2021 Russell Sage Foundation Summer Institute in Social-Science Genomics

Problem Set 2

The purpose of this problem set is to teach some of the core ideas and methods of behavior genetics, as well as the appropriate interpretation of the resulting heritability estimates.

This problem set is due at 9:30am on Thursday, August 12.

1. Adoption studies, twin studies, and the ACE model

This problem is intended to illustrate the logic behind the estimation of heritability based only on family resemblance (i.e., no DNA data). The various proofs and derivations that follow build on Falconer's Formula, a simple moment condition for estimating heritability based on phenotypic correlations between monozygotic (MZ, identical) and dizygotic (DZ, fraternal) twins. Extensions of Falconer's Formula to other types of siblings are also derived. Additional questions scattered throughout emphasize the assumptions that must be made with these data in order to obtain unbiased estimates of heritability, some of which may be untenable. Note that there are many important elements to understand about heritability and behavior-genetic models, which we will cover here and throughout this problem set.

Since we will be studying genetic effects within families, we assume that genotypes are randomly assigned. For some phenotype y, the population regression equation is

$$y_i = \sum_{j=1}^{J} \beta_j x_{ij} + \epsilon_i, \tag{1}$$

where i indexes individuals, j indexes the J causal loci, and ϵ_i is an error term that has mean 0 and is uncorrelated with every genotype score x_{ij}. All variables are de-meaned,

which is why there is no intercept. The uncorrelatedness between ϵ_i and x_{ij} implies that $Cov(\epsilon_i, x_{ij}) = 0$. Throughout, we assume that the population is in Hardy-Weinberg equilibrium. We also assume that genotype scores across loci are independent, which implies that $Cov(x_{ij}, x_{ik}) = 0$ for all $j \neq k$.

Because genotypes are randomly assigned within a family, the regression coefficients have a causal interpretation: β_j is the average causal effect of genotype score j (averaged across any dominance and epistasis effects) within the family. For simplicity, we will further assume that the β_j's are equal across families. (Without this assumption, we could still work with Equation (1) and everything in this problem would be unaffected, but we would need to interpret the β_j's as an average causal effect that is also averaged across families.)

Note that Equation (1) is the multi-locus generalization of the (additive) single-locus model we considered in Problem 2 from Problem Set 1, except that we have assumed that the genotypes are randomly assigned. It is important that the model have a causal interpretation so that we can define and analyze "heritability."

The *heritability* of the phenotype, denoted a^2, is defined as the proportion of total variance explained by the genetic effects in the causal model. It is also the R^2 from regression Equation (1):

$$a^2 \equiv \frac{Var\left(\sum_{j=1}^{J} \beta_j x_{ij}\right)}{Var(y_i)}.$$

(Important note: Heritability is often denoted h^2, but in this problem we use a^2, which stands for the *additive* genetic variance explained, because a^2 is the standard notation in the ACE model that we will derive below.)

The fact that heritability is defined with respect to genetic effects that have a causal interpretation is important; it means, for example, that if genes are not causal and instead

predictive of a phenotype due only to correlation with an environmental factor, then those genes do not contribute to the heritability of the phenotype.

In Problem 2, we will distinguish between "narrow-sense heritability" (the variance explained by additive genetic effects) and "broad-sense heritability" (the variance explained by all genetic effects). What we are calling "heritability" in this problem refers to narrow-sense heritability.

The remaining proportion of phenotypic variance,

$$u^2 \equiv \frac{Var(\epsilon_i)}{Var(y_i)} = 1 - a^2,$$

is often referred to as the variance explained by environmental factors. Although this terminology is standard, we note that there are several reasons that it is potentially misleading. One is that the error term captures all variability in the phenotype measure that is not explained by the genetic loci, including transient measurement error that would not usually be considered an environmental effect. Another reason is that such language can help propagate the common misunderstanding that heritability is the proportion of variance *not* explained by environmental factors—a common and problematic misunderstanding that we will explore in Problem 3 below (see also Jencks (1980)). For these reasons, we prefer referring to the error term as capturing "residual effects."

a. Show that

$$a^2 = \sum_{j=1}^{J} \frac{\beta_j^2 2p_j(1-p_j)}{Var(y_i)},$$

where p_j is the minor allele frequency at locus j.

Hint: Recall that a^2 is also the R^2 from Equation (1). Using this fact, generalize the results from Problem 2(g) of Problem Set 1 to the multiple-locus case. Keep in mind that $Cov(x_{ij}, x_{ik}) = 0$ for all $j \neq k$.

b. Show that the mean value of the phenotype in the population, $E(y_i)$, is equal to $E\left(\sum_{j=1}^{J} \beta_j x_{ij}\right)$.

Note that $\sum_{j=1}^{J} \beta_j x_{ij}$ is the additive variance component in this multi-locus model, the generalization of the additive variance component that you derived in Problem 2(o) from Problem Set 1. To make our notation consistent with standard notation in much of the literature on twin studies, define the following variables:

$$Y_i \equiv \frac{y_i - E(y_i)}{\sqrt{Var(y_i)}},$$

$$X_i \equiv \frac{\sum_{j=1}^{J} \beta_j x_{ij} - E\left(\sum_{j=1}^{J} \beta_j x_{ij}\right)}{\sqrt{Var(y_i)}},$$

$$U_i \equiv \frac{\epsilon_i}{\sqrt{Var(y_i)}}.$$

Y_i is simply the *standardized phenotype*. (Note: In our problem sets, we usually denote standardized variables using double squiggles (i.e. $\widetilde{\widetilde{y}}_i$). However, we omit this symbol here for clarity and simplicity). The additive variance component normalized by the phenotypic standard deviation, X_i, is called the *genetic factor*. The error term normalized by the phenotypic standard deviation, U_i, is called the *environmental factor*.

As is standard, we will drop the i subscript going forward.

c. Show that

$$Y = X + U. \tag{2}$$

d. Show that

$$Var(X) = a^2,$$

$$Var(U) = u^2.$$

Hint: Note that $Var(y_i)$ is a constant.

Define the correlation between X and U as the *gene-environment correlation*:

$$\rho_{XU} \equiv \frac{Cov(X, U)}{\sqrt{Var(X)Var(U)}}.$$

e. Is it possible that $\rho_{XU} = 0$ for some phenotypes but $\rho_{XU} \neq 0$ for other phenotypes?

Some potential causes and implications of non-zero ρ_{XU} are developed in parts (f), (g), and (h).

f. Suppose that the genes that influence scholastic achievement also influence traits and skills that are rewarded in labor markets. Furthermore, suppose that children's scholastic achievement is also influenced by their parents' incomes. Explain why under these conditions, we expect non-zero gene-environment correlation (with $Cov(X, U) > 0$).

g. Next, suppose that some genes influence children's scholastic achievement by causing them to select into environments that positively affect cognitive development. Explain why any correlation between X and the environment induced by such selection is *not* an example of gene-environment correlation for the phenotype of scholastic achievement. (Hint: Any effect on y_i of changing x_{ij} at conception is part of the causal effect of x_{ij}.)

h. By taking the variance of both sides of Equation (2), show that

$$1 = a^2 + u^2 + 2au\rho_{XU}. \tag{3}$$

We will now restrict our attention to individuals in the population who have a sibling (e.g., who could be a half-sister, an identical-twin brother, a non-twin full sibling, etc.). For each sibling pair, we randomly label one sibling as Sibling 1 and the other as Sibling 2. Letting Y, X, and U be defined as above for each Sibling 1, let Y', X', and U' be defined the same way for each Sibling 2. Because siblings are randomly assigned to groups 1 or 2, we can assume that $Var(Y) = Var(Y')$, $Var(X) = Var(X')$, $Var(U) = Var(U')$, and $Cov(X, U') = Cov(U, X')$. Throughout, we maintain the assumption that $Var(X)$, $Var(U)$, and $Cov(X, U)$ are constant across all types of siblings.

