_g = \frac{\mathbb{V}(1_{G=g})}{\mathbb{E}[\mathbb{V}(1_{G=g} \mid X)]}.$$

By Jensen's inequality (or, seeing that $\mathbb{E}[\mathbb{V}(1_{G=g} \mid X)]$ is a variance component of $\mathbb{V}(1_{G=g})$), we know that $\mathbb{V}(1_{G=g}) \geq \mathbb{E}[\mathbb{V}(1_{G=g} \mid X)] \geq 0$. Hence $a_g \geq 1$. Then

$$\begin{aligned}\mathbb{V}((1-\pi_g)\tau_g^{\text{VWT}}) &= \mathbb{V}(a_g\delta_g) = \mathbb{V}(\delta_g + (a_g-1)\delta_g) \\ &= \mathbb{V}(\delta_g) + \mathbb{V}((a_g-1)\delta_g) + 2\text{Cov}(\delta_g, (a_g-1)\delta_g).\end{aligned}$$

The covariance term equals $\mathbb{E}[(a_g-1)\delta_g^2] - \mathbb{E}[\delta_g]\mathbb{E}[(a_g-1)\delta_g]$. Since $a_g \geq 1$, we know $(a_g-1)\delta_g^2 \geq 0$ for all g, so $\mathbb{E}[(a_g-1)\delta_g^2] \geq 0$. The product $\mathbb{E}[\delta_g]\mathbb{E}[(a_g-1)\delta_g]$ drops because

$$\mathbb{E}[\delta_g] = \sum_g \pi_g \mathbb{E}[Y \mid G=g] - \mathbb{E}[\mathbb{E}[Y \mid X] \mid G=g] = \mathbb{E}[Y] - \mathbb{E}[Y] = 0.$$

Thus $\mathbb{V}\left((1-\pi_g)\tau_g^{\text{VWT}}\right) \geq \mathbb{V}(\delta_g)$.

In the linear case described in 2.3, A.2 means that we can similarly write

$$\mathbb{V}((1-\pi_g)\tau_g^L) = \mathbb{V}(\overline{a}_g\delta_g^L)$$

for $\overline{a}_g = \mathbb{V}(1_{G=g})/\mathbb{V}(1_{G=g}^{\perp X}) \geq 1$. In this case, $\mathbb{E}[\delta_g^L] = \sum_g \pi_g \mathbb{E}[Y \mid G=g] - \mathbb{E}[X \mid G=g]'\beta^{\text{p}} = \mathbb{E}[Y] - \mathbb{E}[Y] = 0$ by construction of β^{p}. So $\mathbb{V}((1-\pi_g)\tau_g^L) \geq \mathbb{V}(\delta_g^L)$.

A.4 Ordering of linear VGR and fixed-effects variance (VFE)

The claim here is that the between-group variance calculated using FEs, $\mathbb{V}(\alpha_g)$ where $\alpha_g := \mathbb{E}[Y \mid G=g] - \mathbb{E}[X \mid G=g]\beta^{\text{w}}$, is greater than the VGR calculated using the pooled-slope approach, $\mathbb{V}(\delta_g^L)$ where $\delta_g^L = \mathbb{E}[Y \mid G=g] - \mathbb{E}[X \mid G=g]\beta^{\text{p}}$.

First, observe that the pooled coefficient vector β^{p} can be expressed as a combination of the within-group coefficients β^{w} and the between-group β^{b}: $\beta^{\text{p}} = \mathbf{W}\beta^{\text{w}} + (\mathbf{I}_k - \mathbf{W})\beta^{\text{b}} = \mathbf{W}(\beta^{\text{w}} - \beta^{\text{b}}) + \beta^{\text{b}}$ for $\mathbf{W} = \Sigma^{-1}\tilde{\Sigma}$. Substituting gives

$$\begin{aligned}\text{VGR} &= \mathbb{V}(\mathbb{E}[Y \mid G=g] - \mathbb{E}[X \mid G=g]\beta^{\text{p}}) \\ &= \mathbb{V}\left(\mathbb{E}[Y \mid G=g] - \mathbb{E}[X \mid G=g]\left\{\mathbf{W}(\beta^{\text{w}} - \beta^{\text{b}}) + \beta^{\text{b}}\right\}\right) \\ &= \mathbb{V}\left(\mathbb{E}[Y \mid G=g] - \mathbb{E}[X \mid G=g]\beta^{\text{b}} - \mathbb{E}[X \mid G=g]\mathbf{W}(\beta^{\text{w}} - \beta^{\text{b}})\right) \\ &= \mathbb{V}(\mathbb{E}[Y \mid G=g] - \mathbb{E}[X \mid G=g]\beta^{\text{b}}) + \mathbb{V}(\mathbb{E}[X \mid G=g]\mathbf{W}\Delta\beta) \\ &= \mathbb{V}(\mathbb{E}[Y \mid G=g] - \mathbb{E}[X \mid G=g]\beta^{\text{b}}) + \Delta\beta'\mathbf{W}'\overline{\Sigma}\mathbf{W}\Delta\beta,\end{aligned}$$

where $\Delta\beta = \beta^{\mathrm{w}} - \beta^{\mathrm{b}}$. The fourth line follows from the fact that, by construction of least-squares, $\mathbb{E}[Y \mid G = g] - \mathbb{E}[X \mid G = g]\beta^{\mathrm{b}}$ is orthogonal to any linear function of $\mathbb{E}[X \mid G = g]$. Similarly, β^{w} can be written as $\beta^{\mathrm{b}} + \Delta\beta$, so

$$\mathrm{VFE} = \mathbb{V}(\mathbb{E}[Y \mid G = g] - \mathbb{E}[X \mid G = g]\beta^{\mathrm{b}}) + \Delta\beta'\overline{\Sigma}\Delta\beta.$$

