Bounds on the fixed effects estimand in the presence of heterogeneous assignment propensities

Macartan Humphreys*
WZB, Humboldt University, Trinity College Dublin

February 19, 2025

Abstract

Fixed effects estimation, with linear controls for stratum membership, is often used to estimate treatment effects when assignment propensities differ across strata. In the presence of heterogeneity in treatment effects across strata, this estimator does not target the average treatment effect, however. Indeed the implied estimand can range anywhere from the lowest to the highest stratum-level average effect. To facilitate interpretation of results using this approach, I establish that if stratum-level average effects are monotonic in the shares assigned to treatment, then the fixed effects estimand lies between the average treatment effect for the treated and the average treatment effect for the controls.

*Contact: `macartan.humphreys@wzb.eu`

1 Introduction

Consider a setting in which study units belong to a collection of strata. Share p_j of units in stratum j is randomly assigned—or "as-if" randomly assigned—to receive treatment D_i and outcome y_i is measured for each unit i. In this case, if p_j varies across strata, treatment assignment is ignorable only conditional upon stratum (1).

The need to condition on observables is common in both experimental and observational research. In experimental work, it arises if researchers employ block randomization with different probabilities within blocks or if they employ multiple treatments with correlated probabilities (2). It can also arise if they are interested in spillover or network effects, where the probability of exposure to spillovers can vary across units even though the direct treatment is randomly assigned (3). In observational work, it arises, for instance, if individuals self-select into treatment on the basis of observable characteristics (4).

In such settings—if assignment propensities are known—there are multiple procedures for generating unbiased estimates of average treatment effects. Effects can be estimated within each stratum and then averaged (5: section 6.1). Unbiased estimates can also be generated using matching (6), or using treatment interactions (7), propensity weighting (3), or doubly robust approaches (8).

In practice, however, a common strategy is to use ordinary least squares (OLS) to estimate

$$y_i = \beta d_i + \gamma_{j[i]} + \epsilon_i, \tag{1}$$

where, d_i is the realized treatment assignment, β represents the effect of the treatment and $\gamma_{j[i]}$ represents the fixed effect for the stratum j to which i belongs. The key feature here is not the use of least squares but rather the fact that intercepts are used to account for stratum effects. Including intercepts for each stratum can be thought of as a flexible strategy for including observable covariates, although, critically, in this form it does not allow for effect heterogeneity across strata. The approach is common in observational studies (see (9) for examples) and has been recommended as a simple approach for experimental work also (5). One recent contribution (10) replicates eight influential economics papers to highlight how common this approach is.

If there are heterogeneous effects, however, estimates from this procedure are prone to bias (11). Less well understood is when these biases arise and how important they are likely to be, with contributions by (9), discussed below, a notable exception.

In this paper I address this interpretive challenge. I identify conditions under which the fixed effects estimand—the quantity implicitly targeted by least squares estimation of equation 1—is 'close' to causal quantities of interest.

In addition I provide a proposition that establishes that if the share of units assigned to treatment in each stratum is monotonic in stratum average treatment effects, then the fixed effects estimand is bounded by the expected average treatment effect for the controls and the expected average treatment effect for the treated.

The utility of this result depends on the plausibility of monotonicity between assignments and treatment effects.

Monotonic relations are guaranteed if there are just two strata. They may also arise however if both treatment effects and assignment propensities reflect some systematic feature of units. For instance, under Roy selection (12), units are more likely to opt into treatment if they expect benefits. Indeed experimental design might deliberately select assignment probabilities to reflect expected benefits (13). More subtle logics might also imply monotonicity. For instance, relatively popular children—with more network connections—might be more likely to be *indirectly* exposed to an antibullying treatment that has been randomly assigned to children, yet less likely to benefit from it (14). In experiments to study in-group cooperation that use random pairing between individuals, individuals from larger groups have a larger propensity to be matched with in-group partners, but larger groups might also display different levels of in-group cooperation on average (15).

Absent monotonicity, the fixed effects estimator may be shooting at an estimand very far from standard estimands of interest.

2 Setup

Let $\mathcal{N} = \{1, 2, \ldots, n\}$ denote a collection of units and $\mathcal{X} = \{X_1, X_2, \ldots, X_s\}$ a collection of strata. I let i denote an arbitrary unit in $\mathcal{N}$ and, when there is no risk of ambiguity, let j indicate an arbitrary stratum X_j. Similarly I use expressions such as $\sum_j p_j w_j$ as shorthand for $\sum_{j=1}^{s} p_{X_j} w_{X_j}$. Let n_j denote the number of units, $w_j = n_j/n$ the share of units, and $p_j \in (0,1)$ the share receiving treatment, in stratum j. I consider w_j and p_j to be known and fixed, as might arise, for instance from blocked random assignment. Let D_i denote a random variable that indicates whether unit i is assigned to treatment. Assume that within-stratum assignment to treatment is ignorable.

