Economic Theory and Genetic Associations

Andrew Caplin, David Cesarini, Magnus Johannesson and Kevin Thom

May 19 2012

- Traditional approach to discovering and quantifying gene-behavior relationships:

$$Y = \beta_0 + \beta_1 \cdot SNP + PC \cdot \beta_2 + X \cdot \beta_3 + \varepsilon, \qquad (1)$$

- Limitations:
  - Looks specifically for a relationship with the mean.
  - Describes relationship between SNP and behavior Y after potentially endogenous responses to genotype.

Beyond the Mean

- *n=9,617* Swedish twins born 1926- 1958.
- Illumina HumanOmniExpress BeadChip ˜600,000 genetic markers.
- SALT survey in 1998: coffee, alcohol, smoking, BMI, health
- Roughly half similar survey in 1973.

- Coffee results for the T-allele on rs2472297, located near *CYP1A2*
- An additional copy of the allele is associated with:
  - *0.38* more cups of coffee per day ($p = 10^{-18}$).
  - Mean relationship replicates findings from previous studies (Sulem *et al.*, 2011, Cornelis et al., 2011, *CYP1A2*)
  - Increase of 0.2 in the growth of cups per day ($p = 10^{-4}$), 1973 - 1998.

Table : Relationship between rs2472297 and Coffee Consumption

rs2472297	0	1	2
1973	3.68 (2.50)	3.97 (2.67)	4.12 (2.96)
N	2500	2015	385
1998	3.62 (2.41)	4.02 (2.60)	4.35 (2.94)
N	4962	3956	758
Δ1973-1998	-.06 (2.54)	0.09 (2.81)	0.37 (2.79)
N	2487	2008	384

So What?

- Genetic effects are often small in magnitude, and precision is a concern.
- Ignoring other features of the distribution of Y is inefficient.
- Points towards mechanisms - Economic theory can unify relationships between genes and multiple behaviors or multiple features of a behavior.

So What Instead?

- Look for associations between genes and multiple features of the data (variance, median and other quantiles, skewness, growth over time)
- Problematic! - Exacerbates problems with multiple hypothesis testing
- Alternative Approach - use all of the information contained in the distribution of Y and time path to estimate structural parameters of a behavioral model.

- Instead of specifying a regression equation as data generating process, we can specify an economic model.
- In spirit of Becker and Murphy (1988). Period utility:

$$U_t(C_t, A_t, \epsilon_t, H_t) = (\alpha_1 + \epsilon_t)\left(\frac{C_t}{1+A_t}\right) + \alpha_2\left(\frac{C_t}{1+A_t}\right)^2 - H_t$$

- **Addiction stock** evolves according to:

$$A_{t+1} = (1-\delta_1)A_t + \delta_2 C_t$$

- **Health shock**: H_t takes the value h with probability $\frac{\exp(\phi_1+\phi_2 C_t)}{1+\exp(\phi_1+\phi_2 C_t)}$, and the value 0 otherwise
- **Taste shocks**: ϵ_t may be serially correlated.
- **Addiction formation / extinction**: δ_1 governs depreciation, δ_2 affects formation

- In spirit of Becker and Murphy (1988). Period utility:

$$U_t(C_t, A_t, \epsilon_t, H_t) = (\alpha_1 + \epsilon_t)\left(\frac{C_t}{1+A_t}\right) + \alpha_2\left(\frac{C_t}{1+A_t}\right)^2 - H_t$$

- The vector of parameters $\theta =< \alpha, \sigma_\epsilon^2, h, \phi, \delta >$ determines the the joint distribution of the life-cycle vector of consumption choices $< C_1, C_2, ... >$.
- θ can be genotype specific, $\theta(G)$
- Choose $\theta(G)$ to simultaneously match multiple features of the data.

Endogenous Response to Genotype

- A real benefit of introducing economic theory explicitly into estimation - allows us to estimate genetic effects in the presence of compensatory behavior in response to genotype.
- One Example:
  - Individuals may be (imperfectly) aware of their own genotype.
  - Such awareness can result in behaviors that mask the gene-behavior interactions in the mean, even if they show up in other features of the distribution of Y.

Endogenous Response to Genotype

- Consider the following extension of the basic model.
- Let $G \in \{0, 1\}$ represent an individual's genetic type.
- Genotype-specific habituation: $\delta_2 = \underline{\delta_2}$ if $G = 0$, and $\delta_2 = \overline{\delta_2}$ if $G = 1$.
- Individuals of type $G = 1$ have a more persistent habit.
- Individuals receive informative signal (belief about the probability of being type $G = 1$):
- Numerical simulations generate some insights into how this can affect inference.

- If the signal is correlated with type:

	G=1	G=0	Diff or Ratio
Mean	0.65	0.62	0.03
Variance	0.36	0.29	1.27**
Simulations	1000	1000	

- But if the signal is not correlated with type:

	G=1	G=0	Diff or Ratio
Mean	0.67	0.62	0.05*
Variance	0.37	0.29	1.26**
Simulations	1000	1000	

Implications for Replication

- Many gene-behavior associations fail to replicate in other samples.
- Often due to false positives.
- Replication failure might also stem from variation in the decision environment across countries or samples.
  - Primitive preference parameters $\theta(G)$ might be stable across populations.
  - If institutions vary (e.g. tax rates), we could see different gene-behavior relationships even with identical gene-parameter relationships.

Theory Based Estimation Offers the following potential advantages

- Efficiency - Uses all information
- Detects genetic effects that do not operate at the mean in a parametrically parsimonious way.
- Allows us to identify genetic effects in the presence of endogenous compensatory behavior.
- Allows us to identify genetic effects on primitives of the utility function and production technologies.
  - Welfare Statements
  - Counterfactuals