An examination of parameter recovery for an integrated approach to investigating the behavioral and genetic components of health behaviors

Most health behavior outcomes of interest to social demographers including smoking, alcohol consumption, and obesity are strongly affected by both environmental and genetic factors. Research in the gene-environment (GxE) interaction literature has made it clear that a full understanding of these complex phenomena requires some information about genetic risks but also a clear accounting of the social context in which individuals work, play, and eat (Boardman et al., 2008; Boardman et al., 2012). Emerging evidence for the GxE perspective has led to calls from social scientists and genetic epidemiologists for a cogent and testable framework for GxE research (Bookman et al., 2011) from interdisciplinary teams (Mabry et al., 2008) that has been echoed by the National Academy of Sciences, (Hernandez and Blazer, 2006), the National Science Foundation through its IGERT program, and multiple RFPs from NIH with an explicit emphasis on gene-environment research.

One of the most limiting factors with respect to integrating genetic information into social demographic research remains the large differences in study designs and the diversity of statistical methods that enable genetic information to be included in the analyses. For example, some studies only sample twins and siblings (e.g. the Longitudinal Twin Study), others have candidate gene information for siblings (e.g., Wisconsin Longitudinal Study), unrelated individuals (e.g., Rochester Youth Development Study), or both (e.g., National Longitudinal Study of Adolescent Health [Add Health]) and at times, candidate gene data are available for both parents and their children (Framingham Heart Study or the Families in Transition Project). The specific methods that are used for these specific research designs may inadvertently create barriers between the different groups of researchers that make collaborative efforts overly difficult.

This limitation is particularly pressing as more studies add genome-wide data from respondents of large ongoing cohort studies such as AddHealth and other ongoing longitudinal studies. Social demographers will soon be "drinking from the firehose" (Hunter and Kraft, 2007) and yet there are very few agreed upon methods to facilitate the incorporation of this vast and growing source of information for respondents into the standard methods used in traditional demographic inquiry for different sample designs. This proposal outlines a general approach for simultaneously estimating the degree of genetic influence alongside the effect of environmental mediators and moderators (such as education in the case of tobacco use). We do this by considering state of the art methodological improvements to the models used in demographic research that utilizes genome-wide and sequenced data. The focus of this paper would be an evaluation of this new method via a simulated investigation of the quality of parameter recovery under a variety of conditions chosen specifically to mimic the types of data demographers may use in practice.

Methodology

There is a long history of comparing phenotypic similarities of individuals *within* families as the primary mechanism to understand the genetic component of phenotypic variation. These methods typically rely on structural equation based (Neale and Cardon, 1992) or regression based methodologies (DeFries and Fulker, 1985; Rodgers and McGue, 1994). These methods are generally specific to a certain design such as studies that include identical and fraternal twins (Kohler, Behrman, and Schnittker 2011). While these approaches are collectively considered the standard toolkit of behavioral genetic analysis, they are somewhat limited in their application to social demographic inquiry. More recently, scholars have begun to use variants on the multi-level model as an analytical tool for solving these problems. Guo and Wang (2002) proposed this approach which has been used extensively and their approach was extended by Rabe-Hesketh, Skrondal, and Gjessing (2008) who developed an alternative estimation strategy that allowed for relationship structure. Both approaches, however, require simplifying assumptions about relationship status and can not use relatedness based on measured genotype. Since genetic similarity between unrelated people can be used to estimate heritabilites (e.g., Yang et al., 2011) it would also be useful to be able to incorporate this information into other demographic inquiries.

For a phenotype of interest y_i, consider the ACE model:

$$y_i = \mu + a_i + c_i + \varepsilon_i$$

for individual i. This model may be used to estimate the proportion of the variance in y_i that is accounted for by additive genetic component, a_i, and shared-environment, c_i, components (ε_i is white noise random error). The effects a_i and c_i are assumed to be random and identification for this model clearly requires distributional

assumptions so that they can be distinguished from the error term. One possibility would be on the basis of family structure (e.g., Rabe-Hesketh, Skrondal, and Gjessing, 2008) but this excludes the use of relatedness based on measured genotype. To allow for such relatedness, we focus on models of the form:

$$y_i = \beta X_i + a_i + e_i,$$
$$a \sim N[0, \sigma_A^2 A],$$

where X_i are predictive covariates, a is again the vector of the individual a_i random effects, and e_i are individual level errors. The A matrix is at the heart of our approach and requires additional comment. For individuals i and j, $A_{i,j}$ is still the genetic relatedness of two individuals. It could be know based on pedigree (e.g., twins, parent-child, etc.) or it could be determined via measured genotype. We discuss estimation of the A matrix and the full model in the subsequent discussion of the two simulation studies.