Employing the potential outcomes framework (1), let $Y_i(1)$ and $Y_i(0)$ denote the value on some outcome variable that unit i would take if allocated to treatment and control conditions respectively. The causal effect of the treatment on unit i is given by $\tau_i = Y_i(1) - Y_i(0)$. Letting $\mathbb{E}_j$ denote averages over the set of units in stratum j, define stratum-level average treatment effects:

$$\tau_j \equiv \mathbb{E}_j[\tau_i]. \tag{2}$$

The outcome for a given unit is a random variable given by $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$. Then under conditions described in (1), the average treatment effect for units in stratum j, τ_j, can be estimated without bias by the difference in average outcomes in treatment and control groups. Letting lower case letters denote realizations of random variables, we have:

$$\widehat{\tau}_j = \mathbb{E}_j[y_i | d_i = 1] - \mathbb{E}_j[y_i | d_i = 0] \tag{3}$$

I consider the following (sample) estimands:

$$\begin{aligned}
\tau_{ATE} &\equiv \mathbb{E}_N\left[\tau_i\right] &= \frac{\sum_j w_j \tau_j}{\sum_j w_j} && (4)\\
\tau_{ATT} &\equiv \mathbb{E}_D\left[\mathbb{E}_{\{i:d_i=1\}}\left[\tau_i\right]\right] &= \frac{\sum_j p_j w_j \tau_j}{\sum_j p_j w_j} && (5)\\
\tau_{ATC} &\equiv \mathbb{E}_D\left[\mathbb{E}_{\{i:d_i=0\}}\left[\tau_i\right]\right] &= \frac{\sum_j (1-p_j) w_j \tau_j}{\sum_j (1-p_j) w_j} && (6)
\end{aligned}$$

where $\mathbb{E}_N$ (similarly: $\mathbb{E}_{\{i:d_i=0\}}$, $\mathbb{E}_{\{i:d_i=1\}}$) averages over sets of units, and $\mathbb{E}_D$ takes expectations with respect to assignments to treatment.

Here τ_{ATE} corresponds to the average treatment effect across all units. Quantity τ_{ATT} (resp. τ_{ATC}) is the *expected* average treatment effect on the treated (resp. controls) with expectations taken over realizations of D. Each of these estimands can be thought of as weighted averages of the stratum-level treatment effects, τ_j. What differs is the weighting: τ_{ATT} (resp. τ_{ATC}) places more weight on the treatment effect of strata with high (resp. low) propensity of treatment.

Now consider an estimate of treatment effects resulting from using OLS to regress the outcome on treatment and a set of indicator variables for each of the strata. In this case, fixed effects estimation returns a weighted average of the estimates of stratum-level treatment effects.

$$\hat{\tau}_{FE} = \frac{\sum_j p_j(1-p_j)w_j\widehat{\tau}_j}{\sum_j p_j(1-p_j)w_j} \tag{7}$$

Derivations for this expression are provided in Theorem 5 in (16) and Equation 2 in (2), both using Frisch–Waugh–Lovell theorem. In addition I provide a direct proof in supplementary materials (A).

Observe that the weights in Equation 7 reflect the *variance* in treatment assignment within strata, not the share treated, within each stratum, and may be increasing or decreasing in the share treated.

The estimator is unbiased for the following estimand (see equation 9 in (11) for the two stratum case):

$$\tau_{FE} \equiv \mathbb{E}_D(\hat{\tau}_{FE}) = \frac{\sum_j p_j(1-p_j)w_j\tau_j}{\sum_j p_j(1-p_j)w_j} \tag{8}$$

Here the second equality follows from the assumption that p_j and w_j are fixed.

We can see from this that since least squares weights can take any value between 0 and 1 for any stratum, depending only on the values taken by the collection (p_j), τ_{FE} can take any value between $\min(\tau_j)$ and $\max(\tau_j)$.

Thus, as a general matter, there is no reason to expect that the least squares estimand is close to τ_{ATC}, τ_{ATT}, or τ_{ATE}, and although τ_{ATE} lies between τ_{ATC} and τ_{ATT}, there is no guarantee that τ_{FE} will.

