ENDOGENOUS VARIABLES AND WEAK INSTRUMENTS IN CROSS-SECTIONAL NUTRIENT DEMAND AND HEALTH INFORMATION ANALYSIS: A COMPARISON OF SOLUTIONS

A Thesis

by

RAFAEL G. BAKHTAVORYAN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

May 2004

Major Subject: Agricultural Economics

ENDOGENOUS VARIABLES AND WEAK INSTRUMENTS IN CROSS-SECTIONAL NUTRIENT DEMAND AND HEALTH INFORMATION ANALYSIS: A COMPARISON OF SOLUTIONS

A Thesis

by

RAFAEL G. BAKHTAVORYAN

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Approved as to style and content by:

George C. Davis (Chair of Committee)	John P. Nichols (Member)
W. Alex McIntosh (Member)	A. Gene Nelson (Head of Department)

May 2004

Major Subject: Agricultural Economics

ABSTRACT

Endogenous Variables and Weak Instruments in Cross-Sectional Nutrient Demand and Health Information Analysis: A Comparison of Solutions. (May 2004)

Rafael G. Bakhtavoryan, B.S., Armenian Agricultural Academy

Chair of Advisory Committee: Dr. George C. Davis

In recent years, increasing attention has turned toward the effect of health information or health knowledge on nutrient intake. In determining the effect of health information on nutrient demand, researchers face the estimation problem of dealing with the endogeneity of health information knowledge. The standard approach for dealing with this problem is an instrumental variables (IV) procedure. Unfortunately, recent research has demonstrated that the IV procedure may not be reliable in the types of data sets that contain health information and nutrient intakes because the instruments are not sufficiently correlated with the endogenous variables (i.e., instruments are weak).

This thesis compares the reliability of the IV procedure (and the Hausman test) with a relatively new procedure, directed graphs, given weak instruments. The goal is to determine if the method of directed graphs performs better in identifying an endogenous variable and also relevant instruments. The performance of the Hausman test and directed graphs are first assessed through conducting a Monte-Carlo sampling experiment containing weak instruments. Because the structure of the model is known in the Monte-Carlo experiment, these results are used as a guideline to determine which procedure would be more reliable in a real world setting. The procedures are then

applied to a real-world cross-sectional dataset on nutrient intake. This thesis provides empirical evidence that neither the IV estimator (and Hausman test) or the directed graphs are reliable when instruments are weak, as in a cross-sectional dataset.

DEDICATION

This thesis is dedicated to my parents, Avetisova Maya and Bakhtavoryan Gagik.

ACKNOWLEDGEMENTS

I would like to express gratitude to my committee members, Dr. Nichols and Dr. McIntosh, for their ready assistance whenever it was necessary. My big indebtedness lies with Dr. Davis, chair of my advisory committee, for being such a huge assistance along the way. I would also like to thank Dr. Davis for his great mentoring and guidance. Without his help and persistence this thesis would have never been written. I thank Dr. Infanger (USDA MAP's former director & coordinator) and Dr. Dunn (ATC director) for having faith in me and granting the funding for my education. I express my genuine gratitude to my best friend and an incredible person, Susan Livingston, who surrounded me with support and inspiration. Also, I would like to express genuine gratitude to my friends Tom, Armen, Haik, and the rest of the Armenian students, for being true friends throughout my academic life. And last, but not least, I would like to eternally thank my Mom and Dad, my sister and Grandma, for all the accomplishments in my life. It has been their unconditional love and support that has gotten me through difficult times.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

According to Variyam and Golan, over the past century consumption patterns of many food commodities have changed in the face of changing consumer demand due to such major economic factors as constantly changing relative price and income levels (USDA, 2002). But many studies have shown that another important factor that also affects consumers' consumption patterns is health information knowledge [e.g., Brown and Schrader 1990; Capps and Schmitz 1991; Carlson and Gould 1994; Variyam, Blaylock, and Smallwood 1996, 1998]. Many consumers have adjusted their food choices according to the health information that they receive from different sources, such as government education programs, nutrition facts labels, product health claims, and the popular media (USDA, 2002). Consumers can also get nutrition information through the sources such as the food guide pyramid and the formulation of quantitative recommendations in the dietary guidelines for Americans provided by the US Department of Agriculture (USDA).

Information will continue to be one of the key determinants in affecting consumers' food consumption patterns. Even though many people spend considerable time watching programs on TV and listening to the radio, there is also another significant source of information, which is the Internet. It creates an additional opportunity to inform more and more consumers about the relationship between health information knowledge

---

This thesis follows the style and format of the *American Journal of Agricultural Economics.*

and nutrient intake and the results of that relationship through numerous publications, articles, papers, and discussions available on the web. Consumers may keep adjusting their product preferences when they become more and more aware of health benefits of specific products. They may increase their spending on relatively nutritional and healthy products, and cut down on the consumption of relatively less nutritional food. This consistent consumption of relatively nutritional and healthy food will benefit consumers by eventually reducing their costs on health care (Fries, Koop, and Beadle 1993). Of course, in order for this argument to work it is implicitly assumed that the effect from the health information is significant. Furthermore, health care costs are a major public concern and their reduction (or at least reduction in rate of increase) is the focus for many public policy interventions. For policy to be efficiently designed and implemented it becomes important to compute accurate parameter estimates of the health information knowledge and the nutrient intake relationship.

The accurate estimation of the relationship between health information knowledge and nutrient intake requires using the appropriate procedure. It is expected that health information is an endogenous variable. That is true for two reasons. First of all, nutrient intake is a matter of choice, and it is up to consumers to decide on the amount of nutrient to consume. Second of all, consumers can also choose and control health information. That is, they can increase the level of their health information knowledge to some extent through reading, television, and the Internet (Park and Davis, 2001). Thus, there is expected to be a correlation between the regressor of interest (health information) and the disturbance term in a nutrient demand equation. Yet, one of the several assumptions in the theory of ordinary least squares (OLS) is that the explanatory variables and the

disturbance term must not be correlated. If they are correlated then it becomes impossible to isolate the impact of the explanatory variable on dependent variable by the OLS. This problem of correlation between an explanatory variable and disturbance term can be due to endogeneity or measurement error. When we have the endogeneity or measurement error problem then the parameter estimate yielded by the ordinary least squares is biased. In this thesis these problems will just be called the endogeneity problem.