This approach allows for a common analytic framework for a variety of study types (e.g., unrelated individuals versus family-based), but it also accommodates the inclusion of predictors that may be of substantive interest to demographers. This ability to control for a variety of predictors within a common framework is important since genetic, individual, family, and environmental predictors may all be of interest. Consider obesity. A study may be simultaneously interested in the influence on obesity of the amount of exercise an individual engages in each week, the number of fast food outlets in an individual's neighborhood, and a single nucleotide polymorphism (SNP) or genetic risk score. In our suggested approach, all of these predictors could be treated within the X_i matrix of predictors and their influence could be considered *controlling* for the genome-wide influence (through a_i) on a phenotype.

Simulation Study 1

The first simulation study in this paper is meant to examine the ability of the method to recover truth based on several manipulated characteristics of a particular data set. Models are estimated using state of the art Bayesian techniques. Estimation will be done using the recently released Stan software (The Stan Development Team, 2012a, 2012b). This software performs Bayesian estimation via Hamiltonian Monte Carlo (Hoffman and Gelman, in press). In initial work, we have focused on a simplification of the equation introduced above:

$$y_i = \beta_0 + \beta_1 X_i + a_i + e_i,$$
$$a \sim N[0, \sigma_A^2 A],$$
$$e_i \sim N[0, \sigma_e^2].$$

The simulation focused on this simple setting where the phenotype for an individual is normally distributed conditional on the effect of a covariate and an individual-level random effect, a_i, related to genetic influence. These individual-level effects are distributed as described earlier, with σ_A^2 varying across the different iterations of the simulation and A, which defines the genetic relatedness of two individuals, containing draws from a uniform distribution on 0 to 0.05 (e.g., individuals are simulated to be unrelated). This model only contains an additive genetic component, it does not contain a shared environmental or a dominant genetic component. We feel that this is a reasonable simplification since additive effects are frequently the prime objects of interest and shared environment will not be an issue with unrelated individuals. The simulation has focused on manipulating three parameters: (1) the size of the sample of unrelated individuals, (2) the ratio of the genetic contribution to phenotypic variance to overall phenotypic variance, and (3) the predictive strength of a covariate (the value of β_1).

In our initial work, recovery of the β_1 parameter is good. This is not surprising since the predictor is not associated with genotype in this simple simulation. However, the quality of the recovery of the ICC ($\sigma_A^2/(\sigma_A^2 + \sigma_e^2)$) heritability is much more dependent on the conditions of the simulation (see how the blue line tracks the black line, truth, in Figure 1). Critically, our simulation suggests that the proposed approach is a feasible method of calculating the genetic contribution to phenotypic variance if the sample size is large enough (e.g., accurate results were obtained for simulations with 750 individuals but not for those with 250 individuals). Moreover, we have found that the accuracy in the estimation of the ratio of genetically explained variance to overall phenotypic variance is a function of the true value of this ratio. When the genetic contribution is low, our method tends to over-estimate this ratio in simulation. This is an important bias (that also potentially applies to other methodological approaches) given that this variance will naturally vary as a choice of phenotype.

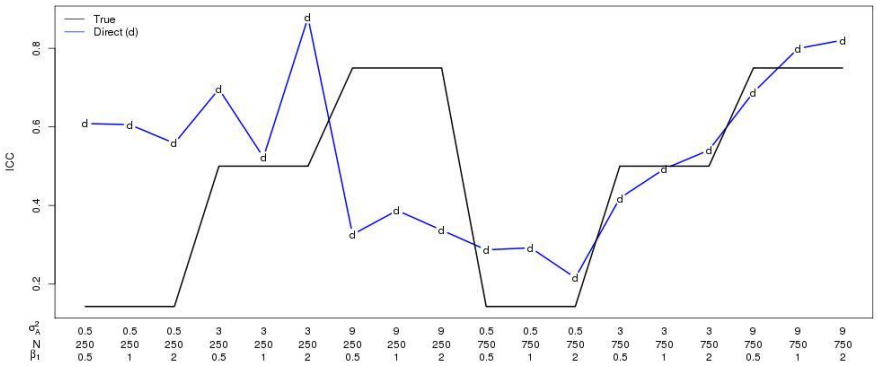

Figure 1: Recovery of ICC for various simulation conditions.