To show VGR < VFE, we must show that $\Delta\beta'\mathbf{W}'\overline{\Sigma}\mathbf{W}\Delta\beta < \Delta\beta'\overline{\Sigma}\Delta\beta$. This will follow, for any coefficient gap $\Delta\beta$, if we can show that

$$\overline{\Sigma} \succ \mathbf{W}'\overline{\Sigma}\mathbf{W},$$

i.e. that $\overline{\Sigma} - \mathbf{W}'\overline{\Sigma}\mathbf{W} \succ 0$ (is positive definite).

To show this, note first that $\mathbf{W}'\overline{\Sigma} = \left[\Sigma^{-1}(\Sigma - \bar{\Sigma})\right]'\overline{\Sigma} = (\Sigma - \overline{\Sigma})\Sigma^{-1}\overline{\Sigma}$ since Σ and $\overline{\Sigma}$ are symmetric pd matrices, hence have symmetric inverses. This can be expressed as $\overline{\Sigma} - \overline{\Sigma}\Sigma^{-1}\overline{\Sigma}$, which can also be factored as $\overline{\Sigma}\Sigma^{-1}(\Sigma - \overline{\Sigma}) = \overline{\Sigma}'\mathbf{W}$. This means that

$$\overline{\Sigma} - \mathbf{W}'\overline{\Sigma}\mathbf{W} = (\mathbf{I}_k - \mathbf{W})'\,\overline{\Sigma}\,(\mathbf{I}_k + \mathbf{W})\,.$$

Substituting in the definition of $\mathbf{W} = \Sigma^{-1}\tilde{\Sigma}$,

$$\begin{aligned}
\overline{\Sigma} - \mathbf{W}'\overline{\Sigma}\mathbf{W} &= (\Sigma^{-1}\overline{\Sigma})'\overline{\Sigma}(2\mathbf{I}_k - \Sigma^{-1}\overline{\Sigma}) \\
&= 2\overline{\Sigma}\Sigma^{-1}\overline{\Sigma} - \overline{\Sigma}\Sigma^{-1}(\Sigma - \tilde{\Sigma})\Sigma^{-1}\overline{\Sigma} \\
&= 2\overline{\Sigma}\Sigma^{-1}\overline{\Sigma} - \overline{\Sigma}\Sigma^{-1}\overline{\Sigma} + \overline{\Sigma}\Sigma^{-1}\tilde{\Sigma}\Sigma^{-1}\overline{\Sigma} \\
&= \overline{\Sigma}\Sigma^{-1}\overline{\Sigma} + (\Sigma^{-1}\overline{\Sigma})'\tilde{\Sigma}(\Sigma^{-1}\overline{\Sigma}),
\end{aligned}$$

which, as the sum of two positive definite matrices, is also positive definite. Hence $\overline{\Sigma} \succ \mathbf{W}'\overline{\Sigma}\mathbf{W}$, which means that $\Delta\beta'\overline{\Sigma}\Delta\beta > \Delta\beta'\mathbf{W}'\overline{\Sigma}\mathbf{W}\Delta\beta$ and so the FE variance is greater than the VGR.

A.5 Incidental double robustness of δ_g

I've noted that δ_g (defined as $\mathbb{E}[Y \mid G = g] - \mathbb{E}[\mathbb{E}[Y \mid X] \mid G = g]$) is identified via $\mathbb{E}[Y \mid G = g] - \mathbb{E}[m(X) \mid G = g]$ for a model $m(x)$ of $x \mapsto \mathbb{E}[Y \mid X = x]$. This appendix illustrates that δ_g calculated using m may still be identified even if m is incorrectly specified, as long as the conditional group prevalences $p_g : x \mapsto \mathbb{E}[1_{G=g} \mid X = x]$ are contained within the linear space of functions $\mathcal{X} \to \mathbb{R}$ from which we have selected m.

Say we have a linear space $\mathcal{M}$ of maps $\mathcal{X} \to \mathbb{R}$. For instance, $\mathcal{M}$ can include nonlinear maps of $\mathcal{X}$. We will pick the best m for modeling $\mathbb{E}[Y \mid X = x]$:

$$m := \mathrm{argmin}_{m' \in \mathcal{M}} \mathbb{E}\left[(\mathbb{E}[Y \mid X] - m'(X))^2\right]. \tag{22}$$

Say one of two conditions is met: either the conditional expectation is in $\mathcal{M}$ (i.e. the map $x \mapsto \mathbb{E}[Y \mid X = x]$ is in $\mathcal{M}$); or the conditional group prevalence is in $\mathcal{M}$ (i.e. $x \mapsto p_g(x) = \mathbb{E}[1_{G=g} \mid X = x]$ is in $\mathcal{M}$).