One of the solutions to this endogeneity problem is the method of instrumental variables (IV). The objective of instrumental variables, or simply, instrument, estimation is to obtain a theory-consistent estimator by finding an instrument that is highly correlated with the explanatory variable of interest and uncorrelated with the disturbance term, using the method of moments. The method of instrumental variables provides consistent parameter estimates when there is a correlation between an explanatory variable and the disturbance term. However, for the IV estimator to provide consistent parameter estimates, the instrumental variables must satisfy two conditions:

1. The instrumental variables must be highly correlated with the explanatory variable of interest;
2. The instrumental variables must be uncorrelated with the disturbance term.

The failure to meet either of these two assumptions will result in the estimation of biased and inconsistent parameter estimates. A common approach for determining whether the OLS or IV is an appropriate estimation technique is to conduct the Hausman specification test (Hausman 1978). However, the Hausman specification test also depends on the degree of correlation between the instruments and explanatory variables. That is, the higher the correlation between the variable of interest and the instrument, the better

the performance of the Hausman test. Conversely, as the correlation between the variable of interest and the instrument decreases then the performance of the Hausman test deteriorates. It would be appealing to have an alternative to the Hausman specification test, because condition 1, and possibly condition 2, may not be satisfied, especially in cross-sectional data, where the instruments are often weak.

A possible alternative to the Hausman specification test is the method of the directed graphs. The method of the directed graphs is designed to assign 'causal' flows between variables that will indicate what variables are endogenous or exogenous. However, unlike the Hausman test, we do not know if the directed graphs are a reliable method when the correlation between variables is low.

Objectives

The major purpose of the thesis is the assessment of the Hausman test versus the method of the directed graphs in determining the endogeneity of the health information, given weak instruments. More specifically the objectives are:

1. To determine if the method of the ordinary least squares outperforms the method of the instrumental variables in terms of yielding unbiased and consistent parameter estimates given the low correlation between variables;
2. To determine how well the Hausman specification test performs when instruments are not highly correlated with regressors, or when correlations are low;
3. To determine how well the method of the directed graphs performs, as an alternative to the Hausman specification test, in identifying which variables are endogenous and which are exogenous when correlations are low;

4. To apply the methods to the real world health information and nutrient intake data and contrast the results.

For the Directed Graph analysis Tetrad II (Scheines, Spirtes, Glymour, and Meek 1994) software was used. Also, TSP and Microsoft Excel software were used in the thesis.

Rationale and Significance

The successful completion of the objectives will be significant in two aspects: econometric and applied. The research will suggest whether the Hausman specification test or the method of the directed graphs is more reliable for constructing models when instruments are weak. The applied significance will be obtaining an accurate measure of the impact of health information knowledge on nutrient intake. If policy makers have more accurate estimate of the parameter on health information obtained from the most reliable estimation technique, they will know better as to what level of health information consumers would need to lead a healthy lifestyle that would eventually lead to the reductions in the consumers' health care costs. The policy makers can influence the desired level of information, having the accurate results of the research at hand, through expansion of the level of health education programs, thus encouraging the consumption of those products and thereby accomplishing the ultimate objective, which is the reduction in the health care costs.

Consumers' consistent consumption of nutritional and healthy food might result in the reduction of the health care costs that will also reduce the cost on government. As such, it can be concluded that government has an interest in obtaining accurate parameter

estimates on the health information and nutrient demand relationship. In addition to all the private sources, such as friends, neighbors, personal physicians, and commercial advertising, some of the health information comes from the government regulatory role. Through its regulatory role government can mandate that health information be provided various ways, or at least foster a clear schematic, such as the Food Pyramid, so that people can understand optimal food choices. Also, government spends money on numerous health education programs (expanded nutrition program, extension service). So it becomes clear that government has a vested interest in obtaining accurate parameter estimates on the health information and nutrient intake relationship. As such, a successful completion of the thesis will improve the decision-making of government, which will be reflected in more effective design and implementation of all types of nutrition intervention programs.

This thesis proceeds in the following manner. A review of the literature will be provided in Chapter II. Chapter III will discuss the empirical results from a Monte-Carlo sampling experiment and the Hausman test, and the results from the method of the directed graphs will also be provided. Chapter IV employs a real-world cross-sectional dataset and presents the results from the Hausman Test and the method of the directed graphs. Thesis summary and conclusions will be presented in Chapter V.

CHAPTER II

LITERATURE REVIEW

The literature reviewed in this thesis is for the purpose of demonstrating the relationship between health information knowledge and nutrient demand, and also for presenting some auxiliary procedures for handling the endogeneity issue present in that particular relationship. This section consists of two different subsections. Presented first is literature showing the relationship between health information knowledge and nutrient intake, while the second subsection focuses on the improvement of statistical procedures for tackling the health information endogeneity issue.

Literature on Health Information and Nutrient Intake

Identification of the different economic and socio-demographic factors that influence the demand for various nutrients or foods is becoming an increasingly important issue. However, mere identification of those factors is just a part of the overall picture. Another significant issue that needs to be considered is the technique used in measuring the impacts of the economic and socio-demographic factors on the demand for different nutrients or foods. Growing scientific evidence suggests that a relationship exists between health information knowledge and the nutrient intake [e.g., Brown and Schrader 1990; Capps and Schmitz 1991; Carlson and Gould 1994; Variyam, Blaylock, and Smallwood 1996, 1998]. However, discussions in some of these papers were done subject to some caveats that might possibly arise due to the choosing inappropriate

technique (Park and Davis, 2001), or neglecting an important problem, such as endogeneity issue.

The endogeneity issue may be caused by measurement error, meaning that health information cannot be measured without error. Also, the endogeneity issue may be caused by the fact that consumers may have control over the level of their knowledge. That is, consumers can dedicate as much time to becoming more knowledgeable as they want, because it is completely up to them. These two problems lead to the endogeneity issue. And, given the importance of the accurate estimation of the relationship between the health information knowledge and nutrient intake, it becomes crucial to use appropriate technique that would take into account the endogeneity issue and yield legitimate results from the estimation procedure. In this chapter some of the literature will be reviewed in terms of what is the relationship between health information knowledge and nutrient intake, or food demand. However, major focus will be on the approach used during the estimation procedure.

