A well understood concern in social research is the difficulty of assessing causality due to endogenous relationships between social variables. Many solutions to account for possible endogeneity of social predictors have been put forward by researchers. Methods use either design or analytic techniques in an attempt to clean the relationship of interest of possible contamination by endogeneity. Longitudinal approaches, instrumental variables, fixed effects models, and quasi-experimental designs have all been used as possible solutions to this problem, with varying amounts of success.

With almost 26,000 papers to date (Google Scholar search 2010), instrumental variables are a common treatment of this endogenous variable problem. This strategy replaces the endogenous independent variable with a factor exogenous to the relationship of interest. The instrumental variable is a unique factor correlated with the endogenous independent variable but uncorrelated with the dependent variable. These requirements, while providing the underlying power of the method, also showcase the greatest difficulty with this approach: finding an acceptable instrument that meets these criteria. Despite the high volume of studies using this method, the difficulty of identifying credible instruments has led to few convincing papers. While previous studies have used such varied instruments as election year police levels (Levitt 1997) and rainfall levels (Munshi 2003), assessing the acceptability of these instruments has been a difficult task. To make a strong case for its use as an instrumental variable, a variable must be found that has no possible connection to the dependent variable, save through the independent variable.

Thanks to advances in biological science, genetic variants are now a potential source of credible instruments. A few previous studies (Norton and Han 2008, Fletcher and Lehrer 2009) have attempted to do just this. A key weakness of these previous studies is that the genetics instruments rely almost exclusively on behavior-related genes such as DRD4, DAT1, and MAOA. The products of these genes are involved in neurochemical responses and each has a known association with a number of social behavior and other outcomes, rendering them problematic as potential instruments. However, other genetic

relationships are available. Instead of using genes with a broad range of effects, we can use genes with more targeted and specific physiological effects of gene expression that may, in turn, be associated with specific behavior. Such a relationship lends itself to use as an instrumental variable system. The gene with a physiological effect (instrumental variable) is associated with the behavior of interest (endogenous independent variable) through that gene's expression, but because of the nature of that expression, is otherwise uncorrelated with other outcomes (such as the dependent variable).

Our project draws data from the College Roommate Study carried out in 2008 at a large, public university (Guo et al 2010). The study was designed to investigate joint peer and genetic effects on health behaviors and attitudes in a college campus setting. The study consists of a survey component and a saliva-based DNA component; 2,664 (79.5%) students in the targeted sample completed a web survey and 2,080 (78.7% of the survey completers) provided a saliva sample for genetic analysis. We also make use of a genome wide association study conducted with over 1,000 SNPs in the genetic sample of the National Longitudinal Study of Adolescent Health (Add Health) Wave 3.

This paper develops the theoretical strengths of using genes with targeted effects as a source for instrumental variables and thus a possible approach to addressing endogeneity before proceeding to explore our model systems. For these model systems, we show a 2 stage least squares analysis that tests the instrument's fit, interprets the results and compares them to results generated without taking into account the endogeneity of the independent variable. The first model system finds a weak instrument, according to commonly used standards. In contrast, the second model system provides a strong and valid instrument. In the second stage of the model, this instrument negates a previously significant result, suggesting that the originally observed relationship is possibly spurious or the result of reverse causation. This finding is replicated in the Add Health data set.

This project showcases the possible use of genetic variation to satisfy the assumptions of the instrumental variable method. As the collection of biological information is increasing within social

research, this approach could provide researchers with new tools for grappling with endogenous relationships.

Fletcher, Jason M. and Lehrer, Steven F., Using Genetic Lotteries within Families to Examine the Causal Impact of Poor Health on Academic Achievement (July 2009). NBER Working Paper No. w15148. Available at SSRN: http://ssrn.com/abstract=1434663

Guo, Guang, Jessica Halliday Hardie, Craig Owen, Jonathan K. Daw, Yilan Fu, Hedwig Lee, Amy Lucas, Emily McKendry-Smith, Greg Duncan (2009). DNA Collection in a Randomized Social Science Study of College Peer Effects. Sociological Methodology 39:1-29.

Levitt, Steven D (1997). Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime. American Economic Review 87: 270-290.

Munshi, Kaivan (2003). Networks in the Modern Economy: Mexican Migrants in the U.S. Labor Market. Quarterly Journal of Economics 118: 549-599.

Norton, Edward C. and Euna Han (2008). Genetic Information, Obesity and Labor Market Outcomes. Health Economics 17: 1089-1104.