ACE in the Hole: Correlation and Classical Twin Studies

Mark M. Fredrickson

December 20, 2009

1 Basics of the ACE Model

Like many models in social science, the ACE model strives to explain observed population variance into smaller components. Like ANOVA techniques, the ACE technique compares the variance within groups to the variance across groups. Unlike the more general ANOVA method, the ACE model uses the same two groups, monozygotic (MZ) and dizygotic (DZ) twins, in all calculations and random variables measured on the same scale. This allows a focus on differences in correlation of MZ and DZ twins. While more complex structural equation models have become the vogue in twin studies, as Alford et al. (2005) observe, the basic logic is intuitive and easy to understand.

Classical twin studies attempt to partition observed variance into three components: the additive genetic component (A), the shared environment of the twin pairs (C), and a residual category of environment unshared by the pairs (E). Since these values are assumed to sum to unity, only two quantities are required to estimate the model for a given variable of interest: the observed correlation of monozygotic twins (r_{MZ}) and the correlation of dizyogtic twins

(r_{DZ}). Explicitly, the model is estimated using the following three equations:

$$A = 2(r_{\mathrm{MZ}} - r_{\mathrm{DZ}}) \tag{1}$$

$$C = 2r_{\mathrm{DZ}} - r_{\mathrm{MZ}} \tag{2}$$

$$E = 1 - r_{\mathrm{MZ}} \tag{3}$$

For traits that are driven exclusively by genes, MZ twins will correlate perfectly, but DZ twins, which are assumed to share half their genetic material on average, will only correlate at the $r_{\mathrm{DZ}} = 0.5$ level. For traits that are environmental in nature, we would expect to see no difference in MZ and DZ correlation. These expectations are encapsulated in the three ACE equations and make for a parsimonious model. Many modern twin studies employ structural equation modeling or Bayesian methods to estimate these quantities (Medland and Hatemi, 2009), but the underlying logic is based on comparing the covariances of MZ and DZ, and the arguments in this paper, while employing these specific estimators, apply equally to other techniques.

Within political science, the classical twin study design has been applied to questions of political ideology (Alford et al., 2005), partisanship (Settle et al., 2009), political behavior (Fowler et al., 2008), and political attitudes and tolerance (Eaves and Hatemi, 2008). At a minimum, these studies make a strong case for the presence of a genetic component to political behavior. Even the most grudging critic must allow that the "sharp null" of no genetic influence can be safely rejected. Of course, these studies also attempt to quantify the effect size, the percent of observed variation "explained" by genes. While this is a noble goal, are the blunt tools of the classical twin study powerful enough to make this goal? If so, we would expect that different patterns in correlation would lead to different ACE estimates. We might also expect the ACE estimates at the macro level to reflect differences at lower levels in the data. In this paper, I demonstrate a small subset of possible data generating processes that could all lead to the same ACE estimates. It is my position that at best ACE estimates consistently exaggerate actual effects and at worst are substantially misleading.

2 Many roads to the same destination

Given the reliance on correlation, there has been little discussion of what could be generating the correlations we observe. As a summary statistic, provides a scale free measure of the linear relationship of two variables, but such a summary can obscur the wide variety of possible distributions that give rise to the same correlation value. Anscombe (1973)'s famous data sets provide a warning against relying on summary statistics without understanding the underlying distributions. Reproduced in Figure 2, these four data sets all have the same means, variance, and correlation, but produce vastly different scatterplots. While some of these datasets seem implausible for the subject matter of twin studies (e.g. a quadratic relationship between twins), Anscombe's work does beg the quesiton, "What meaningful differences could the correlations in classical studies be hiding?"

Consider a bivariate mixture distribution $f(x, y)$ composed of k components $f_k(x, y)$, each a bivariate joint distribution with correlation $\rho_i, 1 \leq i \leq k$. Let w be the weights of the components, such that $w_i > 0$ and $\Sigma w_i = 1$. The correlation of the overall joint distribution is then the weighted sum of the component correlations[1]:

$$\rho = \Sigma_1^k w_i \rho_i \tag{4}$$

A single value of ρ could be derived from several different distributions. For example, a correlation of 0.5 could come from pair of component distributions with weights $w_1 = 0.5, w_2 = 0.5$ and $\rho_1 = 0.4, \rho_2 = 0.6$ It could also arrise if $w_1 = 0.8, w_2 = 0.2$, and $\rho_1 = 0.7, \rho_2 = -0.3$. Figure 2 shows simulated data from these two mixture distributions. For 100