Simulation Study 2

The second study examines the sensitivity of the method to population stratification, particularly as it affects different methods for measuring genetic relatedness based on measured genotype. Our initial work on estimation of genetic relatedness suggests that these estimates are extremely sensitive to population stratification. Since they underlie the entire approach, understanding this sensitivity is of paramount importance. There are a variety of methods available for measuring genetic relatedness. One approach is to use the genetic relatedness values computed by GCTA (Yang et al. 2011). However, we have found these to be quite sensitive to the racial diversity of the sample in which they are computed and intend to compare them from alternative possible estimates such as those from KING (Manichaikul et al. 2010). Measured genetic data from a nationally representative sample that, crucially, contains individuals from diverse racial backgrounds, will be used as the basis for examining the sensitivity of inferences to the method while still using simulated phenotypes. The use of simulated phenotypes is important since it will allow us to determine under what conditions certain indices of relatedness, perhaps restricted to only certain subsamples of our overall sample, lead to accurate inference regarding the genetic influence on a phenotype.

Discussion

The need for demographers to consider genes alongside environmental and physical characteristics in understanding physical and mental health behaviors is clear but doing so can be challenging. The approach explored in this paper allows demographers to include genetic information into the statistical models that are standard in the literature using the recently developed Bayesian techniques. The simulation studies described here are necessary to describe conditions under which accurate inferences can be reasonably obtained.

References

Boardman JD, Saint Onge JM, Haberstick BC, Timberlake DS, and Hewitt JK (2008). Do schools moderate the genetic determinants of smoking? *Behavior Genetics*, 38(3):234-46.

Boardman JD, Roettger ME, Domingue BW, McQueen MB, Haberstick BC, and Harris KM (2012). Gene-environment interactions related to body mass: School policies and social context as environmental moderators. *Journal of Theoretical Politics*, 24(3):370-388.

Bookman EB, McAllister K, Gillanders E, Wanke K, Balshaw D, et al. (2011), Gene-environment interplay in common complex diseases: forging an integrative model—recommendations from an NIH workshop. *Genetic.Epidemiology*., 35: 217–225.

DeFries J and Fulker D (1985).Multiple regression analysis of twin data.*Behavior Genetics*, 315: 467–473.

Guo G and Wang J (2002).The mixed or multilevel model for behavior genetic analysis. *Behavior Genetics*, 32(1): 37–49.

Hernandez LM and Blazer DG (Eds.). (2006). *Moving beyond the Nature/Nurture Debate*. Washington DC: National Academies Press.

Hoffman MD, and Gelman A (In press). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo.*Journal of Machine Learning Research.*

Hunter D.J. and Kraft P. (2007).Drinking from the fire hose---Statistical issues in genomewide association studies. *New England Journal of Medicine*, 357(5): 436-439.

Kohler H, Behrman JR, and Schnittker J (2011). Social science methods for twins data: Integrating causality, endowments, and heritability. *Biodemography and Social Biology*, 57(1): 88-141.

Mabry PL, Olster DH, Morgan GD, and Abrams DB (2008). Interdisciplinarity and systems science to improve population health: a view from the NIH Office of Behavioral and Social Sciences Research. *American Journal of Preventative Medicine*, 35: S211–S224

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics* 26(22):2867-2873.

Neale MC and Cardon LR (1992).*Methodology for genetics studies of twins and families*. Boston, MA: Kluwer Academic Publishers.

Rabe-Hesketh S, Skrondal A, Gjessing HK (2008). Biometrical modeling of twin and family data using standard mixed model software. *Biometrics*, 64: 280-288.

Rodgers JL and McGue M (1994).A simple algebraic demonstration of the validity of DeFries-Fulker analysis in unselected samples with multiple kinship levels.*Behavior Genetics*, 24(3): 259-262.

Stan Development Team (2012a). Stan: A C++ library for probability and sampling [Computer software manual]. Retrieved from http://mc-stan.org/ (Version 1.0).

Stan Development Team(2012b). Stan modeling language user's guide and reference manual [Computer software manual].Retrieved from http://mc-stan.org/ (Version1.0).

Yang, J, Lee S, Goddard M, and Visscher PM. (2011). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, 88(1): 76–82