2019 Russell Sage Foundation Summer Institute in Social-Science Genomics

Problem Set 4

This problem set focuses on polygenic scores, gene-environment interaction, and Mendelian randomization. It is due at 9:30am on Wednesday, June 19.

1. Interpreting gene-environment interaction regressions when genetic and environmental factors are correlated

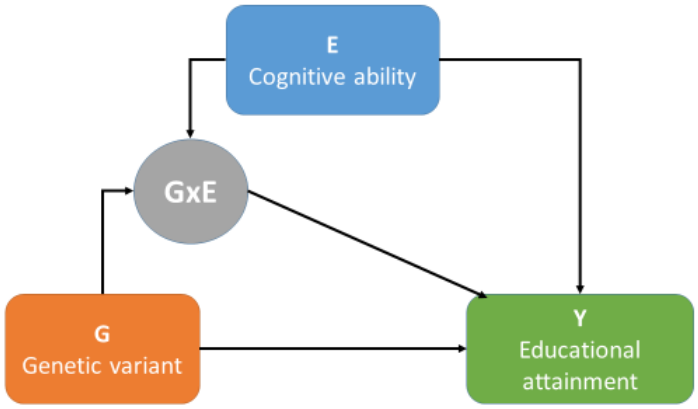
Suppose we wish to estimate the model

$$y_i = \beta_0 + \beta_1 e_i + \beta_2 x_i + \beta_3 e_i \times x_i + \epsilon_i, \quad (1)$$

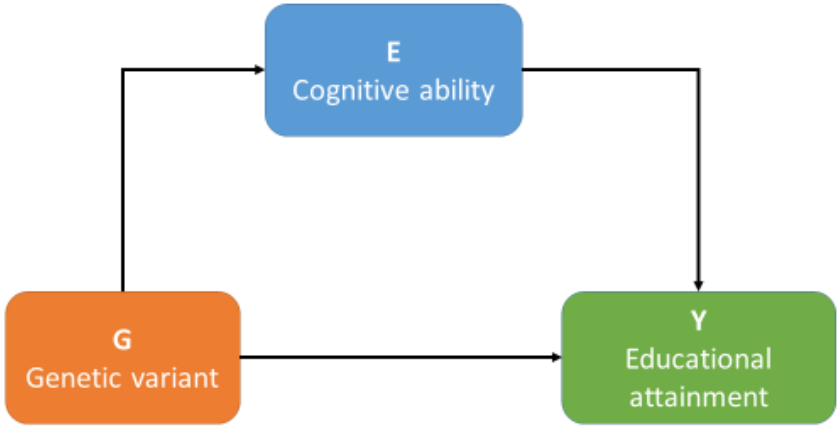
where x_i is individual *i*'s genotype, e_i is individuals *i*'s environment (as measured by a specific environmental variable), and where we ignore control variables for simplicity. We would like to test H_0: $\beta_3 = 0$ vs. H_1: $\beta_3 \neq 0$.

There are several scenarios where we can observe a non-zero correlation between x_i and e_i: $\text{Cov}(x_i, e_i) \neq 0$.

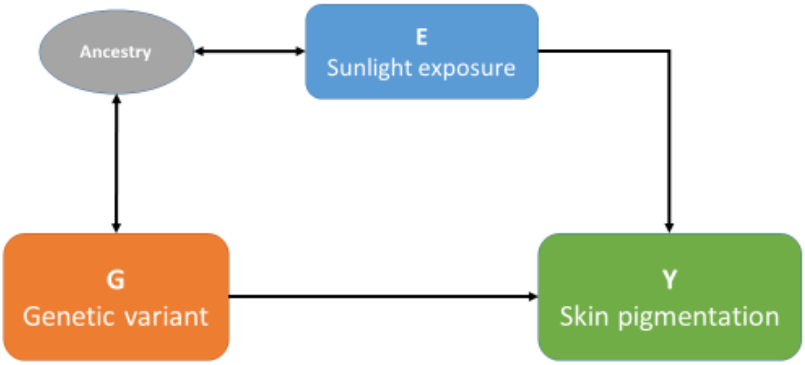
1) We might want to study how a genetic variant's effect on years of schooling (y_i) is modified by cognitive ability (e_i) – this is called the "moderating effect" of cognitive ability on educational attainment.