[1] I have not proved this yet, but I have convinced myself, at least for normal X and Y

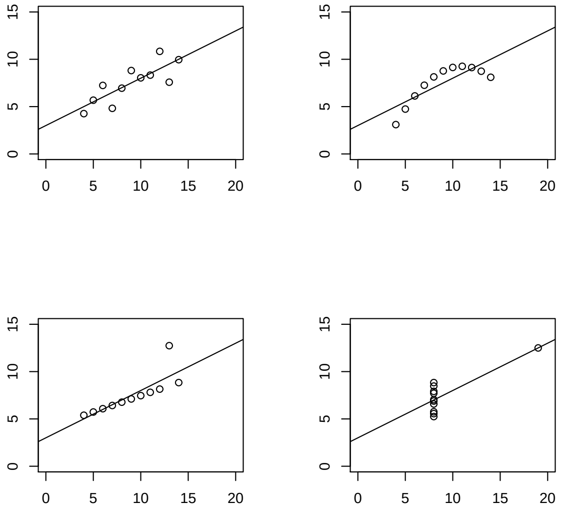

Figure 1: Bivariate Distributions with Identical Correlation (Anscombe, 1973)

samples each, a pair (x, y) is from the first bivariate normal with ρ_1 with probability w_1 and from the second distribution with ρ_2 otherwise. All of the univariate variables are standard normal. Analyzing these data results in $r = 0.569$ for the first distribution and $r = 0.459$ for the second distribution, well within expected deviations of the true population correlation. But unlike Anscombe's data, these scatter plots do not immediately suggest that there are different distributions at work.

Of course, we should only care if mixtures exist in twin data if these mixtures obscure important information. Consider a world in which the political ideology is linked to one environmental factor and one Mendelian gene with a recessive genotype and a dominant genotype. We can then abbreviate individuals with the dominant genotype as G and those with the recessive as g. Similarly, let E represent the presence of the environmental stimulus and e its absence. For any individual there are 4 possible pairings: GE, gE, Ge, gg. For pairs of individuals (where order is not important), there are 10 possible joint distributions: $GEGE, gEgE, GeGe, gggg, GEgE, GEGe, GEgg, gEGe, gEgg, Gegg$. It would be quite be-

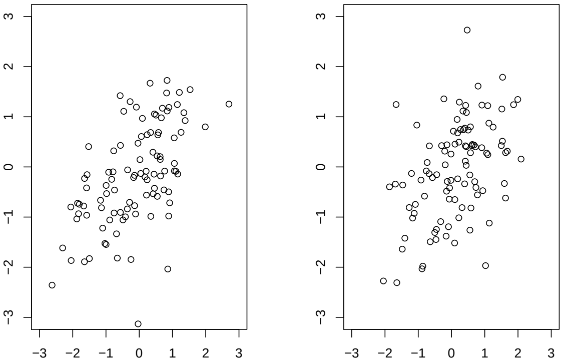

Figure 2: Two bivariate mixtures distributions with overall $\rho = 0.5$

lievable that the correlation of each of these distributions could vary, with some showing higher correlation and some lower. As in the above examples, different combinations of a mixture of these distributions could lead to the same overall population correlation, which might not be immediately apparent from a highlevel understanding of the data.

As an example set of correlations for the overall distributions f_{MZ} and f_{DZ}, consider Alford et al. (2005)'s estimate of conservative ideology that found "for the overall index of political conservatism, genetics accounts for approximately half of the variance in ideology, while shared environment including parental influence accounts for only 11%." In terms of the ACE model this represents $A = 0.53, C = 0.11, E = 0.36$. Solving for r_{MZ} and r_{DZ} from the ACE equations yields:

$$r_{\text{MZ}} = 1 - E = 0.64 \tag{5}$$

$$r_{\text{DZ}} = r_{\text{MZ}} - \frac{A}{2} = .64 - .265 = 0.375 \tag{6}$$

In the next five examples, I show how these correlations can be the product of mixture distributions, specifically subsets of the 10 distributions in the previous paragraph.