The claim here is that if either of these conditions holds, then $\mathbb{E}[m(X) \mid G = g]$ identifies $\mathbb{E}[\mathbb{E}[Y \mid X] \mid G = g]$.

First note that

$$\begin{aligned}\mathbb{E}\left[m(X) \mid G = g\right] &= \mathbb{E}\left[m(X)\frac{p_g(X)}{\pi_g}\right] \quad \text{by Bayes} \\ &= \mathbb{E}\left[\mathbb{E}[Y \mid X]\frac{p_g(X)}{\pi_g}\right] + \mathbb{E}\left[\left(m(X) - \mathbb{E}[Y \mid X]\right)\frac{p_g(X)}{\pi_g}\right].\end{aligned}$$

If the first condition is met, then it is clear that the second term vanishes and so $\mathbb{E}[m(X) \mid G = g]$ identifies the target.

Now suppose that the second conditions holds – that is $p_g \in \mathcal{M}$. Then p_g is orthogonal to the difference $\mathbb{E}[Y \mid X] - m(x)$, so $\mathbb{E}\left[\left(m(X) - \mathbb{E}[Y \mid X]\right)\frac{p_g(X)}{\pi_g}\right] = 0$. To see this, note that if it were not the case, then $\mathbb{E}\left[\left(m(X) - \mathbb{E}[Y \mid X]\right)p_g(X)\right] = c$ for some $c \neq 0$. We could then form a new estimator $m' := m - cp_g \in \mathcal{M}$ since $m, p_g \in \mathcal{M}$ and $\mathcal{M}$ is a linear space by hypothesis. m' would then out-perform m in the sense of (22), violating the hypothesis.

Thus $p_g \in \mathcal{M}$ suffices to identify $\mathbb{E}[\mathbb{E}[Y \mid X] \mid G = g]$, and hence to identify δ_g – even if the map $x \mapsto \mathbb{E}[Y \mid X = x]$ is *not* in $\mathcal{M}$, meaning that m does not capture $\mathbb{E}[Y \mid X = x]$. This property does not require us to actually model $p_g(x)$ as we will do in A.6; it comes simply from using a model $m(x)$ of $\mathbb{E}[Y \mid X = x]$ that we have selected according to (22). The linear version above is a special case of this, where $\mathcal{M}$ is the space of linear maps $x \mapsto x'b$ with $b \in \mathbb{R}^k$. This is related to the observation of Kline (2011) that the Oaxaca-Blinder-Kitagawa decomposition can also be seen as a reweighting estimator.

A.6 Consistency of the $\hat{\delta}_g$ terms

This section deals with the three types of estimators of the $\hat{\delta}_g$ terms: those that use estimated conditional Y-means $\hat{m}(x) = \hat{\mathbb{E}}[Y \mid X = x]$; those that use estimated group prevalences $\hat{p}_g(x) = \hat{\mathbb{E}}[1_{G=g} \mid X = x]$; and the doubly-robust estimator that combines them. In all cases, I consider the setting where sample size $n \to \infty$ and group sizes $n_g \to \infty$ for all groups g, while $\frac{n_g}{n}$ converges to some π_g bounded away from 0.

A.6.1 Average (m) based estimation

I show that if $\hat{m}$ is uniformly consistent for m, then we can consistently estimate δ_g. That is, I require that $\sup_{x \in \mathcal{X}} |\hat{m}(x) - \mathbb{E}[Y \mid X = x]| \to_P 0$ where $A \to_P 0$ denotes convergence in probability to 0 of a sequence $\{A_n\}$ indexed by n.

$$\begin{aligned}\hat{\delta}_h^m &:= \frac{1}{n_h}\sum_{g_i=h} y_i - \hat{m}(x_i) \\ &= \frac{1}{n_h}\sum_{g_i=h} y_i - \mathbb{E}[Y \mid X = x_i] + \frac{1}{n_h}\sum_{g_i=h} \mathbb{E}[Y \mid X = x_i] - \hat{m}(x_i) \to_P \delta_h,\end{aligned}$$

where the second term disappears by the hypothesis of uniform consistency of $\hat{m}$ and the first term converges to δ_h by the Law of Large Numbers.

