Winkelmann, Rainer

Working Paper

Neglected heterogeneity and the algebra of least squares

Working Paper, No. 426

Provided in Cooperation with:
Department of Economics, University of Zurich

Suggested Citation: Winkelmann, Rainer (2023) : Neglected heterogeneity and the algebra of least squares, Working Paper, No. 426, University of Zurich, Department of Economics, Zurich, https://doi.org/10.5167/uzh-229123

This Version is available at:
https://hdl.handle.net/10419/268860

University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 426

Neglected Heterogeneity and the Algebra of Least Squares

Rainer Winkelmann

January 2023

Neglected heterogeneity and the algebra of least squares*

Rainer Winkelmann
University of Zurich

January 2023

Abstract

This paper explores an algebraic relationship between two types of coefficients for a regression with several predictors and an additive binary group variable. In a general regression, the regression coefficients are allowed to be group-specific, the restricted regression imposes constant coefficients. The key result is that the restricted coefficients imposing homogeneity are not necessarily a convex average of the unrestricted coefficients obtained from the more general regression. In the context of treatment effect estimation with several treatment arms and group-level controls, this means that the estimated effect of a specific treatment can be non-zero, and statistically significant, even if the estimated unrestricted effects are zero in each group.

Keywords: Ordinary least squares, subsample heterogeneity, variance-weighting, average treatment effect.

JEL classification: C21

* Department of Economics, University of Zurich, E-mail: rainer.winkelmann@econ.uzh.ch. I am grateful to Joshua Angrist, Joao Santos Silva and Julius Schäper for helpful comments on an earlier draft of the paper.

1 Introduction

The computation of regression coefficients by ordinary least squares (OLS) has the elementary algebraic property that if the data are split into two subsets

$$y = (y_0, y_1) \text{ and } X = (X_0, X_1),$$

then

$$\begin{aligned} b &= (X'X)^{-1}X'y \\ &= (X_0'X_0 + X_1'X_1)^{-1}(X_0'\,y_0 + X_1'\,y_1) \\ &= (H^0 + H^1)^{-1}H^0b^0 + (H^0 + H^1)^{-1}H^1b^1 \end{aligned}$$

where $H^0 = X_0'X_0$, $H^1 = X_1'X_1$, $b^0 = (H^0)^{-1}X_0'\,y_0$ and $b^1 = (H^1)^{-1}X_1'\,y_1$.

Hence, the overall vector of regression coefficients is a X-variance-weighted average of subset specific regression coefficients. This is a mechanical property of OLS. Whether or not it is a "desirable" feature depends on context. For instance, when considering Bayesian updating for the normal linear model with homoskedastic errors, the property says that the posterior mean b is obtained as a precision-matrix weighted average of the prior location vector b^0 and the sample location vector b^1 (e.g. Zellner, 1973), which seems like an intuitively reasonable way of combining information.[1]

In other contexts, such variance-weighting can lead to unexpected, and perhaps undesirable, results. One such case is neglected slope heterogeneity in linear regression models. Suppose for illustrative purposes that there are two groups only. b^0 denotes the coefficients when a regression is fit to group-0 data, whereas b^1 are the group-1 coefficients. In general, $b^0 \neq b^1$. We know from above that pooling over both groups yields an overall coefficient vector that is a variance-weighted average of the group specific coefficients. But a matrix-weighted average of vectors does not mean that element by element, the pooled coefficients lie algebraically between b^0 and b^1, except for very specific circumstances. For example, element by element convex averages are obtained if both weighting matrices are diagonal, or if one matrix is proportional to the other (See Chamberlain and Leamer, 1976). In general, however, each coefficient is a mixture of all group-specific coefficients, with weights that can be negative. I demonstrate this property of least squares by deriving the exact averaging weights in a regression where there are two predictors and two groups.

[1] Chamberlain and Leamer (1976) provide a general characterization of matrix weighted averages of pairs of vectors in the context of a Bayesian analysis of the linear regression model, deriving bounds for the posterior location parameter when the prior precision matrix is unknown, and thus a kind of sensitivity analysis.

The problem addressed in this paper is related to, but different from a sizeable literature on heterogeneous partial effects and the question what linear constant-coefficients regression estimates in such a case. For example, Stoker (1986) shows that the regression slope coincides with the average partial effect for a general class of non-linear conditional expectation functions as long as the regressor is normally distributed. Angrist (1998) considers a scalar binary treatment and a single discrete confounder, where OLS gives a variance-weighted average (see also Yitzhaki, 1996). Apart from least squares regression, numerous covariate adjustments methods exist that actually recover the average treatment effect under the conditional independence assumption (see Wooldridge and Imbens, 2009, for a survey of such methods).

The extension to multiple treatment arms has been considered by Goldsmith-Pinkham et al. (2021). These authors point to the close connection with the recent literature on heterogenous effects in difference-in-differences regressions (e.g., Goodman-Bacon, 2021). For continuous, and potentially multiple, regressors of interest, methods for consistent estimation of average partial effects are discussed in Wooldridge (2004) and Graham and Pinto (2022), among others.

While most of this literature is model-based (Goodman-Bacon, 2021, being a notable exception), the current paper solely exploits algebraic properties of ordinary least squares. Thus, results hold regardless of model assumptions, which may or may not be valid. Also, they hold for purely descriptive regressions, and for any sample size as they do not rely on asymptotic properties. On the other hand, this is not a framework to address questions of causality, population estimands and efficiency.