2.1 Example 1: Genetic Effects

The simplest example is actually simpler than the world of one gene and one environmental stimulus. We start with a world in only the gene matters, with any necessary environmental conditions common to all individuals. We can then consider only the following three joint distributions: GG, gg, Gg. For the overall distribution for MZ twins, f_{MZ}, note that only the GG and gg distributions are possible, given that identical twins share 100% of their genetic material. Given a gene that is prevalent at a rate of 50%, the weights for the component distributions are immediate: $w_{GG} = 0.5, w_{gg} = 0.5$. Since DZ twins each have a 50% chance of getting the dominant form and can differ, the weights for f_{DZ} vary more: $w_{GG} = 0.25, w_{gg} = 0.25, w_{Gg} = 0.5$. Given these weights, we have a range of possible of values for the correlations. Here is an example:

$$r_{\mathrm{MZ}} = 0.64 = 0.5(r_{GG} + r_{gg}) = 0.5(0.88 + 0.4) \tag{7}$$

$$r_{\mathrm{DZ}} = 0.375 = 0.25(r_{GG} + r_{gg}) + 0.5r_{Gg} = 0.25(0.88 + 0.4) + 0.5 \cdot 0.11 \tag{8}$$

This example is fairly consistent with the ACE assumptions. Twin pairs with identical genetics correlate more than pairs with dissimilar genes. This is true for both the macro-level correlation comparing r_{MZ} and r_{DZ} and also the mixture level comparisons where $r_{GG} > r_{gg} > r_{Gg}$. What is most interesting about this example is the lack of variation in environmental stimulus (we have assumed all twin pairs experience the same environment), yet the ACE estimate from these data would ascribe 11% of the population variance to shared environmental factors. Interestingly, this value is the same as the correlation of heterogeneous DZ twins ($r_{Gg} = .11$). But this relationship only holds when the distribution of dominant and recessive genes sits at 50%/50%. For example, let the dominant form be found in 75% of the population. Weights for f_{MZ} are then $w_{GG} = 0.75, w_{gg} = 0.25$ and for f_{DZ}, $w_{GG} = 0.5625, w_{gg} = 0.0625, w_{Gg} = 0.375$. Given these weights we could see an entirely different pattern of correlations:

$$r_{\mathrm{MZ}} = 0.64 = 0.75 \cdot 0.5333 + 0.25 \cdot 0.9 \tag{9}$$

$$r_{\mathrm{DZ}} = 0.375 = 0.5625 \cdot 0.5333 + 0.0625 \cdot 0.9 + 0.375 \cdot 0.05 \tag{10}$$

This is example shows a critical flaw in the assumptions of the ACE model. The ACE model assumes that DZ twins share on average half of their genetic material. In the aggregate, this assumption generally holds. But when specific genes are not evenly distributed over the population, we can see divergences from this assumption. In the previous example, 62.5% of DZ twin pairs had identical genotypes (i.e. $w_{GG} + w_{gg} = 0.5625 + 0.0625 = .625$), not the assumed 50%. Interestingly, it seems that as prevalence of a gene increases, we are more likely to ascribe variation to environmental factors, a rather counterintuitive result.

2.2 Example 2: Common Environment

The previous example assumed that all environmental factors were common to all twin pairs. In this example, consider environmental factors that are common to both individuals in a twin pair. For example, for twins raised together, we might consider parents' political beliefs to be an environmental stimulus. Using the earlier notation, if parents are "conservative," the twin pair is labeled as E, otherwise labeled as e. This leads to the following possible mixture distributions: $GGE, GGe, ggE, gge, GgE, Gge$.

Assume that environment is independent of genetics (usually assumed in ACE studies, a point to which we shall return later), and both the gene and environmental factor have 50% prevalence in the population. For f_{MZ}, the mixture weights would be $w_{GGE} = w_{ggE} = w_{GGe} = w_{gge} = 0.25$. For f_{DZ}, the mixture weights would be $w_{GGE} = w_{ggE} = w_{GGe} = w_{gge} = 0.125, w_{GgE} = w_{Gge} = 0.25$. Constraining the correlations such that the presence of E always leads to greater similarity of twin pairs:

$$r_{\mathrm{MZ}} = 0.64 = 0.25(GGE + GGe + ggE + gge) \tag{11}$$

$$= 0.25(0.94 + 0.74 + 0.54 + 0.34) \tag{12}$$

$$r_{\mathrm{DZ}} = 0.375 = 0.125(GGE + GGe + ggE + gge) + .25(GgE + Gge) \tag{13}$$

$$= 0.125(0.94 + 0.74 + 0.54 + 0.34) + .25(0.21 + .01) \tag{14}$$

In this example, the presence of the environmental stimulus added a constant factor (0.2) compared to twin pairs of the same genotypes without the environmental stimulus. It is not clear how to square the contribution of the environmental factor in this example with the ACE estimate of $C = .11$.

