ASSOCIATION OF POPULATION CENTERS ARXIV
WORKING PAPER SERIES

No. 2303

The Phenotype Differences Model Reveals Genetic Effects on Mortality Using Incomplete Sibling Data

Sam Trejo and Klint Kanopka

June 2023

ASSOCIATION OF POPULATION CENTERS

The Phenotype Differences Model Reveals Genetic Effects on Mortality Using Incomplete Sibling Data

Sam Trejo[1,†] and Klint Kanopka[2]

[1]Princeton University, Department of Sociology & Office of Population Research
[2]Stanford University, Graduate School of Education
[†]samtrejo@princeton.edu

June 10, 2023

Abstract

The identification of causal relationships between specific genes and social, behavioral, and health outcomes is challenging due to environmental confounding from population stratification and dynastic genetic effects. Numerous existing methods leverage the random genetic differences between parents and their children induced by genetic recombination to estimate effect that are free from environmental confounding. However, such methods require dyadic genetic data within families (i.e. parent-child pairs and/or sibling pairs) and therefore can only be applied in relatively small and selected samples. We introduce the *phenotype differences* model to compare siblings and estimate the causal effect of genetic predictors using just a single individual's genotype. We show that, under plausible assumptions, the phenotype differences model provides unbiased and consistent estimates of genetic effects. We then utilize the phenotype differences model to estimate the effects of 40 polygenic scores on premature mortality using asymmetrically genotyped sibling pairs in the Wisconsin Longitudinal Study. We find that twelve polygenic scores related to self-rated health, body mass index, education, cognition, depression, life satisfaction, smoking behavior, and chronic obstructive pulmonary disease have a meaningful impact on mortality outcomes. When we combine information across multiple polygenic scores, the sibling in a pair who inherited more longevity-increasing DNA from their parents on average lived 9 months longer and was 7 pp (12%) more likely to survive until age 75 than their brother/sister.

Acknowledgements. The authors would like to thank Benjamin Domingue, Dalton Conley, Jason Fletcher, Qiongshi Lu, and Elliot Tucker-Drob for helpful comments on early versions of this manuscript. We are also grateful to Pamela Herd and the Wisconsin Longitudinal Study staff for access to the restricted-use data. Our research has benefited from the use of the Social Science Genetic Association Consortium's Polygenic Index Repository (`https://www.thessgac.org/pgi-repository`). This work has been supported, in part, by the Institute of Education Sciences under Grant No. R305B140009. All opinions expressed are those of the authors alone and should not be construed as representing the opinions of any institution.

1 INTRODUCTION

Understanding whether and how variation in individual DNA sequence produces variation in life outcomes is a key goal of the field of human genomics. Over the last decade, researchers have been very successful at assembling large genome-wide association study (GWAS) samples of unrelated individuals to precisely estimate genetic associations for a wide ranch of traits [1]. However, the identification of causal relationships between specific genes and social, behavioral, and health outcomes – as well as the mechanisms that such effects operate through – is challenging due to between-family environmental confounding from population structure [2, 3, 4] and dynastic genetic effects [5, 6]. Thus, it difficult to know how well current between-family GWAS discoveries and existing polygenic scores index the causal effects of genes [7].

While naive genetic associations are often environmentally confounded, there exist promising solutions. In the case of DNA, we have the ultimate 'natural' experiment; conditional on their parents' genes, a child's genes are quasi-randomly assigned via genetic recombination. Numerous strategies have been developed to leverage these quasi-random genetic differences between parents and their children in order to isolate the causal effects of genes.[1] These strategies include both [i] trio methods, which explicitly conditions on parental genotype, and [ii] sibling methods, which difference out all shared family-level factors, indirectly conditioning on parental genotype. However, existing methods require the use of dyadic genetic data within families (i.e., parent-child pairs and/or sibling pairs) and therefore can only be applied in relatively small and selected samples. A dearth of such data exists; for instance, though the UK Biobank has roughly 500,000 genotyped individuals, it has only about 22,000 sibling pairs (and even fewer parent-child pairs) [9]. At present, researchers are left with estimates of genetic effects that are either precise but environmentally biased or quite imprecise but environmentally unbiased.

Typical sibling methods, such as the *fixed effects* model,[2] require four pieces of information: the genotype of both siblings and the phenotype of both siblings. We introduce a new within-family regression specification to compare siblings and estimate direct genetic effects, which we call the *phenotype differences* model. Importantly, the phenotype differences model provides, in expectation, the same estimates as fixed effects models but instead using just a single individual's genotype (as well as the phenotype of that individual and one of their siblings).[3] In doing so, the phenotype differences model can increase statistical power (by increasing the size of analytic samples) and improve external validity (by increasing the representativeness of samples) in within-family genetic analyses. While the phenotype differences method can potentially be applied when studying the effects of non-genetic variables, it is especially well-suited for genetic predictors because of our strong prior about the correlation of genes within families ($\rho^{G1,G2}$).

We show that, under plausible assumptions, the phenotype differences model provides unbiased and consistent estimates of genetic effects and that, when genetic effects are small, phenotype differences provides the same precision as fixed effects per genotype. Additionally, we show that the phenotype differences model works for analyses which use individual-level genetic characteristics as an instrumental variable, also known as Mendelian randomization; conducting Mendelian randomization within-families helps reduce biases that result from violations of the exclusion restriction [12].[4] The key mathematical intuition and assumptions of the phenotype differences model can be found in the *Online Methods* section, and full derivations are available in the SI.

[1] While within-family approaches successfully eliminate environmental confounding, recent work suggests that genetic confounding may still be an issue [8]. That is, such models may suffer from confounding of the relationship between g_{ij} and y_{ij} by other correlated genetic variants not encompassed in g_{ij}.

[2] In general, fixed effects and first differences models are slightly different statistical approaches for making for within-group comparisons. However, in our case, where the number of observations i in each group j is equal to two (e.g., sibling pairs), the fixed effects and first differences specifications are algebraically identical (see chapter 5.1. of Angrist & Pischke 2009) [10].

[3] Both fixed effects and phenotypes difference models may suffer from bias in the presence of indirect effects between siblings [11]. However, little evidence of meaningful siblings effects has been detected, thus far [5].

[4] The phenotype differences model may be adapted for use in the study of polygenic score-by-environment interactions [13, 14], but such an extension are beyond the scope of the present paper.

$$\text{Fixed Effects Model:} \quad y_{1j} - y_{2j} = \hat{\beta}^{\text{FE}}(g_{1j} - g_{2j}) + \hat{\varepsilon}^{*}_{ij}$$

$$\text{Phenotype Differences Model (General):} \quad y_{1j} - y_{2j} = \hat{\alpha} + \hat{\beta}^{\text{PD}}\left(g_{1j}(1 - \rho^{G1,G2})\right) + \hat{\varepsilon}_{ij}$$

$$\text{Phenotype Differences Model (Reduced):} \quad y_{1j} - y_{2j} = \hat{\alpha} + \hat{\beta}^{\text{PD}}\frac{g_{1j}}{2} + \hat{\varepsilon}_{ij}$$

Where:

y_{ij}: Outcome for individual i in family j

g_{ij}: Genotype of individual i in family j

$\hat{\alpha}$: Intercept

$\rho^{G1,G2}$: Population correlation between between g_{1j} and g_{2j}

While the phenotype model differences model is valid for both variant-level (i.e., in GWAS, where g_{ij} is a single SNP) and genome-wide (i.e., when g_{ij} is a polygenic score or some other summary measure) analyses, in our empirical analyses we focus on the genome-wide case. As an example application, we consider the case of the Wisconsin Longitudinal Study (WLS); the WLS is a longitudinal survey based on a $\frac{1}{3}$ sample of all 1957 Wisconsin high school graduates ($N = 10,317$) and a randomly selected sibling of these graduates [15]. The graduates were originally empanelled with an in-person questionnaire at age 18; both WLS graduates and the randomly selected siblings were re-interviewed periodically across the life course. The WLS began collecting genotype data in the early 2000s, meaning that respondents must have survived and remained empanelled in the study in order to be included in the genetic sample. Panel A of Figure 1 provides a binned scatter plot of the likelihood that a WLS respondent is genotyped as a function of their year of death. As can be seen, no WLS respondents who died before 2006 were genotyped. On the other hand, among those WLS respondents who survived past 2015, nearly 80% were genotyped.

[**Insert Figure 1 Here**]

Ignoring WLS siblings pairs where neither sibling is genotyped, we are left with two mutually exclusive samples of siblings pairs. There there is [i] the One Genotype Sample, where only a single sibling is genotyped, and [ii] the Two Genotypes Sample, where both siblings are genotyped. Table 1 displays summary statistics for these two samples of WLS sibling pairs. To date, all sibling analyses using the WLS data have focused on the 2,088 siblings pairs that comprise the Two Genotypes Sample. In this paper, we show how the phenotype differences model allows us to extend our analyses to the additional 3,548 siblings pairs in the One Genotype Sample.

[**Insert Table 1 Here**]

Importantly, conducting within-family genetic analyses in the WLS using only the Two Genotypes Sample fundamentally limits the kinds of causal inferences that can be made. Panel B of Figure 1 displays overlayed histograms of the within-family difference in lifespan for both the One Genotype Sample (in red) and the Two Genotypes Sample (in blue). In order for a pair to be included in the Two Genotypes Sample, both siblings must have survived until genotyping; this mechanically restricts the within-family variation in lifespan, limiting our ability to use siblings comparisons to understand genetic effects on premature mortality outcomes in this sample. However, as can be seen, the One Genotype Sample does not face the same limitation. Therefore, the phenotype differences model also allows us to leverage such variation and to explore genetic effects on premature mortality.[5]

[5]Notably, we are forced to focus on *premature* mortality due to the fact that our lifespan variable is right-censored. The most recent time that the WLS collected mortality data from the National Death Index was 2018. At that time, 79% of the members of the combined sibling sample were still alive.

2 RESULTS

We begin by fitting a series of regressions to empirically validate the performance of the phenotype differences model compared to the fixed effects model using WLS sibling pairs. Figure 2 compares the estimated β coefficients from fixed effects and phenotype differences regressions of 30 phenotypes on their respective polygenic score. All polygenic scores are drawn from the recent Social Science Genomics Association Consortium Polygenic Index Repository [16]. Though the repository contains 47 distinct polygenic scores, in the WLS there only exists phenotype data for a subset of 30 traits. Because both the fixed effects and phenotype differences model leverage only sibling comparisons, the β coefficients hold a causal interpretation.

[**Insert Figure 2 Here**]

All three panels of Figure 2 display the same fixed effects estimates, which are derived from the full Two Genotypes Sample. In Panels A and B, the phenotype differences estimates come from a procedure using the Two Genotypes Sample in which the genetic (but not phenotypic) data of a randomly selected sibling in each pair is discarded. Panel A displays the mean phenotype differences estimate from 1000 iterations of this procedure whereas Panel B displays the estimates from a single iteration. In Panel C, the phenotype differences estimates are derived from the One Genotype Sample, meaning the two estimates are fit on entirely non-overlapping samples.

For all of the phenotype differences models, $\rho^{G1,G2}$ is estimated using the Two Genotypes Sample. While empirical estimates of $\rho^{G1,G2}$ were close to 0.5 for most phenotypes, we found the greatest evidence of positive assortative mating for the height phenotype ($\rho^{G1,G2} = 0.62$); this value is notably higher than the phenotype with the next highest within-family correlation, chronic obstructive pulmonary disorder ($\rho^{G1,G2} = 0.56$). We found little evidence for negative assortative mating, with the lowest empirical estimates of $\rho^{G1,G2}$ being for hayfever ($\rho^{G1,G2} = 0.49$). A table with the sibling correlations for all 47 polygenic scores can be found in the SI.

As can be seen, there is a high correspondence (moving from left to right: $\rho = 0.99$, $\rho = 0.87$, and $\rho = 0.88$) between the fixed effects and phenotype differences estimates, even when fit on non-overlapping samples. In Panel A, the two estimates are virtually identical; this suggests that differences between the fixed effects and phenotype differences estimates in Panel B are simply the result of sampling variance. Importantly, the correspondence between fixed effects and phenotype differences estimates is very similar across Panel B and Panel C, implying that the key phenotype differences assumptions are, indeed, largely met in the WLS One Genotypes Sample. To achieve a single estimate with the greatest precision, the two estimates displayed in Panel C can be pooled using inverse variance-weighted meta-analysis.

Next, we use the phenotype difference model to estimate the causal effects of 40 polygenic scores on mortality outcomes. While previous studies have shown that both specific genetic variants [17, 18] and polygenic scores [19, 20] are associated with mortality outcomes, the extent to which these associations represent the causal effects of genes versus environmental confounding is largely unknown. Figure 3 displays the β coefficients of theses polygenic score on lifespan in years (Panel A) and a dichotomous indicator for surviving to age 75 (Panel B).[6] To reduce the extent to which the regressions fit on noise induced by the right-censoring of the lifespan variable, only sibling pairs (N=2,191) where at least one of the siblings is deceased are included in the regression sample.[7] See the SI for a table describing the analytic sample used in our analysis of premature mortality.

[**Insert Figure 3 Here**]

In Panel A of Figure 3, 7 polygenic scores have statistically significant effects on lifespan. The polygenic scores for self-rated health, life satisfaction–family, educational attainment, and cognitive

[6] These coefficient estimates are pooled via inverse variance-weighted meta-analysis across the One and Two Genotype Samples using phenotype differences and fixed effects, respectively (though, as can be seen in Figure 1, much of the identifying outcome variation comes from the phenotype differences estimates on the One Genotype Sample.)

[7] Of the 2,191 sibling pairs used in the lifespan analyses, just 1,789 pairs included only siblings who were born before 1943 (and were therefore able reach age 75 by the time of data collection in 2018). These 1,789 pairs become the analytic sample for our survival to 75 analyses.

ability increase lifespan and the polygenic scores for chronic obstructive pulmonary disease and body mass index decrease lifespan. In Panel B, a similar, although distinct, set of polygenic scores have statistically significant effects on surviving until age 75. The polygenic scores for self-rated health, life satisfaction–family, educational attainment, and cognitive ability increase the likelihood of surviving until age 75 and the polygenic scores for depressive symptoms, loneliness, body mass index, and smoking initiation decrease the likelihood of surviving until age 75. A figure containing all of the statistically significant β coefficients from Figure 3 corrected for measurement error using the procedure described in Becker et al. 2021 [16] can be found in the SI.

In addition, we summarize genetic risk information across traits and create a so-called meta-polygenic score for each of our two mortality outcomes. Each meta-polygenic score is a weighted average of the statistically significant polygenic scores. The weights are derived from between-family ridge regression of these polygenic scores on mortality. A table with the weights used to compute each meta-polygenic score can be found in the SI.

A 1 SD change in meta-polygenic score caused a 0.93 year and 8.8 pp (16%) increase in lifespan and the probability of surviving to 75, respectively. In our sample, the mean absolute difference between sibling pairs for a difference in the meta-polygenic score is approximately 0.8 SD (for two randomly selected unrelated individuals, the mean absolute difference is approximately 1.1 SD). This entails that, on average, the sibling who inherited the higher lifespan meta-polygenic score lived 9 months longer than their brother/sister as a result. Similarly, the sibling who inherited the higher survive-to-75 meta-polygenic score was, on average, 7 pp (12%) more likely to survive until the age of 75 than their brother/sister.