2) We might want to study how the genetic variant also influences years of schooling (y_i) *via* its effect on cognitive ability (e_i) – this is called the "mediating effect".



3) We might want to study how a genetic variant's effect on skin pigmentation (y_i) is modified by a daily sunlight exposure (e_i), but both the allele frequencies and daily sunlight exposure are correlated with ancestry.



a. Now suppose that we estimate Equation 1 and find that β_3 is significantly different from 0. Can we then conclude there are gene-environment interactions? (Hint: You can define $e_i = \theta x_i + \varepsilon_i$ because $\text{Cov}(x_i, e_i) \neq 0$.)

b. Suppose x_i and e_i are *not* correlated, and we estimate Equation 1 and find that β_3 is significantly different from 0. Can we then conclude that there is a gene-environment interaction? (Hint: While $\text{Cov}(x_i, e_i) = 0$, the genotype at a *different* locus ($x_{i,2}$) might be correlated with e_i. In this case, we could write $e_i = \theta x_{i,2} + \varepsilon_i$.)

Thus, one must establish that an interaction is really driven by the environmental component (not the genetic component) of the e_i variable. This is why, when possible, it is desirable to use random variation in the environmental variable from either an actual randomization or a natural experiment.

c. Many gene-environment interaction studies test whether family environment interacts with a genetic variant of interest. For instance, in their study of aggressive behavior, Caspi et al. (2002) examine if the *MAOA* gene interacts with childhood maltreatment.

They find a significant interaction. Are they right to conclude that their data suggest there is a gene-environment interaction? What is an alternative explanation?

(Hint (to help you think of one possible alternative explanation): What if the childhood maltreatment variable is genetically correlated with aggressiveness, and the childhood maltreatment variable and aggressiveness are heritable?)

2. Mendelian randomization, the exclusion restriction, and MR-Egger regression

Suppose we want to estimate the causal effect of an environmental variable e on some outcome y. Suppose the true causal model is:

$$y = e\beta + \epsilon, \tag{2}$$

$$e = x_j\gamma_j + \zeta_j, \tag{3}$$

where β is the causal effect that we want to estimate, x_j is the genotype of genetic variant $j \in \{1,2, \dots, J\}$ that has causal effect γ_j on e and that is available in our data, and ϵ and ζ_j are residuals. To conserve notation, we define y, e, and x_j as de-meaned variables, and we omit the *i* subscript (that would index individuals) from y, e, ϵ, x_j, and ζ_j. We code the alleles for each genetic variant such that $\gamma_j > 0$ for all j. We assume that ϵ and every x_j are independent of every ζ_j, and we assume that all the x_j's are independent of each other. We denote our sample size of individuals by N, which we will assume is a large number.

The problem we face is that e is not exogenous: $\text{Cov}(e, \epsilon) \neq 0$. As a result, simply estimating regression Equation 2 in our data would give a biased estimate of the true causal effect β. The idea of *Mendelian randomization*, often abbreviated *MR*, is to exploit the fact that we have data on genetic variants (the x_j's) that we know affect e—as described in Equation 3—in order to get a consistent estimate of β. Specifically, the idea is that if we can treat the x_j's as randomly assigned independent of ϵ, then the changes in e that are caused by different values of the x_j's are exogenous, and we can restrict our attention to the effect on y of those differences in e that are caused by differences in the x_j's. In essence, we treat the x_j's as a natural experiment that gives us random variation in e, and then we use that random variation to estimate the causal effect of e on y. We will refer to the x_j's as *instrumental variables*, or *instruments* (because they are not our direct objects of interest but are used only instrumentally to help us estimate the causal effect of e on y).