Brown and Schrader (1990) investigated the impact of cholesterol information on the consumption of the shell eggs. The quarterly time-series data used in the study came from 1955-87 and 1966-87 from various sources, such as US Department of Agriculture, Bureau of Labor Statistics. Also, data from *Livestock and Poultry*: *Situation and Outlook*, and *Survey of Current Business*, *Grain and Feed Market News*, *Fats and Oils Situation*, and *Monthly Labor Review* were used. A demand equation included the shell egg consumption as a dependent variable, and the price for grade A large eggs to retailers, the real price of meat as a substitute, per capita income, cholesterol information index, the percentage of women in the labor force, and time (quarterly) as independent variables.

The cholesterol index was constructed after reviewing almost 3,200 journals from Medline database that contained materials on basic health research. The cholesterol index was developed based on the running total of the number of articles that either supported or questioned a link between diet cholesterol or serum cholesterol and heart disease. Each article that supported a link between cholesterol and heart disease added one unit to the running total (lagged two quarters) and each article questioning such a link subtracted one unit from the running total.

A variant of a double log demand equation was estimated using fixed coefficients weighted two-stage least-squares (instrumental variables procedure). At first, the demand equation for the shell eggs was estimated using the quarterly time-series data from 1966-87. They were able to obtain the cholesterol information index during that time period. However, because of high multicolinearity between some variables, they had to reestimate the demand equation using the quarterly time-series data that came from 1955-87 setting the value of the cholesterol information index to zero before 1966. The results of the study employing the data from 1955-87 showed that cholesterol index was statistically significant at the 1% significance level. Also, according to the results of the study using the data from 1955-87, cholesterol information led to a decrease in a per capita shell egg consumption of about 16% by the end of 1986. According to the results of the study using 1966-87 data, cholesterol index was significant at the 5% significance level, and cholesterol information decreased per capita shell egg consumption by 25% by the end of 1986.

Obviously, in this study the endogeneity problem was handled through implementing the instrumental variables procedure, however, the authors did not point

out the instruments throughout their discussion, nor did they mention the possibility of weak instruments.

Capps and Schmitz (1991) measured the impact of the health information on the consumption of beef, pork, poultry and fish. In their study they employed annual data from 1966-1988 from the USDA series Food Consumption, Prices, and Expenditures. The dependent variables of the demand equations were the quantity of beef, pork, poultry, and fish consumed, and the independent variables included in the demand equations were prices of beef, pork, poultry, and fish, total meat expenditure, and the cholesterol information index developed by Brown and Schrader.

After building a Rotterdam model the iterative Zellner estimation procedure was used to estimate the coefficients of the demand equation at the 0.10 significance level. The estimation results showed that cholesterol information index had a negative effect on beef and pork consumption, and had a positive effect on poultry and fish consumption, all else held constant. The cholesterol index coefficient estimates of beef, pork, poultry, and fish were -0.000219, -0.000884, 0.000892, and 0.00021, respectively. However, the cholesterol index coefficient of beef was statistically insignificant. As a result the conclusion can be drawn that the cholesterol information had very small effects on the consumption of pork, poultry and fish.

The endogeneity issue was not considered in this study, or, at least no discussion regarding checking for the endogeneity was presented in their work, indicating that the authors implicitly assumed there was no endogeneity problem.

Variyam, Blaylock, and Smallwood (1998) estimated the effect of nutrition information knowledge on cholesterol consumption. The cross-sectional data used in this

study came from the USDA's 1989-91 Continuing Survey of Food Intakes of Individuals (CSFII) and the companion Diet and Health Knowledge Survey (DHKS). Cholesterol intake was included in the model as a dependent variable. As independent variables they included nutrition information, income, household size, age, sex, race, ethnicity, schooling, presence of children in the family, place of living (regional or urbanization), employment (part employed, not employed), program participation (e.g., Food Stamp Program), body mass index (BMI), being vegetarian, being smoker, dieting status, disease, TV hours watched, comparing product nutrition on labels (always, sometimes, never). Two variables were included in the model for capturing the nutrition information effects, $INFO_1$ and $INFO_2$. While the nutrition variable $INFO_1$ examined consumers' opinions about cholesterol control, without considering any specific diet-health link, the nutrition variable $INFO_2$ measured consumers' ability to name health problems caused by excessive cholesterol intake.

The reduced-form intake and information equations were estimated using a generalized probit minimum distance estimator (MDE) to measure the impact of consumers' socio-demographic and attitudinal factors on cholesterol intake. The estimation results indicated that the estimate of the $INFO_1$ coefficient was – 0.066, and it was significant at the 1% level. The estimate of the $INFO_2$ was -0.243, and it was significant at the 10% level. Both results indicate that greater nutrition information leads to the reduction in the consumption of cholesterol, ceteris paribus. The results of the estimation also showed that, everything else constant, income and BMI variables had a significant positive total effect on cholesterol consumption with 10 mg total income effect and 44 mg total BMI effect at 1% significance level. The study also found that

variables such as age, sex (female), being vegetarian, and being on low-calorie diet had a negative significant effect. For example, total age effect was 0.62 mg, total sex (female) effect was 91 mg, ceteris paribus, at 1 % significance level. The other independent variables included in the model turned out to be insignificant at the 1% significance level.

The authors handled the endogeneity problem in their analysis by separating direct and indirect informational effects of independent variables that had an impact on both information and cholesterol intake simultaneously, by treating those variables as endogenous. Because of the problem of endogeneity the OLS estimation was considered inappropriate. As a result, they used instruments for the information variables included in the structural model, thus tackling the problem of endogeneity. Also, it is worth pointing out that no weak instrument possibility was mentioned.