3 DISCUSSION

Our results demonstrate that the phenotype differences model is a robust estimator of genetic effects in the presence of environmental confounding. Crucially, the comparatively less genetic data required by the phenotype differences model has the potential to increase precision and generalizability of within-family genomic studies. The potential applications of the phenotype differences model extend far beyond addressing mortality selection into genotyping in a longitudinal study, as we have done here. More broadly, our work highlights the value of collecting sibling phenotype data, even when genotype data is unavailable. Indeed, perhaps the most beneficial use cases of the phenotype differences model will come through new data collection efforts and/or creative uses of existing data resources; for instance, in the case of phenotypes that are easily reported (like height and educational attainment), through [i] surveying unrelated individuals on the phenotypes of all or a randomly selected sibling; or, when studying rare and sensitive phenotypes (such as severe mental disorders), through [ii] leveraging population registries and other administrative data bases. Table 2 describes various applications of the phenotype differences model in more detail.

[**Insert Table 2 Here**]

This study's example application of the phenotype differences model provided a glimpse into the so-called genetic lottery [21, 22] for premature mortality in mid-century Wisconsin. We find that twelve polygenic scores related to self-rated health, body mass index, education, cognition, depression, life satisfaction, smoking behavior, and chronic obstructive pulmonary disease have meaningful effects on an individual's premature mortality outcomes. That is, holding the other circumstances of one's birth constant, if a person were to have inherited a different DNA sequence (and, in turn, a different polygenic score), their expected lifespan and probability of surviving to 75 would have also changed [23]. Nonetheless, the precise pathways through which the genetic effects observed in this study operate remain largely unknown. While some mistakenly believe that genetic effects exist strictly within the body, the effects of genes instead often operate through long, complex causal chains meditated by social and environmental aspects of our world [24].

One polygenic score which stands out in our mortality analyses is the score for self-rated health, a less-frequently studied phenotype in genomic research. The self-rated health polygenic

score had the largest in magnitude estimated effect on lifespan and the second largest estimated effect on survival to age 75 (behind the body mass index polygenic score). After measurement error correction, a 1 SD increase in the self-rated health polygenic score causally increased lifespan by 1.4 years and increased the probably of survival to age 75 by 15 pp (26%).[8]

While self-rated health is a well-known and robust predictor of mortality [25, 26], and common genetic variants have been shown to explain 13% of the variation in self-rated health [27], we provide the first evidence that the genetic variants associated with self-rated health themselves have a causal effect on mortality outcomes. Interestingly, self-rated health does not have a high genetic correlation with any of the other statistically significant predictors of mortality,[9] suggesting that its effects operate through relatively unique pathways; this is consistent with the observation in numerous non-genomic studies that self-rated health measures aspects of health relevant to survival which are not captured by other health indicators[26]. In sum, our results provide evidence for the idea that individuals are capable of subjectively indexing meaningful information about genetic influences on their own health, and that genetically-influenced variation in subjective health status, in turn, shapes individual risk for premature mortality. Future genomic studies may benefit from considering novel ways to integrate subjective measures of health status.

Broadly, our results show that certain recent between-family genetic discoveries – as summarized by polygenic scores – have causal effects on premature mortality, an outcome of substantive interest other than the traits these scores were trained to predict. If, in fact, between-family GWAS were overwhelmingly capturing environmental confounding from population structure and dynastic effect, it would be unlikely to see such a result. As genetic indices continue to become increasingly powerful causal predictors, policymakers may need to increase regulations of the uses of genomic predictors in order to protect citizens [28] and prevent adverse outcomes in insurance markets [20].

4 ONLINE METHODS

4.1 The Phenotype Differences Model

Mathematical Intuition

If some family-level environment (e_j) is associated with *both* genotype (g_{ij}) and phenotype (y_{ij}), any naive estimate of the relationship between genotype and phenotype will be subject to confounding. One way to address environmental confounding is the use of the fixed effects model fit on sibling pairs. The fixed effects model eliminates both the environment-phenotype and environment-genotype relationships simultaneously by leveraging sibling differences of all regressors:

$$y_{1j} - y_{2j} = \hat{\beta}^{FE}(g_{1j} - g_{2j}) + \hat{\varepsilon}_{ij}. \tag{1}$$

Notice that, because the environmental effect, e_j, does not vary within-families, it is mechanically uncorrelated with the sibling difference in phenotypes, $y_{1j} - y_{2j}$. In addition, because genotype is quasi-randomly assigned within-families, e_j is uncorrelated with the sibling difference in phenotypes, $g_{1j} - g_{2j}$ in expectation. Thus, regressing $y_{1j} - y_{2j}$ on $g_{1j} - g_{2j}$ allows the fixed effects model to recover estimates of direct that are free from environmental confounding.

Recall that the family-level environment, e_j, must be associated with genotype *and* phenotype in order to confound estimates of genetic effects. Thus, the fixed effects model, which breaks the link between e_j and both e_{ij} and g_{ij}, *exceeds* the requirements for eliminating environmental confounding. This fact forms the conceptual basis for the phenotype differences model, which breaks only the link between environment and phenotype, and thereby requires using genetic

[8] It is difficult to say how we would expect effect sizes from our premature mortality analytic sample to generalize to other populations. On one hand, effects on premature mortality may be larger than in the general population, because we were forced to focus on an analytic sample of siblings pairs that had experienced at least one death by 2018 (and are therefore likely at higher risk). On the other hand, the effects we observe may also be too small, a result of our right-censored lifespan variable and our focus on *premature* mortality.

[9] That is, none, have a genetic correlation over 0.6, the threshold used in Becker et al. 2021 [16]. The top four genetic correlations with self-rated health are: life satisfaction–finance =0.61, age first birth=0.56, body mass index=0.54, and physical activity 0.52.

information from just a single sibling. As above, while e_j is correlated with both y_{1j} and y_{2j}, it is uncorrelated with $y_{1j} - y_{2j}$. Thus, even though a correlation between e_j and g_{1j} persists, it does not introduce environmental confounding. However, because the covariance of $y_{1j} - y_{2j}$ and g_{1j} becomes distorted, we must use the within-family correlation of genetic predictors, $\rho^{G1,G2}$, to re-inflate our genetic effect estimate. This leaves us with the following phenotype differences equation:

$$y_{1j} - y_{2j} = \hat{\alpha} + \hat{\beta}^{PD}\Big(g_{1j}(1 - \rho^{G1,G2})\Big) + \hat{\varepsilon}_{ij}. \tag{2}$$

Section 2.4 of the SI shows a proof of the unbiasedness of the phenotype differences estimator when two key assumptions hold: quasi-random assignment of genotype within families, an assumption also required by fixed effects, and equal variance of the genetic predictor in the genotyped and ungenotyped siblings, an assumption unique to phenotype differences. For genetic predictors, in the absence of assortative mating, $\rho^{G1,G2} = 0.5$. In this case, one can plug in and use a simplified version of the phenotype differences equation:

$$y_{1j} - y_{2j} = \hat{\alpha} + \hat{\beta}^{\mathrm{PD}}\frac{g_{1j}}{2} + \hat{\varepsilon}_{ij} \tag{3}$$

When meaningful assortative mating exists, mother and father genotypes become correlated with one another and $\rho^{G1,G2} \neq .5$ (under positive assortative mating, $\rho^{G1,G2} > .5$). In such a case, an estimate of $\rho^{G1,G2}$ can be derived from a sub-sample of fully genotyped sibling pairs, as we do in our WLS application.

Assumption of Equal Variance of Genetic Predictor

The key assumption required by the phenotype differences model is that the population variance be equal for the genotyped sibling's genetic predictor and the ungenotyped sibling's genetic predictor. Let g_{1j} be drawn from the random variable $G1$ with variance σ_1^2, and let g_{2j} be drawn from the random variable $G2$ with variance σ_2^2. Thus, the assumption of phenotype differences is that $\sigma_1^2 = \sigma_2^2$. Because we don't observe g_{2j}, this assumption is inherently untestable. However, if we know that we observe the genotype of a random sibling (e.g., random sampling), then $G1$ are $G2$ identical random variables and our assumption is trivially met.

When g_{ij} is a polygenic score, there is no inherent mean-variance dependence. Thus, average differences in genetic characteristics will not themselves create a problem. Put differently, the existence of genetic differences between individuals that linearly increase or decrease likelihood of being the genotyped (versus ungenotyped) sibling does not distort the *variance* of $G1$ (compared to $G2$) and therefore will not violate our assumption. However, non-linear and/or non-monotonic selection into genotyping may impact the variance of $G1$ (compared to $G2$) and therefore induce violations of this assumption.

When g_{ij} is the number of major alleles at a single-nucleotide polymorphism (i.e., taking the value of 0, 1, or 2), such as in GWAS, mean differences likely entail variance differences; therefore, systematic differences in the allele frequencies across the genotyped and ungenotyped sibling may induce violatations of the equal variance assumption. When this key assumption is not met, estimates of the true genetic effect become biased as a function of how extreme the variance discordance is. In general, we observe that:

$$\mathbb{E}\big[\hat{\beta}^{PD}\big] = \beta \cdot \frac{1 - \rho^{G1,G2}\sqrt{\frac{\mathrm{var}(G2)}{\mathrm{var}(G1)}}}{1 - \rho^{G1,G2}} \tag{4}$$

When $\rho^{G1,G2} = 0.5$, this reduces to:

$$\mathbb{E}\big[\hat{\beta}^{PD}\big] = \beta \cdot \left(2 - \sqrt{\frac{\mathrm{var}(G2)}{\mathrm{var}(G1)}}\right) \tag{5}$$

While we cannot explicitly test the equal variance assumption in our One Genotype Sample (as we do note observe g_{2j}), we do observe y_{2j}. Thus, we can test for variance differences using

phenotypic data. See the SI for a table that tests the equal variance assumption (specifically, the the Brown-Forsythe test [29] [10]) using phenotypic data. We restrict our tests to phenotypes that are largely continuous, to reduce the presence of mean-variance dependancies (which do not exist for normally distributed polygenic scores). In general, the magnitude of the phenotypic variance discordance that we observe is small and insignificant.

Comparative Precision

When the genetic predictor of interest explains only a small portion of the within-family variation in the outcome variable, fixed effects and phenotype differences yield approximately equally precise effect estimates *per genotype*. That is, when the effects of g_{ij} are small, fixed effects estimates will have asymptotically identical standard errors to phenotype differences estimates derived from the same number genotypes (although, when using phenotype differences, one typically has half as many genotypes per family). However, as the fraction of within-family outcome variation explained by g_{ij} grows, phenotype differences provides less precise effect estimates than fixed effects. Specifically, this decrease in precision is governed by the following formula:

$$\underset{N\to\infty}{\text{p } lim}\ \frac{var(\hat{\beta}^{\text{FE}})}{var(\hat{\beta}^{\text{PD}})} = \frac{1-\phi}{4-\phi} \tag{6}$$

Here, ϕ is the fraction of within-family variation in the y_{ij} that is explained by g_{ij}:

$$\phi = \frac{cov(g_{1j} - g_{2j}, y_{1j} - y_{2j})}{var(y_{1j} - y_{2j})} \tag{7}$$

The within-R^2 of the fixed effects model provides a useful estimate of ϕ. For currently available genetic predictors, this reduction in precision as ϕ increases is, in practice, relatively trivial. For example, even when g_{ij} is the polygenic score for height – the most predictive score available – $\hat{\phi} = 0.16$ (estimated using the WLS Two Genotypes Sample). Therefore, the comparative precision *per genotype* is $2 \times \sqrt{\frac{1-0.16}{4-0.16}} = 0.86$. That is, when fit on the same number of genotypes, fixed effects estimates of the effect of the height polygenic score on phenotypic height will have standard errors that are 0.86 the size of the standard errors of phenotype difference estimates. Nonetheless, the smaller amount of genetic data required by phenotype differences compared to fixed effects can increase sample sizes available for within-family genetic analyses. See the SI for a figure the visualizes the relationship between ϕ and the asymptotic ratio of standard errors of the fixed effects and phenotype differences models.

Non-Paternity Events

Occasionally, siblings pairs believe themselves to be full biological siblings but are, in actuality, only half siblings. Because, under random mating, half siblings have a within-family correlation of genetic predictor of 0.25, sufficient prevalence of non-paternity events can produce $\rho^{G1,G2} < 0.5$. In such a case, phenotype differences estimates become biased away from 0. In the WLS Two Genotypes sample, approximately one-in-fifty sibling pairs who self-report being full biological siblings are actually half siblings. If the variance of the overall distribution of g_{ij} is identical for full and half siblings, we would expect this frequency of non-paternity events to induce $\rho \approx 0.495$. Such a small departure from $\rho = 0.5$ will have only a trivial impact on phenotype differences estimates. Importantly, because misreporting individuals are unaware that they are half siblings, it is unlikely that there exist meaningful and systemic environmental differences between such half siblings that are correlated with paternal genotype, so confounding from population stratification and/or dynastic effects is unlikely.

[10] For a more thorough discussion of tests of equal variance, see [30].

4.2 Empirical Application

The Wisconsin Longitudinal Study

The Wisconsin Longitudinal Study (WLS) is a survey based on a $\frac{1}{3}$ sample of all 1957 Wisconsin high school graduates and a randomly-selected sibling of these graduates (Herd, Carr, & Roan, 2014). The graduate respondents were originally empaneled with an in-person questionnaire at age 18 in 1957, which was followed with data collection at ages 25, 36, 54, 65, and finally 72 in 2012. The WLS includes a wide range of administrative and prospectively collected data from early life, adolescence, and early adulthood. Genetic samples were assayed from saliva for a subsample of consenting WLS graduates and siblings. Genotyping was performed using the Illumina HumanOmniExpress 24 BeadChip arrays (Version 1/1.1; Illumina). We restrict our analytic sample to only individuals of European ancestries, as polygenic score analyses in diverse ancestries are both methodologically [31] and conceptually [32] fraught.

Polygenic Scores

The polygenic scores used in this study are drawn from version 1.1 of the Social Science Genomics Association Consortium Polygenic Index Repository [16]. All polygenic scores are standardized over the full sample of genotyped WLS graduates. We removed 3 polygenic scores – attention deficit hyperactivity disorder, pollen allergy, and risk tolerance – from both our phenotype analyses (Figure 2) and mortality (Figure 3) analyses because these scores failed to statistically significantly predict their phenotype in between-family analyses. We further removed five sex-specific polygenic scores – age at first menses, age voice deepened, number ever born: men, and number ever born: women – from our mortality analyses, as the sample size of same-sex sibling pairs is too small to achieve adequate statistical power. A table describing which polygenic scores are utilized in each analysis presented in this study can be found in the SI.

Phenotype Variables

We generate phenotypes identically to Becker et al. 2021 [16]; see Supplementary Table 12 of that paper for a list of the specific WLS survey items used. Though the repository contains 47 distinct polygenic scores, in the WLS their only exists phenotype data for as subset of 30 traits. When only a single phenotypic measurement was available for a given trait, we residualized the phenotype on a second-degree polynomial of age, sex, and their interactions. When multiple measurements were available, for variables such as depression, we residualized on age, sex, and their interactions within each wave and that took the average for an individual across waves. For variables like educational attainment with multiple measurements, we took the maximum value across waves and then residualized on birth year, sex, and their interactions. All phenotypes are standardized over the full sample of genotyped WLS graduates.