Note the following definitions carefully:

- The *phenotypic correlation* across siblings is the correlation between Y and Y', denoted $\rho_{YY'}$
- The *genetic correlation* across siblings is the correlation between X and X', denoted $\rho_{XX'}$
- The *environmental correlation* across siblings is the correlation between U and U', denoted $\rho_{UU'}$
- The correlation between Sibling 1's genetic factor and Sibling 2's environmental factor is denoted $\rho_{XU'}$

i. Now, justify the following steps of algebra:

$$\rho_{YY'} \equiv \frac{Cov(Y, Y')}{\sqrt{Var(Y)Var(Y')}} = Cov(Y, Y') = Cov(X + U, X' + U')$$

$$= Cov(X, X') + Cov(U, U') + Cov(X, U') + Cov(U, X')$$

$$= \left(\frac{Cov(X, X')}{\sqrt{Var(X)Var(X')}}\right) Var(X) + \left(\frac{Cov(U, U')}{\sqrt{Var(U)Var(U')}}\right) Var(U)$$

$$+2\left(\frac{Cov(X, U')}{\sqrt{Var(X)Var(U')}}\right)\sqrt{Var(X)Var(U')}$$

$$= \rho_{XX'}a^2 + \rho_{UU'}u^2 + 2au\rho_{XU'}.$$

Note that on the left-hand side of the equation, we have the phenotypic correlation, a variable that can be estimated given suitable sibling data. The methodology of studies in *behavior genetics* is fundamentally about using estimated phenotypic correlations $\hat{\rho}_{YY'}$ to try infer population parameters such as a^2 and u^2 by making assumptions about unobserved correlations $\rho_{XX'}$, $\rho_{UU'}$, and $\rho_{XU'}$ in different types of relatives. The assumptions about $\rho_{XX'}$ are typically derived from quantitative genetic theory.

We now illustrate how behavior-genetic approaches work using the conventional twin-study approach. There are two types of twins: *monozygotic twins* (colloquially called identical twins) and *dizygotic twins* (colloquially called fraternal twins).

In a monozygotic (MZ) twin pair, the two individuals result from a single zygote (i.e., fertilized egg) that splits into two embryos. As a result, the MZ twins are identical at virtually all of their genetic loci (the exceptions being *de novo* mutations—i.e., mutations that are new to the individual rather than inherited—but these comprise a negligible fraction of loci). Thus, for MZ twin pairs, $\rho_{XX'} \approx 1$.

To build up intuition for the more common twin-study methods that we will explore below, we begin with the ideal behavior-genetic experiment:

Suppose that each member of a MZ twin pair is adopted at birth into a different randomly assigned family.

j. Explain why in that case, $\rho_{UU'} \approx 0$ and $\rho_{XU'} \approx 0$, and therefore the phenotypic correlation for MZ twins reared apart is $\rho_{YY'}^{MZ-apart} \approx a^2$. Conclude that in this case a reasonable estimate of heritability, $\hat{h}$, would be simply the phenotypic correlation: $\hat{a}^2 = \hat{\rho}_{YY'}^{MZ-apart}$.

k. Which key assumption underlying this conclusion would be violated if the pre-birth (uterine) environment is important for the specific outcome we are studying? Explain why this is the case.

Note that the equation $\rho_{YY'}^{MZ-apart} = a^2$ is called a *moment condition* because it relates a population quantity that can be estimated (in this case, $\rho_{YY'}^{MZ-apart}$) to parameters of the model (in this case, a^2). The *method of moments* is an approach for deriving estimators from a set of moment conditions: (1) the moment conditions are solved to give the parameters as a function of the population quantities—in this case, simply the moment condition itself, and then (2) the population quantities are replaced by their sample estimates. In this part of the problem, you derived a very simple method-of-moments estimator; below, you will see further examples (some a bit more complicated).

(Under general conditions, method-of-moments estimators are *consistent*, meaning that the probability limit of the parameter's estimate (as the sample size goes to infinity) is the true value of the parameter. In the example here, $plim(\hat{a}^2) = a^2$. However, method-of-moments estimators may be biased in small samples.)

In a dizygotic (DZ) twin pair, the two individuals result from two distinct zygotes (that develop in the same uterus). As a result, the DZ twins have the same genetic relationship to each other as non-twin full siblings.

In Problem 2(t) from Problem Set 1, you showed that in the single-locus case, the parent-offspring correlation $Corr(A_m, A_i)$, equals $\frac{1}{2}$. A similar calculation shows that for a single locus, the full-sibling correlation is also equal to $\frac{1}{2}$. Define the (multi-locus) additive genetic variance component as $A_i \equiv \sum_{j=1}^{J} \beta_j x_{ij} = \sum_{j=1}^{J} A_{ij}$, where A_{ij} is the single-locus additive variance component.

Given our assumption of independent loci, it follows from the single-locus result that $\rho_{XX'} = Corr(A, A') = \frac{1}{2}$. The proof is provided, for your reference, here:

We have that:

$$Cov(A, A') \equiv Cov\left(\sum_{j=1}^{J} A_{ij}\,, \sum_{j=1}^{J} A'_{ij}\right) = \sum_{j=1}^{J} Cov\left(A_{ij}, A'_{ij}\right)$$

The last step follows under the assumption of the independence of the loci. These same assumptions also imply that:

$$Var(A) = Var(A') = Var\left(\sum_{j=1}^{J} A_{ij}\right) = \sum_{j=1}^{J} Var\left(A_j\right)$$

Hence the correlation between the two factors is:

$$\rho_{XX'} \equiv \frac{Cov(A, A')}{\sqrt{Var(A) \times Var(A')}} = \frac{Cov(A, A')}{Var(A)} = \frac{\sum_{j=1}^{J} Cov\left(A_{ij}, A'_{ij}\right)}{\sum_{j=1}^{J} Var\left(A_j\right)}$$

$$= \frac{Cov(A_{i1}, A'_{i1}) + Cov(A_{i2}, A'_{i2}) + \cdots + Cov\left(A_{iJ}, A'_{iJ}\right)}{\sum_{j=1}^{J} Var\left(A_j\right)}$$

$$= \frac{Var(A_1)\frac{Cov(A_1, A_1')}{Var(A_1)} + Var(A_2)\frac{Cov(A_2, A_2')}{Var(A_2)} + \cdots + Var(A_J)\frac{Cov(A_J, A_J')}{Var(A_J)}}{\sum_{j=1}^{J} Var(A_j)}$$

$$= \frac{\frac{1}{2}[Var(A_1) + Var(A_2) + \cdots + Var(A_J)]}{\sum_{j=1}^{J} Var(A_j)} = \frac{1}{2}$$

l. Suppose that each individual in each DZ twin pair is adopted into a different randomly assigned family. Using a logic similar to that in part (j) and the result of the derivation above, explain (or show) why a natural method-of-moments estimator to consider in this setting is $\hat{a}^2 = 2\hat{\rho}_{YY'}^{DZ-apart}$.

m. By the same logic used in part (l), if we instead ran the same experiment in full siblings who are not twins, a natural method-of-moments estimator is $2\hat{\rho}_{YY'}^{FS-apart}$. Explain why this is no different from the estimator for DZ twins found in part (l).

n. With reference to the discussion in part (k) above about potential biases due to positive correlations in the pre-birth environments, explain under what conditions the estimator from part (m) may actually be better than the estimator from part (l).