In a study by Carlson and Gould (1994), the authors examined the impact of health information on determining dietary fat intake. The cross-sectional data employed in this study came from 1989 to 1990 and 1990 to 1991 Continuing Survey of Food Intakes of Individuals (CSFII) and companion Diet and Health Knowledge Surveys (DHKS). The binary dependent variables in the model were total fat and saturated fat. Independent variables included in the models were socio-demographic and attitudinal characteristics of the meal planners, such as health awareness, income, age, education (less than 12 years education, one to three years of college, four years or more of college), race (black), ethnicity (Hispanic), region of residence, disease (having high blood pressure or high cholesterol), pregnancy, TV watching (watched no TV yesterday, watched about an hour or less of TV yesterday, watched about two or three hours of TV yesterday) , the presence of children in the family, the importance of nutrition when

buying food (always, sometimes), dieting status (low-fat diet). Dummy variables were used to represent health awareness for both total fat and saturated fat: if a meal planner was knowledgeable about the relationship between fat (total and saturated) intake and health effects the dummy variable took on the value of 1, otherwise 0.

The probit model was estimated implementing Heckman's two step estimation procedure. In the study they split dietary fat intake into two categories: total fat and saturated fat. "Each fat type required estimation of a first-stage probit equation relating health knowledge status to a set of meal planner characteristics, and a pair of second-stage nutrient intake equations, one for each health knowledge regime" (p.377).

The study found that coefficients associated with health awareness for both total fat and saturated fat for the health "aware" regime (when the dummy variable took on the value of 1) were -8.981 and -3.465, respectively. Although, both coefficients were negative, they appeared to be significant at the 5% significance level. The coefficients associated with health awareness for both total and saturated fat for the health "unaware" regime (when the dummy variable took on the value of 0) were -5.863 and -1.658, respectively. Both coefficients were negative, as well as insignificant at the 5% significance level.

The results also showed that such variables as income, education (one to three years of college, four years or more of college), TV watching (watched about an hour or less of TV yesterday, watched about two or three hours of TV yesterday), and the importance of nutrition when buying food (always, sometimes) had a positive significant effect on total fat intake, whereas the effect of race was significant, though negative, at the 5% significance level. The effects of the rest of the variables were inconclusive at the

5% significance level. For the saturated fat, among the variables that appeared to have a positive significant effect was income, the importance of the nutrition when buying food (always, sometimes), education (four years or more of college), and TV watching (watched no television yesterday, watched about two or three hours of television yesterday) at the 5% significance level. Such variables as education (less than 12 years education), race (black), and ethnicity (Hispanic) had negatively influenced the saturated fat intake at the 5% significance level. The rest of the variables were statistically insignificant.

The information endogeneity issue was addressed by using a switching regression models. Another thing that needs to be pointed out is that no weak instrument possibility was mentioned in the study. Considering the results of the 1994 study, Carlson and Gould made several suggestions regarding implementing education programs that would enhance consumer's awareness on the link between the health knowledge and the fat intake specifically involving people with less education, less income, and those who watch more hours of television.

Variyam, Blaylock, and Smallwood (1996) investigated the influence of fiber-related nutrition information on the dietary fiber intake. The cross-sectional data used in the study came from the 1989 Continuing Survey of Food Intake of Individuals (CSFII) and the companion Diet and Health Knowledge Survey (DHKS). The dependent variable in the model was total dietary fiber intake, whereas a wide set of independent variables was divided into three categories:

1. Health information knowledge, household characteristics, which include income, program participation, household size, presence of children in the family, region

of the household (Midwest, South, West), location of the household (Suburban, Nonmetro), meal planners;

2. Socio-demographic characteristics, such as gender (female), race (black), ethnicity (Hispanic), education (High school, College, Postgraduate), age, employment (not employed);
3. Meal-planner health-diet characteristics, such as smoker, vegetarian, special diet, fiber supplement, body mass index.

Fiber information was represented by three variables knowledge, awareness, and attitude.

A probit latent variable model was estimated using computationally tractable estimation procedures, based on the minimum distance estimator. The results of the study confirmed the expectation that fiber information variables did affect the dietary fiber intake. The coefficients of the fiber information variables, knowledge, awareness, and attitude were 0.067, 0.126, and 0.071, respectively. Even though the coefficients of all the three fiber information variables had a positive direct effect on fiber intake, however, the knowledge coefficient was insignificant, whereas the awareness coefficient was significant at the 1% level and the attitude coefficient was significant at the 10% level. Also, according to the results of the estimation, such variables as income, program participation, race (black), smoker, and body mass index had a negative significant effect on the dietary fiber intake. However, region of the household (West), gender (female), education (High school, College, Postgraduate), age, vegetarian, special diet, and fiber supplement positively influenced the intake of the dietary fiber intake.

Isolating direct and indirect effects of exogenous variables on the dietary fiber intake and treating the information variables as endogenous determinants of intake handled the endogeneity problem. Also, it should be pointed out that no weak instrument possibility was mentioned in the study.

The authors directed readers' attention to the fact that all the analysis of the results had to be done with caution. They pointed out two problems: mismeasurement and generalizability. The first problem might be caused by the question format and wording, yet the second problem called for the questioning every member of the household.

Literature on Statistical Procedures and Methods

All the work discussed above implements procedures to analyze the health information knowledge and nutrient demand relationship. However, the appropriate procedures usually depend on underlying statistical assumptions, which may or may not be satisfied. This is especially true of the IV procedures in cross-sectional datasets, because the correlations between variables are relatively low, as opposed to the correlations in the time-series datasets.

As was already mentioned above, the high correlation between the regressor and the instrument is an important condition that needs to be satisfied, so that the technique of the IV estimation can be considered legitimate. Instrumental variables are said to be relevant if they are highly correlated with explanatory variables. The strength of correlation (relevance) is extremely important. This is important because in cross-sectional datasets there are normally low correlations between variables. Low degree of relevance results in inconsistent IV parameter estimates. In fact, in the paper by

Nakamura and Nakamura the authors attach a great deal of importance to the relevance issue (Nakamura and Nakamura 1998). They argue that failure of the instruments to capture the part of variation in the endogenous regressors of interest that is crucial from application standpoint can be one of the major reasons strongly advocating against the procedure of instrumentation. One of the major findings in the study by Staiger and Stock (1997) was that when there is a low correlation between the instrument and the endogenous explanatory variable the conventional asymptotic property of the instrumental variable estimator cannot be considered as legitimate even in the case of a large sample size. These conclusions bring up motivation for developing an accurate measure of the instrument relevance.