Mortality Variables

The mortality data used in this study is derived from the National Death Index [33] – importantly, such data is available for all members of the WLS, regardless of how long they remained empaneled in the study. Mortality data was last collected by the WLS staff in 2018. We utilize two mortality variables; the first is lifespan in years, which is calculated as the difference between death date and birth date (for individuals who are still alive, we use 2018 as their death date). Although this mortality variable is right-censored, because genotype is quasi-randomly assigned within families, we would differences in polygenic scores between siblings will not be correlated with differences in birth date; therefore, genetic effect estimates will not be confounded by birth year as a result of this censoring. Our second mortality outcome is a dichotomous indicator for survival to age 75; this threshold was chosen both because of it's substantive meaning with respect to the phenomenon of premature mortality and also to maximize statistical power in our sample (e.g., an earlier dichotomous cutoff would decrease the fraction of deaths, but a later cutoff would exclude more sibling pairs who were born later and therefore have a restricted potential lifespan). Both variables

were residualized on sex and second-degree polynomial for age; note, it is useful to residualize outcomes before implementing phenotype differences regression (rather than including them as covariates, which may influence the conditional within-family correlation of genetic predictor).

References

[1] Abdel Abdellaoui, Loic Yengo, Karin JH Verweij, and Peter M Visscher. 15 years of gwas discovery: Realizing the promise. *The American Journal of Human Genetics*, 2023.

[2] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008.

[3] Arslan A Zaidi and Iain Mathieson. Demographic history mediates the effect of stratification on polygenic scores. *Elife*, 9:e61548, 2020.

[4] Abdel Abdellaoui, David Hugh-Jones, Loic Yengo, Kathryn E Kemper, Michel G Nivard, Laura Veul, Yan Holtz, Brendan P Zietsch, Timothy M Frayling, Naomi R Wray, et al. Genetic correlates of social stratification in great britain. *Nature human behaviour*, 3(12):1332–1342, 2019.

[5] Augustine Kong, Gudmar Thorleifsson, Michael L Frigge, Bjarni J Vilhjalmsson, Alexander I Young, Thorgeir E Thorgeirsson, Stefania Benonisdottir, Asmundur Oddsson, Bjarni V Halldorsson, Gisli Masson, et al. The nature of nurture: Effects of parental genotypes. *Science*, 359(6374):424–428, 2018.

[6] Sam Trejo and Benjamin W Domingue. Genetic nature or genetic nurture? introducing social genetic parameters to quantify bias in polygenic score analyses. *Biodemography and Social Biology*, 64(3-4):187–215, 2018.

[7] Laurence J Howe, Michel G Nivard, Tim T Morris, Ailin F Hansen, Humaira Rasheed, Yoonsu Cho, Geetha Chittoor, Rafael Ahlskog, Penelope A Lind, Teemu Palviainen, et al. Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nature genetics*, 54(5):581–592, 2022.

[8] Carl Veller and Graham Coop. Interpreting population and family-based genome-wide association studies in the presence of confounding. *bioRxiv*, pages 2023–02, 2023.

[9] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[10] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

[11] Alexander I Young, Seyed Moeen Nehzati, Stefania Benonisdottir, Aysu Okbay, Hariharan Jayashankar, Chanwook Lee, David Cesarini, Daniel J Benjamin, Patrick Turley, and Augustine Kong. Mendelian imputation of parental genotypes improves estimates of direct genetic effects. *Nature genetics*, 54(6):897–905, 2022.

[12] Ben Brumpton, Eleanor Sanderson, Karl Heilbron, Fernando Pires Hartwig, Sean Harrison, Gunnhild Åberge Vie, Yoonsu Cho, Laura D Howe, Amanda Hughes, Dorret I Boomsma, et al. Avoiding dynastic, assortative mating, and population stratification biases in mendelian randomization through within-family analyses. *Nature communications*, 11(1):3519, 2020.

[13] Benjamin W Domingue, Sam Trejo, Emma Armstrong-Carter, and Elliot M Tucker-Drob. Interactions between polygenic scores and environments: Methodological and conceptual challenges. *Sociological Science*, 7:465–486, 2020.

[14] Rebecca Johnson, Ramina Sotoudeh, and Dalton Conley. Polygenic scores for plasticity: a new tool for studying gene–environment interplay. *Demography*, 59(3):1045–1070, 2022.

[15] Pamela Herd, Deborah Carr, and Carol Roan. Cohort profile: Wisconsin longitudinal study (wls). *International journal of epidemiology*, 43(1):34–41, 2014.

[16] Joel Becker, Casper AP Burik, Grant Goldman, Nancy Wang, Hariharan Jayashankar, Michael Bennett, Daniel W Belsky, Richard Karlsson Linnér, Rafael Ahlskog, Aaron Kleinman, et al. Resource profile and user guide of the polygenic index repository. *Nature human behaviour*, 5(12):1744–1758, 2021.

[17] Paul RHJ Timmers, Ninon Mounier, Kristi Lall, Krista Fischer, Zheng Ning, Xiao Feng, Andrew D Bretherick, David W Clark, Xia Shen, et al. Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *elife*, 8:e39856, 2019.

[18] Joris Deelen, Daniel S Evans, Dan E Arking, Niccolò Tesi, Marianne Nygaard, Xiaomin Liu, Mary K Wojczynski, Mary L Biggs, Ashley van Der Spek, Gil Atzmon, et al. A meta-analysis of genome-wide association studies identifies multiple longevity genes. *Nature communications*, 10(1):3669, 2019.

[19] Riccardo E Marioni, Stuart J Ritchie, Peter K Joshi, Saskia P Hagenaars, Aysu Okbay, Krista Fischer, Mark J Adams, W David Hill, Gail Davies, Social Science Genetic Association Consortium, et al. Genetic variants linked to education predict longevity. *Proceedings of the National Academy of Sciences*, 113(47):13366–13371, 2016.

[20] Richard Karlsson Linnér and Philipp D Koellinger. Genetic risk scores in life insurance underwriting. *Journal of Health Economics*, 81:102556, 2022.

[21] Jason M Fletcher and Steven F Lehrer. Genetic lotteries within families. *Journal of health economics*, 30(4):647–659, 2011.

[22] Kathryn Paige Harden. *The genetic lottery: why DNA matters for social equality*. Princeton University Press, 2021.

[23] Sam Trejo and Daphne Oluwaseun Martschenko. Beware the phony horserace between genes and environments. *Behavioral and Brain Sciences*, 2023.

[24] Sam Trejo and Daphne Oluwaseun Martschenko. Expanding the conceptual boundaries of social and behavioral genomics. *Under review*, 2023.

[25] Ellen L Idler and Yael Benyamini. Self-rated health and mortality: a review of twenty-seven community studies. *Journal of health and social behavior*, pages 21–37, 1997.

[26] Marja Jylhä. What is self-rated health and why does it predict mortality? towards a unified conceptual model. *Social science & medicine*, 69(3):307–316, 2009.

[27] Sarah E Harris, Saskia P Hagenaars, Gail Davies, W David Hill, David CM Liewald, Stuart J Ritchie, Riccardo E Marioni, International Consortium for Blood Pressure Genome-Wide Association Studies METASTROKE Consortium, International Consortium for Blood Pressure Genome-Wide Association Studies, CHARGE Consortium Aging, Longevity Group, et al. Molecular genetic contributions to self-rated health. *International journal of epidemiology*, 46(3):994–1009, 2017.

[28] Laurie H Seaver, George Khushf, Nancy MP King, Dena R Matalon, Kunal Sanghavi, Matteo Vatta, Kristi Wees, ACMG Social, Ethical, Legal Issues Committee, et al. Points to consider to avoid unfair discrimination and the misuse of genetic information: A statement of the american college of medical genetics and genomics (acmg), 2022.

[29] Morton B Brown and Alan B Forsythe. Robust tests for the equality of variances. *Journal of the American statistical association*, 69(346):364–367, 1974.

[30] Ralph G O'Brien. Robust techniques for testing heterogeneity of variance effects in factorial designs. *Psychometrika*, 43(3):327–342, 1978.

[31] Yi Ding, Kangcheng Hou, Ziqi Xu, Aditya Pimplaskar, Ella Petter, Kristin Boulier, Florian Privé, Bjarni J Vilhjálmsson, Loes M Olde Loohuis, and Bogdan Pasaniuc. Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature*, pages 1–8, 2023.

[32] Jeremy Freese, Ben Domingue, Sam Trejo, Kamil Sicinski, and Pamela Herd. Problems with a causal interpretation of polygenic score differences between jewish and non-jewish respondents in the wisconsin longitudinal study. 2019.

[33] National Center for Health Statistics et al. What is the ndi?, 1999.

[34] Carsten Boecker Pedersen, Jonas Bybjerg-Grauholm, Marianne Gioertz Pedersen, Jakob Grove, Esben Agerbo, Marie Baekvad-Hansen, Jesper Buchhave Poulsen, Christine Soeholm Hansen, John J McGrath, Thomas D Als, et al. The ipsych2012 case–cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Molecular psychiatry*, 23(1):6–14, 2018.

FIGURES

Mortality Selection into Genotyping

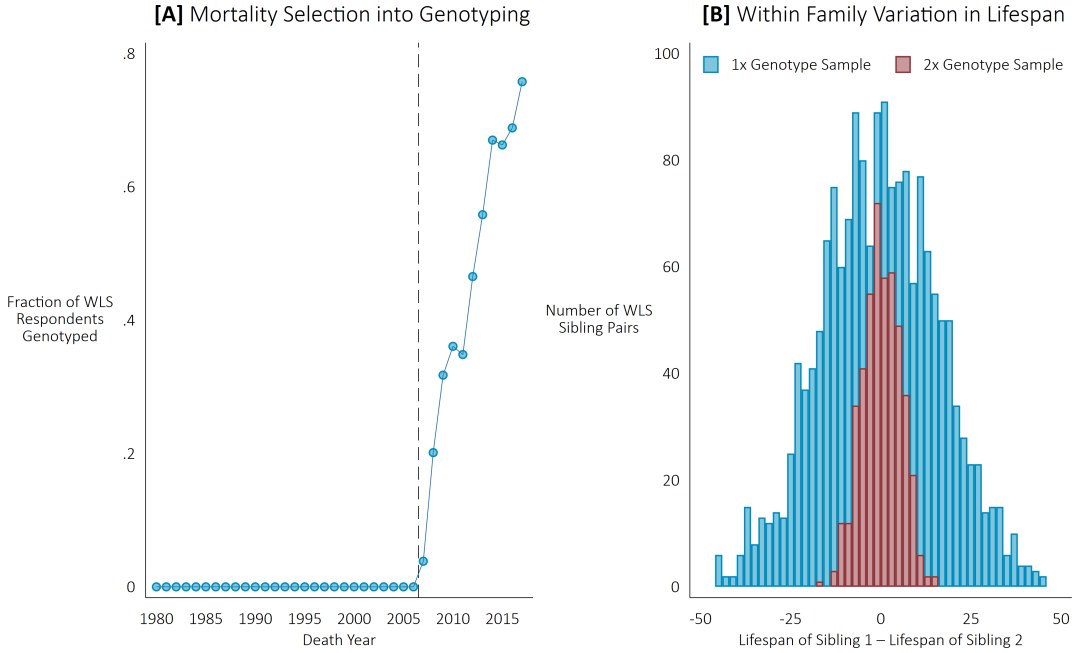

Figure 1: This figure shows the impact of mortality selection on the sample of genotype individuals in the WLS. Panel A displays the fraction of WLS respondents (graduates and randomly selected siblings) who have valid genotypic data as a function of their year of death, with a vertical line drawn at 2006. Each marker represents one year and contains, on average, 500 individuals. Panel B displays a histogram showing the within-family variation in lifespan for the One Genotype Sample (blue bars) and Two Genotypes Sample (red bars). Which sibling is Sibling 1 vs. Sibling 2 is randomly assigned.

Comparing Fixed Effects to Phenotype Differences Estimates

Figure 2: This figure compares effects of various polygenic scores on their target phenotype estimated using fixed effects (FE) and phenotype differences (PD). All panels display the same FE estimates, which are derived from the full Two Genotypes Sample. In Panels A and B, the PD estimates come from a procedure using the Two Genotypes Sample in which the genetic data of a randomly selected sibling in each pair is discarded; Panel A displays the mean from 1000 iterations whereas Panel B displays the estimates from a single iteration. Panel C displays PD estimates derived from the One Genotype Sample (phenotypes with insufficient sample size in this sample are removed). Both polygenic scores and phenotypes are standardized over the full sample of genotyped WLS graduates. A table containing a list of names and labels of all the phenotypes use in this figure can be found in the SI.

Effects of Polygenic Indices on Mortality

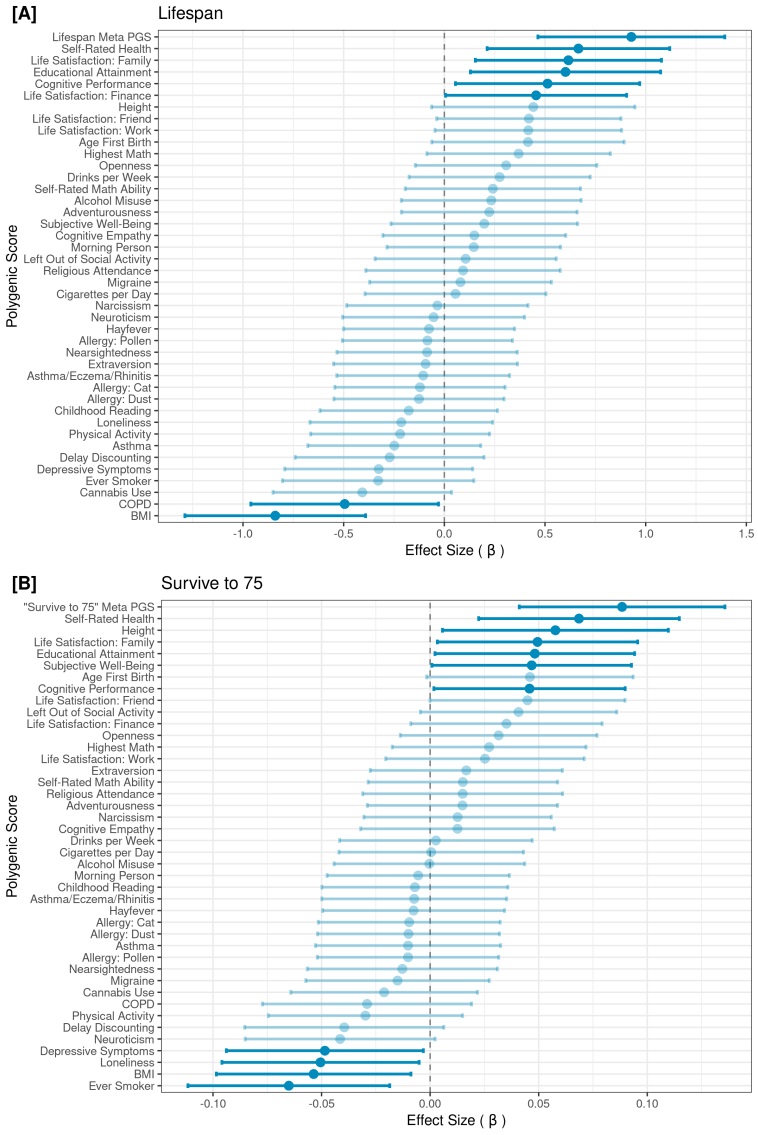

Figure 3: This figure displays estimates of the causal effect of 40 polygenic scores on premature mortality outcomes. Panel A displays effects on lifespan in years, and Panel B displays the effects on the probability of survival to age 75.