Understanding the algebra of least squares can be helpful for applied research, as the set-up discussed in this paper is encountered quite frequently in practice, both in the context of causal analysis or that of descriptive regression. An early example is Griliches (1977) who considers a regression of earnings on years of schooling and experience as main predictors, with a person's IQ as an additional control. Schooling and experience coefficients are not allowed to vary by IQ. As another example from education research, the Project STAR trial randomized students within, but not across, schools to either a small classroom treatment, a teaching aide treatment, or a control condition. The results in this paper speak to the consequences of regressing the outcome on treatment status and a school dummy, when there are heterogeneous, school specific treatment effects (this example is borrowed from Goldsmith-Pinkham et al. 2022).

2 Linear regression with group-specific heterogeneity

2.1 A single regressor

Let X denote the single regressor and $W \in \{0, 1\}$ denote the binary group variable. The regression

$$Y = b_0 + b_x X + b_w W + e \tag{1}$$

can then be solved for b_x by first partialling out W. The auxiliary regression of X on W gives predicted values

$$\bar{X}_0 = \frac{1}{N_0}\sum_{i=1}^{N}(1 - W_i)X_i, \qquad N_0 = \sum_{i=1}^{N}(1 - W_i)$$

$$\bar{X}_1 = \frac{1}{N_1}\sum_{i=1}^{N} W_i X_i, \qquad N_1 = \sum_{i=1}^{N} W_i$$

and therefore, we can obtain b_x from the bivariate regression of y on $X - \bar{X}_w$. Using the partitioning of y and X as described in the introduction, we get the scalar least squares expression

$$\begin{aligned} b_x &= \frac{(X_0 - \iota_{N_0}\bar{X}_0)'(X_0 - \iota_{N_0}\bar{X}_0)\, b_x^0 + (X_1 - \iota_{N_1}\bar{X}_1)'(X_1 - \iota_{N_1}\bar{X}_1)\, b_x^1}{(X_0 - \iota_{N_0}\bar{X}_0)'(X_0 - \iota_{N_0}\bar{X}_0) + (X_1 - \iota_{N_1}\bar{X}_1)'(X_1 - \iota_{N_1}\bar{X}_1)} \\ &= \frac{S_{xx}^0}{S_{xx}^0 + S_{xx}^1} b_x^0 + \frac{S_{xx}^1}{S_{xx}^0 + S_{xx}^1} b_x^1 \end{aligned}$$

where ι_N is a $(N \times 1)$ vector of ones, $S_{xx}^0 = \sum_{i=1}^{N_0}(X_i - \bar{X}_0)^2 = N_0\, \hat{\sigma}^2_{x,w=0}$ and $S_{xx}^1 = \sum_{i=1}^{N_1}(X_i - \bar{X}_1)^2 = N_1\, \hat{\sigma}^2_{x,w=1}$. b_x^0 is the least squares coefficient in a separate group-0 regression (of y_0 on X_0), b_x^1 that in a separate group-1 regression. Hence, the pooled coefficient b_x is a convex average: it lies between b_x^0 and b_x^1. The weights depend on relative group sizes, N_0 and N_1, as well as on the within-group variances, $\hat{\sigma}^2_{x,w=0}$ and $\hat{\sigma}^2_{x,w=1}$. A population version of this result is given in Angrist (1998).[2] As I will show, it unfortunately does not generalize when there are two or more regressors.

2.2 Extension to two regressors

With two regressors of interest, from now on labeled X and Z, and one group indicator $W \in \{0, 1\}$, the regression takes the form:

$$Y_i = b_0 + b_x X_i + b_z Z_i + b_w W_i + e_i \qquad \text{for } i = 1, \ldots, N \tag{2}$$

[2] To be precise, this is Angrist (1998) in reverse, as he considers the regression coefficient of a dummy predictor after partialling out a multivalued, discrete or continuous, confounder.

where e_i is a regression residual such that $Cov(e, X) = Cov(e, Z) = Cov(e, W) = 0$. The regressors X and Z can be binary, discrete, or continuous, and it is assumed that there are two groups only. Regression (2) allows the constant to shift depending on W_i but imposes homogeneous slopes b_x and b_z. After partialling out, as before, the constant and W_i, we obtain the trivariate regression

$$Y_i = b_x(X_i - \bar{X}_w) + b_z(Z_i - \bar{Z}_w) + u_i$$

where $w \in \{0, 1\}$, groups do not necessarily need to be of equal size, and $\bar{X}_0$, $\bar{X}_1$, $\bar{Z}_0$ and $\bar{Z}_1$ are group-specific means. Assuming sorted data ($W = 0$ observations first, followed by those for $W = 1$), the $((N_0 + N_1) \times 2)$ matrix of regressors after partialling out can written as

$$\begin{pmatrix} X_0 - \iota_{N_0}\bar{X}_0 & Z_0 - \iota_{N_0}\bar{Z}_0 \\ X_1 - \iota_{N_1}\bar{X}_1 & Z_1 - \iota_{N_1}\bar{Z}_1 \end{pmatrix}$$