In his 1997 paper, John Shea suggests the following procedure for calculating the partial R^2p that measures instrument relevance:

> For a given explanatory variable Xi compute the squared correlation between the component of Xi orthogonal to the other explanatory variables, and the component of Xi's projection on the instruments orthogonal to the projection of the other explanatory variables on the instruments. This partial R^2p can be computed using a series of simple ordinary least squares (OLS) regression (p.348).

However, Godfrey, in his 1999 paper, suggested a simpler way of computing the partial R^2p. Specifically, he proposed the following formula for partial R^2p:

$$R^2p = (VARols/VARiv)*(SSRV/SSRS),$$

where VARols is the variance of the OLS estimate, VARiv is the variance of the IV estimate, SSRV is the error variance estimate of the IV estimate, and SSRS is the error variance estimate of the OLS estimate. The motivation for Godfrey to write a critique on Shea's paper was the assumption made by Shea. Specifically, Shea assumed that common

error variance must be used to get covariance matrix estimates for the OLS and IV. Godfrey disagrees with this assumption by claiming that the variance of error term is unknown and computer software uses estimated error variances of OLS and IV parameter estimates to calculate estimates of covariance matrices. Godfrey goes on to argue that failure to recognize the importance of common error variance can lead to the underestimated instrument relevance.

Park and Davis (2001 use the partial R^2p as a measure of instrument relevance in their paper on the relationship between health information and nutrient demand. One of the objectives of the paper was to identify whether the results through the IV estimation outperformed the results yielded by the OLS with the help of different specification tests employing 1994-1996 Continuing Survey of Food Intakes by Individuals (CSFII) and Diet and Health knowledge survey (DHKS) data. As with the most cross-sectional data settings, they had to face the problem of endogeneity and measurement error. They developed appropriate instruments according to a procedure developed by Lewbel (1997) to handle the problem. To measure the relevance of those instruments they reported values of auxiliary R^2 and partial R^2p. These values implied that instruments were extremely weak, as a result, it was expected that the Hausman test would not be reliable because of the present relevance problem. However, for comparison they conducted the Hausman test and the result of that test indicated the need of using the IV instead of the OLS. That is, the Hausman test rejected the null hypothesis and concluded that the estimate obtained from the IV estimator was different from the estimate yielded by the OLS estimator. Because the Hausman test does not tell us about the strength of the endogeneity but rather the existence of it, Park and Davis followed technique suggested

by Nakamura and Nakamura (1998), which advocates for considering the predictive evaluations of alternative sets of estimation results. They reported results from both OLS and IV estimation and then made out of sample comparisons. According to those out of sample comparisons, the OLS estimation technique clearly outperformed the IV estimation technique, in spite of all the results from the Hausman test. This provides further evidence that the Hausman test may not be reliable when the instruments are weak.

Based on the literature reviewed above, it becomes clear that there might be other procedures to better handle the health endogeneity issue. This becomes more appealing especially considering the fact that in the many of the papers reviewed the authors did not consider that issue in their analysis. In the next chapter, a Monte-Carlo sampling experiment will be conducted, and the Hausman test and the method of the directed graphs will be applied to the artificially created dataset.

CHAPTER III

MONTE–CARLO

Description of Monte-Carlo Experiment

A Monte-Carlo sampling experiment is a useful method to use when evaluating the performance of different estimators. The current econometric literature strongly suggests the use of this method for studying the properties of different estimators for small or finite samples (Kennedy 1998, p. 22; Gujarati 2002, p. 92.). The advantage of doing a Monte-Carlo sampling experiment is that the researcher already knows the underlying model structure because he is the one who makes decisions on the parameter settings. The complete knowledge about the underlying model structure will make it a lot easier to evaluate the performance of various estimators.

The general structure of a Monte-Carlo sampling experiment consists of several steps:

1. We have to come up with a model, which is going to be used for generating the data. Suppose, we have the following linear function:

   $Y=\beta_1 X_1+\beta_2 X_2+\varepsilon,$

   where Y is the dependent variable, X_1 and X_2 are the regressors of interest, β_1 and β_2, are the true values of the parameters, and ε is the disturbance to the equation.
2. We set β_1 and β_2 equal to certain numbers (one number for each β), and hold them constant across experiments.
3. Then we have to specify the sample size (N). It is a prerogative of a researcher to set the sample size at any specific level.

4. Next we set X_1 and X_2 equal to certain numbers. We are going to have N values of X_1 and X_2. These are going to remain constant across all experiments.
5. The next step is to generate N values of ε. Those N values of ε are randomly drawn from a normal distribution with mean zero and known variance σ^2. Nowadays, most statistical packages have built-in random number generators, which enable us to get these N values of ε. A key feature of the study is that all of the (usually unknown) parameter values are known to the person conducting the study, because this person chooses these values.
6. Now that we have N values of all the necessary variables, N values of Y can be created.
7. After creating N values of Y, we regress them on N values of X_1 and X_2 chosen in step 3. Thus, we will get the ordinary least-squares estimates $\hat{\beta}_1$ and $\hat{\beta}_2$.
8. We can repeat this same experiment many times. But, we have to keep in mind that every such experiment will differ from the next one due to the changes in the values of ε, because, as mentioned above, every time N values of ε are going to be randomly drawn from normal distribution with zero mean and known variance σ^2. Usually this experiment is conducted many times, say 1000 or 2000.
9. After getting, say 1000 or 2000 ordinary least-squares estimates $\hat{\beta}_1$ and $\hat{\beta}_2$, we, then, have an empirical distribution for β_1 and β_2, from which we may compute the average and compare those numbers with the true values of β_1 and β_2. If the computed least-squares estimates are about the same as their respective true values, then we conclude that least-squares estimator is performing appropriately.