TABLES

Wisconsin Longitudinal Study Sibling Data

Panel A. Two Genotypes Sample.						
	Graduate			Not Graduate		
	Mean	SD	N	Mean	SD	N
Female	0.52	0.50	2088	0.53	0.50	2088
Birth Year	1939.41	0.46	2088	1941.18	6.82	2088
Deceased by 2018	0.12	0.32	2088	0.11	0.32	2088
Deceased by Age 75	0.06	0.24	2088	0.07	0.25	1346
Lifespan	78.52	1.86	2088	76.78	6.62	2088
Panel B. One Genotype Sample.						
	Genotyped			Not Genotyped		
	Mean	SD	N	Mean	SD	N
Graduate	0.73	0.44	3548	0.27	0.44	3548
Female	0.51	0.50	3548	0.48	0.50	3548
Birth Year	1939.84	3.49	3548	1941.15	7.25	3548
Deceased by 2018	0.12	0.33	3548	0.41	0.49	3548
Deceased by Age 75	0.07	0.25	3218	0.46	0.50	2686
Lifespan	78.03	3.78	3548	70.54	10.32	3548

Table 1: This table uses data from the Wisconsin Longitudinal Study. Graduates are the original members of the WLS, who graduated from high school in 1957. Later, a randomly selected sibling of each graduate was empaneled into the study. The lifespan variable is right-censored, as the last collection of data from the National Death Index was in 2018.

This table describes various potential applications of the phenotype differences model.

Potential Application	**Description**
Collecting Sibling Reported Phenotypes	Imagine researchers decide to expand an existing data repository, like the UK biobank [9], to increase the number of genotyped pairs of first-degree relatives available for causal genomic analyses. However, many phenotypes, like height and educational attainment, are easily reported. Rather than empanel, interview, and genotype multiple members of the same family, it is likely more cost-effective to instead ask the already empaneled members (or an entirely new sample of unrelated individuals) to provide the phenotypes of all or a randomly selected sibling. One does not necessarily have to choose between the two strategies – both types of data can be simultaneously collected and utilized). Importantly, the phenotype differences model provides estimates that are robust to asymmetric bias and classical measurement error (e.g., reporting a sibling's height systemically lower or less accurately than one's own height), so long as such biases are not meaningfully genetically caused.
Leveraging Administrative Data	While collecting large genotyped sampled of siblings when studying common and easily measured phenotypes is a difficult task, it is even more challenging when studying rare and highly-sensitive phenotypes, such as severe mental disorders. Case-control studies, such as iPSYCH [34], have achieved the near-herculean task of assembling a sufficient sample size for molecular genetic analyses, but it is currently difficult to integrate such data sources into a within-family causal framework. However, using the phenotype differences model, the same administrative data that was used to create iPSYCH could be leveraged to collect the phenotypes of siblings without the need to contact participants or to obtain consent for the use of additional biological assays. The use of administrative data is particularly appealing, because utilizing phenotypes of *all* siblings (rather than just a single sibling) increases precision of the the phenotype differences model.
Accounting for Sample Selection	As we have done in this study using the WLS data, the phenotype differences model is useful for addressing selection into genotyping. In our case, such selection resulted largely from premature mortality, though other forms of selection into genotyping, such as an individual's concerns related to privacy or trust in research institutions, may exist.

Table 2: Three potential applications of the phenotype differences model

SUPPLEMENTARY INFORMATION

Trejo & Kanopka 2023

The Phenotype Differences Model Reveals
Genetic Effects on Mortality Using Incomplete Sibling Data

Contents

1 SUPPLEMENTARY FIGURES

1.1 Measurement Error-Corrected Effects of Polygenic Scores on Mortality

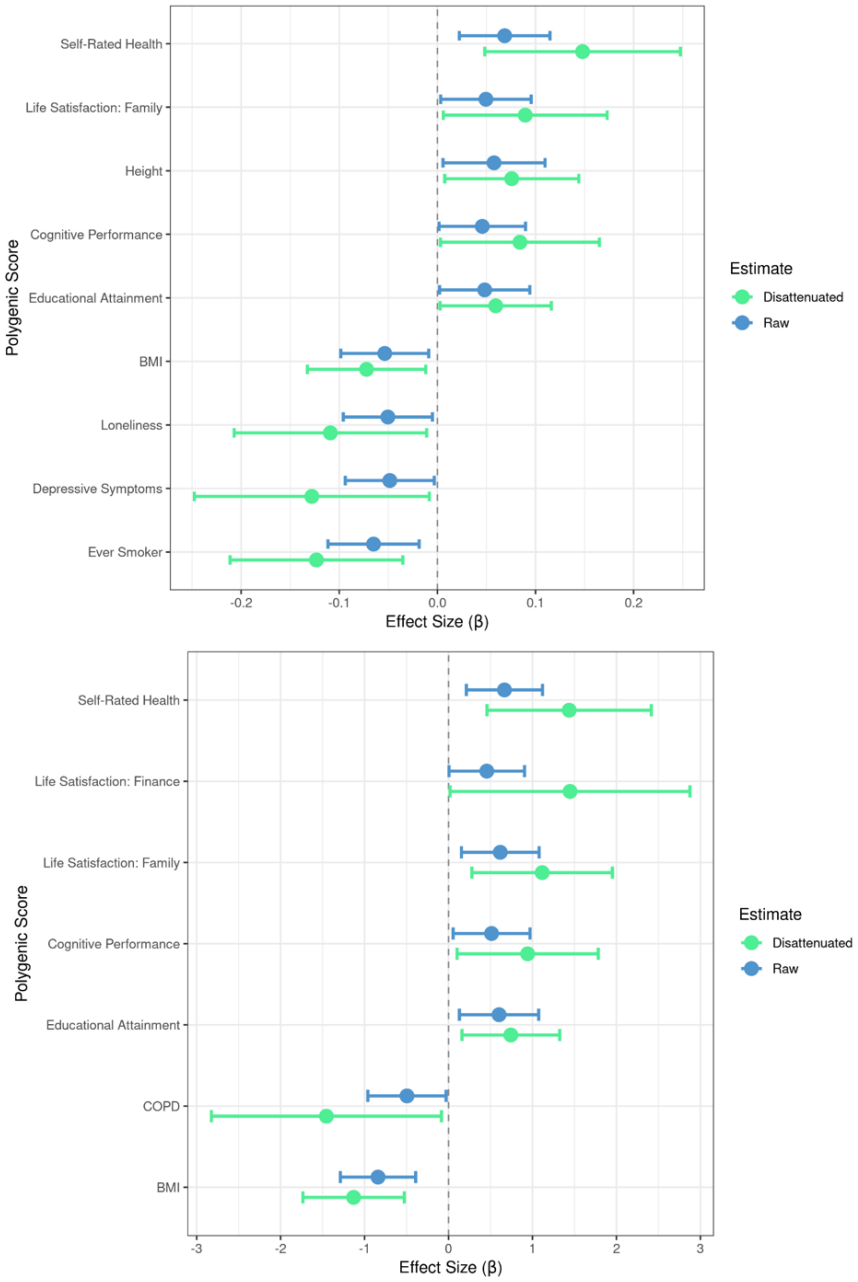

Figure S1: This figure displays estimates of the causal effect of various polygenic scores on premature mortality outcomes. Both raw estimates and estimates that are deattenuated for measurement error are shown. Such measurement error in polygenic scores results from the finite GWAS sample used to estimate the underlying allelic weights. We implemented measurement error correction following the procedure outlined in Becker et al. 2021. Panel A displays effects on lifespan in years, and Panel B displays the effects on the probability of survival to age 75.

1.2 Comparative Precision

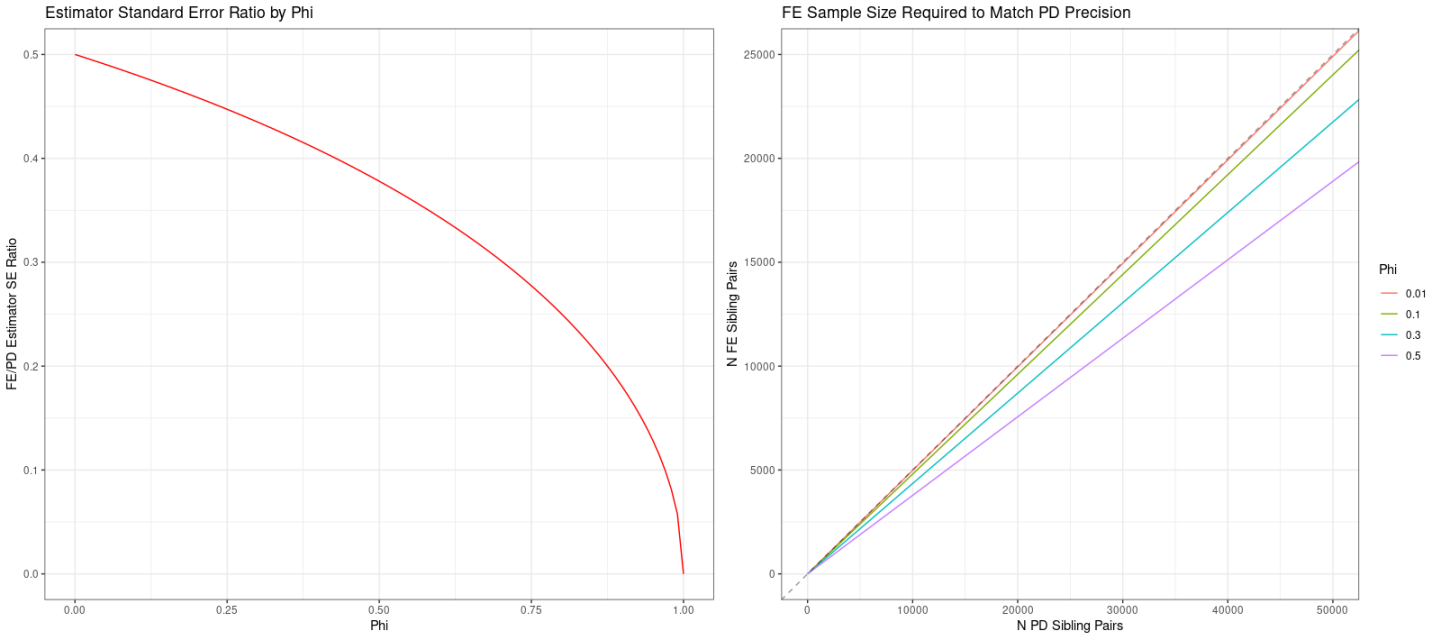

Figure S2: This figure displays the comparative precision of the fixed effects and phenotype differences estimators. The left panel shows the asymptotic ratio of the standard errors of the fixed effects and phenotype differences estimator for a given number of sibling pairs as a function of ϕ, the fraction of within-family outcome variation explained by the genetic predictor. The right panel shows the comparative size of a phenotype differences sample required to match the precision estimates from a fixed effects sample; for values of ϕ near zero, phenotype differences requires double the sibling pairs (i.e., the same number of genotypes) as fixed effects to achieve the equal precision.

2 SUPPLEMENTARY TABLES

2.1 List of Polygenic Scores

Trait	Label	Phenotype Analyses	Mortality Analyses	Lifespan Meta-PGI	Survive to 75 Meta-PGI
Adventurousness	adv		X		
Age at First Birth	birth	X	X		
Age at First Menses	menses	X			
Age Voice Deepened	deep				
Attention Deficit Hyperactivity Disorder	adhd				
Allergy - Cat	cat	X	X		
Allergy - Dust	dust	X	X		
Allergy - Pollen	pollen				
Asthma/Eczema/Rhinitis	aer	X	X		
Asthma	asthma	X	X		
Alcohol Misuse	alcoh	X	X		
Body Mass Index	bmi	X	X	-.23	-.18
Cannabis Use	canna		X		
Cognitive Empathy	cog_emp		X		
Childhood Reading	read		X		
Chronic Obstructive Pulmonary Disease	copd	X	X	-.11	
Cigarettes per Day	cig_day	X	X		
Cognitive Performance	cog	X	X	.04	.02
Delay Discounting	dly_disc		X		
Depressive Symptoms	dep	X	X		-.11
Drinks per Week	drinks	X	X		
Educational Attainment	edu	X	X	.08	.06
Ever Smoker	ever_smk	X	X		-.24
Extraversion	extra	X	X		
Life Satisfaction - Family	sat_fam	X	X	.22	.12
Life Satisfaction - Finance	sat_fin	X	X	.14	
Life Satisfaction - Friend	sat_frnd		X		
Life Satisfaction - Work	sat_job	X	X		
Hayfever	hay	X	X		
Height	hgt	X	X		.01
Highest Math	high_math		X		
Left Out of Social Activity	left_out		X		
Loneliness	lonely	X	X		-.12
Migraine	migrn		X		
Morning Person	chrono		X		
Narcissism	narci		X		
Nearsightedness	near_sgt		X		
Number Ever Born - Men	neb_male	X			
Number Ever Born - Women	neb_fem	X			
Neuroticism	neuro	X	X		
Openness	open	X	X		
Physical Activity	phys_act	X	X		
Religious Attendance	relig	X	X		
Risk Tolerance	risk				
Self-Rated Health	health	X	X	.18	.05
Self-Rated Math Ability	self_math		X		
Subjective Well-Being	swb	X	X		.10

Table S1: This table provides a list of the 47 polygenic scores derived from the SSGAC repository, as well as which of our WLS analyses each score was included in. If a phenotype contributed to the meta-polygenic score for a either lifespan or survival to 75, its weight is also included; the absolute value of meta-polygenic score weights for a given premature mortality outcome sum to 1.

2.2 Observed Sibling Correlations for 47 Polygenic Scores

pgi_phys_act	pgi_bmi	pgi_canna	pgi_cig_day	pgi_edu	pgi_ever_smk
0.512	0.500	0.493	0.459	0.526	0.542
(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.018)
pgi_hgt	pgi_migrn	pgi_chrono	pgi_narci	pgi_near_sgt	pgi_open
0.620	0.487	0.507	0.504	0.493	0.544
(0.017)	(0.019)	(0.019)	(0.019)	(0.019)	(0.018)
pgi_read	pgi_adhd	pgi_adv	pgi_birth	pgi_cat	pgi_dust
0.504	0.540	0.501	0.550	0.494	0.495
(0.019)	(0.018)	(0.019)	(0.018)	(0.019)	(0.019)
pgi_pollen	pgi_aer	pgi_asthma	pgi_alcoh	pgi_cog_emp	pgi_copd
0.494	0.501	0.502	0.504	0.524	0.564
(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.018)
pgi_cog	pgi_dly_disc	pgi_dep	pgi_drinks	pgi_extra	pgi_sat_job
0.496	0.520	0.523	0.508	0.498	0.531
(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
pgi_sat_fin	pgi_sat_fam	pgi_sat_frnd	pgi_hay	pgi_high_math	pgi_left_out
0.510	0.537	0.531	0.486	0.506	0.535
(0.019)	(0.018)	(0.019)	(0.019)	(0.019)	(0.019)
pgi_lonely	pgi_menses	pgi_neb_male	pgi_neb_fem	pgi_neuro	pgi_relig
0.537	0.534	0.534	0.536	0.509	0.522
(0.018)	(0.019)	(0.019)	(0.018)	(0.019)	(0.019)
pgi_risk	pgi_health	pgi_self_math	pgi_swb	pgi_deep	
0.501	0.548	0.507	0.539	0.522	
(0.019)	(0.018)	(0.019)	(0.018)	(0.019)	
N=2088 Sibling Pairs					

Table S2: Empirical sibling correlations ($\rho^{G1,G2}$) for 47 polygenic scores from the WLS Two Genotypes Sample. Sibling 1 vs. Sibling 2 are randomly assigned across 1000 repetitions and we take the average correlation and variance (this variance is then converted to a standard error). Deviations from $\rho^{G1,G2} = 0.5$ suggest the existence of assortative mating.