(MR is a specific application of *instrumental variables estimation* where the instruments are a set of one or more genetic variants. The term "Mendelian randomization" comes from Mendel's law of segregation, which states that one allele from each parent is inherited at random. If we conduct a MR study with sibling data and include family fixed effects in Equation 3, then we reproduce exactly this Mendelian experiment (for an example, see Fletcher and Lehrer (2011)). In the more usual case, we do not have family data, and then the usual concerns about population stratification apply and need to be addressed. Even if we can exploit randomization within a family, however, MR still relies on the assumption of the exclusion restriction, discussed next.)

The key assumption underlying MR is called the *exclusion restriction*:

$$\text{Cov}(x_j, \epsilon) = 0. \tag{4}$$

This assumption states that the genetic variants do *not* affect the outcome y through any channel other than the environmental factor e. (It is called the exclusion restriction because we are "excluding" any other channels.) If the exclusion restriction fails, then we cannot use the random variation from the x_j's to isolate the causal effect of e on y because the variation in the x_j's also affects y for other reasons.

(In practical applications of MR, the other key assumption is that the "instrument is strong," meaning that the x_j's are sufficiently predictive of e (where "sufficiently predictive" depends on the sample size). Throughout this problem, we assume that the analysis is conducted in a sufficiently large sample of individuals that this source of bias is negligible.)

MR estimation involves a two-stage approach called *two-stage least squares*, or *2SLS*. We will consider it in the simplest case of just a single instrument ($J = 1$), denoted x.

In the first stage, you estimate $\hat{\gamma}$ by running the regression of e on x from Equation 3. You then create the predicted value of e from that regression, $\hat{e} = x\hat{\gamma}$; this predicted value isolates the component of the variation in e that is due to x. In the second stage, you run a regression of y on $\hat{e}$. The slope coefficient from this regression is the 2SLS estimator: $\hat{\beta}_{\text{2SLS}} \xrightarrow{N\to\infty} \frac{\text{Cov}(\hat{e},y)}{\text{Var}(\hat{e})}$.

a. Show that under the exclusion restriction, $\hat{\beta}_{\text{2SLS}}$ is a consistent estimate of β. (Hint: Substitute $\hat{e} = x\hat{\gamma}$ into the limit of the 2SLS estimator, and use the fact that $\hat{\gamma} \xrightarrow{N\to\infty} \frac{\text{Cov}(e,x)}{\text{Var}(x)}$.)

Unfortunately, it is impossible to directly test the exclusion restriction using the data at hand. MR studies are most persuasive when we have detailed knowledge of the biology that underlies the effects of the genetic variants—detailed enough that we can be confident not only about what the relevant genes do but also what they do *not* do. Skeptics about MR argue that *pleiotropy*—that is, multiple phenotypic effects from a given gene—is rampant, and that our state of knowledge is currently inadequate to rule out violations of the exclusion restriction. Moreover, even if the genetic variants we use as instruments satisfy the exclusion restriction, they could be in LD with other variants that violate it. For examples of the debate and discussion, see Cawley, Han, and Norton (2011) and Taylor et al. (2015).

In the context of instrumental variables estimation (not MR specifically), Kolesár et al. (2011) recently showed that if we have access to multiple instruments ($J > 1$), then we can weaken the exclusion restriction and still obtain consistent estimates of the causal effect of e on y. To understand the weaker assumption that is needed, we will first define

$$\alpha_j \equiv \frac{\text{Cov}(x_j, \epsilon)}{\text{Var}(x_j)}. \tag{5}$$

α_j is the coefficient you would get from regressing ϵ (the determinants of y other than e) on x_j. In other words, it is the effect of x_j on ϵ. It is a measure of the extent to which the exclusion restriction is violated for instrument j. If the exclusion restriction holds for instrument j, then $\alpha_j = 0$.