Example 1. For a dramatic illustration, consider a case with three equal sized strata (a, b, c) in which for all units $Y_i(0) = 0$ and:

$$\begin{array}{rclll} Y_i(1) & = & 3 & \text{for all } i \in a, & p_a = \frac{1}{2} - \frac{\sqrt{3}}{4} \\ Y_i(1) & = & -3 & \text{for all } i \in b, & p_b = \frac{1}{2} \\ Y_i(1) & = & 3 & \text{for all } i \in c, & p_c = \frac{1}{2} + \frac{\sqrt{3}}{4} \end{array}$$

This case has striking symmetry in treatment and control. Half the units are in treatment and half are in control. The variation in propensities is the same in both groups. And $\tau_{ATE} = \tau_{ATT} = \tau_{ATC} = 1$. However, $\tau_{FE} = -1$. The sharp divergence of τ_{FE} from the other estimands arises from the fact that stratum b has the greatest treatment variance and so τ_b is weighted more heavily by τ_{FE} than by τ_{ATE}, τ_{ATT} and τ_{ATC}. The example highlights that there is no general guarantee that τ_{FE} is close to quantities of interest and that a rule of thumb based on shares in treatment and control can sometimes seriously mislead.

The example can also be used to illustrate a more subtle point: biases can arise even if all units have identical assignment propensities if the *shares* assigned to treatment are nevertheless heterogeneous. Consider a variation of this example induced by a "randomized saturation design" (17) in which there is a prior randomization to determine whether share p_a or share $1 - p_a$ is assigned to treatment. Similarly for stratum c. From an *ex ante* perspective, under this assignment scheme all units are assigned to treatment with probability 0.5 (assessed by combining the probability that a stratum is assigned to a given condition times the probability that a unit is assigned to treatment given the stratum assignment). However, under each stratum assignment, the shares assigned to treatment vary across strata and there is systematically varying *variation* in treatment assignment across the three strata. The result is that τ_{FE} diverges from the other estimands in the same way as in the original example even though (*ex ante*) assignment probabilities are now homogeneous.

3 Results

Inspection of Equations 4 - 5 and 7 suggests three cases in which τ_{FE} can be interpreted in terms of the other estimands. First, as is already well appreciated, τ_{FE} corresponds to τ_{ATE} when treatment effects or shares assigned to treatment are constant across strata. Second τ_{FE} corresponds to τ_{ATE} if propensity variance is constant across strata, for instance if there is some p such that for each j either $p_j = p$ or $p_j = 1 - p$. This might arise in a partial population design in which say, one third are treated in one group and two thirds are treated in another. Third, one can see that $\tau_{FE} \approx \tau_{ATT}$ for 'rare' treatments (p small) and $\tau_{FE} \approx \tau_{ATC}$ for 'common' treatments (p large).

Proposition 1 below establishes that if the shares of units assigned to treatment is monotonic in within-stratum treatment effects, then τ_{FE} lies between τ_{ATC} and τ_{ATT}.

Proposition 1. *If for all j, j', $p_j \geq p_{j'} \leftrightarrow \tau_j \geq \tau_{j'}$, or if for all j, j', $p_j \leq p_{j'} \leftrightarrow \tau_j \geq \tau_{j'}$, then $\tau_{FE} \in [\tau_{ATC}, \tau_{ATT}]$*

Proof. Consider the case in which p_j is monotonically increasing in τ_j and so $\tau_{ATT} \geq \tau_{ATC}$. The proof for the case in which p_j is monotonically decreasing in τ_j is similar.

We have:

$$\tau_{FE} \leq \tau_{ATT} \leftrightarrow \sum_j \frac{p_j(1-p_j)w_j}{\sum_j p_j(1-p_j)w_j}\tau_j \leq \sum_j \frac{p_j w_j}{\sum_j p_j w_j}\tau_j.$$

Equivalently (see supplementary materials B):

$$\sum_j \left(\frac{p_j w_j}{\sum_j p_j w_j} - \frac{p_j^2 w_j}{\sum_j p_j^2 w_j} \right)\tau_j \leq 0. \tag{9}$$

Note that the quantity in parenthesis in Equation 9 can be positive or negative. More specifically, defining $d_j \equiv \frac{p_j w_j}{\sum_j p_j w_j} - \frac{p_j^2 w_j}{\sum_j p_j^2 w_j}$ and $p^* \equiv \frac{\sum_j p_j^2 w_j}{\sum_j p_j w_j}$:

$$d_j \geq 0 \leftrightarrow p_j \leq p^*.$$

Exploiting monotonicity, let τ^* denote a value such that $\tau_j \leq \tau^* \leftrightarrow p_j \leq p^*$. Then, since for any constant c, $\sum_j d_j c = 0$, Equation 9 can be written:

$$\sum\nolimits_j d_j(\tau_j - \tau^*) \leq 0, \tag{10}$$

which we know to be true because $d_j \geq 0 \leftrightarrow p_j \leq p^* \leftrightarrow \tau_j - \tau^* \leq 0$.