Adoption studies rely on comparisons of siblings who are genetically related but reared in different homes. Above, we discussed idealized examples of adoption studies. Such studies can be valuable for making credible inferences about a^2. But like all research strategies, they have limitations. In parts (o) through (r), you are asked to think through potential limitations.

o. Unlike in the idealized examples, in practice adoptees are rarely assigned to families at random (see Sacerdote (2007) for a possible exception). Show how non-random assignment will violate the assumption, made in part (j) above, that $\rho_{UU'}$ is zero, and explain how this will bias the estimator in (j).

p. Can you think of any factors other than non-random placement that may induce positive covariance between the environments of the siblings "reared apart"? (Hint: Adoptees are never assigned to their adoptive families at birth, and sometimes the assignment process can take many years.)

q. Suppose next that random assignment holds, so that $\rho_{UU'} = 0$ in our sample of siblings reared apart. However, only households whose value of U is above some threshold $\underline{U}$ are eligible to adopt a child. Suppose that in the population of adoptees, the marginal distribution of X is the same as in the population as a whole. Without giving a technical answer, how you think the truncation of the distribution of U will impact the estimate of a^2?

r. Suppose that you are interested not in the a^2 parameter in the population as a whole, but a^2 in the population of adoptees (whose environments are left-truncated). Are the method-of-moments estimators from parts (j), (l), and (m) biased?

We now turn to more common study designs in behavior genetics, which do not attempt to mimic the ideal experiment of randomized environments. When the environment is not randomly assigned, environmental differences between siblings are confounded with genetic differences, making it harder to infer heritability from phenotypic correlations. The basic idea behind a *family study* is to compare the phenotypic correlations across different pairs of relatives and make assumptions about $\rho_{XX'}$, $\rho_{UU'}$, and $\rho_{XU'}$ which make it possible to infer a^2 and u^2. The most common type of family study is a *twin study*, which we explore here.

s. Derive these two moment conditions, under the assumption that a^2 and u^2 are the same in the population of DZ twins as in the population of MZ twins:

$$\rho_{YY'}^{MZ} = a^2 + \rho_{UU'}^{MZ} u^2 + 2au\rho_{XU'}^{MZ}. \quad (4)$$

$$\rho_{YY'}^{DZ} = \frac{1}{2}a^2 + \rho_{UU'}^{DZ}u^2 + 2au\rho_{XU'}^{DZ}. \quad (5)$$

Note that—putting together Equations (3), (4), and (5)—we have 3 equations and 6 model parameters ($a^2, u^2, \rho_{UU'}^{MZ}, \rho_{XU'}^{MZ}, \rho_{UU'}^{DZ}, \rho_{XU'}^{DZ}$). It is therefore impossible to solve for the model parameters: there are many possible combinations of parameter values that would satisfy Equations (3)-(5). In such a situation, the parameters are said to be *unidentified*.

In order to *identify* the parameters (i.e., make it possible to solve for a unique value for each parameter), we either need additional, linearly independent equations (which would come from data on the resemblance of other types of relatives) or we need to make assumptions about some of the unknown parameters. In a classical twin study, we identify the parameters by making two assumptions:

$$\rho_{XU'}^{MZ} = \rho_{XU'}^{DZ} = 0. \quad (6)$$

$$\rho_{UU'}^{MZ} = \rho_{UU'}^{DZ} \equiv \rho_{UU'}. \quad (7)$$

Equation (6) is the *assumption of no gene-environment correlation*. (Recall the example of gene-environment correlation provided in part (f).)

IMPORTANT NOTE: Equation (7) is called the *equal-environments assumption* (often abbreviated *EEA*)—but it is poorly named because it does *not* say that MZ and DZ twins have *the same* environments, just that the magnitude of the correlation between environments of MZ twins is the same of that for DZ twins. This assumption is often misinterpreted, as illustrated in part (t).

t. Explain why the following is not a violation of the EEA: Suppose that some genes affect children's scholastic achievement because they evoke parenting behaviors that affect

reading skills. Then, since X and X' are the same in MZ pairs, MZ twins are on average treated more similarly by their parents than DZ pairs.

u. Show that under assumptions (6) and (7), Equations (3)-(5) can be written as:

$$1 = a^2 + u^2, \tag{8}$$

$$\rho_{YY'}^{MZ} = a^2 + \rho_{UU'}u^2, \tag{9}$$

$$\rho_{YY'}^{DZ} = \frac{1}{2}a^2 + \rho_{UU'}u^2. \tag{10}$$

The three model parameters $(a^2, u^2, \rho_{UU'})$ are now identified.

v. We will now solve for the model parameters. Using Equations (8), (9), and (10), show that the method-of-moments estimators are:

$$\hat{a}^2 = 2\left(\hat{\rho}_{YY'}^{MZ} - \hat{\rho}_{YY'}^{DZ}\right), \tag{11}$$

$$\hat{u}^2 = 1 - \hat{a}^2,$$

$$\hat{\rho}_{UU'} = \frac{\hat{\rho}_{YY'}^{MZ} - \hat{a}^2}{\hat{u}^2} = \frac{\hat{\rho}_{YY'}^{DZ} - \frac{1}{2}\hat{a}^2}{\hat{u}^2}.$$

Equation (11) is known as *Falconer's formula* or, colloquially, the *"double-the-difference" estimator*.

The above approach to twin studies is called the *ACE model*. In presentations of the ACE model, it is common to re-parameterize the environmental component so that the variance component that is correlated between siblings reared in the same household is called the *common environmental component* (c^2), or the *shared environmental component*:

$$c^2 \equiv \rho_{UU'} u^2.$$

The remaining part of the environmental component (e^2) is called the *non-shared environmental component*:

$$e^2 \equiv (1 - \rho_{UU'}) u^2.$$

The ACE model is then commonly written as

$$Y = aA + cC + eE,$$

where A, C, and E are random variables that each have mean 0 and variance 1.

w. Explain why this formulation is equivalent to Equation (2), where $X = aA$ and $U = cC + eE$.

x. Show that in the ACE model's notation, Equations (8)-(10) become

$$1 = a^2 + c^2 + e^2,$$

$$\rho_{YY'}^{MZ} = a^2 + c^2,$$

$$\rho_{YY'}^{DZ} = \frac{1}{2} a^2 + c^2.$$

y. Show that the method-of-methods estimators, using the ACE model and its notation (as opposed to that used in part (v)) are:

$$\hat{a}^2 = 2\left(\hat{\rho}_{YY'}^{MZ} - \hat{\rho}_{YY'}^{DZ}\right),$$

$$\hat{c}^2 = \hat{\rho}_{YY'}^{MZ} - \hat{a}^2 = \hat{\rho}_{YY'}^{DZ} - \frac{1}{2}\hat{a}^2,$$

$$\hat{e}^2 = 1 - \hat{a}^2 - \hat{c}^2.$$

The table below, based on Swedish data reported in Table S11 in Rietveld et al. (2013), shows the estimated correlations in years of education for seven different types of brothers: monozygotic twins, dizygotic twins, full siblings reared together (FST), full siblings reared apart (FSA), half-siblings reared together (HST), half-siblings reared apart (HSA), and adopted siblings with no genetic relationship (ADO).