The assumption is that α_j (the effect of x_j on ϵ) is uncorrelated with γ_j (the effect of x_j on e):

$$\text{Cov}(\alpha_j, \gamma_j) = 0. \tag{6}$$

In other words, the assumption states that across the J genetic variants, the effect of a genetic variant's effect on ϵ is uncorrelated with its effect on e. Bowden et al. (2015) named this assumption *InSIDE* (which stands for Instrument Strength Independent of Direct Effect) and explored it in the context of MR. The remainder of this problem largely mirrors Bowden et al.'s analysis.

b. Show that InSIDE is a weaker assumption than the exclusion; that is, if the exclusion restriction holds, then the InSIDE assumption holds, but the converse is not necessarily true.

We will now examine how a consistent estimate of β can be obtained, assuming that the InSIDE assumption holds. The approach is a two-stage procedure. In the first stage, we run two sets of regressions:

1. For each genetic variant (i.e., instrument) x_j, we estimate regression Equation 3, the regression of e on x_j. (This is also the "first-stage" regression of the standard MR estimation procedure.) From these regressions, we obtain a set of coefficient estimates, $\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_J$.

2. For each genetic variant x_j, we estimate the following regression of the phenotype on x_j:

$$y = x_j\Gamma_j + w_j,$$

where Γ_j is the coefficient we are estimating, and w_j is the residual. (In the language of instrumental variables estimation, this regression is called the "reduced-form." It is the regression you would get if you started with the two regression Equations 2 and 3, and then you "reduced" the two equations to a single equation by substituting Equation 3 into the value of e in Equation 2.) From these regressions, we obtain a set of coefficient estimates, $\widehat{\Gamma_1}, \widehat{\Gamma_2}, \ldots, \widehat{\Gamma_J}$.

c. Show that in a large sample of individuals, each of the estimated coefficients, $\hat{\Gamma}_j$, will converge to

$$\widehat{\Gamma_j} \xrightarrow{N\to\infty} \gamma_j\beta + \alpha_j. \qquad \text{(Equation 7)}$$

(Hint: Recall that in a large sample of individuals, the OLS coefficient from a regression of y on x_j converges to $\frac{\mathrm{Cov}(x_j,y)}{\mathrm{Var}(x_j)}$. Substitute Equation 2 for y into this expression and simplify.)

In the second stage, we consider our dataset to be the set of estimates, $\left\{\hat{\Gamma}_j, \hat{\gamma}_j\right\}_{j=1,2,\ldots,J}$, from the first stage, and we regress $\hat{\Gamma}_j$ on $\hat{\gamma}_j$:

$$\hat{\Gamma}_j = \beta_{0E} + \beta_E\hat{\gamma}_j + \xi_j$$

This regression is called the *MR-Egger regression*. (The "MR" is for "Mendelian randomization," and the "Egger" pays homage to the closely analogous idea of Egger regression that is commonly used in meta-analysis research. For the coefficients in the

regression, the subscript "E" stands for "Egger.") From this regression, we're interested in both the estimated intercept, $\hat{\beta}_{0E}$, and the estimated slope, $\hat{\beta}_E$.

d. Show that under the InSIDE assumption, $\hat{\beta}_E$ is a consistent estimator of β. (Hint: Start with the fact that $\hat{\beta}_E \xrightarrow{J\to\infty} \frac{\mathrm{Cov}(\hat{\Gamma}_j,\hat{\gamma}_j)}{\mathrm{Var}(\hat{\gamma}_j)}$. Substitute for $\hat{\Gamma}_j$ using Equation 7, use the fact that $\hat{\gamma}_j \xrightarrow{N\to\infty} \gamma_j$, and simplify.)

(Note: If the exclusion restriction holds, then MR estimation using all the available instruments has two advantages over MR-Egger regression: (1) consistency of $\hat{\beta}_E$ requires both $N \to \infty$ and $J \to \infty$, whereas $\hat{\beta}_{2\mathrm{SLS}}$ relies on only $N \to \infty$; and (2) MR estimation has greater statistical power. One advantage of MR-Egger regression is that individual-level data is not required.)

