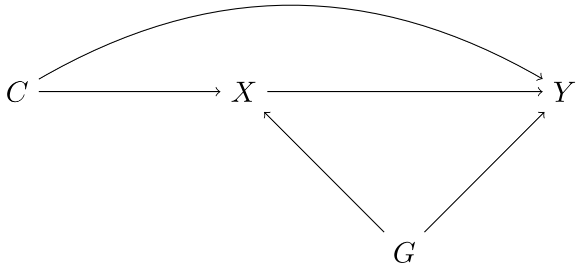
Technical Appendix for: Genetic Confounding of the Relationship Between Father Absence and Age at Menarche

J.C. Barnes[1], Nicole Barbaro[2], Brian B. Boutwell[3], and Todd K. Shackelford[4]

[1]University of Cincinnati, School of Criminal Justice, Cincinnati, OH 45221, jc.barnes@uc.edu
[2]108 Pryale Hall, Oakland University, Department of Psychology, Rochester, MI 48309, nmbarbar@oakland.edu
[3]Saint Louis University, School of Social Work, St. Louis, MO 63103, boutwellb@slu.edu
[4]112 Pryale Hall, Oakland University, Department of Psychology, Rochester, MI 48309, shackelf@oakland.edu

Imagine a researcher has a dataset with information on an outcome Y, a key independent variable X, and a host of covariates C, but s/he does not have a way to control for genetic factors G, as in the diagram below:



This sort of scenario presents itself in many—if not most—behavioral science studies. When this happens, scholars are forced to either: 1) abandon viable ideas for fear of producing biased parameter estimates; 2) expend more resources to collect additional data (e.g., identify and interview MZ twins) so that a genetically sensitive design can be used; or 3) publish potentially biased parameter estimates. The purpose of this discussion is to present a novel alternative.

Specifically, the new tool developed here blends a well-established equation for estimating the degree to which a phenotypic correlation r_p is driven by genetic correlation r_g with modern statistical simulation methods. By combining these two elements, one is able to simulate the degree to which an observed correlation may be sensitive to uncontrolled genetic influences. The degree to which r_p is sensitive to genetic influences will be referred to as h^2_{cov}. In the context of the above diagram, the tool developed here will allow one to estimate the degree to which the $X \rightarrow Y$ association (i.e., r_p) is sensitive to the inclusion of G.

In order to understand the estimation routine, it is first necessary to introduce the various pieces of information that must be supplied by the user. Then, the equation that sits at the center of the estimation routine—the equation for h^2_{cov}—will be introduced.

Necessary Information: r_p, h^2_X, h^2_Y, & r_g

The estimation routine is carried out in several steps. The first step is to estimate the phenotypic correlation between X and Y, referred to as r_p. This step can be carried out using any statistical analysis package and, it is worth pointing out, partial correlations can be used when available. In other words, there is no requirement that the unconditional correlation between X and Y be preferred over a partial correlation that has already accounted for other measured covariates C.

The second step is to arrive at an estimate for the heritability of X, h^2_X. Recognizing that this value is not directly estimable—because if it were, one of several other methods would be preferable to the present approach—the researcher is encouraged to consult the available behavioral genetic literature that has bearing on the heritability of the phenotype of focus (see Polderman et al. [2015] and/or the accompanying webpage: `http://match.ctglab.nl/#/home`).

The same is true for the heritability of Y, h^2_Y. While it is not necessary that the user be an expert in behavioral genetics, the utility of this novel tool is contingent upon the user inputting heritability estimates that are both meaningful and realistic.

We now have three pieces of information necessary for estimating h^2_{cov}, but in order to garner an estimate of h^2_{cov}, we will also need an estimate of the genetic correlation r_g between X and Y. In essence, r_g provides an estimate of the degree to which the genetic factors that affect X also impact Y. Thus, r_g is simply an estimate of the correlation between the genetic factors that influence the phenotypes—X and Y—of interest.

Building a Distribution of h^2_{cov} Estimates

Researchers interested in unpacking the covariance between X and Y often rely on one of several bivariate biometrical models (Loehlin, 1996). What is unique about the bivariate biometrical model is that the covariance between X and Y can be decomposed into a heritability component that we will refer to as h^2_{cov}. This value represents the proportion of the phenotypic correlation r_p that is due to a shared genetic overlap between X and Y.

One can calculate h^2_{cov} as:

$$h^2_{cov} = \frac{\sqrt{h^2_X} * r_g * \sqrt{h^2_Y}}{r_p}$$

where: $\sqrt{h^2_X}$ is the square root of h^2_X; $\sqrt{h^2_Y}$ is the square root of h^2_Y; r_g is the genetic correlation between X and Y; and r_p is the phenotypic correlation between X and Y. Conceptually, the equation provides an estimate of the proportion of r_p that is due to shared genetic influences between X and Y. For this reason, the equation for h^2_{cov} is the centerpiece of the new estimation tool developed here.