A.6.2 Prevalence (p_g) based estimation

Here I show that uniform consistency of the prevalence $\hat{p}_h(x)$ (i.e., $\sup_{x\in\mathcal{X}} |\hat{p}_h(x) - p_h(x)| \to_P 0$) means that the prevalence-based estimator is consistent. I also require that the overall group share $\hat{\pi}_h$ is estimated consistently and is bounded away from zero.

$$\begin{aligned}\hat{\delta}_h^p &:= \frac{1}{n_g}\sum_{g_i=h} y_i - \frac{1}{n}\sum_i y_i \frac{\hat{p}_h(x_i)}{\hat{\pi}_h} \\ &= \frac{1}{n_g}\sum_{g_i=h} y_i - \frac{1}{n}\sum_i y_i \frac{p_h(x_i)}{\hat{\pi}_h} + \frac{1}{n}\sum_i \frac{y_i}{\hat{\pi}_h} p_h(x_i) - \hat{p}_h(x_i).\end{aligned}$$

Application of Cauchy-Schwartz to the third summation, coupled with the fact that

$$\sqrt{\frac{1}{n}\sum_i \left(p_h(x_i) - \hat{p}_h(x_i)\right)^2} \to_P 0$$

by the uniform consistency, means that the third term converges to 0. We are left with the first two terms. The Law of Large Numbers coupled with the consistency of $\hat{\pi}$ means that $\hat{\delta}_h^p \to \delta_h$.

A.6.3 Doubly robust estimator of δ_g

Given estimators $\hat{p}_g$ of the conditional proportions $p_g : x \mapsto \mathbb{E}[1_{G=g} \mid X = x]$ and an estimator $\hat{m}$ of the conditional means $x \mapsto \mathbb{E}[Y \mid X = x]$, Equation 10 described a new estimator of δ_h:

$$\hat{\delta}_h^{DR} = \underbrace{\frac{1}{n_h}\sum_{j:g_j=h} \left\{y_j - \hat{m}(x_j)\right\}}_{A} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{p}_h(x_j)}{\pi_h}\left\{\hat{m}(x_j) - y_j\right\}}_{B}.$$

The claim is that if either $\hat{p}_h$ is uniformly consistent for p_h or $\hat{m}$ is uniformly consistent for $\mathbb{E}[Y \mid X = x]$, then $\hat{\delta}_h^{DR}$ is consistent for $\delta_h := \mathbb{E}[Y \mid h] - \mathbb{E}[\mathbb{E}[Y \mid X] \mid h]$.

I also require that both $\mathbb{E}[\hat{p}_h(X)^2]$ and $\mathbb{E}[\hat{m}(X)^2]$ are finite and nonzero, and that both $x \mapsto \hat{m}(x)$ and $x \mapsto \hat{p}_h(x)$ are continuous on $\mathcal{X}$.

The proof is very similar to proofs of the consistency of doubly-robust estimators of various causal quantities (ATE, etc.).

Case where $\hat{m}$ is uniformly consistent for $\mathbb{E}[Y \mid X = x]$. First note that

$$A = \frac{1}{n_h} \sum_{i:g_i=h} y_i - \mathbb{E}[Y \mid X = x_i] + \frac{1}{n_h} \sum_{i:g_i=h} \mathbb{E}[Y \mid X = x_i] - \hat{m}(x_i). \tag{23}$$

If $\hat{m}$ is uniformly consistent for $\mathbb{E}[Y \mid X = x]$, then the second term above $\to_P 0$. We are left with the first term $\frac{1}{n_h} \sum_{i:g_i=h} y_i - \mathbb{E}[Y \mid X = x_i]$, whence by the Law of Large Numbers,

$$A \to_P \mathbb{E}[Y - \mathbb{E}[Y \mid X] \mid h] = \delta_h. \tag{24}$$

We now must show that $B \to_P 0$. Note that B can be expanded to

$$B = \frac{1}{n} \sum_i [\hat{m}(x_i) - \mathbb{E}[Y \mid X = x_i]] \frac{\hat{p}_h(x_i)}{\pi_h} + \frac{1}{n} \sum_i [\mathbb{E}[Y \mid X = x_i] - y_i] \frac{\hat{p}_h(x_i)}{\pi_h}. \tag{25}$$

By application of Cauchy-Schwartz and the uniform convergence of $\hat{m}$, the first term of (25) converges to 0. This leaves the second term of (25). By the uniform law of large numbers,

$$\frac{1}{n} \sum_i [\mathbb{E}[Y \mid X = x_i] - y_i] \frac{\hat{p}_h(x_i)}{\pi_h} \to_P \mathbb{E}\left[[\mathbb{E}[Y \mid X = x_i] - Y] \frac{\hat{p}_h(X)}{\pi_h} \right], \tag{26}$$

and by iterated expectations,

$$\begin{aligned} \mathbb{E}\left[[\mathbb{E}[Y \mid X] - Y] \frac{\hat{p}_h(X)}{\pi_h} \right] &= \mathbb{E}\left[\mathbb{E}\left[(\mathbb{E}[Y \mid X] - Y) \frac{\hat{p}_h(X)}{\pi_h} \mid X \right] \right] \\ &= \mathbb{E}\left[(\mathbb{E}[Y \mid X] - \mathbb{E}[Y \mid X]) \frac{\hat{p}_h(X)}{\pi_h} \right] = 0. \end{aligned}$$

This means that $|B| \to_P 0$. Hence $\hat{\delta}_h^{DR} \to_P \delta_h$.