Bowden et al. analyze (and conduct simulations to assess) the performance of the $\hat{\beta}_E$ estimator in four pleiotropy scenarios:

1. No pleiotropy: $\alpha_j = 0$ for all j.
2. "Balanced pleiotropy, InSIDE assumption satisfied": $\alpha_j \neq 0$ for all j, but $\mathrm{Cov}(\gamma_j, \alpha_j) = 0$.
3. "Directional pleiotropy, InSIDE assumption satisfied": $\alpha_j > 0$ for all j, but $\mathrm{Cov}(\gamma_j, \alpha_j) = 0$.
4. "Directional pleiotropy, InSIDE assumption not satisfied": $\alpha_j > 0$ for all j, and $\mathrm{Cov}(\gamma_j, \alpha_j) \neq 0$.

e. Explain why the standard MR estimator $\hat{\beta}_{2\mathrm{SLS}}$ yields a consistent estimate of β only in scenario 1, whereas the MR-Egger estimator $\hat{\beta}_E$ yields a consistent estimate of β in scenarios 1-3 but not 4. (Hint: These results follow directly from what you showed in previous parts of the problem.)

f. Bowden et al. show that the intercept estimate from the MR-Egger regression, $\hat{\beta}_{0E}$, can be used to assess whether or not there is pleiotropy that is "directional" on average. Show that, in a large sample, $\hat{\beta}_{0E}$ is a consistent estimator of the average "directionality" of pleiotropy: $\hat{\beta}_{0E} \xrightarrow{N\to\infty, J\to\infty} \mathrm{E}(\alpha_j)$. (Hint: Use the fact that the MR-Egger regression equation converges to Equation 7 for every j as N gets large. Also, note that in OLS, $\hat{\beta}_{0E} = \widehat{\mathrm{E}}[\hat{\Gamma}_j] - \hat{\beta}_E \widehat{\mathrm{E}}[\hat{\gamma}_j]$.)

Just as the exclusion restriction from standard MR cannot be tested using the data at hand and must be assessed with respect to external knowledge about what the genetic variants do and do not do, the InSIDE assumption similarly cannot be tested using the data at hand.

g. Explain why learning whether or not the pleiotropy is directional does not help determine whether the InSIDE assumption is satisfied.

h. When we know that the relationship between some x_j and e is larger, we may believe that this genetic variant is more important biologically in general and may have a larger influence on a number of phenotypes. Explain why in that case we might worry that the InSIDE assumption could be violated, even if the pleiotropy is balanced ($\mathrm{E}(\alpha_j) = 0$).

i. Note that in implementing MR-Egger regression analyses, we may choose to include a small number of genetic variants that are the most strongly associated with e, or a larger number of genetic variants, some of which may not be strongly associated with e. Given the relationship between the number of instruments and power (described in part (d)) and the potential bias of weak instruments (described in the paragraph before part (a)), what are the tradeoffs that one should consider in choosing between these options? If we only have detailed biological knowledge about a limited number of genetic variants, how does that affect the tradeoff?

Computational Problem

4. Constructing Polygenic Scores (PGS)

In this problem, you will create a polygenic score for educational attainment to be used in Add Health, alongside publicly available summary statistics from Lee et al. (2018) (downloaded to here: /home/data/LDSC_sumstats/GWAS_EA3_excl23andMe.txt). In order to avoid a server crash (and to save you time), we will construct the score just for chromosome 22.

As you heard in lecture, polygenic scores can be constructed using several different methods and types of software. We will be using the software PRS-CS (Polygenic Risk Scores – Continuous Shrinkage), a recently developed method that provides substantial computational advantages over existing methods. PRScs utilizes a high-dimensional Bayesian regression framework, by placing a continuous shrinkage prior on SNP effect sizes, which is robust to varying genetic architectures and enables multivariate modeling of local LD patterns (Ge, et al. 2019). Other prediction methods such as LDpred and Pruning and Thresholding (P&T) will not be covered in this tutorial, but we encourage you to explore them yourself and ask us any questions you have during the TAs' office hours!