Experimental Statistics

The above Monte-Carlo experiment framework must be amended if there is intent in comparing an OLS to an IV estimator. This can be done using a data-generating process similar to one discussed by Shea (1997).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \lambda u_1 + (1- \lambda)u_2 \quad (3.1)$$

$$X_1 = \gamma u_1 + (1 - \gamma)e_1 \quad (3.2)$$

$$X_2 = \gamma u_2 + (1 - \gamma)e_2 \quad (3.3)$$

$$Z_1 = \delta e_1 + \varphi v_1 \quad (3.4)$$

$$Z_2 = \delta e_2 + \varphi v_2 \quad (3.5)$$

where u_1, u_2, e_1, e_2, v_1, and v_2 are unobserved disturbances drawn from independent standard normal distributions; and where Y, X_1, X_2, Z_1, and Z_2 are generated observable variables. Equation (3.1) is the structural equation of the interest. From equations (3.2) and (3.3), the OLS estimation is inappropriate, because X_1 and X_2 are correlated with u_1 and u_2, respectively. From equations (3.4) and (3.5), the Z's are correlated with the X's but uncorrelated with the u's, so that the instrumental variables (IV) estimation of equation (3.1) may be warranted using Z's as instruments. The relevance control parameter δ governs correlation among the Z's and X's. When the value of the relevance control parameter δ goes down, so does the correlation between Z's and X's. The parameters λ and γ govern the correlation between the X's and the disturbance to equation (3.1). Increases in the endogeneity control parameter γ raise the endogeneity of both X_1 and X_2, but they also reduce the correlation with Z for a given δ from equations (3.2) and (3.3). Increases in λ raise the endogeneity of X_1 relative to that of X_2. Prior research suggests that the performance of the IV estimator may deteriorate as γ increases,

particularly if relevance is weak (Bound, Jaeger, and Baker 1995; Buse 1992; Shea 1997; Staiger and Stock 1997). The parameter φ, finally, governs the amount of variation in the Z's and is unrelated to the exogenous components of the X's.

In all cases β_0 is set equal to 3, β_1 to 1, β_2 to 2, λ to 0.9, and φ to 0.1; the relevance control parameter δ and the endogeneity control parameter γ vary across experiments. Specifically, three values of 0.3, 0.7, and 0.9 were chosen for the endogeneity control parameter γ, and five values of 0.1, 0.07, 0.014, 0.01, and 0.008 were chosen for the relevance control parameter δ. The values for the endogeneity control parameter γ and the relevance control parameter δ were chosen carefully so that the data generated through Monte-Carlo sampling experiment would mimic our real-world data. For that purpose, the parameter values of the endogeneity control parameter γ and the relevance control parameter δ needed to be set at such levels, so that the correlations between variables in the artificially created dataset would be similar to the correlations between variables in our real-world dataset. That objective was accomplished through several iterations by varying the parameter values of the endogeneity control parameter γ and the relevance control parameter δ until the desired correlations were computed. Only one set of the parameter values of the endogeneity control parameter γ and the relevance control parameters δ, which are 0.7 and 0.014, respectively, generated necessary correlations in the artificial dataset. Then from that one set, the parameter values of the endogeneity control parameter γ and the relevance control parameter δ were increased and decreased trying to stay within a range of numbers that generated correlations similar to the ones in the real-world dataset.

Experiment 1 (200 Observations per Row)

In all, there are 15 combinations of γ and δ (5x3). Each experiment consists of 10 trials. Hence, there are 10 OLS and IV estimates. The OLS and IV estimates are obtained by drawing 200 observations for each trial. For each experiment the median values of the 10 parameter estimates and the absolute median value of the 10 biases of the parameter estimates are reported. Consideration of the absolute median value of the biases will provide an alternative picture of true bias, because the sign of the bias may be of little concern. Also reported are the median of partial correlation between X_1 and Z_1 denoted by R^2p_1 and the median of partial correlation between X_2 and Z_2 denoted by R^2p_2. Partial correlations measure the relevance of the instruments and were calculated using Godfrey's formula (Godfrey 1999, pp.550-52). Low partials correlations imply that the instruments are weak, and most likely they will not perform well bringing up the relevance problem. Finally, the percentage of the times the Hausman test rejects the null hypothesis is reported. The null hypothesis of Hausman test is whether the OLS estimator yields a value that is significantly different from that produced by the IV estimator (Griffiths, Judge, and Hill 1992, pp. 463).

Before presenting the results, prior reasoning suggests that with decreasing relevance control parameter δ the OLS estimator is expected to perform better than the IV estimator in terms of parameter estimation and in terms of bias. The reason the OLS will outperform the IV in terms of parameter estimation and bias is because the relevance control parameter δ governs the correlation between Z's, as well as Z's and X's; lower correlation between Z's and X's violates one of the two assumptions about IVs that they should be highly correlated with the regressors, (i.e. the relevance problem). It should be

pointed out that a monotonic relationship is expected between the median estimates from both estimators and the relevance control parameter δ, as well as between the median absolute biases from both estimators and the relevance control parameter δ. A decline in the values of both partial correlations is expected with the decrease in the relevance control parameter δ as well, with the endogeneity control parameter γ being held constant, because decreasing the relevance control parameter δ lowers correlation between Z's and X's. Also, it should be noted that the value of the partial correlation associated with the independent variable of X_2 is expected to be greater compared to the value of the partial correlation associated with the independent variable of X_1, for each level of the endogeneity control parameter γ, because by the parameter setting for λ, X_1 is more endogenous than X_2. In case of a small value of the partial correlation the Hausman test behaves poorly. In other words, the power of the Hausman test is low (Park and Davis, p. 846), so it is expected that the percentage of the times the Hausman test rejects the null hypothesis will go down given weak instruments when the relevance control parameter δ declines.