Define

$$S^0_{xx} = \sum_{i=1}^{N_0}(X_i - \bar{X}_0)^2 \qquad S^0_{zz} = \sum_{i=1}^{N_0}(Z_i - \bar{Z}_0)^2$$

$$S^0_{xy} = \sum_{i=1}^{N_0}(X_i - \bar{X}_0)y_i \qquad S^0_{zy} = \sum_{i=1}^{N_0}(Z_i - \bar{Z}_0)y_i$$

$$S^0_{zx} = \sum_{i=1}^{N_0}(Z_i - \bar{Z}_0)(X_i - \bar{X}_0)$$

and same for S^1_{xx}, S^1_{zz}, etc. Then the least squares coefficients $b = (b_x, b_z)'$ in (2) are obtained as

$$b = \begin{pmatrix} S^0_{xx} + S^1_{xx} & S^0_{xz} + S^1_{xz} \\ S^0_{zx} + S^1_{zx} & S^0_{zz} + S^1_{zz} \end{pmatrix}^{-1} \begin{pmatrix} S^0_{xy} + S^1_{xy} \\ S^0_{zy} + S^1_{zy} \end{pmatrix} \tag{3}$$

Alternatively, consider the heterogeneous coefficients as computed from two separate regressions of y on X and Z, one using the $W = 0$ observations only, say b^0, and the other one using $W = 1$ observations only, say b^1

$$b^w = \begin{pmatrix} S^w_{xx} & S^w_{xz} \\ S^w_{zx} & S^w_{zz} \end{pmatrix}^{-1} \begin{pmatrix} S^w_{xy} \\ S^w_{zy} \end{pmatrix} \qquad \text{for } w \in \{0, 1\} \tag{4}$$

Comparing (3) and (4), it is clear that the algebraic variance weighting property is satisfied:

$$b = (H^0 + H^1)^{-1}(H^0\, b^0 + H^1\, b^1) \tag{5}$$

where

$$H^w = \begin{pmatrix} S^w_{xx} & S^w_{xz} \\ S^w_{zx} & S^w_{zz} \end{pmatrix} \text{ for } w \in \{0, 1\}$$

is proportional to the within-group variances and covariances of the two regressors. In the following, I will derive the element by element relationship between b_x and the heterogeneous coefficients b_x^0, b_x^1, b_z^0 and b_z^1.

2.3 Decomposing the regression coefficient b_x

With two regressors, computation of least squares coefficients requires the inversion of the (2×2) variance-covariance matrix. Without loss of generality, I will focus on the first element of b, pertaining to X and denoted as b_x^0 and b_x^1 when considering heterogeneity, and as b_x for the pooled regression. The results for b_z follow from symmetry. For the subsamples (conditional on W being either 0 or 1), the coefficients are given by

$$b_x^w = \frac{S_{zz}^w S_{xy}^w - S_{xz}^w S_{zy}^w}{S_{xx}^w S_{zz}^w - S_{xz}^w S_{xz}^w} \qquad w \in \{0,1\} \tag{6}$$

For estimation of a common b_x without heterogeneity, we simply need to replace each term of the right-hand side fraction in (6) with its respective sum over the two groups (see the definition of $(b_x, b_z)'$ in equation (3)), such that

$$b_x = \frac{(S_{zz}^0 + S_{zz}^1)(S_{xy}^0 + S_{xy}^1) - (S_{xz}^0 + S_{xz}^1)(S_{zy}^0 + S_{zy}^1)}{(S_{xx}^0 + S_{xx}^1)(S_{zz}^0 + S_{zz}^1) - (S_{xz}^0 + S_{xz}^1)^2} \tag{7}$$

In the Appendix, I show how to express b_x as a relatively simple function of four group-specific coeffients b_x^0, b_x^1, b_z^0 and b_z^1. In particular, the numerator of (7) can be written as

$$(S_{xx}^0 S_{zz}^0 - S_{xz}^0 S_{xz}^0 + S_{zz}^1 S_{xx}^0 - S_{xz}^1 S_{xz}^0) b_x^0 + (S_{xx}^1 S_{zz}^1 - S_{xz}^1 S_{xz}^1 + S_{zz}^0 S_{xx}^1 - S_{xz}^0 S_{xz}^1) b_x^1$$

$$+ (S_{zz}^1 S_{xz}^0 - S_{xz}^1 S_{zz}^0) b_z^0 + (S_{zz}^0 S_{xz}^1 - S_{xz}^0 S_{zz}^1) b_z^1$$

To obtain the aggregate b_x coefficient, we need to divide this numerator by the original denominator $(S_{xx}^0 + S_{xx}^1)(S_{zz}^0 + S_{zz}^1) - (S_{xz}^0 + S_{xz}^1)^2$. Comparing coefficients, we can see that

$$b_x = \frac{\mathcal{A} b_x^0 + \mathcal{B} b_x^1 + \mathcal{C}(b_z^1 - b_z^0)}{\mathcal{A} + \mathcal{B}} \tag{8}$$

where

$$\mathcal{A} = S_{xx}^0 S_{zz}^0 - S_{xz}^0 S_{xz}^0 + S_{zz}^1 S_{xx}^0 - S_{xz}^1 S_{xz}^0$$

$$\mathcal{B} = S_{xx}^1 S_{zz}^1 - S_{xz}^1 S_{xz}^1 + S_{zz}^0 S_{xx}^1 - S_{xz}^0 S_{xz}^1$$

and

$$\mathcal{C} = S^0_{zz}S^1_{xz} - S^0_{xz}S^1_{zz} = S^0_{zz}S^1_{zz}(g_1 - g_0)$$

where g_1 and g_0 are the slopes of a regression of X on Z in group 1 and group 0, respectively. It follows that $C(b^1_z - b^0_z)$ can be an arbitrarily large positive or negative number.