2.3 Testing the Equal Variance Assumption Using Phenotypic Data

	Not Genotyped		**Genotyped**		**Ratio**	***p*-value**
	Var_2	N_2	Var_1	N_1	$\frac{Var\_2}{Var\_1}$	
Body Mass Index	0.96	1728	1.01	3394	0.95	0.34
Height	1.10	1038	0.87	3279	1.26	0.20
Cognitive Ability	1.10	2881	1.01	3397	1.08	0.088
Years of Schooling	1.13	2432	1.12	3528	1.01	0.44
Extroversion	1.07	1808	0.97	3426	1.11	0.098
Neuroticism	1.06	1804	0.96	3425	1.10	0.046
Openness to Experience	0.98	1804	0.93	3423	1.06	0.30

Table S3: This table displays p-values from a Brown–Forsythe test for equal variance across the genotype and ungenotyped sibling for 7 approximately continuously distributed phenotypes in the WLS One Genotype Sample.

2.4 Descriptive Statistics of Sub-Sample Used in Mortality Analyses

Panel A. Individual Characteristics.						
	No Deaths			1 or 2 Deaths		
	Mean	SD	N	Mean	SD	N
Female	0.53	0.50	6890	0.47	0.50	4382
Birth Year	1941.08	5.19	6890	1939.38	5.67	4382
Deceased by 2018	0.00	0.00	6890	0.54	0.50	4382
Survived to Age 75	1.00	0.00	5358	0.58	0.49	3980
Lifespan*	77.42	5.19	6890	72.56	9.69	4382
Panel B. Sibling Pair Characteristics.						
	No Deaths			1 or 2 Deaths		
	Mean	SD	N	Mean	SD	N
$\lvert \text{Female}_1 - \text{Female}_2 \rvert$	0.50	0.50	3445	0.50	0.50	2191
$\lvert \text{Birth Year}_1 - \text{Birth Year}_2 \rvert$	6.20	4.56	3445	6.52	4.61	2191
$\lvert \text{Survived to Age 75}_1 - \text{Survived to Age 75}_2 \rvert$	0.00	0.00	1913	0.69	0.46	1789
$\lvert \text{Lifespan}_1 - \text{Lifespan}_2 \rvert$	6.20	4.56	3445	11.92	9.79	2191
Two Genotype Sample	0.47	0.50	3445	0.21	0.41	2191
Non-Missing Survived to Age 75_1 & Survived to Age 75_2	0.56	0.50	3445	0.82	0.39	2191

Table S4: This table uses data from the Wisconsin Longitudinal Study. Sibling pairs with 1 or 2 deaths by 2018 comprise the analytic sample for our premature mortality analyses.

3 DERIVATIONS

3.1 Random Variables & Causal Model

In our data, we observe individuals i nested in families j. In particular, we observe one pair of siblings, $i = 1$ and $i = 2$, in each family j from $j = 1, ..., N$. In the notation of the fixed effects portion of the panel data literature, this entails $T = 2$. We call our outcome of interest y_{ij}.

Let g_{ij} be a generic genetic predictor derived from an individual's genome. This genetic predictor g_{ij} can be either an allele count at a particular genetic locus (e.g., the independent variable of interest in a GWAS) or a genome-wide summary measure such as a polygenic score. Thus, the causal effect of g_{ij} on y_{ij} is a direct genetic effect. Empirically observed values of g_{1j} and g_{2j} are draws of random variables $G1$ and $G2$, respectively. $G1$ is distributed with mean μ_1 and some non-zero and finite variance σ_1^2, and $G2$ is distributed with mean μ_2 and some non-zero and finite variance σ_2^2.

Let g_j be the sum of this same genetic predictor derived from the mother in family j's genome and the genetic predictor derived from the father in family j's genome. The causal effect of g_j on y_{ij} is a genetic nurture effect.

Let e_{ij} be an unobserved measure of individual-level environmental influences on y_{ij}. Although unobserved, e_{1j} and e_{2j} are draws of random variables $E1$ and $E2$, respectively. $E1$ is distributed with mean π_1 and non-zero and finite variance ω_1^2, and $E2$ is distributed with mean π_2 and non-zero and finite variance ω_2^2.

Finally, let e_j be an unobserved measure of family-level environmental influences on y_{ij}.

y_{ij}: Outcome (i.e. phenotype) of individual i in family j
g_{ij}: Genetic predictor of individual i in family j
g_j: Genetic predictor of mother in family j + genetic predictor of father in family j
e_{ij}: Individual-level environmental effects for individual i in family j
e_j: Family-level environmental effects for family j

We allow our outcome of interest y_{ij} be a function of linear direct genetic effects, linear genetic nurture effects, and flexible individual-level and family-level environmental effects. This underlying causal model is displayed in in Equation 3.1.

$$y_{ij} = \alpha + \beta g_{ij} + \delta g_j + e_{ij} + e_j \tag{3.1}$$

α: Intercept
β: Magnitude of direct genetic effects
δ: Magnitude of genetic nurture effects

3.2 Within Family Models

Fixed effects models that compare genetic differences in siblings to phenotypic differences in siblings have a common approach for the identification of direct genetic effects. The fixed effects approach involves fitting a linear regression with a unique intercept for each j, as a displayed in Equation 3.2.

$$y_{ij} = \hat{\tau}_j + \hat{\beta}^{\mathrm{FE}} g_{ij} + \hat{\varepsilon}^*_{ij} \tag{3.2}$$

$\hat{\beta}^{\mathrm{FE}}$: Estimated direct genetic effect from fixed effects model

The fixed effects model displayed in Equation 3.2 can be identically expressed by demeaning the dependent and independent variables using the *within* transformation, as shown in Equation 3.3.

$$y_{ij} - \bar{y_j} = \hat{\beta}^{\mathrm{FE}}(g_{ij} - \bar{g_j}) + \hat{\varepsilon}^*_{ij} \tag{3.3}$$

Because there are exactly two individuals i in each family j (i.e. $T = 2$), the fixed effects model from Equation 3.3 is equivalent to the *first differences* model displayed in Equation 3.4. The first differences model involves fitting a linear regression of $y_{1j} - y_{2j}$, the difference in outcomes between siblings, on $g_{1j} - g_{2j}$, the differences in genetic predictors between siblings.

$$y_{1j} - y_{2j} = \hat{\beta}^{\mathrm{FE}}(g_{1j} - g_{2j}) + \hat{\varepsilon}^*_{ij} \tag{3.4}$$

In this paper, we introduce the *phenotype differences* model. The phenotype differences model, shown in Equation 3.5, involves fitting a linear regression of $y_{1j} - y_{2j}$, the differences in outcomes between siblings, on $g_{1j}(1-\rho)$, the genetic predictor of a single sibling multiplied by one minus the correlation of genetic predictors within families.

$$y_{1j} - y_{2j} = \hat{\alpha} + \hat{\beta}^{\mathrm{PD}}\Big(g_{1j}(1 - \rho^{G1,G2})\Big) + \hat{\varepsilon}_{ij} \tag{3.5}$$

$$\text{Where} \qquad \rho^{G1,G2} = \frac{cov(G1, G2)}{var(G1)^{\frac{1}{2}} G2)^{\frac{1}{2}}}$$

$\hat{\beta}^{\text{PD}}$: Estimated direct genetic effect from phenotype differences model
$\rho^{G1,G2}$: Correlation between the genetic predictor of individual 1 and individual 2 in each family j

When individual 2 is randomly selected from of all individual 1's brothers and sisters and mating is random, genetic recombination causes siblings to share on average half of their genomes. In such a case, $\rho^{G1,G2} = \frac{1}{2}$ and the phenotype differences model specializes to Equation 3.6.

$$y_{1j} - y_{2j} = \hat{\alpha} + \hat{\beta}^{\text{PD}} \frac{g_{1j}}{2} + \hat{\varepsilon}_{ij} \tag{3.6}$$

In cases where $\rho^{G1,G2} \neq \frac{1}{2}$, an estimate of $\rho^{G1,G2}$ can be obtained using a representative sub-sample fully genotyped of siblings pairs. Estimates of $\rho^{G1,G2}$ may also be obtained using a representative sample of parent-child pairs. Alternatively, one could infer $\rho^{G1,G2}$ using population-level estimates of genetic assortative mating.

3.3 Assumptions of Phenotype Differences

We show that the phenotype differences model provides unbiased and consistent estimates of β, the true causal effect of g_{ij} on y_{ij}, when two assumptions hold. The first assumption, which is also required for fixed effects models, is the random assignment of genotype within families. That is, we must assume that an individual's genetic predictor is uncorrelated with his or her individual-level environment and the individual-level environment of his or her siblings. This assumption is displayed mathematically in Equation 3.7.

$$cov(G1, E1) = cov(G1, E2) = cov(G2, E1) = cov(G2, E2) = 0 \tag{3.7}$$

The second assumption of phenotype differences is that the population variance of individual 1's genetic predictor is the equal to the population variance of individual 2's genetic predictor. That is, unobserved genetic predictor has the same variance as the observed genetic predictor (which also implies that the unobserved genetic predictor also have the same standard deviation of the observed genetic predictor). This assumption is displayed mathematically in Equation 3.8.

$$\sigma_1^2 = \sigma_2^2 \text{ and, therefore, } \sigma_1 = \sigma_2 \tag{3.8}$$

3.4 Unbiasedness of Phenotype Differences

In this section, we show that the phenotype differences method obtains unbiased estimates of the causal effect of g_{ij} on y_{ij}. However, as is common in the empirical literature, we first standardize g_{1j} within-sample:

$$\tilde{g}_{1j} = \frac{g_{1j} - \bar{g_1}}{var(g_{1j})^{\frac{1}{2}}}$$

We therefore aim to show that phenotype differences is an unbiased estimator of the *standardized* effect; that is, the expected value of $\hat{\beta}^{PD}$ is equal to $\beta\sigma_1$. Note that, by construction, $var(\tilde{g}_{1j}) = 1$ and $\rho^{\tilde{G1},G1} = 1 \Rightarrow \rho^{\tilde{G1},G2} = \rho^{G1,G2}$.

Expected Value of $\hat{\beta}^{PD}$

$$\hat{\beta}^{\text{PD}} = \frac{cov\Big(\tilde{g}_{1j}(1-\rho^{G1,G2}), y_{1j} - y_{2j}\Big)}{var\Big(\tilde{g}_{1j}(1-\rho^{G1,G2})\Big)} \qquad \text{OLS coefficient formula}$$

$$\hat{\beta}^{\text{PD}} = \frac{(1-\rho^{G1,G2})cov(\tilde{g}_{1j}, y_{1j} - y_{2j})}{var\Big(\tilde{g}_{1j}(1-\rho^{G1,G2})\Big)} \qquad cov(yA, zB) = (yz)cov(A, B)$$

$$\hat{\beta}^{\text{PD}} = \frac{(1-\rho^{G1,G2})cov(\tilde{g}_{1j}, y_{1j} - y_{2j})}{(1-\rho^{G1,G2})^2 var(\tilde{g}_{1j})} \qquad var(zA) = z^2 var(A)$$

$$\hat{\beta}^{\text{PD}} = \frac{(1-\rho^{G1,G2})cov(\tilde{g}_{1j}, y_{1j} - y_{2j})}{(1-\rho^{G1,G2})^2} \qquad var(\tilde{g}_{1j}) = 1$$

$$\hat{\beta}^{\text{PD}} = \frac{1}{1-\rho^{G1,G2}} \times cov(\tilde{g}_{1j}, y_{1j} - y_{2j}) \qquad \text{Simplify}$$

$$\hat{\beta}^{\text{PD}} = \frac{1}{1-\rho^{G1,G2}} \times cov\Big(\tilde{g}_{1j}, \beta(g_{1j} - g_{2j}) + (e_{1j} - e_{2j})\Big) \qquad \text{Substitute in Equation ??}$$

$$\hat{\beta}^{\text{PD}} = \frac{1}{1-\rho^{G1,G2}} \times \Big(cov\big(\tilde{g}_{1j}, \beta(g_{1j} - g_{2j})\big) + cov(\tilde{g}_{1j}, e_{1j} - e_{2j})\Big) \qquad cov(A, B+C) = cov(A, B) + cov(A, C)$$

$$\hat{\beta}^{\text{PD}} = \frac{1}{1-\rho^{G1,G2}} \times \Big(cov\big(\tilde{g}_{1j}, \beta(g_{1j} - g_{2j})\big) + cov(\tilde{g}_{1j}, e_{1j}) - cov(\tilde{g}_{1j}, e_{2j})\Big) \qquad cov(A, B-C) = cov(A, B) - cov(A, C)$$

$$\hat{\beta}^{\text{PD}} = \frac{1}{1-\rho^{G1,G2}} \times \Big(\beta cov\big(\tilde{g}_{1j}, (g_{1j} - g_{2j})\big) + cov(\tilde{g}_{1j}, e_{1j}) - cov(\tilde{g}_{1j}, e_{2j})\Big) \qquad cov(yA, zB) = (yz)cov(A, B)$$

$$\hat{\beta}^{\text{PD}} = \frac{1}{1-\rho^{G1,G2}} \times \Big(\beta\big(cov(\tilde{g}_{1j}, g_{1j}) - cov(\tilde{g}_{1j}, g_{2j})\big) + cov(\tilde{g}_{1j}, e_{1j}) - cov(\tilde{g}_{1j}, e_{2j})\Big) \qquad cov(A, B-C) = cov(A, B) - cov(A, C)$$

$$\mathbb{E}[\hat{\beta}^{\text{PD}}] = \mathbb{E}\left[\frac{1}{1-\rho^{G1,G2}} \times \Big(\beta\big(cov(\tilde{g}_{1j}, g_{1j}) - cov(\tilde{g}_{1j}, g_{2j})\big) + cov(\tilde{g}_{1j}, e_{1j}) - cov(\tilde{g}_{1j}, e_{2j})\Big)\right] \qquad \text{Evaluate Expectation}$$

$$\mathbb{E}[\hat{\beta}^{\text{PD}}] = \frac{1}{1-\rho^{G1,G2}}\,\mathbb{E}\left[\beta\Big(cov(\tilde{g}_{1j}, g_{1j}) - cov(\tilde{g}_{1j}, g_{2j})\Big) + cov(\tilde{g}_{1j}, e_{1j}) - cov(\tilde{g}_{1j}, e_{2j})\right] \qquad \mathbb{E}[zA] = z\,\mathbb{E}[A]$$

$$\mathbb{E}[\hat{\beta}^{\mathrm{PD}}] = \frac{1}{1-\rho^{G1,G2}} \times \left(\mathbb{E}\left[\beta\Big(cov(\tilde{g}1j, g_{1j}) - cov(\tilde{g}_{1j}, g_{2j})\Big)\right] + \mathbb{E}\left[cov(\tilde{g}_{1j}, e_{1j})\right] - \mathbb{E}\left[cov(\tilde{g}_{1j}, e_{2j})\right]\right) \qquad \mathbb{E}[A+B] = \mathbb{E}[A] + \mathbb{E}[B]$$

$$\mathbb{E}[\hat{\beta}^{\mathrm{PD}}] = \frac{1}{1-\rho^{G1,G2}} \times \left(\mathbb{E}\left[\beta\Big(cov(\tilde{g}_{1j}, g_{1j}) - cov(\tilde{g}_{1j}, g_{2j})\Big)\right] + \mathbb{E}\left[\rho^{\tilde{g}_{1j},e_{1j}} var(\tilde{g}_{1j})^{\frac{1}{2}} var(e_{1j})^{\frac{1}{2}}\right] - \mathbb{E}\left[\rho^{\tilde{g}_{1j},e_{2j}} var(\tilde{g}_{1j})^{\frac{1}{2}} var(e_{2j})^{\frac{1}{2}}\right]\right) \qquad cov(A,B) = \rho^{A,B} var(A)^{\frac{1}{2}} var(B)^{\frac{1}{2}}$$