The proof for $\tau_{FE} \geq \tau_{ATC}$ proceeds similarly.

□

A number of considerations are of interest with regard to this result.

First, monotonicity is not a necessary condition for $\tau_{FE} \in [\tau_{ATC}, \tau_{ATT}]$, as is easily shown with counterexamples. The necessary and sufficient condition for $\tau_{FE} \leq \tau_{ATT}$ is given in Equation 9.

Second, while monotonicity ensures that τ_{FE} lies between τ_{ATT} and τ_{ATC}, there is no guarantee that τ_{ATT} and τ_{ATC} are close to each other or to τ_{ATE}. Indeed, all else equal, the difference between these two is greatest under monotonicity. In particular, given sets $(\tau_j)_{j=1}^s$ and $(p_j)_{j=1}^s$ for s equal sized strata, the difference $\tau_{ATT} - \tau_{ATC}$ is maximized (resp. minimized) by a (bijective) mapping $h : \{1, 2, ...s\} \rightarrow \{1, 2, ...s\}$ for which $(\tau_j)_{j=1}^s$ is monotonically increasing (resp. decreasing) in $(p_{h(j)})_{j=1}^s$. More positively, whether or not τ_{ATT} and τ_{ATC} are far from τ_{ATE} depends on the variance of the weights used in each case ($p/\sum_j p_j$ and $(1-p)/\sum_j(1-p_j)$, respectively). Ignoring w for simplicity, letting ω denote a set of weights, and using the Cauchy-Schwarz inequality, the difference between the weighted and unweighted means is bounded according to $\sqrt{\left(\sum_j(\omega_j - \frac{1}{s})\tau_j\right)^2} \leq \sqrt{\left(\sum_j(\omega_j - \frac{1}{s})^2\right)}\sqrt{\left(\sum_j \tau_j^2\right)}$. Since $\mathbb{E}[\omega] = \frac{1}{s}$, the term $\sum_j(\omega_j - \frac{1}{s})^2$ corresponds to $sVar(\omega)$ and so the bound scales with the standard deviation of the weights.

Third, an analogous statement holds for sample statistics. Defining $\hat{\tau}_{ATT} \equiv \frac{\sum_j p_j w_j \hat{\tau}_j}{\sum_j p_j w_j}$ and $\hat{\tau}_{ATC} \equiv \frac{\sum_j (1-p_j) w_j \hat{\tau}_j}{\sum_j (1-\hat{p}_j) w_j}$, we have that if p_j is monotonic in the observed within-stratum difference in means ($\hat{\tau}_j$), then $\hat{\tau}_{FE}$ lies between $\hat{\tau}_{ATC}$ and $\hat{\tau}_{ATT}$. The proof exactly parallels that of Proposition 1.

Finally, there are fruitful connections here with findings in (9). (9) identifies τ_{FE} as a weighted average of two quantities. When potential outcomes (and so, effects) are linear in propensities, these correspond to τ_{ATC} and τ_{ATT}. Interestingly, in case of linearity, weights can be also be calculated directly from the expressions in Equation 5, 6, and 8, with a weight on τ_{ATT} given by:

$$\lambda = \frac{\frac{\sum_j p_j^2(1-p_j)w_j}{\sum_j p_j(1-p_j)w_j} - \frac{\sum_j p_j(1-p_j)w_j}{\sum_j(1-p_j)w_j}}{\frac{\sum_j p_j^2 w_j}{\sum_j p_j w_j} - \frac{\sum_j p_j(1-p_j)w_j}{\sum_j(1-p_j)w_j}} \tag{11}$$

See supplementary materials (C) for intermediate steps.