To shed light on the interpretation of weights $\mathcal{A}$ and $\mathcal{B}$, consider the auxiliary regression of X on Z and W for the pooled data. The slope is equal to S_{xz}/S_{zz}, and the covariance between residuals $u_i = (X_i - \bar{X}_w) - S_{xz}/S_{zz}(Z_i - \bar{Z}_w)$ and X_i for group 0 is equal to

$$\begin{aligned} Cov_0(u_i, X_i) &= \frac{1}{N_0}\sum_{i=1}^{N_0}\left((X_i - \bar{X}_0) - \frac{S_{xz}}{S_{zz}}(Z_i - \bar{Z}_0)\right)X_i \\ &= \frac{S^0_{xx}(S^0_{zz} + S^1_{zz}) - (S^0_{xz} + S^1_{xz})S^0_{xz}}{N_0(S^0_{zz} + S^1_{zz})} = \frac{\mathcal{A}}{N_0 S_{zz}} \end{aligned}$$

In this case, $\mathcal{A}$ is proportional to a group-specific covariance, rather than to a group-specific variance. While $\hat{X}_i$ is by definition orthogonal to u_i in the full sample, and hence $Cov(u_i, X_i) = Cov(u_i, X_i - \hat{X}_i) = Var(u_i)$, this equivalence does not hold in the subset. Subset-orthogonality of u_i and $\hat{X}_i$ fails, because the auxiliary regression omits the interaction between Z and W and is thus not "saturated" in W.[3] Adding the interaction is equivalent to separate group-wise regressions of X on Z, which restores orthogonality of u_i and $\hat{X}_i$ in each group. But this would not correspond to the regression equation (2) we are considering.

Since $\mathcal{A}$ is proportional to a covariance, it can be positive or negative. On the other hand, $\mathcal{A} + \mathcal{B}$, the denominator in equation (7), is always positive since it equals the determinant of a positive definite matrix. Hence the least-squares weights for b^0_x and b^1_x, $\mathcal{A}/(\mathcal{A} + \mathcal{B})$ and $\mathcal{B}/(\mathcal{A} + \mathcal{B})$, can be negative or greater than unity.

The decomposition result (8) is quite remarkable. Clearly, b_x is not a convex combination of b^0_x and b^1_x in general. By symmetry, the same argument holds for b_z. The reasons for the non-convexity are twofold. First, weights can be negative even in the absence of the $\mathcal{C}$-term, depending on the within-group covariance between X from a regression of X on Z and W. On top of that, the aggregate X-coefficient in the regression of Y depends directly on the heterogeneous coefficients of the *other* regressor Z. Thus, there can be a spill-over of neglected heterogeneity, or a "contamination" as

[3] Angrist (1998) also noted that the variance-weighting result for the case of a single predictor requires an auxiliary regression saturated in the confounder. By definition, a binary confounder as in (1) satisfies this requirement.

described recently by Goldsmith-Pinkham et al. (2022). Even if both sub-sample coefficients of X are zero, the estimated overall effect b_x can be non-zero, because the Z-coefficients matter as well.

2.4 Determinants of contamination "bias"

We can write

$$\text{"bias"} = \frac{(b_z^1 - b_z^0)(S_{zz}^0 S_{xz}^1 - S_{zz}^1 S_{xz}^0)}{(S_{xx}^0 + S_{xx}^1)(S_{zz}^0 + S_{zz}^1) - (S_{xz}^0 + S_{xz}^1)^2} = \frac{\mathcal{C}(b_z^1 - b_z^0)}{\mathcal{A} + \mathcal{B}}$$

Intuitively, it makes sense to talk about bias here, although in slight abuse of language. No population model is defined and I merely exploit algebraic properties of regression. The key point is that this term would not exist if one were to run instead a regression of Y on X, Z, W and the interaction terms XW and ZW. The bias arises since heterogeneity in the coefficients is neglected.

For example, the thus defined bias is positive, if

$$b_z^1 > b_z^0 \qquad \text{and} \qquad S_{zz}^0 S_{xz}^1 - S_{zz}^1 S_{xz}^0 = S_{zz}^0 S_{zz}^1 (g_z^1 - g_z^0) > 0$$

where g_z^0 and g_z^1 are the group-specific bivariate regression slopes in a regression of X on Z. The heterogeneous effect of Z positively spills over into estimation of the X-effect if in the group with a large direct-Z-effect, changes in Z also predict larger, or less negative, changes in X.

We can now state conditions that make contamination through b_z disappear. For example:

- $b_z^0 = b_z^1 = 0$: In this case, the second regressor Z could be dropped, leading back to the bivariate case.
- $b_z^0 = b_z^1$: the coefficient of Z is homogenous across groups.
- $S_{xz}^0 = S_{xz}^1 = 0$. Note: in the case of mutually exclusive treatment arms (when X and Z are both dummy variables) and $Z = 0$ whenever $X = 1$, X and Z can't be uncorrelated.
- The regressors in both groups are the same, for example $S_{zz}^0 = S_{zz}^1$ etc. In this case, one can show that the overall effect of b_x is simply the arithmetic mean of b_x^0 and b_x^1 (if the two groups have equal size). So it is equal to the "average treatment effect".

Finally, note that contamination does not depend on actual confounding: it arises also if (X, W) and (Z, W) are uncorrelated but the coefficient of X and / or Z varies with W. With a single

regressor, one could simply drop W from the model and obtain an estimate of the average effect. With multiple regressors and effect heterogeneity, this is not the case.

3 Generalizations

In practice, one will rarely encounter an application with two regressors and two groups only, so the question is whether any algebraic results can be derived for more complex regression situations. In particular, what happens as the number of regressors or the number of groups is increased?