One might be tempted to simply solve for h^2_{cov} using estimates that come to mind for h^2_X, h^2_Y, r_p, and r_g. Indeed, one can easily calculate the proportion of the phenotypic correlation that is due to genetic factors (i.e., h^2_{cov}) knowing nothing more than these four values. One key point, however, would be overlooked. Specifically, the inaccuracy of the estimates for h^2_X, h^2_Y, r_p, and r_g are ignored if one solves the equation with just one set of values. Of course, the very foundation of statistical analysis rests on the assumption of random error, meaning that any estimate we receive from this equation is likely to be too high or too low.

Fortunately, drawing on certain principles and techniques that have become commonplace in Bayesian analysis can help solve this problem. The mechanics of modern Bayesian statistical analysis is one of "brute force" sampling and simulation (Gelman et al., 2014; Gill, 2013; Jackman, 2000). Recognizing that integrating over the posterior distributions of interest—even for very simple problems—is often too complicated to calculate with closed form integral calculus, contemporary Bayesian statisticians have adopted Markov chain Monte Carlo (MCMC) routines of simulation and sampling as their primary workhorse for generating estimates of the posterior distribution. The logic is straightforward: if you cannot directly calculate a solution to a problem, use MCMC to simulate and estimate the problem a large number of times and create a distribution of posterior estimates.

Thus, the estimation tool developed herein will allow for the uncertainty of the estimates provided by the user to be taken into account when calculating the posterior distribution of h^2_{cov} estimates. This is done by solving the above equation k times, each time including a slightly different configuration of values for the heritability estimates and for r_p. The value k will be supplied by the user—it is recommended that k be set to a large value (e.g., $k = 10,000$) in order to ensure adequate coverage of the parameter space. Rather than force one to calculate the equation for h^2_{cov} for all the possible combinations of heritability estimates and r_p estimates—a procedure that would necessarily ignore the probability distributions of the various statistics—the approach developed herein allows one to randomly sample values from a distribution of heritability estimates and from a distribution of r_p estimates. But first, the user must construct said probability distributions. The beta distribution makes this task tractable.

The Beta Distribution

The beta distribution is appropriate for building a probability distribution of prior estimates for r_p and the heritability estimates because it is bounded at 0 and 1, but can take on any real value between those two integers. The beta distribution is a well-defined univariate distribution that has a direct relationship with the normal distribution and has a probability density function of (Gelman et al., 2014: 578; Leemis, 1986: 146):

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} * x^{a-1}(1-x)^{b-1}$$

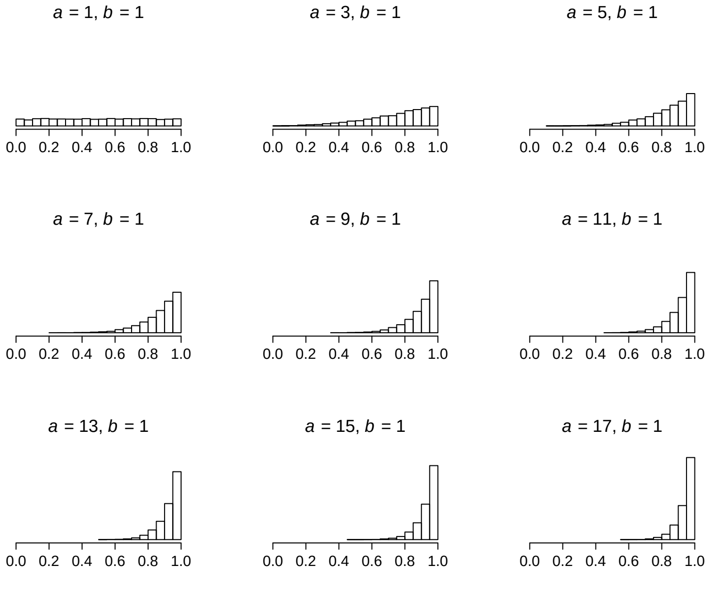
where both a and b are greater than 0 and can be thought of as shape parameters that affect the form and location of the distribution along the support region. The three values of interest—the expected value [$\mathbb{E}(x)$], the mode [$mode(x)$], and the variance [$var(x)$]—are calculated as:

$$\begin{aligned}
\mathbb{E}(x) &= \frac{a}{a+b} \\
mode(x) &= \frac{a-1}{(a-1)+(b-1)} \\
var(x) &= \frac{ab}{(a+b)^2(a+b+1)}
\end{aligned}$$

Thus, the shape parameters can be used to adjust the balance point (i.e., the mean or expected value) of the distribution, the modal value, and the dispersion (i.e., variance) around the expected value. Generally, a can be thought of as the right shape parameter meaning that larger values for a, relative to b, will place more density in the right portion of the support region. The opposite is true for the shape parameter b, which is the left shape parameter. Taken together, this means that the user can adjust the beta distribution to load more density for the heritability estimate distribution and/or the r_p estimate distribution in the right side of the support region if a is increased relative to b and the opposite effect is achieved if b is increased relative to a.