This weight admits a substantive interpretation. Quantity $\frac{\sum_j p_j^2(1-p_j)w_j}{\sum_j p_j(1-p_j)w_j}$ is the variance-weighted average propensity and $\frac{\sum_j p_j(1-p_j)w_j}{\sum_j(1-p_j)w_j}$ and $\frac{\sum_j p_j^2 w_j}{\sum_j p_j w_j}$ give, respectively, the average propensity among units in control and in treatment. The denominator is then the difference in average propensities between treatment and control groups. The numerator is the difference between the variance weighted average propensity and the average propensity in control. We then have $\lambda = 1$ when the variance weighted mean propensity is equal to the average propensity in treatment, and 0 when it equals the average propensity in control.

Linearity is a stronger assumption than monotonicity however, and if only monotonicity can be defended, then the weighted quantities in (9) lose their connection to causal estimands. However Proposition 1 provided here can still be used.

4 Conclusion

Researchers commonly use covariate adjustment to account for known variation in treatment assignment propensities. This situation can arise in both observational and experimental studies.

A common analysis strategy in such cases is to regress outcomes on treatment using a set of controls entered additively. A maximally flexible version of this approach, which I focus on here, is one in which researchers use fixed effects specifications to seek to capture variation in assignment propensities.

This approach, is unfortunately not guaranteed to produce unbiased estimates of the average treatment effect. Moreover, it is not well understood how estimates generated in this manner diverge from τ_{ATE} and so how to interpret these results.

For this reason this approach should, in general, be avoided. And fortunately there are multiple ways to generate estimates of average treatment effects in this setting. Most simply, Equation 3 can be used to estimate within-stratum effects; a weighted average of these will be unbiased for τ_{ATE}. A blocked difference in means estimator is available in (18). Other strategies include inverse propensity weights or regression interacting treatment with demeaned stratum dummy variables. Further strategies are described in (10). Supplementary materials (C) provide code drawing on (19) to illustrate the performance of some of these approaches for a variant of Example 1 above.

Despite the availability of these alternatives, using fixed effects to address assignment heterogeneity remains common, as documented recently in (10). If users are unable to access data and re-estimate effects correctly, rules of thumb become useful to help interpret reported findings. A number are provided here. First for 'rare' treatments the least squares estimand lies close to the average treatment effect for the treated; for 'common' treatments it is close to the treatment effect for the controls. Second, if propensity variance is similar across strata, then the OLS estimand lies close to the ATE, even if actual propensities diverge. Third, when a monotonicity condition is satisfied τ_{FE} lies between the average treatment effect for the treated and the average treatment effect for the controls. Thus when higher values on third variables are associated both with more positive (or more negative) treatment effects and with a higher (or lower) propensity to being assigned to treatment, τ_{FE} is bounded by causal quantities of interest. Under the stronger assumption that effects are linear in propensities, a new intuitive weight is provided to indicate relative proximity to τ_{ATC} and τ_{ATT}.

References

[1] Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika. 1983;70:41-55.

[2] Goldsmith-Pinkham P, Hull P, Kolesár M. Contamination bias in linear regressions. National Bureau of Economic Research; 2022.

[3] Aronow P, Samii C. Estimating average causal effects under general interference, with application to a social network experiment. Annals of Applied Statistics. 2017;11(4):1912-47.

[4] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41-55.

[5] Duflo E, Glennerster R, Kremer M. Chapter 61 Using Randomization in Development Economics Research: A Toolkit. In: Schultz TP, Strauss JA, editors. Handbook of Development Economics. vol. 4 of Handbook of Development Economics. Elsevier; 2007. p. 3895 3962.

[6] Ho DE, Imai K, King G, Stuart EA. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. Political Analysis. 2007;15(3):199-236.

[7] Lin W. Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique. The Annals of Applied Statistics. 2013:295-318.

[8] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005;61(4):962-73.

[9] Słoczyński T. Interpreting OLS estimands when treatment effects are heterogeneous: Smaller groups get larger weights. Review of Economics and Statistics. 2022;104(3):501-9.

[10] Gibbons CE, Suárez Serrato JC, Urbancic MB. Broken or fixed effects? Journal of Econometric Methods. 2019;8(1):20170002.

[11] Angrist JD. Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants. Econometrica. 1998 March;66(2):249-88.

[12] Heckman JJ, Taber C. Roy model. In: Microeconometrics. Springer; 2010. p. 221-8.

[13] Chassang S, Padró i Miquel G, Snowberg E. Selective trials: A principal-agent approach to randomized controlled experiments. American Economic Review. 2012;102(4):1279-309.

[14] Paluck EL, Shepherd H. The salience of social referents: a field experiment on collective norms and harassment behavior in a school social network. Journal of personality and social psychology. 2012;103(6):899.