Increasing the number of groups does not change the nature of the argument. For instance, with three groups instead of two, (5) generalizes to

$$b = (H^0 + H^1 + H^2)^{-1}(H^0\, b^0 + H^1\, b^1 + H^2\, b^2)$$

where the weight matrices H^w and the subset regression coefficients b^w, $w \in \{0, 1, 2\}$ are defined as before. As a practical problem it becomes tedious to derive closed form results on the relation between b_x and the six subset coefficients (for X and Z, respectively), as the number of terms in the numerator and denominator of the b_x equation increases quadratically in the number of groups.

Similar issues arise if the number of groups is kept at two but the dimensionality of the regressor vector is increased. While the trivariate problem studied above was manageable, higher order regressions are less so. General results are given in Chamberlain and Leamer (1976). They show for example that for two arbitrary positive definite weighting matrices H^0 and H^1, and given subset coefficient vectors b^0 ad b^1, the aggregate coefficient vector b can lie essentially anywhere. Only if the weighting matrices H^0 and H^1 are of specific forms (e.g. diagonal or proportional to each other) does the matrix weighted average guarantee an element by element convex combination of the heterogeneous coefficients.

4 Numerical examples

In this section, I present three different illustrations for non-convex weighting of heterogeneous coefficients in pooled regressions. The first two are based on simulated data while the third uses a real dataset on wages and worker characteristics.

4.1 Negative residual covariance

Non-convexity due to a negative $\mathcal{A}$-weight depends entirely on the predictors and thus does not require any knowledge of the regression model for y. Consider the least squares regression of X on Z and W. A negative correlation between residuals $X - \hat{X}$ and X for $W = 0$ can be obtained by judiciously defining the subsets $W \in \{0, 1\}$.

For example, assume that $Z \sim Normal(0, 1)$, $X = \beta_1 Z + u$, where $u \sim Normal(0, \sigma^2)$. Moreover, let

$$W = \begin{cases} 0 & \text{when } X < 0 \text{ and } u \geq 0 \text{ or } X \geq 0 \text{ and } u < 0 \\ 1 & \text{else} \end{cases}$$

Under these assumptions, subset coefficients are heterogenous. For example, for $X < 0$ and $W = 0$, we obtain

$$\mathrm{E}(X|X < 0, W = 0, Z) = \beta_0 + \beta_1 Z + \sigma \mathrm{E}(u|0 < u < -\beta_1 Z))$$

for $Z < 0$. Hence, increasing Z (towards zero) reduces the mean residual in the selected subset, and the least squares coefficient will be smaller than β_1. Nevertheless, it can be shown that the least squares regression of X on Z and W recovers a variance-weighted average (for $W = 0$ and $W = 1$), that is converging to β_1.

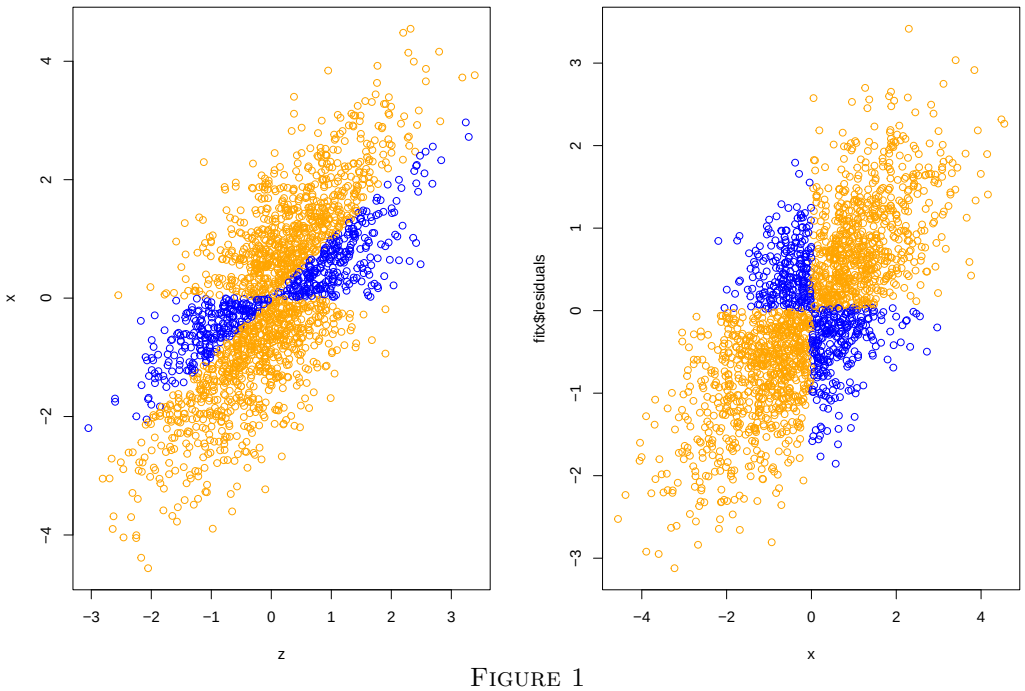

FIGURE 1

The covariance between least squares errors and X conditional on $W = 0$ is given by $\text{Cov}(\text{u}, \text{X}|\text{W} = 0) = \text{E}(\text{uX}|\text{W} = 0) - \text{E}(\text{u}|\text{W} = 0)\text{E}(\text{X}|\text{W} = 0)$. Because of symmetry of the multivariate normal, in this case $\text{E}(u|W = 0) = \text{E}(X|W = 0) = 0$, and $\text{Cov}(\text{uX}|\text{W} = 0) = \text{E}(\text{uX}|\text{W} = 0) < 0$, because there are two cases to consider whenever $W = 0$, and both of them lead to negative products uX. In a similar way, following from the definition of $W = 1$, $\text{E}(uX|W = 1)$ only involves positive products and hence is positive overall.