	MZ	DZ	FST	FSA	HST	HSA	ADO
$\hat{\rho}_{YY'}$	0.71	0.50	0.45	0.20	0.25	0.13	0.21
N	1409	1922	206,518	1,362	6,554	14,713	858

z. Show that if you apply Falconer's formula (Equation (11)) to the two twin correlations in the table, the estimate of narrow heritability is $\hat{a}^2 = 0.42$. Show that if you apply the above method-of-moments estimator for $\hat{c}^2$, the estimate of the common environmental component is $\hat{c}^2 = 0.29$.

aa. Suppose that the siblings reared apart were randomly assigned to families (as we already discussed, this assumption is unlikely to hold in practice, but we make it nevertheless). Give some intuition for why, under this assumption, the ACE model implies the following moment conditions:

$$\rho_{YY'}^{FST} = \frac{1}{2}a^2 + c^2,$$
$$\rho_{YY'}^{FSA} = \frac{1}{2}a^2,$$
$$\rho_{YY'}^{HST} = \frac{1}{4}a^2 + c^2,$$
$$\rho_{YY'}^{HSA} = \frac{1}{4}a^2,$$
$$\rho_{YY'}^{ADO} = c^2.$$

ab. Using your estimates of a^2 and c^2 from part (z) and the moment conditions in part (aa), fill in the bottom row of the below table with the predicted values for the correlations.

	MZ	DZ	FST	FSA	HST	HSA	ADO
$\widehat{\rho}_{YY'}$	0.71	0.50	0.45	0.20	0.25	0.13	0.21
Prediction							

Note: Rather than using only the MZ and DZ correlations to estimate the parameters of the ACE model, it is possible instead to use all of the information contained in the correlations in the above table. However, the method of moments cannot be used because once we combine Equations (8)-(10) with those from part (aa) of this problem, we have too *many* equations relative to the number of parameters! In this situation, we say that the parameters are *overidentified*. One approach to estimation in this case is the *generalized method of moments (GMM)*. Loosely speaking, the basic idea is to find the parameter values so that the equations are jointly as good an approximation as possible. In particular, GMM finds parameter estimates so that the sum of squared deviations from exact equality across all the equations is minimized.

(For more details on the ACE model, see Goldberger (1978), on which our treatment is largely based.)

2. The ACDE model and the misspecified ACE model

In Problem 1, we generalized the additive genetic component of phenotypic variance from the single-locus case to the multi-locus case. We then applied this framework to sibling correlations to derive the ACE model.

Rather than assuming a purely additive model as we did in Problem 1, we can instead begin with a more general multi-locus case that includes dominance effects. We can then decompose the genetic effects into additive and dominance variance components, as we did in Problem 2(n) from Problem Set 1 for the single-locus case. When we apply this framework to sibling data, the result is the ACDE model, which we will explore in this problem.

To be more precise, for phenotype y and J independent biallelic loci, we can begin with the following true causal model:

$$y_i = \alpha + \sum_{j=1}^{J} \beta_j x_{ij} + \sum_{j=1}^{J} \delta_j I\{x_{ij} = 1\} + \epsilon_i,$$

where i indexes individuals, j indexes loci, and ϵ_i is an error term that has mean 0 and is independent of every genotype score x_{ij}. Each δ_j is a dominance deviation for locus j. (In the multi-locus case, there could also be interaction effects between the loci, called *gene-gene interactions*, or *epistasis*. We defer discussing such effects to later in the course.)

Just as in Problem 2 from Problem Set 1, we can decompose the overall genetic variance (the variance explained by the best predictor function) into an additive variance component (the variance explained by the best linear predictor function) and a dominance variance component. The additive and dominance variance components are constructed to have mean zero and to be mutually uncorrelated. (Analogously to the single-locus case, when there are dominance effects, the additive variance component is *not* simply equal to

$\sum_{j=1}^{J} \beta_j x_{ij}$, since the best linear predictor function will also pick up some of the dominance effects.)

Narrow heritability is defined as the fraction of phenotypic variance explained by the additive variance component derived from the true causal model. It is denoted:

$$a^2 \equiv \frac{Var(additive\ component)}{Var(y_i)}.$$

The fraction of phenotypic variance explained by the dominance variance component is denoted:

$$d^2 \equiv \frac{Var(dominance\ component)}{Var(y_i)}.$$

Broad heritability refers to the fraction of phenotypic variance explained by genetic effects as a whole: $a^2 + d^2$.

In the ACDE model (as in the ACE model), we decompose the residual variance, $\frac{Var(\epsilon_i)}{Var(y_i)} = 1 - a^2 - d^2$, into a shared environmental component that is the same for siblings reared in the same household, denoted c^2, and a non-shared component that is distinct for every individual, denoted e^2.

The ACDE model is commonly written as

$$Y = aA + cC + dD + eE,$$

where Y, A, C, D, and E are random variables that each have mean 0 and variance 1.

Since the non-shared environmental component E is constructed (as in Problem 1) to be uncorrelated with the common environmental component C, and uncorrelated across any

two individuals, we know that: $Cov(C, E) = 0$ and $Cov(E, E') = 0$. In addition, it is standard to make the following assumptions, which generalize the corresponding assumptions from Problem 1:

1. No gene-environment correlation: $Cov(A, C) = Cov(D, C) = Cov(A, E) = Cov(D, E) = 0$.
2. Equal-environment assumption: For siblings who grew up in the same household, $Cov(C, C') = 1$.

a. Suppose we restrict attention to the population of MZ twins and the population of DZ twins. Justify the following steps to derive the three bolded moment conditions:

$$Y = aA + cC + dD + eE$$

$$Var(Y) = Var(aA + cC + dD + eE)$$

$$Var(Y) = a^2Var(A) + c^2Var(C) + d^2Var(D) + e^2Var(E) + 2acCov(A, C) + 2adCov(A, D) + 2aeCov(A, E) + 2cdCov(C, D) + 2ceCov(C, E) + 2deCov(D, E)$$

$$1 = a^2 + c^2 + d^2 + e^2$$

$$Cov(Y, Y')^{MZ} = Cov(aA + cC + dD + eE, aA + cC + dD + eE)$$

$$= a^2Cov(A, A') + acCov(A, C') + adCov(A, D') + aeCov(A, E') + c^2Cov(C, C') + cdCov(C, D') + ceCov(C, E') + d^2Cov(D, D') + deCov(D, E') + e^2Cov(E, E')$$

$$= a^2Cov(A, A') + c^2Cov(C, C') + d^2Cov(D, D')$$

$$\rho_{YY'}^{MZ} = a^2 + c^2 + d^2$$

$$Cov(Y, Y')^{DZ} = Cov(aA + cC + dD + eE, aA + cC + dD + eE)$$

$$\begin{aligned} = a^2Cov(A, A') + acCov(A, C') + adCov(A, D') + aeCov(A, E') + c^2Cov(C, C') \\ + cdCov(C, D') + ceCov(C, E') + d^2Cov(D, D') + deCov(D, E') \\ + e^2Cov(E, E') \end{aligned}$$

$$= a^2Cov(A, A') + c^2Cov(C, C') + d^2Cov(D, D')$$

$$\rho_{YY'}^{DZ} = \frac{1}{2}a^2 + c^2 + \frac{1}{4}d^2$$

b. Explain why the parameters of the ACDE model (a^2, c^2, d^2, e^2) cannot be recovered from these three moment conditions.

If we assume that the dominance component is negligible for the phenotype under consideration ($d^2 \approx 0$), then the ACDE model specializes to the ACE model, which we know is identified (from Problem 1(v)). Since the ACE model can be estimated from twin data alone whereas the ACDE model cannot, researchers often estimate the ACE model. In doing so, they are making the identifying assumption that $d^2 = 0$.