2.3 Example 3: G × E

The impacts of environmental stimuli need not be linear in effect. If conservative parents have a multiplicative impact on their children, given the presence of a specific genotype, gene-environment interaction (G × E) will manifest differently in homogeneous and heterogeneous twin pairs. Purcell (2002) shows that when genes and shared environment interact, estimates of the A quantity will be inflated. If genes are interacting with unshared environment, the resulting E estimates will be upwardly biased. The logic of these effects is simple. If the environmental stimulus (E) has no effect outside of the presence of a specific (G), we would observe far more correlation in homogeneous twins. Put another way, twins with identical genes will be made *more* similar by the interaction since both twins will experience the interaction. Only one twin in heterogeneous pairs will experience the interaction, driving down the similarity of the behavior of the two. An example of G × E in practice might look like the following:

$$
\begin{aligned}
r_{\mathrm{MZ}} &= 0.64 = 0.25(GGE + GGe + ggE + gge) && (15)\\
&= 0.25(0.97 + .53 + .53 + .53) && (16)\\
r_{\mathrm{DZ}} &= 0.375 = 0.125(GGE + GGe + ggE + gge) + .25(GgE + Gge) && (17)\\
&= 0.125(0.97 + .53 + .53 + .53) + .25(0.01 + .21) && (18)
\end{aligned}
$$

By itself, the environmental stimulus has little impact on the similarity of behavior of twin pairs. But in the presence of the dominant form, the environmental stimulus becomes far more influential. Similarly, for the GgE and Gge we see the opposite effect: without the environmental stimulus, the DZ twins behave more similarly when neither twin experiences an interactive effect than when the environmental stimulus interacts with the genes of one twin only. Of course, these underlying mixture distributions generate precisely the same ACE estimates as the previous examples.

2.4 Example 4: $r_{\mathbf{GE}}$

The previous two examples assumed that genes and environment take values independently. It seems logical, however, that an environmental stimulus such as conservative parenting will be associated with children with conservative genotypes (and by extension, liberal parents with children with liberal genotypes). Such dependence between genetic and environmental variables is commonly referred to as "gene-environment correlation" (r_{GE}). While the population may still have 50% prevalence for a gene and 50% prevalence for a environmental stimulus, the joint distribution of the two may be decidedly non-independent. Table 2.4 shows a joint distribution of twin pair genotypes with the environmental stimulus for MZ and DZ twins. The rule defining the cells is simple: if one or more twin has the G form of the gene, the probability of also receiving in the environmental stimulus is high. If neither twin has the G form, the probability of receiving the environmental stimulus is very low.

MZ Twins

	E	e	
GG	0.4	0.1	0.5
gg	0.1	0.4	0.5
	0.5	0.5	

DZ Twins

	E	e	
GG	0.20	0.05	0.25
gg	0.05	0.20	0.25
Gg	0.4	0.1	0.5
	.65	0.35	

Table 1: Joint distributions of genes and environment for MZ and DZ twins

These cells form the weights of the mixture distributions for an example of r_{GE} that would result in the same overall correlation as previous examples:

$$r_{\mathrm{MZ}} = 0.64 = 0.4GGE + 0.1GGe + 0.1ggE + 0.4gge \tag{19}$$

$$= 0.4 \cdot 0.6625 + 0.1 \cdot 0.55 + 0.1 \cdot 0.55 + 0.4 \cdot 0.6625 \tag{20}$$

$$r_{\mathrm{DZ}} = 0.375 = 0.2GGE + 0.05GGe + 0.05ggE + 0.2gge + 0.4GgE + 0.1Gge \tag{21}$$

$$= 0.2 \cdot 0.6625 + 0.05 \cdot 0.55 + 0.05 \cdot 0.55 + 0.2 \cdot 0.6625 + 0.4 \cdot 0 + 0.1 \cdot 0.55 \tag{22}$$

What is most interesting about this example is that heterogeneous pairs without the environmental stimulus correlate just as much as identical twins with “mismatched” environmental stimuli, but because they are such a small part of the overall DZ twin population (10%), this finding is masked by the larger groups.