The right panel of Figure 1 illustrates the covariance between residuals and X for $\beta_1 = \sigma = 1$ for a sample of 1000 random draws from the joint (bivariate normal) distribution of X and Z. It is negative for $W = 0$ (blue) and positive for $W = 1$ (orange).

In this case, the weights are $\mathcal{A}/(\mathcal{A} + \mathcal{B}) = -0.064$ and $\mathcal{B}/(\mathcal{A} + \mathcal{B}) = 1.064$. If in the outcome equation, the X-coefficient is zero in group 0 and 1 in group 1, the estimated average effect over both groups will be 1.064, and thus larger than any of the two sub-coefficients.

4.2 Contamination bias

The following numerical example illustrates the potential for contamination bias. In order to generate data, I need to take a stance regarding the full data generating process: there are two normally distributed regressors that are uncorrelated in group 0 but correlated in group 1. The group level binary confounder shifts the intercept of the regression as well as the slopes. The two groups are of equal size of 100, so 200 data points are generated as follows:

Z_0, Z_1, and X_0 are i.i.d standard normal random variables. In group 1, X_1 and Z_1 are related by a regression model

$$X_1 = g_1 Z_1 + \text{ rnorm}(100)$$

Since $g_0 = 0$, i.e. X_0 and Z_0 are uncorrelated, $g_1 = g_1 - g_0$ indicates the differential "response" of X to Z in group 1 relative to group 0. One goal of the simulations is to show the sensitivity of results to changes in g_1, which increases stepwise from -1 to $+1$. By construction, an increase in the absolute value of g_1 also affects the group-1 variance of X, since $\text{Var}(X_1) = g_1^2 + 1$.

Outcome data are generated as

$$y_0 = 1 \qquad \text{and} \qquad y_1 = 2 + 1 \times X_1 + 4 \times Z_1$$

For the fully interacted regression, this means that $b_w = 1$, $b_x^0 = 0$, $b_z^0 = 0$ and $b_x^1 = 1$, $b_z^1 = 4$.

In the full model, there is no estimation uncertainty, as the outcome equation does not have an additional error term.

The regression of Y on X, Z, and W has sampling variation, however, due to the mis-specification and random draws of the regressors. To account for the effect of sampling variation, the estimation is repeated for 1000 different samples.

Figure 2 shows the means and error bands ($\pm$ 2 standard deviations obtained from the repeated samples), for four types of statistics: the aggregate estimate b_x (in green), together with the weights $\mathcal{A}/(\mathcal{A}+\mathcal{B})$ (in black), $\mathcal{B}/(\mathcal{A}+\mathcal{B})$ (in light blue) and $\mathcal{C}/(\mathcal{A}+\mathcal{B})$ (in red) associated with b_x^0, b_x^1 and b_z^1, respectively, for varying degrees of departure from zero-correlation between X and Z in group 1.

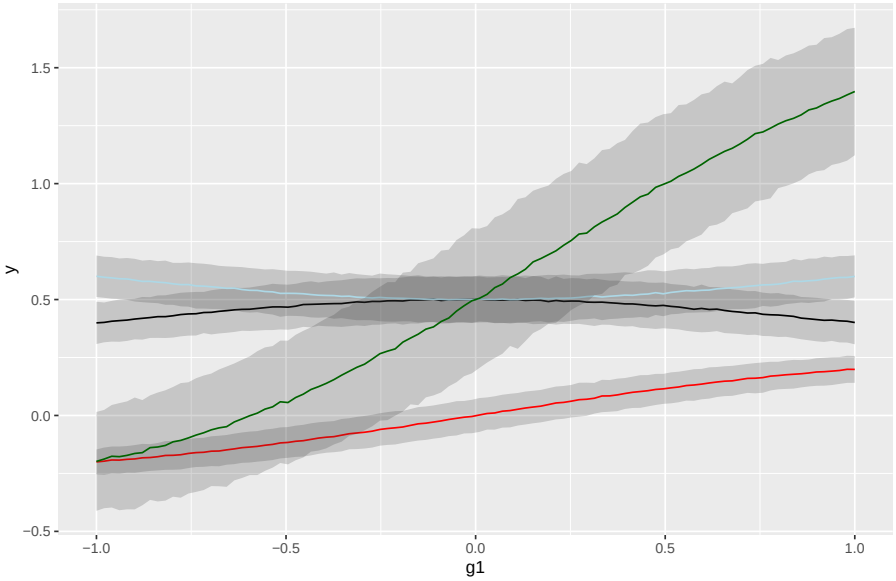

FIGURE 2

We know that the true heterogenous effects of X on Y are 0 in group 0 and 1 in group 1. The groups are of equal size and hence, in the absence of contamination, the non-interacted coefficient would be equal to 0.5 as long as the within-group X-covariances are identical. This is the case in the present DGP only if $g_1 = 0$, because otherwise, the conditional covariance of X_1 is larger than that of X_0 and hence the weight $\mathcal{B}/(\mathcal{A}+\mathcal{B})$ moves quadratically away from 0.5.