Sometimes, researchers attempt to justify the assumption that $d^2 = 0$ on the basis of the results from estimating the ACE model. Here are two examples:

- "Since we observed $\hat{\rho}_{YY'}^{MZ} < 2\hat{\rho}_{YY'}^{DZ}$, there is no evidence of dominance, and an ACE model is appropriate."
- "Since we consistently see that $\hat{\rho}_{YY'}^{MZ} \approx 2\hat{\rho}_{YY'}^{DZ}$, the ACE model organizes the data most parsimoniously."

We will examine why such arguments have less force than it may seem.

c. Show that $\rho_{YY'}^{MZ} < 2\rho_{YY'}^{DZ}$ does not imply that d^2 must be zero. Give example values of the ACDE model parameters such that $d^2 > 0$ but nonetheless $\rho_{YY'}^{MZ} = 2\rho_{YY'}^{DZ}$.

In general, it is problematic to conclude from MZ and DZ twins alone that the ACE model provides a better fit to the data than the ACDE model does. Because the ACDE model parameters are not identified using MZ and DZ twin data, there are an infinite number of parameter values that satisfy the above equations. We cannot test the null hypothesis that $d^2 = 0$ because for any ACDE parameters with $d^2 = 0$, there is some other set of parameters with $d^2 > 0$ that fits the data equally as well.

Note that we *can*, however, reject the ACE model if we can reject the null hypothesis that $2\rho_{YY'}^{DZ} \geq \rho_{YY'}^{MZ}$ because (as we will show next) the alternative hypothesis falls outside the range of correlations that are consistent with the ACE model.

To illustrate, recall that under the ACE model,

$$\rho_{YY'}^{MZ} = a^2 + c^2$$

$$\rho_{YY'}^{DZ} = \frac{1}{2}a^2 + c^2.$$

Indeed, $2\hat{\rho}_{YY'}^{DZ} < \hat{\rho}_{YY'}^{MZ}$ implies that $c^2 < 0$ which is impossible.

Thus, to be consistent with the ACE model, and because variance components like a^2 and c^2 cannot be negative, we need:

- $\rho_{YY'}^{MZ}, \rho_{YY'}^{DZ} \geq 0$
- $2\rho_{YY'}^{DZ} \geq \rho_{YY'}^{MZ}$ with equality only if $c^2 = 0$
- $\rho_{YY'}^{MZ} \geq \rho_{YY'}^{DZ}$ with equality only if $a^2 = 0$

d. Suppose that ACDE model is true (that is, $d^2 > 0$), but we estimate the ACE model. Show that our ACE model estimator for the additive genetic component, $\hat{a}^2 = 2\left(\hat{\rho}_{YY'}^{MZ} - \hat{\rho}_{YY'}^{DZ}\right)$, will yield a biased estimate of a^2 even in a large sample:

$$plim(\hat{a}^2) = 2\left(\rho_{YY'}^{MZ} - \rho_{YY'}^{DZ}\right) = a^2 + \frac{3}{2}d^2.$$

Hint: In the above equation, substitute in the true values of $\rho_{YY'}^{MZ}$ and $\rho_{YY'}^{DZ}$ from the ACDE model, and then simplify.

e. Show that our ACE model estimator for the shared environmental component, $\hat{c}^2 = \hat{\rho}_{YY'}^{MZ} - \hat{a}^2$, will also yield a biased estimate of c^2 even in a large sample:

$$plim(\hat{c}^2) = \rho_{YY'}^{MZ} - plim(\hat{a}^2) = c^2 - \frac{1}{2}d^2.$$

Hint: In the above equation, substitute in the true value of $\rho_{YY'}^{MZ}$ and the value of $plim(\hat{a}^2)$ from above, and then simplify.

In sum, both estimators are asymptotically biased, with dominance causing the additive genetic component to be overestimated and the common environmental component to be underestimated.

f. Consider again the moment conditions from other types of sibling pairs from Problem 1(aa). Explain why the ACDE model parameters would still *not* be identified if (in addition to MZ and DZ twins) we also had data on full siblings reared together. Also, explain why the ACDE model parameters *would* be identified if we also had data on full siblings reared apart.

3. Interpreting heritability: The Jencks and Goldberger critiques

The previous problems were designed to help flesh out some challenges of inference that arise when estimating the parameters of behavior genetic problems. In this problem, we instead turn to some important interpretational challenges that arise once the parameters have been estimated. Although the fallacies were pointed out at least 35 years ago (as we will see), they persist to this day.

The Jencks Critique

We begin with a famous paper by the sociologist Christopher Jencks (1980). To provide some historical context, Jencks was writing at a time when estimates from twin studies were highly controversial, with emotions running high. Some critics argued that the family studies provided no evidence whatsoever that behavioral phenotypes were heritable.[1]

Jenck's contribution was not to question the evidence of heritability *per se*, but to question the usual interpretation that genetic effects on complex phenotypes operate via physiological mechanisms that are not modifiable. The first part of this problem asks you to work through a simple example that illustrates Jencks' point.

Suppose that an individual's scholastic achievement, y, is determined by the number of books read, x, and other environment factors, ϵ. The true causal model can be written as

$$y_i = \beta x_i + \epsilon_i, \tag{12}$$

[1] Although Turkheimer's (Turkheimer 2000) "First Law" of behavior genetics ("Everything is heritable") is not universally acknowledged today, much of the controversy of the 1970s has nonetheless now subdued. Of course, one can accept the "First Law" while at the same time believing that heritability is overestimated by Falconer's formula-type estimators.

where i indexes individuals, and ϵ_i is an i.i.d. error term that has mean 0 and variance σ_ϵ^2 and is independent of x_i. (To streamline notation, we omit the tildas on top of the variables even though they are standardized as in Problem 4 from Problem Set 1.)

Suppose further that the number of books read, x, is determined by additive genetic factors, z, and environment factors, η. The true causal model can be written as

$$x_i = \gamma z_i + \eta_i, \tag{13}$$

where η_i is an i.i.d. error term that has mean 0 and variance σ_η^2 and is independent of z_i. To keep the calculations simple, we will further assume that η_i is independent of ϵ_i.

a. Recall that the narrow heritability of a phenotype is defined as the phenotypic variance explained by the additive genetic component derived from the true causal model. Show that the narrow heritability of the number of books read, denoted h_x^2, is equal to γ^2.

   Hint: Use the result from Problem 4(d) from Problem Set 1.

In the remainder of this problem, we will use the term "heritability" as shorthand for "narrow heritability."

b. Now we'll calculate the heritability of scholastic achievement, h_y^2. First, using Equations (12) and (13), show that

$$y_i = \delta z_i + \xi_i,$$

   where $\delta \equiv \beta\gamma$, and $\xi_i \equiv \epsilon_i + \beta\eta_i$ is an i.i.d. error term that has variance $\sigma_\epsilon^2 + \beta^2\sigma_\eta^2$.

   Second, show that the heritability of scholastic achievement is $h_y^2 = \beta^2\gamma^2$.

It is often asserted that heritability is the fraction of variance that is *not* explained by environmental factors. In his paper, Jencks (1980) pointed out that this assertion is wrong whenever the variance explained by genetic factors (heritability) and the variance explained by environmental factors are not mutually exclusive. We will now illustrate this result.

c. With reference to Equation (12), explain why the fraction of variance in scholastic achievement explained by environmental factors (books together with other environmental factors) is 100%.

d. Show that if we add up the variance in scholastic achievement explained by environmental factors and the variance explained by genetic factors, is equal to 100% + $\beta^2\gamma^2$.

e. Explain what is going on: how is it possible that the fraction of variance explained by environmental and genetic factors adds up to more than 100%?