Case where $\hat{p}_h$ is uniformly consistent for p_h. Application of the uniform law of large numbers to A in (23) gives

$$A \to_P \delta_h + \mathbb{E}[\mathbb{E}[Y \mid X] - \hat{m}(X) \mid h].$$

Further expanding (25) gives

$$\begin{aligned} B = &\frac{1}{n} \sum_i [\hat{m}(x_i) - \mathbb{E}[Y \mid X = x_i]] \frac{p_h(x_i)}{\pi_h} \\ &+ \frac{1}{n} \sum_i [\hat{m}(x_i) - \mathbb{E}[Y \mid X = x_i]] \left(\frac{\hat{p}_h(x_i) - p_h(x_i)}{\pi_h} \right) + \frac{1}{n} \sum_i [\mathbb{E}[Y \mid X = x_i] - y_i] \frac{\hat{p}_h(x_i)}{\pi_h}. \end{aligned} \tag{27}$$

Via Cauchy-Schwartz and the uniform consistency of $\hat{p}_h$, the first term of the second line in (27) $\to_P 0$. Additionally, by the same logic surrounding (26), the second term in the second line of (27) also converges in probability to 0.

Coupled with the uniform law of large numbers, this means that

$$B \to_P \mathbb{E}\left[(\hat{m}(X) - \mathbb{E}[Y \mid X])\,\frac{p_h(X)}{\pi_h}\right] = \mathbb{E}[\hat{m}(X) - \mathbb{E}[Y \mid X] \mid h],$$

the second equality by Bayes. Hence $\hat{\delta}_h^{DR} \to_P \delta_h$.

A.7 Quicker computation of VGD

Calculating the VGD requires us to first obtain the variance-weighted group effects/differences τ_g for each group g. As mentioned in the text, the effect for group g equals the coefficient on an indicator for group g in the following regression:

$$Y = \lambda(g)'X + \tau_g \mathbf{1}_{G=g} + U.$$

The issue with this approach is we would have to run a separate regression for each group g, each featuring different coefficients on X – this is why the coefficients λ on X are indexed by $g = 1, 2, ..., M$.

This section explains how we can avoid computing M separate regressions. First applying Frisch-Waugh-Lovell to partial out X, note that we can write the τ_g as

$$\tau_g = \frac{\mathbb{E}[Y\tilde{\mathbf{1}}_{G=g}]}{\mathbb{E}[\tilde{\mathbf{1}}_{G=g}^2]},$$

where $\tilde{\mathbf{1}}_{G=g}$ is the residual of $\mathbf{1}_{G=g}$ from regression on X, i.e. $\mathbf{1}_{G=g} = \theta(g)'X + V$. Expanding with $\tilde{\mathbf{1}}_{G=g} = \mathbf{1}_{G=g} - \theta(g)'X$ gives

$$\tau_g = \frac{\mathbb{E}[Y\mathbf{1}_{G=g}] - \mathbb{E}[YX]'\theta(g)}{\mathbb{E}[\mathbf{1}_{G=g}^2] - \theta(g)'\mathbb{E}[XX']\theta(g)}.$$

For any group g, the coefficient vectors $\theta(g)$ are obtained by

$$\theta(g) = \mathbb{E}[XX']^{-1}\mathbb{E}[X\mathbf{1}_{G=g}].$$

Thus, τ_g can be expressed as:

$$\tau_g = \frac{\mathbb{E}[Y\mathbf{1}_{G=g}] - \mathbb{E}[YX]\mathbb{E}[XX']^{-1}\mathbb{E}[X\mathbf{1}_{G=g}]}{\mathbb{E}[\mathbf{1}_{G=g}] - \mathbb{E}[X\mathbf{1}_{G=g}]'\mathbb{E}[XX']^{-1}\mathbb{E}[X\mathbf{1}_{G=g}]}.$$

This can be estimated from the data using (12) in the unweighted case. If sample weights are to be included, then a vector $\mathbf{W}$ of weights can be incorporated. The crucial observation here is that all the group effects can be obtained using only a single matrix inversion – that

of $\mathbb{E}[XX']$, estimated proportional to $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ in the unweighted case. This greatly speeds up computation. This puts it on par with computation of the VFE or the VGR, both of which require fitting only one regression. In some cases it can even be faster than the fixed effects regression, which typically involves inverting the covariance matrix of the vector $(X_1, ..., X_K, \mathbf{1}_1, ..., \mathbf{1}_G)'$ instead of just $(X_1, ..., X_K)'$.

A.8 Bias of plug-in VGR estimator

The goal here is to formalize conditions for the finite bias and asymptotic consistency of the plug-in VGR and VGD estimators, as well as their finite-sample bias. This exposition makes no assumptions about the specific type of estimators used for the δ or τ terms (whether pooled regression, classification, etc.).