[15] Habyarimana J, Humphreys M, Posner DN, Weinstein JM. Why does ethnic diversity undermine public goods provision? American political science review. 2007;101(4):709-25.

[16] Ding P. The Frisch–Waugh–Lovell theorem for standard errors. Statistics & Probability Letters. 2021;168:108945.

[17] Baird S, Bohren JA, McIntosh C, Özler B. Optimal design of experiments in the presence of interference. Review of Economics and Statistics. 2018;100(5):844-60.

[18] Blair G, Cooper J, Coppock A, Humphreys M, Sonnet L. estimatr: Fast Estimators for Design-Based Inference; 2024. R package version 1.0.2, https://github.com/DeclareDesign/estimatr. Available from: `https://declaredesign.org/r/estimatr/`.

[19] Blair G, Coppock A, Humphreys M. The trouble with 'controlling for blocks'; 2018. Accessed: 2024-10-14. Available from: `https://declaredesign.org/blog/posts/biased-fixed-effects.html`.

[20] Bernstein DS. Matrix mathematics: theory, facts, and formulas. Princeton university press; 2009.

MSC 2020 Codes

62D20

Acknowledgments

My thanks to Craig McIntosh, Marion Dumas, Andy Gelman, Joshua Angrist, Kosuke Imai, Laura Paler, Neelan Sircar, and Guido Imbens for generous comments on an earlier version of this paper.

Funding information

No funding was used for this research.

Author statement

The author declares no conflict of interest.

Data availability statement

No data is used in this research.

Supplementary Material for 'Bounds on the fixed effects estimand in the presence of heterogeneous assignment propensities'

Macartan Humphreys
WZB, Humboldt University, Trinity College Dublin

A Derivation of $\hat{\tau}_{FE}$

Proposition 2. *Let y denote a vector of outcomes and X a matrix in which the first column is the treatment assignment and columns 2 to $s+1$ are dummy variables for each of s strata. Let $n_j, p_j, n_j^1, \overline{y}_j, \overline{y}_j^0$ and $\overline{y}_j^1$ denote, respectively, the size of stratum j, the share of units in stratum j in treatment, the number of units in stratum j in treatment, the average outcome of units in stratum j and the stratum j average (observed) outcome among treated and control units respectively.*

Then, OLS regression of y on X yields:

$$\begin{bmatrix} \hat{\tau}_{FE} \\ \hat{\alpha}_{FE}^1 \\ \vdots \\ \hat{\alpha}_{FE}^s \end{bmatrix} = (X'X)^{-1}X'y = \begin{bmatrix} \frac{\sum_j n_j p_j (1-p_j)(\overline{y}_j^1 - \overline{y}_j^0)}{\sum_j n_j p_j (1-p_j)} \\ \overline{y}_1 - p_1 \frac{\sum_j n_j p_j (1-p_j)(\overline{y}_j^1 - \overline{y}_j^0)}{\sum_j n_j p_j (1-p_j)} \\ \vdots \\ \overline{y}_s - p_s \frac{\sum_j n_j p_j (1-p_j)(\overline{y}_j^1 - \overline{y}_j^0)}{\sum_j n_j p_j (1-p_j)} \end{bmatrix}.$$

Proof. To establish the result, note first that the matrix $X'X$ can be represented as a block matrix:

$$X'X = \begin{bmatrix} n^1 & \mathbf{n^1}' \\ \mathbf{n^1} & M \end{bmatrix}$$

where n^1 is the number of units in treatment, $\mathbf{n^1} = [n_1^1, n_2^1, \ldots, n_s^1]'$ is the number treated in each stratum, and M is a diagonal matrix reporting the number of units in each stratum.

From the inversion of block matrices (see Eqn 2.8.17 in (20)):

$$(X'X)^{-1} = \begin{bmatrix} (n^1 - \mathbf{n^1}' M^{-1}\mathbf{n^1})^{-1} & -(n^1 - \mathbf{n^1}' M^{-1}\mathbf{n^1})^{-1}\mathbf{n^1}' M^{-1} \\ -M^{-1}\mathbf{n^1}(n^1 - \mathbf{n^1}' M^{-1}\mathbf{n^1})^{-1} & M^{-1} + M^{-1}\mathbf{n^1}(n^1 - \mathbf{n^1}' M^{-1}\mathbf{n^1})^{-1}\mathbf{n^1}' M^{-1} \end{bmatrix}$$