However, this effect alone would not lead to substantial changes in the obtained average coefficient. In the numerical example, the first order issue is contamination with the large group-1 specific effect of Z (in combination with no effect in group 0), that can enter the estimation of b_x with

positive or negative weight, depending on the sign of g_1. Overall, the estimated b_x coefficient can be negative or greater than one, thus falling outside of the actual effects sizes of b_x^0 and b_x^1 and illustrating the issue of non-convexity in this set-up.

4.3 Wages of men and women

In this example I use a textbook dataset on wages and worker's characteristics from Wooldridge (2012), an extract from the 1976 Current Population Survey.[4] The dependent variable is the logarithm of average hourly earnings. Explanatory variables include years of education, years of potential experience and its square, years with current employer and its square, number of dependents, indicators for being nonwhite, married, living in a metropolitan area, as well as three regional and nine industry dummies. Thus, there is a total of 21 regressors and the dataset provides 526 observations. As group variable of interest, I consider here the gender of the worker. This choice is of course somewhat arbitrary, but wage related regression analyses that do not stratify by gender have been conducted in the literature (e.g. Oreopolous, 2006).

Table 1: Wages of U.S. workers

Dependent variable: *logarithmic hourly wage*				
	b	b^{men}	b^{women}	weight
years of education	0.047	0.052	0.043	0.41
experience	0.025	0.032	0.020	0.44
experience squared	-0.001	-0.001	0.000	0.48
tenure	0.022	0.025	0.024	-4.82
tenure squared	0.000	0.000	-0.001	1.06
nonwhite	-0.004	0.051	-0.093	0.62
married	0.056	0.159	-0.054	0.52
number of dependents	-0.022	-0.032	-0.022	-0.09
lives in SMSA	0.139	0.142	0.101	0.91
lives in north central U.S	-0.058	-0.118	-0.023	0.37
lives in southern region	-0.044	-0.112	0.013	0.46
lives in western region	0.055	0.018	0.067	0.26
construc. indus.	-0.053	0.026	-0.081	0.26
nondur. manuf. indus.	-0.107	-0.060	-0.109	0.03
trans, commun, pub ut	-0.096	-0.073	-0.142	0.67
trade (wholesale or retail)	-0.303	-0.271	-0.271	-771.9
services indus.	-0.309	-0.255	-0.236	3.83
prof. serv. indus.	-0.095	-0.172	0.013	0.58
profess. occupation	0.225	0.215	0.193	1.44
clerical occupation	0.038	0.115	0.022	0.18
service occupation	-0.094	-0.087	-0.149	0.88
female	-0.268			

[4]The data file "wage1" can be obtained from the R-repository, package name "wooldridge": https://cran.r-project.org/web/packages/wooldridge/wooldridge.pdf.

To obtain heterogeneous effects, the sample is split and two regressions are conducted, one using the subset of 274 men and one the using the subset of 252 women. Results are shown in Table 1. For 6 out of 21 coefficients, the aggregation weights are non-convex. This is seen in the last column of Table 1, where the equation $b_x = \alpha \times b_x^{men} + (1-\alpha) \times b_x^{women}$ is solved for α. In three instances, this weight is negative (for tenure, for the number of dependents and for the wholesale dummy). In another three instances, it is greater than one (for tenure squared, services and professional occupations). This can lead to quite counterintuitive results. For instance, if one were to use these results to rank industries by their wage differentials, some reversals would occur. For example, for both genders, services pays higher wages than trade, ceteris paribus. Yet, in the aggregate, trade wages are estimated to lie above those of service workers.

Note that this crude assessment of non-convex weighting cannot discriminate between the two sources of non-convexity. So we do not know whether it is primarily due to contamination, or to covariance-weighting of own heterogeneous coefficient contrasts, or both. Since the formulae derived in this paper only dealt with the two-regressor-case, and not with a high-dimensional regressor vector as presently, such a decomposition is not feasible.

5 Discussion

There are substantial perils of ignoring group-level heterogeneity in the context of a regression with multiple regressors: if one wants to estimate the effect of multiple regressors, adding a group dummy to allow for shifts in the constant can lead to counterintuitive results, as the estimated coefficients are not necessarily convex combinations of the group-level coefficients.

The non-convexity result follows directly from regression algebra: subgroup regression coefficients are aggregated using matrix level variance-covariance weighting, but this does not imply element by element convex aggregation. This result mirrors recent findings by Goldsmith-Pinkham et al. (2022), although their set-up is different: They consider a partially linear model with a set of mutually exclusive (binary) treatment variables and an additive function of a continuous confounder. Moreover, they provide population-level identification results and do not exploit the regression algebra as I do here.

Both approaches provide the same key insight: with several regressors of interest, non-convex weighting can arise due to two reasons. The first one is technical, since residuals in the partialling

out equation cannot be mean independent, implying covariance-weighted averaging; the second reason is substantive, as coefficients in general suffer from spill-overs, or contamination, from heterogeneous coefficients of *other* regressors.

As a remedy to these problems, one should better conduct group-wise (i.e. fully interacted) regressions, from where one can obtain "average coefficients", for example by weighting the heterogeneous coefficients by the relative group sizes.

An application to estimating a wage regression illustrated that non-convex weights arise quite commonly in practice. When enforcing common coefficients on 21 regressors, rather than letting them vary by gender, six out of these 21 estimates did not lie between the female and the male estimates. This is not a "paradox" but rather a fluke of regression algebra in connection with neglected heterogeneity.

6 References

Angrist, J.D. (1998) Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants, *Econometrica* 66, 249-288.