Suppose we estimate δ_g with $\hat{\delta}_g$ from one of the methods described in Section 3. We also estimate the group proportion π_g as $\hat{\pi}_g$ – for instance, using the sample proportion $\hat{\pi}_g = n_g/n$ where n is the number of total observations and n_g the number in group g. We can create a plug-in estimator of the VGR:

$$\hat{\text{VGR}} = \sum_g \hat{\pi}_g \hat{\delta}_g^2.$$

Even if $\hat{\pi}$ and $\hat{\delta}$ are both unbiased – that is, $\mathbb{E}[\hat{\pi}_g] = \pi_g$ and $\mathbb{E}[\hat{\delta}_g] = \delta_g$ – this plug-in estimator can still be biased in finite samples:

$$\begin{aligned}
\mathbb{E}\left[\hat{\text{VGR}}\right] &= \mathbb{E}\left[\sum_g \hat{\pi}_g \hat{\delta}_g^2\right] = \sum_g \mathbb{E}\left[\hat{\pi}_g \hat{\delta}_g^2\right] \\
&= \sum_g \mathbb{E}\left[\hat{\pi}_g\right] \mathbb{E}\left[\hat{\delta}_g^2\right] + \text{Cov}\left(\hat{\pi}_g, \quad \hat{\delta}_g^2\right) \\
&= \sum_g \pi_g \left\{\mathbb{E}\left[\hat{\delta}_g\right]^2 + \mathbb{V}\left(\hat{\delta}_g\right)\right\} + \text{Cov}\left(\hat{\pi}_g, \quad \hat{\delta}_g^2\right) \\
&= \text{VGR} + \sum_g \pi_g \mathbb{V}\left(\hat{\delta}_g\right) + \text{Cov}\left(\hat{\pi}_g, \quad \hat{\delta}_g^2\right). \qquad (28)
\end{aligned}$$

If group assignment G is nonrandom and π_g is known, the covariance drops and we get the bias expression in (13). If G is random and π_g is estimated as $\hat{\pi}_g = n_g/n$, then $\text{Cov}(\hat{\pi}_g, \hat{\delta}_g^2)^2 \leq \mathbb{V}(\hat{\pi}_g)\mathbb{V}(\hat{\delta}_g^2)$. Given minimal regularity conditions – that $\hat{\pi}$ bounded away from 0 and $\mathbb{V}(\hat{\delta}^2)$ is finite – we see that $\mathbb{V}(\hat{\pi}_g) \to 0$ relatively quickly as $n \to \infty$. This leaves the middle term, which captures the average sampling variance of the $\hat{\delta}$ terms. The finite-sample bias of the plug-in VGD is similar to (28).

Now, take the asymptotic setting where sample size $n \to \infty$ and group size $n_g \to \infty$ for each group g, and where $\hat{\pi}_g = n_g/n$ converges to a share π_g bounded away from 0. I assume the consistency of the $\hat{\delta}_g$ or $\hat{\tau}_g^w$; that is, $\hat{\delta}_g \to_P \delta_g$ or $\hat{\tau}_g^w \to_P \tau_g^w$ for each group g,

with $\to_P$ denoting convergence in probability. Then application of Slutsky's Theorem and the Continuous Mapping Theorem show that

$$\hat{\text{VGR}} := \sum_g \hat{\pi}_g \hat{\delta}_g^2 \longrightarrow_P \sum_g \pi_g \delta_g^2 =: \text{VGR},$$

$$\text{and} \quad \hat{\text{VGD}}^w := \sum_g \hat{\pi}_g \hat{\tau_g^w}^2 - \left(\sum_g \hat{\pi}_g \hat{\tau}_g^w\right)^2 \longrightarrow_P \sum_g \pi_g \tau_g^{w2} - \left(\sum_g \pi_g \tau_g^w\right)^2 =: \text{VGD}.$$

B Causal interpretation of the VGD

I have noted that the VGD can be interpreted either causally or descriptively. In the causal case, it measures the spread in the average *effect* of group membership; in the descriptive case, it measures the spread in covariate-adjusted differences between groups. Here, I focus on the causal interpretation and formulate the VGD in potential outcomes notation.

Assumptions & interpretation. I suppose individuals i have potential outcomes for Y under each group assignment $Y_i(1), Y_i(2), ..., Y_i(M)$. I make the following assumptions:

- Group assignment is ignorable conditional on x_i, that is, $(Y_i(1), ..., Y_i(M)) \perp g_i \mid x_i$.
- Potential outcomes are consistent conditional on X, that is, $\mathbb{E}[Y_i(g) \mid g_i = g, x_i = x] = \mathbb{E}[Y(g) \mid X = x]$.
- There is no interference or spillover: i's potential outcomes $Y_i(g)$ are independent of the treatment status g_j of others $j \neq i$.

In general, we can consider a weighted average

$$\tau_g^\psi = \frac{1}{\mathbb{E}[\psi(p_g(X))]} \mathbb{E}\left[\psi(p_g(X)) \Delta_g(X)\right]$$

of X-specific contrasts between g and its complement,

$$\Delta_g(x) := \mathbb{E}[Y(g) \mid X = x] - \sum_{h \neq g} \frac{p_h(x)}{1 - p_g(x)} \mathbb{E}[Y(h) \mid X = x],$$

using a weight ψ that is a function of $p_h(x) = \mathbb{E}[1_{G=h} \mid X = x]$. The term $\Delta_g(x)$ measures the average effect of switching into group g at $X = x$ from another group h, the choice of which is proportional to the prevalence of h at $X = x$.