Observing that

$$\mathbf{n^1}' M^{-1} = (p_1, p_2 \ldots p_s)$$

and defining

$$w := (n^1 - \mathbf{n^1}' M^{-1}\mathbf{n^1})^{-1} = \frac{1}{\sum_j p_j n_j - \sum_j p_j^2 n_j} = \frac{1}{\sum_j n_j p_j (1 - p_j)}$$

we have:

$$(X'X)^{-1} = w \cdot \begin{bmatrix} 1 & -p_1 & -p_2 & \cdots & -p_s \\ -p_1 & \frac{1}{n_1 w} + p_1^2 & p_1 p_2 & \cdots & p_1 p_s \\ -p_2 & p_2 p_1 & \frac{1}{n_2 w} + p_2^2 & \cdots & p_2 p_s \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -p_s & p_s p_1 & p_s p_2 & \cdots & \frac{1}{n_s w} + p_s^2 \end{bmatrix}$$

Similarly:

$$X'y = \left(\sum_{i:d_i=1} y_i, \sum_{i \in X_1} y_i, \ldots, \sum_{i \in X_s} y_i \right)$$

$\hat{\tau}_{FE}$ is then the inner product of the first row of $(X'X)^{-1}$ and $X'y$:

$$\hat{\tau}_{FE} \;\; = \;\; w \left(\sum_{i:d_i=1} y_i - p_1 \sum_{i \in X_1} y_i - p_2 \sum_{i \in X_2} y_i - \cdots - p_s \sum_{i \in X_s} y_i \right)$$

To simplify, observe that $\sum_{i:d_i=1} y_i = \sum_j n_j p_j \overline{y}_j^1$ and $\sum_{i \in j} y_i = n_j (p_j \overline{y}_j^1 + (1 - p_j)\overline{y}_j^0)$, and so:

$$
\begin{aligned}
\hat{\tau}_{FE} &= w\left(\sum_j n_j p_j \overline{y}_j^1 - \sum_j n_j p_j (p_j \overline{y}_j^1 + (1-p_j)\overline{y}_j^0)\right) \\
&= \frac{\sum_j n_j p_j (1-p_j)(\overline{y}_j^1 - \overline{y}_j^0)}{\sum_j n_j p_j (1-p_j)}
\end{aligned}
$$

In the same way:

$$
\begin{aligned}
\hat{\alpha}_{FE}^j &= -wp_j\left(\sum_{i:d_i=1} y_i - p_1 \sum_{i\in X_1} y_i - p_2 \sum_{i\in X_2} y_i - \cdots - p_s \sum_{i\in X_s} y_i\right) - w\left(-\frac{1}{n_j w}\right)\sum_{i\in X_j} y_i \\
&= \overline{y}_j - p_j \hat{\tau}_{FE}
\end{aligned}
$$

□

B Omitted steps in proof of Proposition 1

The proof for Proposition 1 relies on an equivalence between:

$$\sum\nolimits_j \frac{p_j(1-p_j)w_j}{\sum_j p_j(1-p_j)w_j}\tau_j \leq \sum\nolimits_j \frac{p_j w_j}{\sum_j p_j w_j}\tau_j \tag{12}$$

and

$$\sum\nolimits_j \left(\frac{p_j w_j}{\sum_j p_j w_j} - \frac{p_j^2 w_j}{\sum_j p_j^2 w_j}\right)\tau_j \leq 0. \tag{13}$$

To see this equivalence, define $\alpha = \sum_j p_j w_j$ and $\beta = \sum_j p_j^2 w_j$. Condition 12 can then be written:

$$\begin{array}{rcl}
\sum\nolimits_j p_j w_j \tau_j \dfrac{1-p_j}{\alpha-\beta} & \leq & \displaystyle\sum_j p_j w_j \tau_j \frac{1}{\alpha} \\
& \leftrightarrow & \\
\sum\nolimits_j p_j w_j \tau_j \left(\dfrac{1-p_j}{\alpha-\beta} - \dfrac{1}{\alpha}\right) & \leq & 0 \\
& \leftrightarrow & \\
\sum\nolimits_j p_j w_j \tau_j \left(\dfrac{\beta-\alpha p_j}{(\alpha-\beta)\alpha}\right) & \leq & 0 \\
& \leftrightarrow & \\
\sum\nolimits_j p_j w_j \tau_j \left(\dfrac{1}{\alpha} - \dfrac{p_j}{\beta}\right) & \leq & 0
\end{array}$$

Where the last step results from multiplying across by $\frac{\alpha-\beta}{\beta} > 0$. Resubstituting for α and β yields the result.