This problem will challenge you to apply the computational skillset you've been developing in the command line, in R and in software developed for genetic data analysis. We are happy to answer questions if you get stuck, but do spend time Googling or browsing documentation files first. This is the way these skills are built!

a. **LD reference file**:

   The first step in the creation of a polygenic score is to prepare an LD reference panel that matches the ancestry of the discovery and prediction cohort. We will use the pre-computed LD reference files based on the European samples in the 1000 Genome Project, downloaded from here: https://github.com/getian107/PRScs. To avoid

redundant copies of identical data on the server, you can use the files here: /home/data/LD_reference/PRS_ref_panel/ldblk_1kg_eur/.

b. **Format discovery statistics**:

Next, you will need to format the GWAS summary statistics into the required input format for PRS-CS: https://github.com/getian107/PRScs#using-prs-cs. We encourage you to do this step yourself using R or Python (both software packages are available on the server).

c. **Weights generation**:

Now that we have all the necessary input to build the polygenic scores, we can run PRS-CS to obtain the "weights" for each SNP, using: (1) the reference LD panel; (2) the genotypic data of the Add Health samples; (3) the formatted GWAS summary statistics.

We have downloaded the software for you: /home/tools/PRScs/. Try to write your own bash script or command line codes, in reference to: https://github.com/getian107/PRScs#prs-cs.

Note that the sample size of the summary statistics is $N = 766{,}345$, which can be found in our README file: http://ssgac.org/documents/README_EA3.txt.

e. **Scores generation**:

Using these weights and the genotypic data of the Add Health samples, we can build the polygenic scores using PLINK. Try to write your own bash script or command line codes. (Hint: Look up the option "--score" on the PLINK index page. PLINK requires the SNP weights to be in a particular format. We have converted the file into the proper format, which can be directly used as input to the flag "--score": /home/huili/PGS/phi_auto/EDUYR_AH_chr22_plinkformat.txt).

5. Evaluating the predictive power of PGS

Now that you have learnt how to build a PGS using PRS-CS, we can evaluate the predictive power of the score, using the Add Health sample. We have prepared the genome-wide PGS for you, and the script can be found here: /home/TA_sample_scripts/5A.PRS_genome.sh. (**Please do not run this script,** again because generating this is relatively time-consuming and running the job all together will crash the server). The genome-wide PGS of a subset of the Add Health individuals is here: /home/data/PGS_prediction/EDUYR_AH_score.txt

Take this score and merge it with the phenotypic and covariates data of the Add Heatlh sample (/home/data/AH_cleaned/ah_ea_sex_byear_pcs_euros.csv) using R or Python. Using the merged dataset, calculate the predictive power (incremental R^2) of your score, using EduYears as the dependent variable. Make sure you control for the standard covariates (sex, birth year, birth year squared, sex * birth year, sex * birth year squared, and the first 10 PCs). Write up a short description of your results, being sure to discuss the predictive power of your score. (Bonus points: Use bootstrap to build the confidence interval for the incremental R^2. If you use R, try the "boot()" function in the boot library.)

References

Bowden, Jack, George Davey Smith, and Stephen Burgess (2015). "Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression." *International Journal of Epidemiology*, 44(2), 512-525.

Caspi, Avshalom (2003). "Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene." *Science*, 301(5631), 386-9.

Cawley, John,EunaHan, and Edward C. Norton (2011). "The validity of genes related to neurotransmitters as instrumental variables." *Health Economics*, 20(8), 884–888.

Fletcher, Jason, and Steven Lehrer (2011). "Genetic Lotteries within Families." *Journal of Health Economics*, 30(4), 647-659.

Kolesár, Michal, Raj Chetty, John N. Friedman, Edward L. Glaeser, and Guido W. Imbens (2011). "Identification and Inference with Many Invalid Instruments." *NBER Working Paper* No. 17519.

Okbay, Aysu, et al. (2016). "Genome-wide association study identifies 74 loci associated with educational attainment". *Nature* 533:539-42.

Taylor et al. (2015). "Using molecular genetic information to infer causality in observational data: Mendelian randomization." *Current Opinion in Behavioral Sciences*, 2, 39–45.