Then, τ_g^ψ tells us the effect of switching into group g, averaged across the Xs using the specified weighting scheme ψ (to obtain the average treatment effect on the treated, the average treatment effect, etc.). The VGD then measures the variation in these averaged effects.

Identification. We can identify the X-specific contrasts $\Delta_g(x)$ using

$$\mathbb{E}[Y_i \mid x_i = x, g_i = g] - \mathbb{E}[Y_i \mid x_i = x, g_i \neq g]$$

due to the assumptions of conditional ignorability, consistency, and non-interference. Given a normalized weighting function $\psi(g, x)$ that we can estimate consistently, then we can identify (and estimate, using the sample analog) the average τ_g^{ψ} using

$$\frac{1}{\mathbb{E}[\psi(p_g(X))]}\mathbb{E}\big[\psi(g, X)\left\{\mathbb{E}[Y_i \mid x_i = x, g_i = g] - \mathbb{E}[Y_i \mid x_i = x, g_i \neq g]\right\}\big].$$

This paper concentrates on a particular, variance-weighted effect. Note that the variance-weighted effect τ_g can also be expressed, in potential-outcomes notation, as

$$\tau_g = \frac{\mathbb{E}\left[Y(G)\ \tilde{1}_{G=g}\right]}{\mathbb{E}\left[\tilde{1}_{G=g}^2\right]},$$

where $\tilde{1}_{G=g} = 1_{G=g} - p_g(X)$. The numerator can be written

$$\mathbb{E}\left[Y(G)\ \tilde{1}_{G=g}\right] = \mathbb{E}\big[p_g(X)\mathbb{E}\left[Y(g) \mid X\right] - p_g(X)\mathbb{E}\left[Y(G) \mid X\right]\big].$$

We can identify $\mathbb{E}[Y(g) \mid X]$ by $\mathbb{E}[Y_i \mid X, g_i = g]$ using conditional ignorability and consistency. This means we can identify the variance-weighted τ_g, using

$$\frac{\mathbb{E}\left[Y_i\tilde{1}_{g_i=g}\right]}{\mathbb{E}\left[\tilde{1}_{g_i=g}^2\right]}.$$

Of course, consistently estimating this involves also consistently estimating the $p_g(x)$.

C Additional results from replication

Here I present additional measurements of between-occupation earnings spread, to supplement the results presented in Figure 4b of the main text. Neither sample-splitting nor alternative modeling strategies substantially changes the main conclusions.

In section 3, I noted that plug-in estimates of any of the variance terms – the variance of fixed effects, VGD, or VGR – can be biased even when the underlying group comparisons/effects are estimated without bias. I proposed randomly splitting the data into two samples; estimating the full set of group comparisons/effects on each sample; and using these to construct unbiased estimates of the variance terms. Figure 6 shows the results of full-sample (solid lines) and split-sample (dashed) approaches to the variance of fixed effects (VFE), the VGD, and the VGR. Regardless of the variance estimand, sample-splitting results in only minimally lower estimates.

Section 4.1 compares the VGR to the VRE. The key differences are that the VGR is defined for a known number of groups, possibly of unequal size. These differences mean

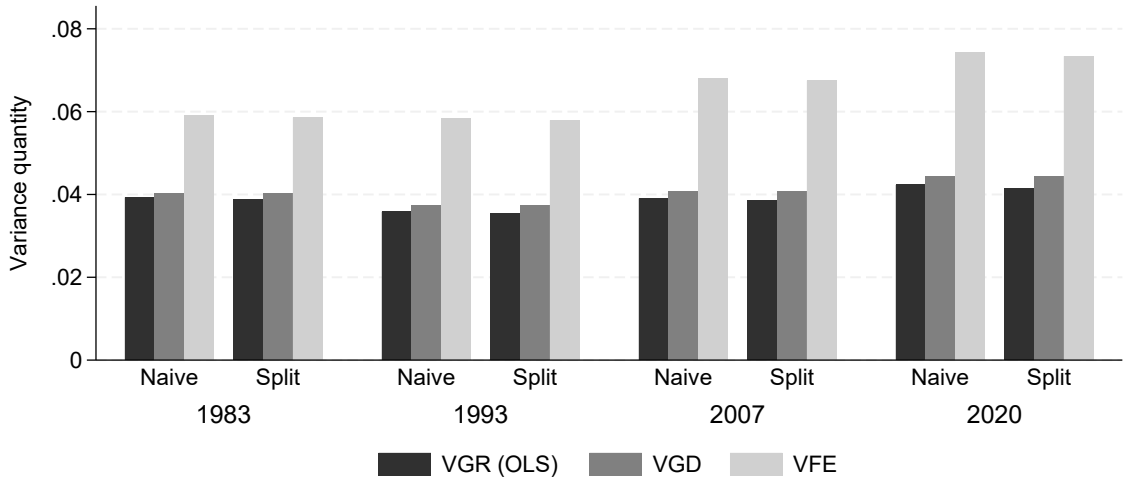


Figure 6: Comparison between naive plug-in and split-sample estimates of between-group variance terms.

the VRE may be considerably different in magnitude. For completeness, I also fit a mixed-effects model of earnings, using occupation random intercepts. I present the results in 7. The VRE is considerably higher than the VGR in all years.