$$\mathbb{E}[\hat{\beta}^{\mathrm{PD}}] = \frac{1}{1-\rho^{G1,G2}} \times \left(\mathbb{E}\left[\beta\Big(cov(\tilde{g}_{1j}, g_{1j}) - cov(\tilde{g}_{1j}, g_{2j})\Big)\right] + \mathbb{E}\left[\rho^{\tilde{g}_{1j},e_{1j}} var(e_{1j})^{\frac{1}{2}}\right] - \mathbb{E}\left[\rho^{\tilde{g}_{1j},e_{2j}} var(e_{2j})^{\frac{1}{2}}\right]\right) \qquad var(\tilde{g}_{1j}) = 1$$

$$\mathbb{E}[\hat{\beta}^{\mathrm{PD}}] = \frac{1}{1-\rho^{G1,G2}} \times \left(\mathbb{E}\left[\beta\Big(cov(\tilde{g}_{1j}, g_{1j}) - cov(\tilde{g}_{1j}, g_{2j})\Big)\right] + \mathbb{E}[\rho^{\tilde{g}_{1j},e_{1j}}]\,\mathbb{E}\left[var(e_{1j})^{\frac{1}{2}})\right] - \mathbb{E}[\rho^{\tilde{g}_{1j},e_{2j}}]\,\mathbb{E}\left[var(e_{2j})^{\frac{1}{2}})\right]\right) \qquad A \perp\!\!\!\perp B \Rightarrow \mathbb{E}[AB] = \mathbb{E}[A]\,\mathbb{E}[B]$$

$$\mathbb{E}[\hat{\beta}^{\mathrm{PD}}] = \frac{1}{1-\rho^{G1,G2}} \times \left(\mathbb{E}\left[\beta\Big(cov(\tilde{g}_{1j}, g_{1j}) - cov(\tilde{g}_{1j}, g_{2j})\Big)\right] + \rho^{\tilde{G}1,E1}\,\mathbb{E}\left[var(e_{1j})^{\frac{1}{2}})\right] - \rho^{\tilde{G}1,E2}\,\mathbb{E}\left[var(e_{2j})^{\frac{1}{2}})\right]\right) \qquad \text{Evaluate Expectation}$$

$$\mathbb{E}[\hat{\beta}^{\mathrm{PD}}] = \frac{1}{1-\rho^{G1,G2}} \times \left(\mathbb{E}\left[\beta\Big(cov(\tilde{g}_{1j}, g_{1j}) - cov(\tilde{g}_{1j}, g_{2j})\Big)\right] + 0\,\mathbb{E}\left[var(e_{1j})^{\frac{1}{2}})\right] - 0\,\mathbb{E}\left[var(e_{2j})^{\frac{1}{2}})\right]\right) \qquad \text{Substitute in Equation 3.7}$$

$$\mathbb{E}[\hat{\beta}^{\mathrm{PD}}] = \frac{1}{1-\rho^{G1,G2}} \times \mathbb{E}\left[\beta\Big(cov(\tilde{g}_{1j}, g_{1j}) - cov(\tilde{g}_{1j}, g_{2j})\Big)\right] \qquad \text{Simplify}$$

$$\mathbb{E}[\hat{\beta}^{\mathrm{PD}}] = \frac{\beta}{1-\rho^{G1,G2}} \times \mathbb{E}\left[cov(\tilde{g}_{1j}, g_{1j}) - cov(\tilde{g}_{1j}, g_{2j})\right] \qquad \mathbb{E}[zA] = z\,\mathbb{E}[A]$$

$$\mathbb{E}[\hat{\beta}^{\mathrm{PD}}] = \frac{\beta}{1-\rho^{G1,G2}} \times \left(\mathbb{E}\left[cov(\tilde{g}_{1j}, g_{1j})\right] - \mathbb{E}\left[cov(\tilde{g}_{1j}, g_{2j})\right]\right) \qquad \mathbb{E}[A+B] = \mathbb{E}[A] + \mathbb{E}[B]$$

$$\mathbb{E}[\hat{\beta}^{\mathrm{PD}}] = \frac{\beta}{1-\rho^{G1,G2}} \times \left(\mathbb{E}\left[\rho^{\tilde{g}_{1j},g_{1j}} var(\tilde{g}_{1j})^{\frac{1}{2}} var(g_{1j})^{\frac{1}{2}}\right] - \mathbb{E}\left[\rho^{\tilde{g}_{1j},g_{2j}} var(\tilde{g}_{1j})^{\frac{1}{2}} var(g_{2j})^{\frac{1}{2}}\right]\right) \qquad cov(A,B) = \rho^{A,B} var(A)^{\frac{1}{2}} var(B)^{\frac{1}{2}}$$

$$\mathbb{E}[\hat{\beta}^{\mathrm{PD}}] = \frac{\beta}{1-\rho^{G1,G2}} \times \left(\mathbb{E}\left[\rho^{\tilde{g}_{1j},g_{1j}} var(g_{1j})^{\frac{1}{2}}\right] - \mathbb{E}\left[\rho^{\tilde{g}_{1j},g_{2j}} var(g_{2j})^{\frac{1}{2}}\right]\right) \qquad var(\tilde{g}_{1j}) = 1$$

$$\mathbb{E}[\hat{\beta}^{\mathrm{PD}}] = \frac{\beta}{1-\rho^{G1,G2}} \times \left(\mathbb{E}[\rho^{\tilde{g}_{1j},g_{1j}}]\,\mathbb{E}\left[var(g_{1j})^{\frac{1}{2}}\right] - \mathbb{E}[\rho^{\tilde{g}_{1j},g_{2j}}]\,\mathbb{E}\left[var(g_{2j})^{\frac{1}{2}}\right]\right) \qquad A \perp\!\!\!\perp B \Rightarrow \mathbb{E}[AB] = \mathbb{E}[A]\,\mathbb{E}[B]$$

$$\mathbb{E}[\hat{\beta}^{\mathrm{PD}}] = \frac{\beta}{1-\rho^{G1,G2}} \times \left(\rho^{\tilde{G}1,G1}\,\mathbb{E}\left[var(g_{1j})^{\frac{1}{2}}\right] - \rho^{\tilde{G}1,G2}\,\mathbb{E}\left[var(g_{2j})^{\frac{1}{2}}\right]\right) \qquad \text{Evaluate Expectation}$$

$$\mathbb{E}[\hat{\beta}^{\mathrm{PD}}] = \frac{\beta}{1-\rho^{G1,G2}} \times \left(\mathbb{E}\left[var(g_{1j})^{\frac{1}{2}}\right] - \rho^{\tilde{G}1,G2}\,\mathbb{E}\left[var(g_{2j})^{\frac{1}{2}}\right]\right) \qquad \text{Simplify}$$

$$\mathbb{E}[\hat{\beta}^{\mathrm{PD}}] = \frac{\beta}{1-\rho^{G1,G2}} \times (\sigma_1 - \rho^{G1,G2}\sigma_2) \qquad \text{Evaluate Expectation}$$

$$\mathbb{E}[\hat{\beta}^{\mathrm{PD}}] = \frac{\beta}{1-\rho^{G1,G2}} \times (\sigma_1 - \rho^{G1,G2}\sigma_1) \qquad \text{Substitute in Assumption 3.8}$$

$$\mathbb{E}[\hat{\beta}^{\mathrm{PD}}] = \frac{\beta}{1-\rho^{G1,G2}} \times (1-\rho^{G1,G2})\sigma_1 \qquad \text{Simplify}$$

$$\mathbb{E}[\hat{\beta}^{\mathrm{PD}}] = \beta\sigma_1 \qquad \text{Simplify} \tag{3.9}$$

3.5 Consistency of Phenotype Differences

In this section, we show that the phenotype differences method is a consistent estimator of the causal effect of g_{ij} on y_{ij}. That is, the probability limit of $\hat{\beta}^{PD}$ as N goes to infinity is equal to β, the true causal effect.

Asymptotic Expected Value of β^{PD}

$$\begin{aligned}
\operatorname*{p\,lim}_{N\to\infty} \mathbb{E}[\hat{\beta}^{\mathrm{PD}}] &= \operatorname*{p\,lim}_{N\to\infty} \mathbb{E}[\hat{\beta}^{\mathrm{PD}}] && \text{Identity} \\
\operatorname*{p\,lim}_{N\to\infty} \mathbb{E}[\hat{\beta}^{\mathrm{PD}}] &= \operatorname*{p\,lim}_{N\to\infty} \beta && \text{Substitute in Equation 3.9} \\
\operatorname*{p\,lim}_{N\to\infty} \mathbb{E}[\hat{\beta}^{\mathrm{PD}}] &= \beta && \operatorname*{p\,lim}_{N\to\infty} z = z \qquad (3.10)
\end{aligned}$$

Variance of $\hat{\beta}^{PD}$

$$\begin{aligned}
var(\hat{\beta}^{\mathrm{PD}}) &= \frac{1}{N-K-1} \cdot \frac{var(\hat{\varepsilon}_{ij})}{var\Big((1-\rho)g_{1j}\Big)} && \text{Bivariate OLS Coefficient Variance Formula} \\
var(\hat{\beta}^{\mathrm{PD}}) &= \frac{1}{N-1-1} \cdot \frac{var(\hat{\varepsilon}_{ij})}{(1-\rho)^2 var(g_{1j})} && \text{Substitute in K=1 and } var(zA) = z^2 var(A) \\
var(\hat{\beta}^{\mathrm{PD}}) &= \frac{1}{N-2} \cdot \frac{var(\hat{\varepsilon}_{ij})}{(1-\rho)^2 var(g_{1j})} && \text{Simplify} \qquad (3.11)
\end{aligned}$$

Asymptotic Expected Value of Variance of $\hat{\beta}^{PD}$

$$\begin{aligned}
\operatorname*{p\,lim}_{N\to\infty} \mathbb{E}[var(\hat{\beta}^{\mathrm{PD}})] &= \operatorname*{p\,lim}_{N\to\infty} \mathbb{E}[var(\hat{\beta}^{\mathrm{PD}})] && \text{Identity} \\
\operatorname*{p\,lim}_{N\to\infty} \mathbb{E}[var(\hat{\beta}^{\mathrm{PD}})] &= \operatorname*{p\,lim}_{N\to\infty} \mathbb{E}\left[\frac{1}{N-2} \cdot \frac{var(\hat{\varepsilon}_{ij})}{(1-\rho)^2 var(g_{1j})}\right] && \text{Substitute in Equation 3.11} \\
\operatorname*{p\,lim}_{N\to\infty} [var(\hat{\beta}^{\mathrm{PD}})] &= \operatorname*{p\,lim}_{N\to\infty} \frac{1}{N-2}\, \mathbb{E}\left[\frac{var(\hat{\varepsilon}_{ij})}{(1-\rho)^2 var(g_{1j})}\right] && \mathbb{E}[zA] = z\,\mathbb{E}[A] \\
\operatorname*{p\,lim}_{N\to\infty} [var(\hat{\beta}^{\mathrm{PD}})] &= \operatorname*{p\,lim}_{N\to\infty} \frac{1}{N-2} \frac{\mathbb{E}[var(\hat{\varepsilon}_{ij})]}{(1-\rho)^2\, \mathbb{E}[var(g_{1j})]} && A \perp\!\!\!\perp B \Rightarrow \mathbb{E}[\frac{A}{B}] = \frac{\mathbb{E}[A]}{\mathbb{E}[B]} \\
\operatorname*{p\,lim}_{N\to\infty} [var(\hat{\beta}^{\mathrm{PD}})] &= \frac{1}{\infty-2} \frac{\mathbb{E}[var(\hat{\varepsilon}_{ij})]}{(1-\rho)^2\, \mathbb{E}[var(g_{1j})]} && \text{Evaluate p } lim \\
\operatorname*{p\,lim}_{N\to\infty} \mathbb{E}[var(\hat{\beta}^{\mathrm{PD}})] &= 0 && \mathbb{E}[var(\hat{\varepsilon}_{ij})] < \infty, \mathbb{E}[var(g_{1j})] > 0 \qquad (3.12)
\end{aligned}$$

$\hat{\beta}^{PD}$ Converges in Quadratic Mean

From above Equations 3.10 and 3.12, we have that $\underset{N\to\infty}{\text{p}\,lim}\ \mathbb{E}[\hat{\beta}^{\text{PD}})] = \beta$ and $\underset{N\to\infty}{\text{p}\,lim}\ \mathbb{E}[var(\hat{\beta}^{\text{PD}})] = 0$. Thus, $\hat{\beta}^{PD}$ converges in quadratic mean to β. This entails that the probability limit of $\hat{\beta}^{PD}$ is β and that $\hat{\beta}^{PD}$ is a consistent estimator of β.

$$\underset{N\to\infty}{\text{p}\,lim}\ \mathbb{E}[\hat{\beta}^{\text{PD}}] = \beta,\ \underset{N\to\infty}{\text{p}\,lim}\ \mathbb{E}[var(\hat{\beta}^{\text{PD}})] = 0 \Rightarrow \underset{N\to\infty}{\text{p}\,lim}\ \hat{\beta}^{\text{PD}} = \beta \tag{3.13}$$

3.6 Comparative Precision of Phenotype Differences & Fixed Effects

In this section, we derive the comparative precision of phenotype differences and fixed effects in large samples.

Additional Assumptions

To simplify exposition, for this section we further assume:

$$\rho^{G1,G2} = 0.5 \tag{3.14}$$

Unbiasedness & Consistency of β^{FE}

It has previously been shown that strict exogeneity of the independent variable within group implies that the fixed effects estimator is unbiased and consistent (for example, see Chapter 10 of Wooldridge 2010, *Econometric Analysis of Cross Section and Panel Data*). In our case, given the random assignment of g_{ij} within families, these past results hold.

$$\mathbb{E}[\hat{\beta}^{\text{FE}}] = \beta \tag{3.15}$$

$$\underset{N\to\infty}{\text{p}\,lim}\ \hat{\beta}^{\text{FE}} = \beta \tag{3.16}$$

Variance of $\hat{\beta}^{FE}$

$$\begin{aligned}
var(\hat{\beta}^{\text{FE}}) &= \frac{1}{N(T-1)-K}\cdot\frac{var(\hat{\varepsilon}^{*}_{ij})}{var(g_{1j}-g_{2j})} && \text{Fixed Effects Coefficient Variance Formula}\\
var(\hat{\beta}^{\text{FE}}) &= \frac{1}{N(2-1)-1}\cdot\frac{var(\hat{\varepsilon}^{*}_{ij})}{var(g_{1j}-g_{2j})} && \text{Substitute in T=2 and K=1}\\
var(\hat{\beta}^{\text{FE}}) &= \frac{1}{N-1}\cdot\frac{var(\hat{\varepsilon}^{*}_{ij})}{var(g_{1j}-g_{2j})} && \text{Simplify}\\
var(\hat{\beta}^{\text{FE}}) &= \frac{1}{N-1}\cdot\frac{var(\hat{\varepsilon}^{*}_{ij})}{var(g_{1j})+var(g_{2j})-2cov(g_{1j},g_{1j})} && var(A-B)=var(A)+var(B)-2cov(A,B)\\
var(\hat{\beta}^{\text{FE}}) &= \frac{1}{N-1}\cdot\frac{var(\hat{\varepsilon}^{*}_{ij})}{var(g_{1j})+var(g_{2j})-(2)(.5)var(g_{1j})^{\frac{1}{2}}var(g_{2j})^{\frac{1}{2}}} && cov(A,B)=\rho_{A,B}var(A)^{\frac{1}{2}}var(B)^{\frac{1}{2}}\\
var(\hat{\beta}^{\text{FE}}) &= \frac{1}{N-1}\cdot\frac{var(\hat{\varepsilon}^{*}_{ij})}{var(g_{1j})+var(g_{2j})-var(g_{1j})^{\frac{1}{2}}var(g_{2j})^{\frac{1}{2}}} && \text{Simplify}
\end{aligned} \tag{3.17}$$