2.5 Example 5: Equal Environments

Throughout these examples, we have treated environmental stimuli exactly the same for MZ and DZ pairs. In the terminology of classical twin studies, this consistency between MZ and

DZ twins is known as the "Equal Environments Assumption" (EEA). If we were to notate the shared environmental stimulus (or lack thereof) for MZ twins as E_{MZ} and e_{MZ} and the shared environment for DZ twins as E_{DZ} and e_{DZ}, the EEA states $E_{\mathrm{MZ}} = E_{\mathrm{DZ}}$ and $e_{\mathrm{MZ}} = e_{\mathrm{DZ}}$. If MZ and DZ twins are systematically raised in different environments, this assumption does not hold. In an extreme example, DZ twins may never be exposed to a necessary environmental stimulus and weights for any distribution with E would be zero (e.g. $w_{GgE} = w_{GGE} = 0$), while MZ twins are always exposed to the stimulus:

$$r_{\mathrm{MZ}} = 0.64 = 0.5GGE + 0GGe + 0.5ggE + 0gge \tag{23}$$

$$= 0.5 \cdot 0.905 + 0.5 \cdot 0.375 \tag{24}$$

$$r_{\mathrm{DZ}} = 0.375 = 0GGE + 0.25GGe + 0ggE + 0.25gge + 0GgE + 0.5Gge \tag{25}$$

$$= 0.25 \cdot .375 + 0.25 \cdot 0.375 + 0.5 \cdot 0.375 \tag{26}$$

This example shows some severe gene-environment interaction. All the component distributions, except for GGE, show precisely the same amount of correlation. Because no DZ twins with the GG genotype pair are ever exposed to the environmental stimulus, we cannot make a meaningful comparison between the two groups.

3 All Models are Wrong, Some are Useful

Despite our concerns about misspecification, r_{GE}, G × E, and equal environments, the basic logic of classical twin studies is seductive. Like an experiment, we imagine that twins have been randomly assigned to treatment (being monozygotic) and control (being dizygotic). In so far as twin zygosity is random (or could be made "as-if" using matching or similar technique (Rosenbaum and Rubin, 1983)), this logic is sound. Any observed in difference cannot be the result of a confounding variable, and we have a host of analysis techniques to test null hypotheses of no effect. We can even estimate effect size as the difference between MZ and DZ twins on any given test statistic.

But can we generalize our findings to a larger population? Even leaving aside the not inconsiderable question of whether twins form a representative sample, does the ACE model create an accurate picture of forces at work in the general population? The five examples in this paper covered a wide range of situations, but they all generated the same ACE estimates: $A = 0.53, C = 0.11, E = 0.36$. If genes are not evenly distributed, if environmental factors correlate with genes, if genes interact with the environment, or if environmental factors are not shared by MZ and DZ twins, the underlying data generating process can deviate widely, and statements of a variation "explained" by genes or environment do not seem to match the actual process.

There are at least two solutions that are within our immediate grasp. First, researchers could be content to test the sharp null of no effect of genetics. Understanding when and where genes play a role is a valuable pursuit. Second, for cases where researchers want to know more about the strength of genetic and environmental effects, explicitly modeling both genes and environment is a necessary action. For example, Purcell (2002) demonstrates regression techniques that explicitly model G × E. These techniques require gathering data on specific environmental stimuli, which has typically been avoided in political science applications. To some degree, this may be the result of the nature of twin registries. With political science only recently showing an interest in twin data, few studies have been conducted that gather information on politically relevant variables. As data emerge, adding it to the model should become common place. If we intend to make claims about genes and environmental factors we will need to measure both, not simply assume them away.

References

Alford, J. R., Funk, C. L., and Hibbing, J. R. (2005). Are political orientations genetically transmitted? *American Political Science Review*, 99(02):153–167.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27(1):17–21.

Eaves, L. and Hatemi, P. (2008). Transmission of attitudes toward abortion and gay rights: Effects of genes, social learning and mate selection. *Behavior Genetics*, 38(3):247–256.

Fowler, J. H., Baker, L. A., and Dawes, C. T. (2008). Genetic variation in political participation. *American Political Science Review*, 102(02):233–248.

Medland, S. E. and Hatemi, P. K. (2009). Political science, biometric theory and twin studies: A methodological introduction. *Political Analysis*, 17:191–214.

Purcell, S. (2002). Variance components models for gene-environment interaction in twin analysis. *Twin Studies*, 5(6):554 – 571.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Settle, J. E., Dawes, C. T., and Fowler, J. H. (2009). The Heritability of Partisan Attachment. *Political Research Quarterly*, 62(3):601–613.