Changes in the value of the endogeneity control parameter γ will also have their implications. An increase in the value of the endogeneity control parameter γ will raise the level of endogeneity of both variables X_1 and X_2. However, the endogeneity level of X_1 will increase by more relative to the endogeneity level of X_2, because the coefficient of the λ is 0.9. So, when there is an increase in the level of the endogeneity, the OLS estimator is expected to perform better relative to the IV estimator in terms of bias for the regressor of X_2 for each level of the relevance control parameter δ. However, the IV estimator is expected to perform better relative to the OLS estimator in terms of bias for

the regressor of X_1 for each level of the relevance control parameter δ. One thing that needs to be pointed out is that there is expected to be a monotonic relationship between the median estimates from both estimators and the endogeneity control parameter γ, as well as between the median absolute biases from both estimators and the endogeneity control parameter γ. It is expected that both partial correlations will decline when there is an increase in the endogeneity control parameter γ. Increased endogeneity in the regressors indirectly influences the instruments (through their correlation with the regressors) by increasing their level of endogeneity, thus making them less relevant. This is reflected in the declining values of both partial correlations (Shea 1997). Furthermore, because increases in the endogeneity control parameter γ raises the endogeneity of both explanatory variables, we expect the value of the percentage of the times the Hausman test rejects the null hypothesis to go up indicating the need to instrument.

Table 3.1 presents results from a series of experiments generated by equations (3.1)-(3.5). In the discussion below table 3.1 is broken into three blocks: block 1, block 2, and block 3. Within each block the value of the endogeneity control parameter γ is held constant at a certain level and the value of the relevance control parameter δ is decreasing moving down the block.

Table 3.1. Results from Monte-Carlo Experiment with the Number of Observations 200 per Row

	δ	γ	Median β_{1OLS}	Median β_{2OLS}	Median β_{1IV}	Median β_{2IV}	Median abs bias β_{1OLS}	Median abs bias β_{2OLS}	Median abs bias β_{1IV}	Median abs bias β_{2IV}	Median R^2p_1	Median R^2p_2	% of times Hausman Test rejects null hypotheses
	.1	.3	1.46588	2.05946	.91672	1.97835	.46588	.059456	.17304	.10374	.39873	.4461	1
	.07	.3	1.51816	2.06325	1.01377	2.07522	.51816	.080627	.11253	.075218	.27599	.26566	.3
	.014	.3	1.44731	2.06145	1.36434	1.51608	.44731	.061449	.43686	.93486	.016896	.0084154	0
	.01	.3	1.47191	1.99966	1.21005	2.061	.47191	.03943	.33178	.28079	.014333	.0093178	0
	.008	.3	1.49279	2.04646	1.22311	2.07468	.49279	.047629	.93873	1.03514	.0068119	.0021077	0
Average			**1.47921**	**2.046056**	**1.145598**	**1.941066**	**.47921**	**.0577182**	**.398588**	**.4859496**	**.14255218**	**.14632018**	**.26**
	.1	.7	2.09308	2.11633	.73248	2.00593	1.09308	.11633	.33657	.14127	.064108	.067393	1
	.07	.7	2.0861	2.11432	.93233	2.02962	1.0861	.11432	.19671	.40117	.042013	.047401	1
	.014	.7	2.08995	2.11537	1.46361	1.28488	1.08995	.11537	.61093	1.07874	.0019344	.0011909	.1
	.01	.7	2.06464	2.11077	1.82733	2.03075	1.06464	.11077	1.00102	.43842	.0019756	.0071541	0
	.008	.7	2.07788	2.10646	2.42251	2.18857	1.07788	.10646	1.42251	.36584	.0015688	.0036815	0
Average			**2.08233**	**2.11265**	**1.475652**	**1.90795**	**1.08233**	**.11265**	**.713548**	**.485088**	**.02231996**	**.0253641**	**.42**
	.1	.9	1.98408	2.10926	1.39899	2.03503	.98408	.10926	.61192	.43277	.009436	.0051607	1
	.07	.9	1.98669	2.1105	1.37524	2.3599	.98669	.1105	.64648	1.14158	.0062251	.0051371	1
	.014	.9	1.98424	2.1122	1.87034	2.30635	.98424	.1122	1.01286	.58626	.00041689	.00055417	0
	.01	.9	1.98793	2.10677	1.97493	2.05395	.98793	.10677	.97493	.3751	.0028262	.00088341	0
	.008	.9	1.98618	2.10974	2.02086	2.21645	.98618	.10974	1.02086	.25188	.0027157	.0020875	0
Average			**1.985824**	**2.109694**	**1.728072**	**2.194336**	**.985824**	**.109694**	**.85341**	**.557518**	**.004323978**	**.002764576**	**.4**

Block 1 ($\gamma = 0.3$)

In block 1 the relevance control parameter δ decreases, whereas the endogeneity control parameter γ is held constant at 0.3 moving down the block. Because the relevance control parameter δ does not influence the OLS estimator, these changes are not expected to directly affect the OLS estimates. Alternatively, the IV estimator will depend on the relevance control parameter δ.

The median of the OLS estimator of β_1 is overall exhibiting an increasing pattern, except when the relevance control parameter δ decreases from 0.07 to 0.014. The median of the IV estimator of β_1, overall, exhibits increasing pattern, except when the relevance control parameter δ decreases from 0.014 to 0.01. The median of the OLS estimator of β_2 displays an increasing pattern in the first and in the last rows of block 1 (δ equals to 0.1 and 0.008). In the second and in the third rows it displays decreasing pattern (δ equals to 0.07 and 0.014). The median of the IV estimator of β_2, overall, exhibits increasing pattern, except when the relevance control parameter δ decreases from 0.07 to 0.014. Given the random patterns of the medians of the OLS estimator of β_1 and β_2 as well as the medians of the IV estimator of β_1 and β_2, the conclusion can be drawn that there is not a monotonic relationship between the median estimates from both estimators and the relevance control parameter δ for the given low levels of correlations. This implies that at these low levels of correlations there is a lot of uncertainty involved, which makes it difficult to see a clear-cut pattern as to which way the median estimates of both estimators vary given the continuous decline in the relevance control parameter δ.

Changes in median absolute biases of both OLS estimators exhibit a very subtle pattern of slight increases in the first and in the last rows of block 1 when the relevance

control parameter δ equals to 0.1 and 0.008, respectively. When the relevance control parameter δ is decreasing the instruments are getting weaker and both median absolute biases of the IV estimators respond in a somewhat similar manner: they both decrease, then increase, and then decrease and increase again. The results of the median absolute biases of both estimators imply that at these low levels of correlations there is not a monotonic relationship between the median absolute biases from the OLS and the IV estimators and the relevance control parameter δ. Also, it should be pointed out that there is a high level of uncertainty involved due to the low levels of correlations.