<u>The Goldberger Critique</u>

We turn next to another common interpretational error, which is to assert that if the heritability of a trait is high, then it cannot be changed through policy interventions.

To provide some historical context, in 1976, the *American Economic Review* published a twin study that reported estimates of the heritabilities of earnings and educational attainment (Taubman 1976). Asked to comment on the study's findings, a famous scientist (not involved in the study) told the *Times of London* that the results "really tell the [Royal] Commission [on the Distribution of Income] that they might as well pack up." In response to this comment, Goldberger sarcastically remarked:

> In the same vein, if it were shown that a large proportion of the variance in eyesight were due to genetic causes, then the Commission on the Distribution

> of Glasses might as well pack up. And if it were shown that most of the variation in rainfall is due to natural causes, then the Commission on the Distribution of Umbrellas could pack up too.

Goldberger's retort has become famous because it elucidates using obvious examples why (i) heritability estimates are not indices of policy effectiveness, and (ii) it is absurd to assert that there is never room for policy to remedy inequalities that are caused by "natural" factors. (For a discussion, see p. 338 in Goldberger (1979).)

We will work through a simple example intended to illustrate Goldberger's first point about policy effectiveness.

Recall from part a that the heritability of the number of books read, h_x^2, is equal to γ^2. Decompose the error in Equation (13), η_i, as:

$$\eta_i = \eta_i^{policy} + \eta_i^{non-policy},$$

where η_i^{policy} is the environmental component influenced by policy, and $\eta_i^{non-policy}$ captures the remaining environmental factors. Suppose that initially, $\eta_i^{policy} = 0$ for all i, but then the government adopts a policy of giving out free books to all families with children, so that now $\eta_i^{policy} = 1$ for all i (and the distribution of $\eta_i^{non-policy}$ is unaffected).

f. With reference to the model, explain why the following two facts are true:

   i. The heritability of the number of books read is equal to γ^2 (which could be 100%) both before and after the policy is instituted.

   ii. The policy increases the scholastic achievement of all individuals by β standard deviations.

4. Meta-Analysis

Meta-analysis is a common procedure for boosting statistical power to detect associations by combining effect-size estimates from multiple independent datasets or studies. Meta-analysis is particularly useful in genome-wide association studies of complex traits (i.e., genome-wide association studies, or GWAS). In GWAS, the individual datasets are often referred to as "cohorts." Typically, very large sample sizes are required for GWAS to be well powered—much larger than the number of subjects in any one cohort. The purpose of this problem set is to familiarize you with some basic theoretical and statistical issues in meta-analysis.

In a typical GWAS meta-analysis, investigators start by formulating an analysis plan. This plan specifies the association analyses to be conducted separately by each of the participating cohorts. Summary statistics from the cohort-level analyses (e.g., $\hat{\beta}'s$) are subsequently stored in a results file, one for each cohort, and uploaded to a central server. After it has been verified that the uploaded files pass certain quality-control checks, the association statistics from the C cohort-level files are combined in a meta-analysis. (We will discuss the many quality checks involved in meta-analysis of genetic associations in detail in class, but an excellent resource is Winkler et al. (2014)).

a. An alternative to meta-analysis is so-called mega-analysis, in which participating cohorts upload the *individual-level* genotype and phenotype data (rather than summary statistics) to the central server. The cohort-level association analyses are then performed on the central server, and the association summary statistics from each of these analyses are subsequently meta-analyzed. Describe one potential advantage and one potential disadvantage of mega-analysis vis-à-vis conventional meta-analysis.

In what follows, suppose the phenotype is height (in cm), and suppose that our meta-analysis is based on association statistics from C non-overlapping (independent) cohorts with sample sizes $N_1, \ldots, N_C$. Each row in a results file contains information about association statistics for a specific genetic variant or SNP (single-nucleotide polymorphism). Thus, it is

straightforward to merge the results files to get, for each SNP, information about the estimated effect size $(\hat{\beta}_1, \dots, \hat{\beta}_C)$ and the variance of each effect-size estimate $\left(Var(\hat{\beta}_1), \dots, Var(\hat{\beta}_C)\right)$ in each cohort (along with other information, such as the cohort-level minor allele frequency).

To avoid cluttering the notation below, we drop SNP subscripts and assume that each cohort only uploaded association results for a single SNP. In practice, association results are typically provided for millions of SNPs, but the procedures are the same.

In meta-analysis, we wish to combine association statistics from participating cohorts using some appropriate procedure. Most estimators have the form:

$$\hat{\beta}_{meta} = \sum_{c=1}^{C} w_c \hat{\beta}_c,$$

where w_c is the weight assigned to the coefficient estimate from cohort $c \in \{1,2, \dots, C\}$.

b. Assuming that each cohort-level estimate is unbiased and $\sum_{c=1}^{C} w_c = 1$, show that $\hat{\beta}_{meta}$ is unbiased. Note that the weights are not random variables; they are fixed constants.

c. Show that, given our assumption of no sample overlap between our cohorts,

$$SE\left(\hat{\beta}_{meta}\right) = \sqrt{\sum_{c=1}^{C} w_c^2 Var(\hat{\beta}_c)}\,.$$

The resulting meta-analysis *t*-statistic is

$$Z_{meta} \equiv \frac{\hat{\beta}_{meta}}{SE\left(\hat{\beta}_{meta}\right)}.$$

If the sample size of the meta-analysis is large, then this *t*-statistic can be treated as a Z-statistic (i.e., assumed to follow a normal distribution). It can then be used to calculate the *p*-value of the null hypothesis of no association $\left(\beta_{meta} = 0\right)$.

We have not yet defined how the weights, w_c, are constructed. Suppose we are interested in finding the unbiased estimator with the smallest variance. That is, we wish to minimize $SE\left(\hat{\beta}_{meta}\right)$ subject to the constraint that $\sum_{c=1}^{C} w_c = 1$.

d. Show that, for $C = 2$, the solution is

$$w_c = \frac{\frac{1}{Var\left(\hat{\beta}_c\right)}}{\left(\sum_{c=1}^{C} \frac{1}{Var\left(\hat{\beta}_c\right)}\right)}.$$

Note that this solution generalizes to $C > 2$. As an optional problem, interested students may prove this themselves. One method is to use the method of Lagrange multipliers, which allows one to find the minimum or maximum of a function subject to some constraint. For example, if we wish to minimize or maximize $f(x, y)$ subject to $g(x, y) = c$, the "Lagrangian" is $L(x, y, \lambda) = f(x, y) - \lambda(g(x, y) - c)$. We then find the derivative of the Lagrangian with respect to x, set it equal to 0, and solve for x.

This weighting scheme is known as *inverse-variance weighted meta-analysis* because the effect size estimated in a cohort ($\hat{\beta}_c$) is weighted by the inverse of its variance, which is then standardized to ensure that the weights sum to one ($\sum_{c=1}^{C} w_c = 1$).

For a number of reasons, some of which we illustrate below, it is sometimes preferable to use a different weighting scheme called *sample-size weighted meta-analysis*. Under this approach, we set

$$w_c = \frac{N_c}{N},$$

where N_c is the sample size in cohort c and $N = \sum_{c=1}^{C} N_c$ is the total sample size across all cohorts considered in the meta-analysis. We will now show that this sample-size weighted meta-analysis is roughly equivalent to inverse-variance weighted meta-analysis.