Finally, since the between-group variance quantities are defined nonparametrically, they can be estimated using methods other than OLS. I demonstrate this with both the VGR and VGD. For the VGR, I model $\mathbb{E}[Y \mid X]$ using Lasso regression allowing for interactions between all covariates in X; I then use the resulting predictions to estimate the VGR using the mean-imputation procedure described in Section 3.1. For the VGD, I fit a logistic regression for each occupation, modeling the probability of being in that occupation given the covariates. I use this to calculate the residuals $\tilde{1}_{\text{occ}_i=\text{occ}} = 1_{\text{occ}_i=\text{occ}} - \hat{\Pr}(\text{occ}_i = \text{occ} \mid x_i)$, effectively partialing out the share of observations at x_i that are in occupation 'occ'. I then recover the variance-weighted τ_{occ} from the coefficient on $\tilde{1}_{\text{occ}_i=\text{occ}}$ in the regression

$$y_i = \tau_g^{\text{logit}} \tilde{1}_{\text{occ}_i=\text{occ}} + v_i.$$

Figures 8 and 9 show the results for the VGR and VGD, respectively. These are very close to the results from the main text (Fig 4b), indicating that the choice of modeling does not affect the results so much as the choice of estimands.

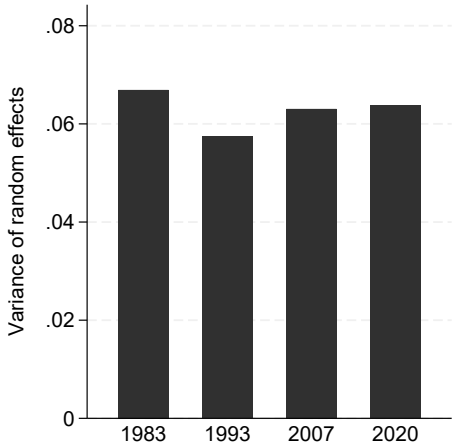


Figure 7: Between-group variance from a mixed-effects model with group random effects/intercepts. Model is estimated by maximum likelihood.

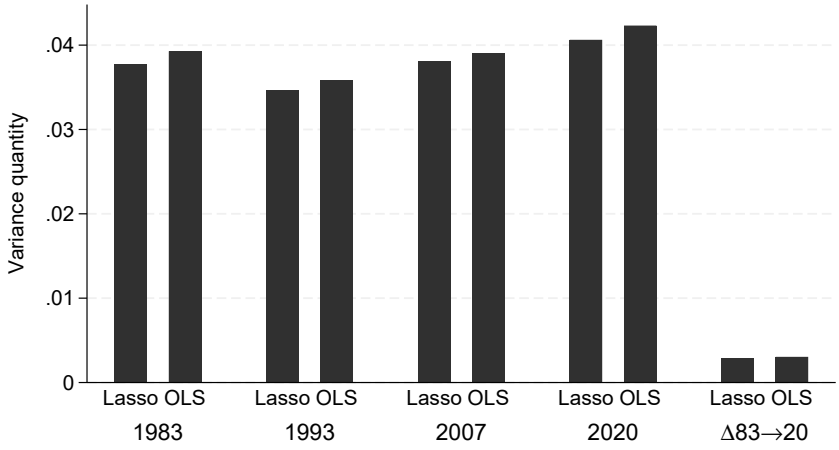


Figure 8: Comparison between VGR based on two strategies to estimate $\mathbb{E}[Y \mid X]$: pooled OLS – presented in the main text – and Lasso.

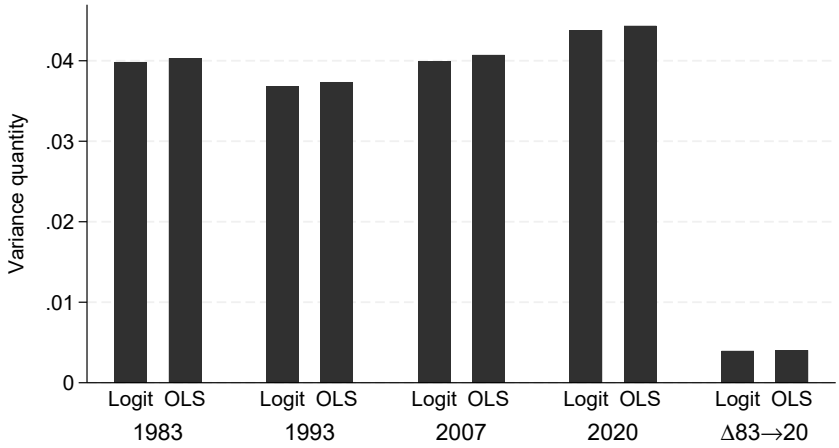


Figure 9: Comparison between VGD based on two strategies to estimate the group effects (τ): OLS, as presented in the main text, and logistic regression.