C Derivation of Equation 11

In the linear case, monotonicity is satisfied and we can therefore write:

$$\tau_{FE} = \lambda \tau_{ATT} + (1-\lambda)\tau_{ATC}$$

Substituting from Equations 5, 6, and 8:

$$\frac{\sum_j p_j(1-p_j)w_j\tau_j}{\sum_j p_j(1-p_j)w_j} = \lambda \frac{\sum_j p_j w_j \tau_j}{\sum_j p_j w_j} + (1-\lambda)\frac{\sum_j (1-p_j)w_j\tau_j}{\sum_j (1-p_j)w_j}.$$

With treatment effects linear in p_j we have for some β:

$$\frac{\sum_j p_j(1-p_j)w_j\beta p_j}{\sum_j p_j(1-p_j)w_j} = \lambda \frac{\sum_j p_j w_j \beta p_j}{\sum_j p_j w_j} + (1-\lambda)\frac{\sum_j (1-p_j)w_j\beta p_j}{\sum_j (1-p_j)w_j}$$

Dividing across by β, and gathering terms gives:

$$\frac{\sum_j p_j^2(1-p_j)w_j}{\sum_j p_j(1-p_j)w_j} = \lambda \frac{\sum_j p_j^2 w_j}{\sum_j p_j w_j} + (1-\lambda)\frac{\sum_j p_j(1-p_j)w_j}{\sum_j (1-p_j)w_j}$$

Solving for λ then yields:

$$\lambda = \frac{\frac{\sum_j p_j^2(1-p_j)w_j}{\sum_j p_j(1-p_j)w_j} - \frac{\sum_j p_j(1-p_j)w_j}{\sum_j (1-p_j)w_j}}{\frac{\sum_j p_j^2 w_j}{\sum_j p_j w_j} - \frac{\sum_j p_j(1-p_j)w_j}{\sum_j (1-p_j)w_j}}$$

D Code illustration

There are many approaches that can be used to generate unbiased estimates in the presence of heterogeneous but known propensities across strata. I illustrate by using the `R` package `DeclareDesign` to simulate data from a version of Example 1 and show the performance of five estimation strategies: pooled OLS, OLS with stratum dummies (fixed effects), Inverse Propensity Weighting (IPW), OLS but with interactions between treatment and demeaned stratum dummies, following (7), and blocked differences in means. This code draws from material in (19).

Code to compare estimators for Example 1

```
library(DeclareDesign)

prob <- c(.067, .5, .933)

design <-
  declare_model(
    block = add_level(N = 3, p = prob, tau = c(3, -3, 3)),
    unit = add_level(N = 1000, Y0 = 10*(p + rnorm(N)), Y1 = Y0 + tau)) +
  declare_inquiry(ATE = mean(Y1 - Y0)) +
  declare_assignment(Z = block_ra(blocks = block, block_prob = prob)) +
  declare_measurement(
    ipw = 1/(Z*p + (1-Z)*(1-p)),
    Y = Z*Y1 + (1-Z)*Y0) +
  declare_estimator(Y ~ Z, .method =  lm_robust,
    label = "Pooled") +
  declare_estimator(Y ~ Z + block, .method =  lm_robust,
    label = "Fixed effects") +
  declare_estimator(Y ~ Z, blocks = block, .method =  difference_in_means,
    label = "Blocked differences in means") +
  declare_estimator(Y ~ Z, covariates = ~ block, .method = lm_lin,
    label = "Interactions (Lin approach)") +
  declare_estimator(Y ~ Z, .method = lm_robust, weight = ipw,
    label = "Inverse propensity weights")

simulate_design(design) |>
 group_by(estimator) |>
 summarize(
  SE_bias = mean(std.error - sd(estimate)),
  ATE_bias = mean(estimate - estimand) )
```

The code is shown in Box D. Table 1 shows the results, highlighting the poor performance of both the pooled approach and the fixed effects approach

Table 1: Performance of five strategies to estimate average treatment effects

Estimator	Bias of standard errors	Bias of estimates
Pooled	0.02	5
Fixed Effects	0.00	-2
Inverse propensity weights	0.04	0
Interactions (Lin approach)	0.00	0
Blocked differences in means	0.00	0

in this setting. IPW, the interaction model, and blocked differences in means are all unbiased, though they differ in the performance of standard errors—assessed here as the difference between the estimated average standard error and the estimated standard deviation of the sampling distribution of estimates under each estimator.