Phenotype Differences Residual

$$
\begin{aligned}
y_{1j} - y_{2j} &= \hat{\alpha} + \hat{\beta}^{\mathrm{PD}}\left(g_{1j}(1-\rho^{G1,G2})\right) + \hat{\varepsilon}_{ij} && \text{Equation 3.5}\\
\beta(g_{1j} - g_{2j}) + (e_{1j} - e_{2j}) &= \hat{\alpha} + \hat{\beta}^{\mathrm{PD}}\left(g_{1j}(1-\rho^{G1,G2})\right) + \hat{\varepsilon}_{ij} && \text{Substitute in Equation ??}\\
\hat{\varepsilon}_{ij} &= \beta(g_{1j} - g_{2j}) - \hat{\beta}^{\mathrm{PD}}\left(g_{1j}(1-\rho^{G1,G2})\right) + (e_{1j} - e_{2j}) - \hat{\alpha} && \text{Rearrange to isolate } \hat{\varepsilon}_{ij}\\
\hat{\varepsilon}_{ij} &= \beta(g_{1j} - g_{2j}) - \hat{\beta}^{\mathrm{PD}}\frac{g_{1j}}{2} + (e_{1j} - e_{2j}) - \hat{\alpha} && \text{Substitute in Assumption 3.14}\\
\hat{\varepsilon}_{ij} &= \beta g_{1j} - \beta g_{2j} - \hat{\beta}^{\mathrm{PD}}\frac{g_{1j}}{2} + (e_{1j} - e_{2j}) - \hat{\alpha} && \text{Distribute } \beta\\
\hat{\varepsilon}_{ij} &= (\beta - \hat{\beta}^{\mathrm{PD}})(\frac{g_{1j}}{2}) + \beta(\frac{g_{1j}}{2} - g_{2j}) + (e_{1j} - e_{2j}) - \hat{\alpha} && \text{Undistribute } (\beta - \hat{\beta}^{\mathrm{PD}}) \text{ \& } \beta
\end{aligned}
\tag{3.18}
$$

Fixed Effects Residual

$$
\begin{aligned}
y_{1j} - y_{2j} &= \hat{\beta}^{\mathrm{FE}}(g_{1j} - g_{2j}) + \hat{\varepsilon}^{*}_{ij} && \text{Equation 3.4}\\
\beta(g_{1j} - g_{2j}) + (e_{1j} - e_{2j}) &= \hat{\beta}^{\mathrm{FE}}(g_{1j} - g_{2j}) + \hat{\varepsilon}^{*}_{ij} && \text{Substitute in Equation 3.1}\\
\hat{\varepsilon}^{*}_{ij} &= (\beta - \hat{\beta}^{\mathrm{FE}})(g_{1j} - g_{2j}) + (e_{1j} - e_{2j}) && \text{Rearrange to isolate } \hat{\varepsilon}_{ij}
\end{aligned}
\tag{3.19}
$$

Variance of $\hat{\varepsilon}_{ij}$

$$
\begin{aligned}
var(\hat{\varepsilon}_{ij}) &= var(\hat{\varepsilon}_{ij}) && \text{Identity}\\
var(\hat{\varepsilon}_{ij}) &= var\left((\beta - \hat{\beta}^{\mathrm{PD}})(\frac{g_{1j}}{2}) + \beta(\frac{g_{1j}}{2} - g_{2j}) + (e_{1j} - e_{2j})\right) && \text{Substitute in 3.18}\\
var(\hat{\varepsilon}_{ij}) &= var\left((\beta - \hat{\beta}^{\mathrm{PD}})(\frac{g_{1j}}{2}) + \beta(\frac{g_{1j}}{2} - g_{2j})\right) + var(e_{1j} - e_{2j}) && cov(A,B) = 0 \Rightarrow var(A+B) = var(A) + var(B)\\
var(\hat{\varepsilon}_{ij}) &= (\beta - \hat{\beta}^{\mathrm{PD}})^2 var(\frac{g_{1j}}{2}) + \beta(\beta - \hat{\beta}^{\mathrm{PD}})cov(g_{1j}, \frac{g_{1j}}{2} - g_{2j}) + \beta^2 var(\frac{g_{1j}}{2} - g_{2j}) + var(e_{1j}) + var(e_{2j}) && cov(A,B) = 0 \Rightarrow var(A-B) = var(A) + var(B)\\
var(\hat{\varepsilon}_{ij}) &= var\left((\beta - \hat{\beta}^{\mathrm{PD}})(\frac{g_{1j}}{2})\right) + var\left(\beta(\frac{g_{1j}}{2} - g_{2j})\right) + 2cov\left((\beta - \hat{\beta}^{\mathrm{PD}})(\frac{g_{1j}}{2}), \beta(\frac{g_{1j}}{2} - g_{2j})\right) + var(e_{1j}) + var(e_{2j}) && var(A+B) = var(A) + var(B) + 2cov(A,B)\\
var(\hat{\varepsilon}_{ij}) &= var\left((\beta - \hat{\beta}^{\mathrm{PD}})(\frac{g_{1j}}{2})\right) + var\left(\beta(\frac{g_{1j}}{2} - g_{2j})\right) + \beta(\beta - \hat{\beta}^{\mathrm{PD}})cov(g_{1j}, \frac{g_{1j}}{2} - g_{2j}) + var(e_{1j}) + var(e_{2j}) && cov(yA, zB) = (yz)cov(A,B)\\
var(\hat{\varepsilon}_{ij}) &= (\beta - \hat{\beta}^{\mathrm{PD}})^2 var(\frac{g_{1j}}{2}) + \beta^2 var(\frac{g_{1j}}{2} - g_{2j}) + \beta(\beta - \hat{\beta}^{\mathrm{PD}})cov(g_{1j}, \frac{g_{1j}}{2} - g_{2j}) + var(e_{1j}) + var(e_{2j}) && var(zA) = z^2 var(A)\\
var(\hat{\varepsilon}_{ij}) &= (\beta - \hat{\beta}^{\mathrm{PD}})^2 var(\frac{g_{1j}}{2}) + \beta(\beta - \hat{\beta}^{\mathrm{PD}})cov(g_{1j}, \frac{g_{1j}}{2} - g_{2j}) + \beta^2 var(\frac{g_{1j}}{2} - g_{2j}) + var(e_{1j}) + var(e_{2j}) && \text{Rearrange}\\
var(\hat{\varepsilon}_{ij}) &= (\beta - \hat{\beta}^{\mathrm{PD}})\left((\beta - \hat{\beta}^{\mathrm{PD}})var(\frac{g_{1j}}{2}) + \beta cov(g_{1j}, \frac{g_{1j}}{2} - g_{2j})\right) + \beta^2 var(\frac{g_{1j}}{2} - g_{2j}) + var(e_{1j}) + var(e_{2j}) && \text{Undistribute } (\beta - \hat{\beta}^{\mathrm{PD}})
\end{aligned}
\tag{3.20}
$$

Variance of $\hat{\varepsilon}^*_{ij}$

$$
\begin{aligned}
var(\hat{\varepsilon}^*_{ij}) &= var(\hat{\varepsilon}^*_{ij}) && \text{Identity} \\
var(\hat{\varepsilon}^*_{ij}) &= var\Big((\beta - \hat{\beta}^{\text{FE}})(g_{1j} - g_{2j}) + (\varepsilon_{1j} - \varepsilon_{2j})\Big) && \text{Substitute in 3.19} \\
var(\hat{\varepsilon}^*_{ij}) &= var\Big((\beta - \hat{\beta}^{\text{FE}})(g_{1j} - g_{2j})\Big) + var(\varepsilon_{1j} - \varepsilon_{2j}) && cov(A,B) = 2 \Rightarrow var(A+B) = var(A) + var(B) \\
var(\hat{\varepsilon}^*_{ij}) &= var\Big((\beta - \hat{\beta}^{\text{FE}})(g_{1j} - g_{2j})\Big) + var(\varepsilon_{1j}) + var(\varepsilon_{2j}) && cov(A,B) = 2 \Rightarrow var(A-B) = var(A) + var(B) \\
var(\hat{\varepsilon}^*_{ij}) &= (\beta - \hat{\beta}^{\text{FE}})^2 var(g_{1j} - g_{2j}) + var(\varepsilon_{1j}) + var(\varepsilon_{2j}) && var(zA) = z^2 var(A) \qquad (3.21)
\end{aligned}
$$

Ratio of Variances

$$
\begin{aligned}
\frac{var(\hat{\beta}^{\text{FE}})}{var(\hat{\beta}^{\text{PD}})} &= \frac{var(\hat{\beta}^{\text{FE}})}{var(\hat{\beta}^{\text{PD}})} && \text{Identity} \\
\frac{var(\hat{\beta}^{\text{FE}})}{var(\hat{\beta}^{\text{PD}})} &= \left(\frac{1}{N-1} \cdot \frac{var(\hat{\varepsilon}^*_{ij})}{var(g_{1j}) + var(g_{2j}) - var(g_{1j})^{\frac{1}{2}} var(g_{2j})^{\frac{1}{2}}}\right)\left(\frac{N-2}{1} \cdot \frac{(\frac{1}{2})^2 var(g_{1j})}{var(\hat{\varepsilon}_{ij})}\right) && \text{Substitute in Equations 3.11 and 3.17, recall } \rho^{G1,G2} = 0.5 \\
\frac{var(\hat{\beta}^{\text{FE}})}{var(\hat{\beta}^{\text{PD}})} &= \frac{N-2}{N-1} \cdot \frac{var(\hat{\varepsilon}^*_{ij})}{var(\hat{\varepsilon}_{ij})} \cdot \frac{\frac{1}{4} var(g_{1j})}{var(g_{1j}) + var(g_{2j}) - var(g_{1j})^{\frac{1}{2}} var(g_{2j})^{\frac{1}{2}}} && \text{Rearrange} \qquad (3.22)
\end{aligned}
$$

Asymptotic Ratio of Variances

$$\operatorname*{p\,lim}_{N\to\infty}\frac{var(\hat{\beta}^{\mathrm{FE}})}{var(\hat{\beta}^{\mathrm{PD}})}=\operatorname*{p\,lim}_{N\to\infty}\frac{var(\hat{\beta}^{\mathrm{FE}})}{var(\hat{\beta}^{\mathrm{PD}})} \qquad \text{Identity}$$

$$\operatorname*{p\,lim}_{N\to\infty}\frac{var(\hat{\beta}^{\mathrm{FE}})}{var(\hat{\beta}^{\mathrm{PD}})}=\operatorname*{p\,lim}_{N\to\infty}\left(\frac{N-2}{N-1}\cdot\frac{var(\hat{\varepsilon}^{*}_{ij})}{var(\hat{\varepsilon}_{ij})}\cdot\frac{\frac{1}{4}var(g_{1j})}{var(g_{1j})+var(g_{2j})-var(g_{1j})^{\frac{1}{2}}var(g_{2j})^{\frac{1}{2}}}\right) \qquad \text{Substitute in Equations 3.22}$$

$$\operatorname*{p\,lim}_{N\to\infty}\frac{var(\hat{\beta}^{\mathrm{FE}})}{var(\hat{\beta}^{\mathrm{PD}})}=\operatorname*{p\,lim}_{N\to\infty}\left(\frac{N-2}{N-1}\right)\operatorname*{p\,lim}_{N\to\infty}\left(\frac{var(\hat{\varepsilon}^{*}_{ij})}{var(\hat{\varepsilon}_{ij})}\cdot\frac{\frac{1}{4}var(g_{1j})}{var(g_{1j})+var(g_{2j})-var(g_{1j})^{\frac{1}{2}}var(g_{2j})^{\frac{1}{2}}}\right) \qquad \operatorname*{p\,lim}_{n\to\infty}A_nB_n=\operatorname*{p\,lim}_{n\to\infty}A_n\cdot\operatorname*{p\,lim}_{n\to\infty}B_n$$

$$\frac{var(\hat{\beta}^{\mathrm{FE}})}{var(\hat{\beta}^{\mathrm{PD}})}=\frac{\infty-2}{\infty-1}\operatorname*{p\,lim}_{N\to\infty}\left(\frac{var(\hat{\varepsilon}^{*}_{ij})}{var(\hat{\varepsilon}_{ij})}\cdot\frac{\frac{1}{4}var(g_{1j})}{var(g_{1j})+var(g_{2j})-var(g_{1j})^{\frac{1}{2}}var(g_{2j})^{\frac{1}{2}}}\right) \qquad \text{Evaluate p } lim$$

$$\operatorname*{p\,lim}_{N\to\infty}\frac{var(\hat{\beta}^{\mathrm{FE}})}{var(\hat{\beta}^{\mathrm{PD}})}=\operatorname*{p\,lim}_{N\to\infty}\left(\frac{var(\hat{\varepsilon}^{*}_{ij})}{var(\hat{\varepsilon}_{ij})}\cdot\frac{\frac{1}{4}var(g_{1j})}{var(g_{1j})+var(g_{2j})-var(g_{1j})^{\frac{1}{2}}var(g_{2j})^{\frac{1}{2}}}\right) \qquad \text{Simplify}$$

$$\operatorname*{p\,lim}_{N\to\infty}\frac{var(\hat{\beta}^{\mathrm{FE}})}{var(\hat{\beta}^{\mathrm{PD}})}=\operatorname*{p\,lim}_{N\to\infty}\left(\frac{var(\hat{\varepsilon}^{*}_{ij})}{var(\hat{\varepsilon}_{ij})}\right)\operatorname*{p\,lim}_{N\to\infty}\left(\frac{\frac{1}{4}var(g_{1j})}{var(g_{1j})+var(g_{2j})-var(g_{1j})^{\frac{1}{2}}var(g_{2j})^{\frac{1}{2}}}\right) \qquad \operatorname*{p\,lim}_{n\to\infty}A_nB_n=\operatorname*{p\,lim}_{n\to\infty}A_n\cdot\operatorname*{p\,lim}_{n\to\infty}B_n$$

$$\operatorname*{p\,lim}_{N\to\infty}\frac{var(\hat{\beta}^{\mathrm{FE}})}{var(\hat{\beta}^{\mathrm{PD}})}=\operatorname*{p\,lim}_{N\to\infty}\left(\frac{var(\hat{\varepsilon}^{*}_{ij})}{var(\hat{\varepsilon}_{ij})}\right)\left(\frac{\frac{1}{4}var(G1)}{var(G1)+var(G2)-var(G2)^{\frac{1}{2}}var(G2)^{\frac{1}{2}}}\right) \qquad \text{Evaluate p } lim$$

$$\operatorname*{p\,lim}_{N\to\infty}\frac{var(\hat{\beta}^{\mathrm{FE}})}{var(\hat{\beta}^{\mathrm{PD}})}=\operatorname*{p\,lim}_{N\to\infty}\left(\frac{var(\hat{\varepsilon}^{*}_{ij})}{var(\hat{\varepsilon}_{ij})}\right)\left(\frac{\frac{1}{4}\sigma^2}{\sigma^2+\sigma^2-\sigma^2}\right) \qquad \text{Change of Notation}$$

$$\operatorname*{p\,lim}_{N\to\infty}\frac{var(\hat{\beta}^{\mathrm{FE}})}{var(\hat{\beta}^{\mathrm{PD}})}=\operatorname*{p\,lim}_{N\to\infty}\left(\frac{var(\hat{\varepsilon}^{*}_{ij})}{var(\hat{\varepsilon}_{ij})}\right)\left(\frac{\sigma^2}{4\sigma^2}\right) \qquad \text{Simplify}$$