We will use a key fact about the sampling variance of the regression coefficient in a bivariate regression. Specifically:

$$Var(\hat{\beta}_c) = \frac{(\sigma_\varepsilon^2 / \sigma_X^2)}{N_c}, \tag{14}$$

where σ_ε^2 is the variance of the residual and σ_X^2 is the variance of the explanatory variable. We assume that each cohort is equivalent in every way except for sample size, and therefore that $\sigma_\varepsilon^2 / \sigma_X^2$ is the same across cohorts (i.e., doesn't depend on c).

e. Using Equation 14, show that the inverse-variance weights, $w_c = \frac{\frac{1}{Var(\hat{\beta}_c)}}{\left(\sum_{c=1}^{C} \frac{1}{Var(\hat{\beta}_c)}\right)}$, are equivalent to the sample-size weights, $w_c = \frac{N_c}{N}$. Conclude that we could conduct our meta-analysis as: $\hat{\beta}_{meta} = \sum_{c=1}^{C} \left(\frac{N_c}{N}\right) \hat{\beta}_c$.

Note that in practice, these two methods of constructing weights, $w_c = \frac{\frac{1}{Var(\hat{\beta}_c)}}{\left(\sum_{c=1}^{C} \frac{1}{Var(\hat{\beta}_c)}\right)}$ and $w_c = \frac{N_c}{N}$, will not lead to exactly the same weights (although they will be similar). There are two reasons. First, the sample-size weights use the formula in Equation 14, which describes the *true* sampling variance of $\hat{\beta}_c$. But since in practice, neither σ_ε^2 nor σ_X^2 is known, this true sampling variance is unknown. The inverse-variance weights therefore use the *estimated* sampling variance of $\hat{\beta}_c$ (i.e., the squared standard error), which will differ from the true sampling variance in finite samples. Second, the sample-size weights were derived under

the assumption that the cohorts are equivalent in every way except for sample size, but in practice they may differ, for example, in σ_ε^2. (For example, a cohort with a noisier measure of the phenotype will have a larger residual variance σ_ε^2.) Because inverse-variance weights use the estimated sampling variance, they account for such cohort differences.

f. In part (e), you derived one version of sample-sized weighted meta-analysis. Now we consider another version, a meta-analysis of the *z*-statistics instead of the $\hat{\beta}_c$ coefficients. That is, we aim to calculate the *z*-statistic for the meta-analysis, $Z_{meta} = \frac{\hat{\beta}_{meta}}{SE(\hat{\beta}_{meta})}$, as a function of the *z*-statistics of the cohorts, $Z_c = \frac{\hat{\beta}_c}{SE(\hat{\beta}_c)}$. Starting from $\hat{\beta}_{meta} = \sum_{c=1}^{C} \left(\frac{N_c}{N}\right) \hat{\beta}_c$, show that: $Z_{meta} = \sum_{c=1}^{C} \left(\sqrt{\frac{N_c}{N}}\right) Z_c$. Hint: From Equation 14 above, we know that $SE(\hat{\beta}_c) = \sqrt{Var(\hat{\beta}_c)} = \sqrt{\frac{(\sigma_\varepsilon^2/\sigma_X^2)}{N_c}} = \frac{(\sigma_\varepsilon/\sigma_X)}{\sqrt{N_c}}$ and similarly $SE(\hat{\beta}_{meta}) = \frac{(\sigma_\varepsilon/\sigma_X)}{\sqrt{N}}$.

This result is important for GWAS meta-analysis because often, the GWAS software used by cohorts outputs *z*-statistics rather than $\hat{\beta}_c$'s. This formula means that we can do sample-size weighted meta-analysis even though we observe the *z*-statistics rather than the $\hat{\beta}_c$'s. But there is an important subtlety: notice that in this case, the weights are the *square roots* of the cohort sample proportions (and their sum, $\sum_{c=1}^{C} \left(\sqrt{\frac{N_c}{N}}\right)$, is *not* equal to one).

Suppose in what follows that <u>in each cohort</u>, we tested a SNP for association with a phenotype by estimating the parameters of the following regression by ordinary least squares:

$$Y = \beta_0 + \beta_1 SNP + Z\theta + \epsilon,$$

where Y is the dependent variable, SNP is the number of reference alleles (0, 1, or 2) and Z is a vector of controls included to guard against stratification biases. For expositional simplicity, we assume all variables are de-meaned in what follows (and hence $\hat{\beta}_0 = 0$).

g. In quality-control and data analysis (e.g., for inverse-variance weighted meta-analysis), it is common to use the following approximation to the OLS standard error for the estimate of β_1:

$$\sqrt{Var(\hat{\beta}_1)} \approx \frac{\hat{\sigma}_Y}{\sqrt{N}} \times \frac{1}{\sqrt{2 \times MAF \times (1 - MAF)}}, \quad (15)$$

where $\hat{\sigma}_Y$ is the sample standard deviation of the dependent variable. In this problem, you are asked to derive this approximation formula from Equation 14 above. The hope is that the derivation will help you appreciate the underlying assumptions and help you use the approximation formula appropriately in your own research. Justify the following steps:

1. We begin by applying Equation 14 to a bivariate regression of Y on SNP. Using estimates of the quantities in the equation instead of their true values, show that:

$$\sqrt{\widehat{Var}(\hat{\beta}_1)} \approx \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - SNP_i\hat{\beta}_{1,bivariate}\right)^2}{\sum_{i=1}^{N} SNP_i^2}},$$

where i indexes individuals in the cohort under consideration, and $\hat{\beta}_{1,bivariate}$ is the estimated coefficient from a bivariate regression of Y on SNP.

2. Assuming that standard OLS assumptions about the error term hold, explain why when $corr(SNP, Z)$ and $corr(Y, Z)$ are very small, the coefficient on SNP will be approximately equal to each other in the following two regressions: (i) a regression of Y on SNP, and (ii) a regression of Y on SNP *and* Z. Feel free to make this

argument intuitively; if you want to make the argument formally, one simple approach is to use the Frisch-Waugh theorem. Conclude that

$$\sqrt{\widehat{Var}(\hat{\beta}_1)} \approx \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - SNP_i\hat{\beta}_1\right)^2}{\sum_{i=1}^{N} SNP_i^2}},$$

where $\hat{\beta}_1$ is the estimated coefficient from a multivariate regression of Y on SNP and Z. That substitution is important because in a GWAS meta-analysis, the cohort-level results will be from such a multivariate regression.

3. Rearrange the expression from part 2 to get:

$$\sqrt{\widehat{Var}(\hat{\beta})} \approx \sqrt{\frac{1}{N}\frac{\hat{\sigma}_y^2 - \hat{\sigma}_{SNP}^2\hat{\beta}_1^2}{\hat{\sigma}_{SNP}^2}}.$$

Optional question: Can you think of a realistic GWAS setting in which the assumptions required in the derivations above do not hold?

4. When you are working with cohort-level summary statistics, $\hat{\sigma}_{SNP}^2$ is often unknown. However, cohorts usually supply the minor allele frequency (MAF), so we can solve for $\hat{\sigma}_{SNP}^2$ as a function of MAF under Hardy-Weinberg equilibrium using the formula derived in Problem 2(f) of Problem Set 1. Show that making this substitution gives:

$$\sqrt{\widehat{Var}(\hat{\beta}_1)} \approx \sqrt{\frac{1}{N}\frac{\hat{\sigma}_y^2 - 2 \times MAF \times (1 - MAF) \times \hat{\beta}_1^2}{2 \times MAF \times (1 - MAF)}}.$$

Optional question: What happens to the approximation formula above if the genotype distribution deviates from the HWE prediction?