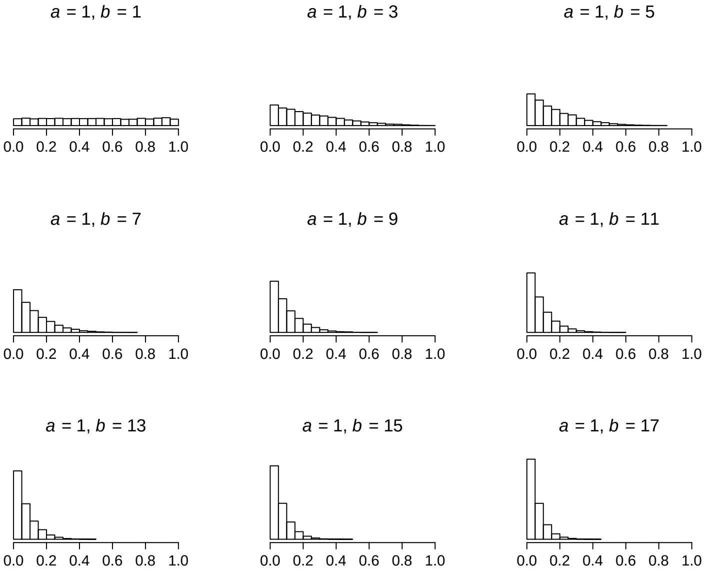
These points are demonstrated graphically in Figure 1 and Figure 2. Note also that the user can set the beta distribution to reflect his/her level of confidence in the estimates by setting the shape parameters to higher or lower values. Higher values for the shape parameters will load more density in increasingly smaller regions of the distribution, meaning the variance approaches its lower limit as a and b approach ∞.

Figure 1: The Effect of Changing a



Several other useful features of the beta distribution are worth pointing out. Imagine a scenario where the researcher is unsure what the heritability estimate(s) and/or the r_p estimate should be. In this case, the researcher would benefit from relying on something similar to the Bayesian diffuse/uninformative prior. This can be achieved by setting both shape parameters to equal 1 (i.e.,

Figure 2: The Effect of Changing b



$a = 1$ and $b = 1$). The panel in the top-left of Figure 1 and Figure 2 reveals the beta distribution is uniform under this condition.

Imagine another case where the researcher believes the heritability estimate(s) and/or the r_p estimate is approximately 0.50. This is an especially important value for the former (i.e., heritability estimates) because much of the behavioral genetic literature converges on heritability estimates that are approximately 0.50 (Polderman et al., 2015). These types of estimates can be modeled with the beta distribution by simultaneously increasing both shape parameters in equal magnitude (i.e., $a = b$). This relationship is revealed graphically in Figure 3.

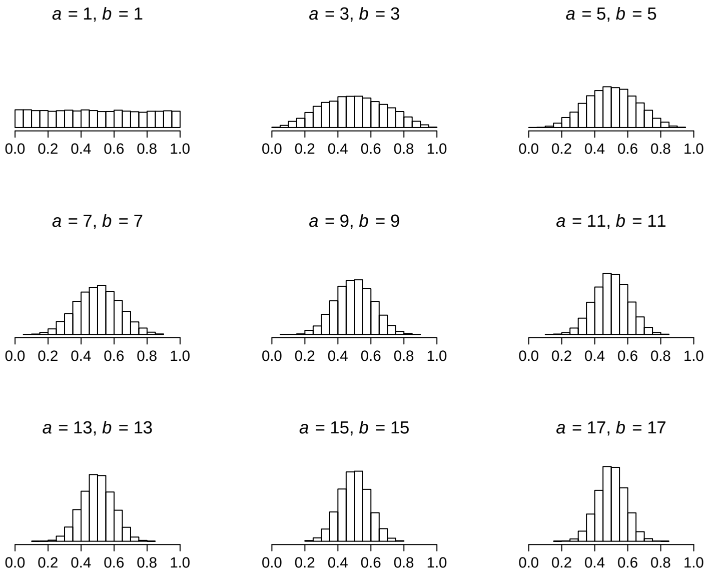
Recognizing that the true population parameter is an unknown that is only estimated in any given study, the beta distribution will capture the uncertainty in the estimates by building a range of values that will be fed through the equation for h^2_{cov} k times. In the end, a posterior distribution of estimates for h^2_{cov}—the degree to which r_p may be biased due to uncontrolled genetic influences—is retrieved.