$$\operatorname*{p\,lim}_{N\to\infty}\frac{var(\hat{\beta}^{\mathrm{FE}})}{var(\hat{\beta}^{\mathrm{PD}})}=\operatorname*{p\,lim}_{N\to\infty}\frac{1}{4}\frac{var(\hat{\varepsilon}^{*}_{ij})}{var(\hat{\varepsilon}_{ij})} \qquad \text{Simplify}$$

$$\operatorname*{p\,lim}_{N\to\infty}\frac{var(\hat{\beta}^{\mathrm{FE}})}{var(\hat{\beta}^{\mathrm{PD}})}=\frac{1}{4}\operatorname*{p\,lim}_{N\to\infty}\frac{(\beta-\hat{\beta}^{\mathrm{FE}})^2var(g_{1j}-g_{2j})+var(e_{1j})+var(e_{2j})}{(\beta-\hat{\beta}^{\mathrm{PD}})\left((\beta-\hat{\beta}^{\mathrm{PD}})var(\frac{g_{1j}}{2})+\beta cov(g_{1j},\frac{g_{1j}}{2}-g_{2j})\right)+\beta^2var(\frac{g_{1j}}{2}-g_{2j})+var(e_{1j})+var(e_{2j})} \qquad \text{Substitute Equations 3.20 \& 3.21}$$

$$\operatorname*{p\,lim}_{N\to\infty}\frac{var(\hat{\beta}^{\mathrm{FE}})}{var(\hat{\beta}^{\mathrm{PD}})}=$$

$$\frac{\operatorname*{p\,lim}_{N\to\infty}(\beta-\hat{\beta}^{\mathrm{FE}})^2var(g_{1j}-g_{2j})+\operatorname*{p\,lim}_{N\to\infty}\left(var(e_{1j})+var(e_{2j})\right)}{4\left[\operatorname*{p\,lim}_{N\to\infty}(\beta-\hat{\beta}^{\mathrm{PD}})\left((\beta-\hat{\beta}^{\mathrm{PD}})var(\frac{g_{1j}}{2})+\beta cov(g_{1j},\frac{g_{1j}}{2}-g_{2j})\right)+\operatorname*{p\,lim}_{N\to\infty}\left(\beta^2var(\frac{g_{1j}}{2}-g_{2j})+var(e_{1j})+var(e_{2j})\right)\right]} \qquad \operatorname*{p\,lim}_{n\to\infty}\frac{A_n}{B_n}=\frac{\operatorname*{p\,lim}_{n\to\infty}A_n}{\operatorname*{p\,lim}_{n\to\infty}B_n}$$

$$\operatorname*{p\,lim}_{N\to\infty}\frac{var(\hat{\beta}^{\mathrm{FE}})}{var(\hat{\beta}^{\mathrm{PD}})}=\frac{(\beta-\beta)^2var(G1-G2)+\operatorname*{p\,lim}_{N\to\infty}\left(var(e_{1j})+var(e_{2j})\right)}{4(\beta-\beta)\left((\beta-\beta)var(\frac{G1}{2})+\beta cov(G1,\frac{G1}{2}-G2)\right)+\operatorname*{p\,lim}_{N\to\infty}4\cdot\left(\beta^2var(\frac{g_{1j}}{2}-g_{2j})+var(e_{1j})+var(e_{2j})\right)} \qquad \text{Evaluate p } lim$$

$$\operatorname*{p\,lim}_{N\to\infty}\frac{var(\hat{\beta}^{\mathrm{FE}})}{var(\hat{\beta}^{\mathrm{PD}})}=\frac{\operatorname*{p\,lim}_{N\to\infty}\left(var(e_{1j})+var(e_{2j})\right)}{\operatorname*{p\,lim}_{N\to\infty}4\cdot\left(\beta^2var(\frac{g_{1j}}{2}-g_{2j})+var(e_{1j})+var(e_{2j})\right)} \qquad \text{Simplify}$$

$$\underset{N\to\infty}{\text{p lim}}\ \frac{var(\hat{\beta}^{\text{FE}})}{var(\hat{\beta}^{\text{PD}})} = \frac{var(E1)+var(E2)}{4\Big(\beta^2 var(\frac{G1}{2}-G2)+var(E1)+var(E2)\Big)} \qquad \text{Evaluate p } lim$$

$$\underset{N\to\infty}{\text{p lim}}\ \frac{var(\hat{\beta}^{\text{FE}})}{var(\hat{\beta}^{\text{PD}})} = \frac{var(E1)+var(E2)}{4\beta^2\Big(var(\frac{G1}{2})+var(G2)-2cov(\frac{G1}{2},G2)\Big)+4var(E1)+4var(E2)} \qquad var(A-B)=var(B)+var(B)-2cov(A,B)$$

$$\underset{N\to\infty}{\text{p lim}}\ \frac{var(\hat{\beta}^{\text{FE}})}{var(\hat{\beta}^{\text{PD}})} = \frac{var(E1)+var(E2)}{4\beta^2\Big(var(\frac{G1}{2})+var(G2)-cov(G1,G2)\Big)+4var(E1)+4var(E2)} \qquad cov(yA,zB)=(yz)cov(A,B)$$

$$\underset{N\to\infty}{\text{p lim}}\ \frac{var(\hat{\beta}^{\text{FE}})}{var(\hat{\beta}^{\text{PD}})} = \frac{var(E1)+var(E2)}{\beta^2\Big(var(G1)+4var(G2)-4cov(G1,G2)\Big)+4var(E1)+4var(E2)} \qquad var(zA)=z^2var(A)$$

$$\underset{N\to\infty}{\text{p lim}}\ \frac{var(\hat{\beta}^{\text{FE}})}{var(\hat{\beta}^{\text{PD}})} = \frac{var(E1)+var(E2)}{\beta^2\Big(var(G1)+4var(G2)-4\rho^{G1,G2}var(G1)^{\frac{1}{2}}var(G2)^{\frac{1}{2}}\Big)+4var(E1)+4var(E2)} \qquad cov(A,B)=\rho^{A,B}var(A)^{\frac{1}{2}}var(B)^{\frac{1}{2}}$$

$$\underset{N\to\infty}{\text{p lim}}\ \frac{var(\hat{\beta}^{\text{FE}})}{var(\hat{\beta}^{\text{PD}})} = \frac{\omega^2+\omega^2}{\beta^2(\sigma^2+4\sigma^2-2\sigma^2)+4\omega^2+4\omega^2} \qquad \text{Change of Notation}$$

$$\underset{N\to\infty}{\text{p lim}}\ \frac{var(\hat{\beta}^{\text{FE}})}{var(\hat{\beta}^{\text{PD}})} = \frac{2\omega^2}{3\beta^2\sigma^2+8\omega^2} \qquad \text{Simplify} \qquad (3.23)$$

Variance of y_{ij} Within Families

$$var(y_{1j}-y_{2j}) = var(y_{1j}-y_{2j}) \qquad \text{Identity}$$

$$var(y_{1j}-y_{2j}) = var\Big(\beta(g_{1j}-g_{2j})+(e_{1j}-e_{2j})\Big) \qquad \text{Substitute in Equation 3.4}$$

$$var(y_{1j}-y_{2j}) = var\Big(\beta(g_{1j}-g_{2j})\Big)+var(e_{1j}-e_{2j}) \qquad cov(A,B)=0 \Rightarrow var(A+B)=var(A)+var(B)$$

$$var(y_{1j}-y_{2j}) = \beta^2 var(g_{1j}-g_{2j})+var(e_{1j}-e_{2j}) \qquad var(zA)=z^2var(A)$$

$$var(y_{1j}-y_{2j}) = \beta^2 var(g_{1j}-g_{2j})+var(e_{1j})+var(e_{2j}) \qquad cov(A,B)=0 \Rightarrow var(A-B)=var(A)+var(B)$$

$$var(y_{1j}-y_{2j}) = \beta^2\Big(var(g_{1j})+var(g_{2j})-2cov(g_{1j},g_{2j})\Big)+var(e_{1j})+var(e_{2j}) \qquad var(A-B)=var(A)+var(B)-2cov(A,B)$$

(3.24)

Expected Value of Variance of y_{ij} Within Families

$$
\begin{aligned}
\mathbb{E}[var(y_{1j} - y_{2j})] &= \mathbb{E}[var(y_{1j} - y_{2j})] && \text{Identity} \\
\mathbb{E}[var(y_{1j} - y_{2j})] &= \mathbb{E}\left[\beta^2\Big(var(g_{1j}) + var(g_{2j}) - 2cov(g_{1j}, g_{2j})\Big) + var(e_{1j}) + var(e_{2j})\right] && \text{Take Expectation} \\
\mathbb{E}[var(y_{1j} - y_{2j})] &= \mathbb{E}\left[\beta^2\Big(var(g_{1j}) + var(g_{2j}) - 2cov(g_{1j}, g_{2j})\Big)\right] + \mathbb{E}[var(e_{1j}) + var(e_{2j})] && \mathbb{E}[A+B] = \mathbb{E}[A] + \mathbb{E}[B] \\
\mathbb{E}[var(y_{1j} - y_{2j})] &= \beta^2\Big(var(G1) + var(G2) - 2cov(G1, G2)\Big) + var(E1) + var(E2) && \text{Evaluate Expectation} \\
\mathbb{E}[var(y_{1j} - y_{2j})] &= \beta^2\Big(var(G1) + var(G2) - 2\rho^{G1,G2} var(G1)^{\frac{1}{2}} var(G2)^{\frac{1}{2}}\Big) + var(E1) + var(E2) && cov(A,B) = \rho^{A,B} var(A)^{\frac{1}{2}} var(B)^{\frac{1}{2}} \\
\mathbb{E}[var(y_{1j} - y_{2j})] &= \beta^2\Big(\sigma^2 + \sigma^2 - (2)(.5)\sigma^2\Big) + \omega^2 + \omega^2 && \text{Change of Notation} \\
\mathbb{E}[var(y_{1j} - y_{2j})] &= \beta^2\sigma^2 + 2\omega^2 && \text{Simplify}
\end{aligned} \tag{3.25}
$$

Within Family Explanatory Power of g_{ij}

Let us call the fraction of within family variation in the outcome y_{ij} that is explained by g_{ij}, our genetic predictor, ϕ. The within R^2 of the fixed effects model displayed in Equation 3.3 provides an estimate of ϕ.

$$\phi = \frac{\beta^2\sigma^2}{\beta^2\sigma^2 + 2\omega^2} \tag{3.26}$$

Isolating $\beta^2\sigma^2$

$$
\begin{aligned}
\phi &= \frac{\beta^2\sigma^2}{\beta^2\sigma^2 + 2\omega^2} && \text{Equation 3.26} \\
\beta^2\sigma^2\phi + 2\omega^2\phi &= \beta^2\sigma^2 && \text{Rearrange} \\
2\omega^2\phi &= \beta^2\sigma^2 - \beta^2\sigma^2\phi && \text{Collect } \sigma^2 \text{ terms} \\
2\omega^2\phi &= (1-\phi)\beta^2\sigma^2 && \text{Factor out } \phi \\
\frac{2\omega^2\phi}{(1-\phi)} &= \beta^2\sigma^2 && \text{Rearrange}
\end{aligned} \tag{3.27}
$$

Ratio of Variances as a Function of ϕ

We can now express the asymptotic variance ratio of the fixed effects and phenotype differences estimators in terms of ϕ.

$$\underset{N\to\infty}{\text{p lim}}\ \frac{var(\hat{\beta}^{\text{FE}})}{var(\hat{\beta}^{\text{PD}})} = \frac{2\omega^2}{3\beta^2\sigma^2 + 8\omega^2} \quad \text{Equation 3.23}$$

$$\underset{N\to\infty}{\text{p lim}}\ \frac{var(\hat{\beta}^{\text{FE}})}{var(\hat{\beta}^{\text{PD}})} = \frac{2\omega^2}{3\frac{2\omega^2\phi}{(1-\phi)} + 8\omega^2} \quad \text{Substitute in Equation 3.27}$$

$$\underset{N\to\infty}{\text{p lim}}\ \frac{var(\hat{\beta}^{\text{FE}})}{var(\hat{\beta}^{\text{PD}})} = \frac{1}{3\frac{\phi}{(1-\phi)} + 4} \quad \text{Simplify}$$

$$\underset{N\to\infty}{\text{p lim}}\ \frac{var(\hat{\beta}^{\text{FE}})}{var(\hat{\beta}^{\text{PD}})} = \frac{1-\phi}{3\phi + 4 - 4\phi} \quad \text{Simplify}$$

$$\underset{N\to\infty}{\text{p lim}}\ \frac{var(\hat{\beta}^{\text{FE}})}{var(\hat{\beta}^{\text{PD}})} = \frac{1-\phi}{4-\phi} \quad \text{Simplify} \qquad (3.28)$$

3.7 Instrumental Variables with Phenotype Differences

Suppose that we are interested in using the following family fixed effects model to estimate the causal relationship between two non-genetic individual-level variables:

$$y_{ij} = \alpha_j + \beta x_{ij} + \varepsilon'_{ij}$$

Here, x_{ij} and y_{ij} are both phenotypic variables that we observe for exactly 2 siblings i in each family j and α_j is term representing family fixed effects. Because $T = 2$, the fixed effects equation above can be identically expressed using first differences:

$$y_{1j} - y_{2j} = \beta(x_{1j} - x_{2j}) + (\varepsilon'_{1j} - \varepsilon'_{2j}) \tag{3.29}$$

However, if $x_{1j} - x_{2j}$ is correlated with $\varepsilon'_{1j} - \varepsilon'_{2j}$, this approach would suffer from omitted variable bias. A potential solution is to use the genotype of *one* of the siblings, g_{1j}, as an instrument for $x_{1j} - x_{2j}$ (the use of genetic characteristics as an instrumental variable is often referred to as Mendelian randomization). For compactness, we define:

$$\Delta y_j = y_{1j} - y_{2j}$$
$$\Delta x_j = x_{1j} - x_{2j}$$
$$\Delta \varepsilon'_j = \varepsilon'_{1j} - \varepsilon'_{2j}$$

This allows us to rewrite Equation 3.29 as follows:

$$\Delta y_j = \alpha + \beta(\Delta x_j) + \Delta \varepsilon'_j \tag{3.30}$$

In order for g_{1j} to be a valid instrument for Δx_j, it must meet two conditions:

1. Relevance: $Cov(g_{1j}, \Delta x_j) \neq 0$
2. Exogeneity: $Cov(g_{1j}, \Delta \varepsilon'_j) = 0$

If these conditions are met, we can use a modified form of the phenotype differences model that acts as the first stage in two-stage least-squares IV regression:

$$\Delta x_j = \alpha + \beta^{PD}\left((1 - \rho^{G1,G2})g_{1j}\right) + \varepsilon_{ij}$$
$$\Delta y_j = \beta^{IV}(\Delta \tilde{x}_j) + \Delta \varepsilon'_j$$

Where $\Delta \tilde{x}_j$ is the fitted value of Δx_j estimated from the first stage equation. In fact, because we are only interested in the fitted values, $\Delta \tilde{x}_j$, from the from the first stage – that is, β^{PD} is not of interest – we can simplify the estimating equation by simply omitting $\rho^{G1,G2}$ altogether:

$$\Delta x_j = \alpha + \beta^{PD} g_{1j} + \varepsilon_{ij} \tag{3.31}$$
$$\Delta y_j = \beta^{IV}(\Delta \tilde{x}_j) + \Delta \varepsilon'_j \tag{3.32}$$