Chamberlain, G. and E.E. Leamer (1976) Matrix Weighted Averages and Posterior Bounds, *Journal of the Royal Statistical Society B*, 38, 73-84.

Goldsmith-Pinkham, P., P. Hull and M. Kolesar (2022) Contamination Bias in Linear Regressions, Working Paper, https://arxiv.org/abs/2106.05024

Goodman-Bacon, A. (2021) Difference-in-differences with variation in treatment timing, *Journal of Econometrics* 225, 254-277.

Graham, B.S., and C.C. Pinto (2022) Semiparametrically efficient estimation of the average linear regression function, *Journal of Econometrics* 226 (1), 115-138.

Griliches, Z. (1977) Estimating the Returns to Schooling: Some Econometric Problems, *Econometrica* 45, 1-22.

Imbens, G.W. and J.M. Wooldridge (2009) Recent Developments in the Econometrics of Program Evaluation, *Journal of Economic Literature*, 47, 5-86.

Oreopoulos, P. (2006) Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter, *American Economic Review* 96: 152-175.

Stoker, T.M. (1986) Consistent Estimation of Scaled Coefficients, *Econometrica* 54, 1461-1481.

Wooldridge, J.M. (2004) Estimating average partial effects under conditional moment independence assumptions, CeMMAP working papers CWP03/04.

Wooldridge, J.M. (2012) *Introductory Econometrics: A Modern Approach*, 5th edition.

Yitzhaki, S. (1996) On using linear regressions in welfare economics, *Journal of Business & Economic Statistics*, 14(4), 478-486.

Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics*, Wiley.

Appendix

Deriving equation (8)

All notation is as defined in the main text. It holds that b_x is defined by the following fraction.

$$b_x = \frac{(S^0_{zz} + S^1_{zz})(S^0_{xy} + S^1_{xy}) - (S^0_{xz} + S^1_{xz})(S^0_{zy} + S^1_{zy})}{(S^0_{xx} + S^1_{xx})(S^0_{zz} + S^1_{zz}) - (S^0_{xz} + S^1_{xz})^2} \tag{9}$$

The numerator can be re-written by multiplying out and re-ordering terms first into those involving group 0 and group 1 only, followed by all mixed terms:

$$S^0_{zz}S^0_{xy} - S^0_{xz}S^0_{zy} + S^1_{zz}S^1_{xy} - S^1_{xz}S^1_{zy} + S^0_{zz}S^1_{xy} + S^1_{zz}S^0_{xy} - S^0_{xz}S^1_{zy} - S^1_{xz}S^0_{zy}$$

$$= (S^0_{xx}S^0_{zz} - S^0_{xz}S^0_{xz})b^0_x + (S^1_{xx}S^1_{zz} - S^1_{xz}S^1_{xz})b^1_x + S^0_{zz}S^1_{xy} + S^1_{zz}S^0_{xy} - S^0_{xz}S^1_{zy} - S^1_{xz}S^0_{zy}$$

where we have substituted

$$S^w_{zz}S^w_{xy} - S^w_{xz}S^w_{zy} = (S^w_{xx}S^w_{zz} - S^w_{xz}S^w_{xz})b^w_x$$

using equation (6). Next, consider the mixed terms in the numerator of (9):

$$S^0_{zz}S^1_{xy} + S^1_{zz}S^0_{xy} - S^0_{xz}S^1_{zy} - S^1_{xz}S^0_{zy}$$

We can substitute all covariance terms involving y using short-long regression algebra. For instance

$$S^1_{xy} = S^1_{xx}(b^1_x + S^1_{xz}/S^1_{xx}b^1_z)$$

where the term in parentheses is the coefficient of the bivariate subset regression of y_1 on X_1 expressed in terms of the direct effect of X_1 plus the effect of X_1 on Z_1 times the direct effect of Z_1 in the trivariate regression. Hence, for instance,

$$S^0_{zz}S^1_{xy} = S^0_{zz}S^1_{xx}b^1_x + S^0_{zz}S^1_{xz}b^1_z$$

etc.; In conclusion, the mixed terms can be written as

$$S^0_{zz}S^1_{xx}b^1_x + S^0_{zz}S^1_{xz}b^1_z + S^1_{zz}S^0_{xx}b^0_x + S^1_{zz}S^0_{xz}b^0_z - S^0_{xz}S^1_{zz}b^1_z - S^0_{xz}S^1_{xz}b^1_x - S^1_{xz}S^0_{zz}b^0_z - S^1_{xz}S^0_{xz}b^0_x$$

Putting things back into (9) and collecting terms, we can write the numerator as an explicit function of the four group-specific, heterogeneous effects of X and Z:

$$(S^0_{xx}S^0_{zz} - S^0_{xz}S^0_{xz} + S^1_{zz}S^0_{xx} - S^1_{xz}S^0_{xz})b^0_x + (S^1_{xx}S^1_{zz} - S^1_{xz}S^1_{xz} + S^0_{zz}S^1_{xx} - S^0_{xz}S^1_{xz})b^1_x \qquad (10)$$

$$+(S^1_{zz}S^0_{xz} - S^1_{xz}S^0_{zz})b^0_z + (S^0_{zz}S^1_{xz} - S^0_{xz}S^1_{zz})b^1_z$$

To obtain the aggregate b_x coefficient, simply divide the numerator (10) by the original denominator $(S^0_{xx} + S^1_{xx})(S^0_{zz} + S^1_{zz}) - (S^0_{xz} + S^1_{xz})^2$.