Recommendations for Estimating a Distribution of h^2_{cov}

The above sections introduced a novel estimation tool that can be used by any researcher who is concerned that the relationship between two variables X and Y might be inflated due to uncontrolled genetic factors. All of the codes—in `R`—necessary to carry out the estimation routine have been posted to the following GitHub page: `https://github.com/jcbarnescrim`. Thus, access to genetically sensitive data is no longer necessary to estimate to extent to which a phenotypic correlation is sensitive to omitted genetic factors.

A brief summary of the estimation procedure is outlined here:

Figure 3: $a = b$



1. The researcher observes (whether from a novel data analysis or from the available literature) a relationship between two variables X and Y. The relationship should be measured in the form of a correlation coefficient (r_p), but note that a partial regression coefficient (i.e., an estimate that already accounts for other known confounders) can also be used as long as the value has been standardized.

   - Form a distribution of r_p values using the beta distribution. The expected value (or the mode if the distribution is skewed) of the beta distribution should be set to equal the observed correlation coefficient.

   - The shape parameters, a and b, are used to construct the desired beta distribution.

2. The researcher specifies the heritability estimate for X (h^2_X). This information should be based on the available behavioral genetic literature. Scholars are encouraged to see Polderman et al. (2015) for heritability estimates.

   - Form a distribution of h^2_X values using the beta distribution. The expected value (or the mode if the distribution is skewed) of the beta distribution should be set to equal h^2_X.

   - The shape parameters, a and b, are used to construct the desired beta distribution.

3. The researcher specifies the heritability estimate for Y (h^2_Y). This information should be based on the available behavioral genetic literature. Scholars are encouraged to see Polderman et al. (2015) for heritability estimates.

   - Form a distribution of h^2_Y values using the beta distribution. The expected value (or the mode if the distribution is skewed) of the beta distribution should be set to equal h^2_Y.

- The shape parameters, a and b, are used to construct the desired beta distribution.

4. The researcher specifies the genetic correlation between X and Y (r_g). This information may not always be available. In cases where r_g is unknown, the researcher is encouraged to try a range of potential values.

5. Enter the information from steps 1 through 4 into the program code located at (`https://github.com/jcbarnescrim`) and generate a posterior distribution of h^2_{cov} estimates. This distribution of estimates is calculated by feeding randomly drawn values from the above distributions through the equation for h^2_{cov} k times.

   - k is set by the user and should be a large value (e.g., 10,000) to ensure adequate coverage of the parameter space for the posterior distribution of h^2_{cov} estimates.

Conclusions

While there are many ways researchers could use this tool, the most obvious is to estimate the sensitivity of a parameter estimate of the association between X and Y (i.e., r_p). Rather than simply speculating about the degree to which a relationship might be confounded, a probability distribution of values can now be formed by carrying out the five simple steps outlined above.

But, it is important to caution researchers from blindly estimating a distribution of h^2_{cov} values. In fact, the distribution of h^2_{cov} values is only meaningful if genetic factors G serve as confounding influences. Confounding influences are those that are antecedent to X and Y and have a causal effect on variance in the two measures:

$$\begin{tikzcd}
X \arrow[rr, dashed] & & Y \\
& G \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

It is important to note, however, that confounding variables are statistically indistinguishable from mediator variables. This may complicate the interpretation of the results gleaned from the proposed tool if the true relationship is:

$$\begin{tikzcd}
G \arrow[r] \arrow[rr, dashed, bend left] & X \arrow[r] & Y
\end{tikzcd}$$

where X mediates the influence of G on Y. This chain of causation is quite different from that which is expected from a confounded relationship. Estimates of h^2_{cov} mean something different in this case, so researchers must rely on theory, empirical evidence, and logical deduction to determine whether the genetic correlation (i.e., r_g) between X and Y is due to genetic confounding or something else.

References

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis (3rd edition).* Boca Raton, FL: CRC Press.

Gill, J. (2013). *Bayesian methods: A social and behavioral sciences approach (3rd edition).* Boca Raton, FL: CRC Press.

Jackman, S. (2000). Estimation and inference via Bayesian simulation: An introduction to Markov chain Monte Carlo. *American Journal of Political Science, 44*, 369-98.

Leemis, L. W. (1986). Relationships among common univariate distributions. *American Statistical Association, 40*, 143-46.

Loehlin, J. C. (1996). The Cholesky approach: A cautionary note. *Behavior Genetics, 26*, 65-69.

Polderman, T. J. C., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics, 47*, 702-09.