5. Solve for the approximation formula (Equation 15) by arguing that for realistic (i.e., very small) values of $\hat{\beta}_1$:

$$\sqrt{\frac{1}{N} \times \frac{\hat{\sigma}_y^2 - 2 \times MAF \times (1 - MAF)\hat{\beta}_1^2}{2 \times MAF \times (1 - MAF)}} \approx \sqrt{\frac{1}{N} \times \frac{\hat{\sigma}_y^2}{2 \times MAF \times (1 - MAF)}}$$

$$= \frac{\hat{\sigma}_Y}{\sqrt{N}} \times \frac{1}{\sqrt{2 \times MAF \times (1 - MAF)}}.$$

In practice, people often use sample-size weighted meta-analysis (rather than inverse-variance weighted meta-analysis) because it is more robust to accidental differences in the scaling of the dependent variable. In parts (h) through (j), we illustrate this property by way of a simple and highly stylized example.

Consider a meta-analysis of two cohorts in which the approximation formula above, Equation 15, is accurate. Assume that in both cohorts, the sample size is 10,000, the true effect of the SNP (β) is 2.5 cm (roughly 1 inch) per allele, the variance of height is 7 cm^2 and the MAF of the SNP is 0.5.

h. Derive the weights under sample-size weighting and inverse-variance weighting.

i. Suppose that one of the cohort-level analysts inadvertently uploaded association results from an analysis in which height was measured in inches instead of cm. How will the weights change under inverse-variance and sample-size weighting, given this error? What will $\hat{\beta}_{meta}$ be under sample-size and inverse-variance weighting? How are the cohort-level z-statistics affected by this measurement error?

References cited throughout

Goldberger, AS. 1978. "The Genetic Determination of Income: Comment." *American Economic Review* 68(5):960–69.

Goldberger, AS. 1979. "Heritability." *Economica* 46(184):327–47.

Jencks, C. 1980. "Heredity, Environment, and Public Policy Reconsidered." *American Sociological Review* 45(5):723–36.

Okbay, Aysu et al. 2016. "Genome-Wide Association Study Identifies 74 Loci Associated with Educational Attainment." *Nature* 533:539–42.

Sacerdote, B. 2007. "How Large Are the Effects from Changes in Family Environment? A Study of Korean American Adoptees." *The Quarterly Journal of Economics* 122(1):119–57.

Taubman, P. 1976. "The Determinants of Earnings: Genetics, Family, and Other Environments: A Study of White Male Twins." *American Economic Review* 66(5):858–70.

Turkheimer, Eric. 2000. "Three Laws of Behavior Genetics and What They Mean." *Current Directions in Psychological Science* 9(5):160–64.

Willer, Cristen J., Yun Li, and Gonçalo R. Abecasis. 2010. "METAL: Fast and Efficient Meta-Analysis of Genomewide Association Scans." *Bioinformatics* 26(17):2190–91.

Winkler, Thomas W. et al. 2014. "Quality Control and Conduct of Genome-Wide Association Meta-Analyses." *Nature Protocols* 9(5):1192–1212.

Wood, Andrew R. et al. 2014. "Defining the Role of Common Variation in the Genomic and Biological Architecture of Adult Human Height." *Nature Genetics* 46(11):1173–86. Retrieved (http://dx.doi.org/10.1038/ng.3097).

Wood, A. R. et al. 2014. "Defining the Role of Common Variation in the Genomic and Biological Architecture of Adult Human Height." *Nat. Genet*. 46:1173–86.

Computational Problems

1. Introduction and Background

The purpose of these problems is to give you experience implementing genetic analyses on real data from Add Health on the secure server we have set up for you.

The Add Health data can be found in the /home/data/AH_phenotypes/ and /home/data/AH_genotypes/ folders on the server, located at g03.nber.org. You have read-only access to these folders. For more details, see the module "Data Access Basics".

Original phenotype data

The /home/data/AH_phenotypes/ folder contains four subfolders called "Wave[1-4]" , one for each wave of the Add Health survey. There is a single file within each subfolder; each of these contains the core set of variables ascertained from respondents in that particular wave. You are welcome to explore these files.

- /Wave1/allwave1.xpt -- 1994-5: grades 7-12; in-school, in-home, parents, and school administrators
- /Wave2/wave2.xpt -- 1996: follow-up in-home interviews with adolescents and follow-up school administrator interviews
- /Wave3/wave3.xpt -- 2001-2: in-home interviews with original respondents (now young adults) and in-home interviews with their partners
- /Wave4/wave4.xpt -- 2008: fourth in-home interview with the original Wave I respondents; personal interview that included physical measurements and biospecimen collection

One useful way to explore most non-zipped files on a server is to use the "less" command followed by the name of the file. You can scroll through results by hitting "enter" and then exit back to your previous screen by hitting "q." The "head –n10" command, when followed

by the filename, is also useful; it will display just the first 10 lines of the file specified. See the codebooks distributed to you for more information. Note that the files listed above (ended with ".xpt") cannot be directly viewed via the "less" command. You can read these data in R by typing the following commands:

```
library(Hmisc)
df <- sasxport.get("FILENAME.xpt", lowernames=FALSE)
head(df)
```

Genetic data

The /home/data/AH_genotypes/ folder containing the original set of Plink binary bim, bed, and fam genetic files (directly genotyped, non-imputed data):

- /home/data/AH_genotypes/omni_joined.freeze3.sharedMarkers

Cleaned data

To reduce computing time and intensity, we have done much of the cleaning of the genetic data for you, using a small "test" subset of the data. Relatedly, we have already cleaned and extracted much of the information you'd need from the phenotype files for this problem set. These files can be found in /home/data/AH_cleaned/. We will describe these files as we progress through the computational problems. Note that there are additional files in this folder that we will not use during the summer institute and that will not be discussed.

Most importantly for now: In the /home/data/AH_cleaned/ folder, there is a file titled "ah_ea_sex_byear_pcs_euros.csv". It includes only those respondents who self-reported as "White" during Wave 1 (i.e. H1GI6A = 1). It contains the following variables:

- AID: respondent unique identifier
- H4OD1Y: birth year
- BIO_SEX4: sex

- HEIGHT: height (outliers dropped)
- EA: educational attainment in years (coded as listed below)
- FID: family identifier
- PC1-PC10: 10 principal components (generated from each chip separately on the entire sample available)

EA was recoded as follows from Wave 4's H4ED2 variable:

- 1 (8th grade or less): 8
- 2 (some high school): 10
- 3 (high school graduate): 12
- 4 (some vocational/technical): 13
- 5 (completed vocational/technical): 14
- 6 (some college): 14
- 7 (completed college): 16
- 8 (some graduate school): 17
- 9 (completed a master's): 18
- 10 (some graduate training beyond master's): 20
- 11 (completed a doctoral degree): 22
- 12 (some post bac professional, e.g. law, med, nurse)): 18
- 13 (completed post bac professional education, e.g. law, med, nurse)): 20
- 96 (refused): dropped
- 98 (don't know): dropped

We have saved the codes for cleaning the Add Health data, formatting the input data, etc. for subsequent computational problems in "/home/TA_sample_scripts/". The scripts are written in three languages – bash/shell, Python or R. The purpose of each of these scripts will become clear as we go through the problem sets later. There might be files in this folder that we will not directly discuss or cover in the TA session – those are mostly cleaning scripts or codes that generate input data for the problems. If you have any questions about them, feel free to ask us (Grant and Tami)!