As for the partial correlation between X_1 and Z_1 (R^2p_1), overall, it exhibits a decreasing pattern when the value of δ is decreasing, which is expected. The situation is a little bit different for the partial correlation between X_2 and Z_2 (R^2p_2). Although the partial correlation does exhibit overall decreasing pattern, as expected, however, it slightly goes up by 0.0009024 when the relevance control parameter δ decreases from 0.014 to 0.01.

The results of the percentage of the times the Hausman test rejects the null hypothesis are consistent with prior reasoning. As the relevance control parameter δ decreases, so does the percentage of the times the Hausman test rejects the null hypothesis, which starts out with the value of 1, and gradually decreases becoming 0 as we read down block 1 of table 3.1.

Block 1 - Averages

A summary of block 1 is considered by averaging over the rows. Specifically, in block 1 of table 3.1 the OLS estimator of β_1 is overestimated, on average, by 0.47921.

Alternatively, the IV estimator of β_1 does a slightly better job, because it is only overestimated, on average, by 0.145598. The situation changes for the OLS estimator of β_2 and the IV estimator of β_2. The OLS estimator of β_2 is overestimated again, but only, on average, by 0.046056, whereas the IV estimator of β_2 is underestimated, on average, by 0.058934, which implies that in this particular case OLS does a better job than IV.

The average of the median absolute bias of the OLS estimator of β_1 is 0.47921, whereas the average of the median absolute bias of the IV estimator of β_1 is 0.398588, implying that IV is doing a slightly better job compared to OLS. However, the average of the median absolute bias of the OLS estimator of β_2 is 0.0577182, which is less than the average of the median absolute bias of the IV estimator of β_2 of 0.4859496, meaning that in this case the OLS outperforms the IV. All the abovementioned results are completely expected because X_1 is more endogenous relative to X_2.

In block 1 the average value of the in R^2p_1 is 0.14255218. The average R^2p_2 is equal to 0.14632018 which is just by 0.003768 greater than the average value of R^2p1. Also, in block 1 the average of the percentage of the times the Hausman test rejects the null hypothesis is equal to 0.26.

Block 2 ($\gamma = 0.7$)

In block 2, the level of endogeneity is increased by increasing the endogeneity control parameter γ from 0.3 to 0.7. Due to the setup the relevance control parameter δ does not influence the OLS estimator. Thus, the changes in the value of the relevance control parameter δ are not expected to directly affect the OLS estimates.

The median of the OLS estimator of β_1 displays the following pattern moving down block 2: it decreases, then increases, and then decreases and increases again. The median of the IV estimator of β_1 exhibits increasing pattern reading down block 2. Overall, the median of the OLS estimator of β_2 exhibits decreasing pattern, except when the relevance control parameter decreases from 0.07 to 0.014. The median of the IV estimator of β_2 exhibits increasing pattern, except when the relevance control parameter declines from 0.07 to 0.014. All this points out to the fact that at these low levels of correlations there is not a monotonic relationship between the median estimates from both estimators and the relevance control parameter δ, and also, there is a high level of uncertainty involved due to the low levels of correlations.

Changes in the median absolute bias of the OLS estimator of β_1 exhibit a very subtle pattern of slight decrease and increase, and then decrease and increase again. The median absolute bias of the IV estimator of β_1 exhibits overall increasing pattern. The changes in the median absolute bias of the OLS estimator of β_2 display decreasing pattern moving down block 2, except for the case, when the relevance control parameter δ declines from 0.07 to 0.014. Finally, the median absolute bias of the IV estimator of β_2 increases to the point, where the relevance control parameter δ equals to 0.014, and decreasing thereafter. It is clear that at these low levels of correlations no monotonic relationship exists between the median absolute biases of the estimates from both estimators and the relevance control parameter δ, and due to these low levels of correlations there is a high level of uncertainty involved.

Both partial correlations exhibit the same decreasing pattern which is expected: they keep decreasing down to the point where the relevance control parameter δ declines

from 0.014 to 0.01, after which they slightly go up, but then again go down again in the last row of block 2. The percentage of the times the Hausman test rejects the null hypothesis displays declining pattern, which is expected as the relevance control parameter δ decreases.

Block 2 - Averages

When summarizing block 2 in terms of average values, it should be noted that almost in all cases the OLS estimator of β_1 is overestimated, on average, by 1.08233. Yet the IV estimator of β_1 is only overestimated, on average, by 0.475652, doing a slightly better job than OLS. When comparing the OLS estimator of β_2 and IV estimator of β_2, it becomes clear that in this case IV estimator is doing better job than OLS estimator again. The OLS estimator of β_2 is overestimated, on average, by 0.11265. The IV estimator of β_2 is underestimated, on average, by 0.09205.

The average of the median absolute bias of the OLS estimator of β_1 is 1.08233, and the average of the median absolute bias of the IV estimator of β_1 is 0.713548, meaning that IV does a better job as opposed to OLS. However, the average of the median absolute bias of the OLS estimator of β_2 is 0.11265, and the average of the median absolute bias of the IV estimator of β_2 is 0.485088, meaning that OLS does a better job than IV. These results are in agreement with the fact that X_1 is more endogenous than X_2. In block 2 the average value for the R^2p_1 is 0.02231996 and the average value for the R^2p_2 is 0.0253641. The average of the percentage of the times the Hausman test rejects the null hypothesis is equal to 0.42.

Block 3 ($\gamma = 0.9$)

In block 3, the level of endogeneity is increased by increasing the endogeneity control parameter γ from 0.7 to 0.9. As a result, the median of the OLS estimator of β_1 exhibits the following pattern: it increases, then decreases, and then increases and decreases again. The median of the IV estimator of β_1 exhibits increasing pattern, except when the relevance control parameter δ declines from 0.1 to 0.07. The median of the OLS estimator of β_2, overall, exhibits increasing pattern, except when the relevance control parameter δ declines from 0.014 to 0.01. The median of the IV estimator of β_2 increases at the beginning and at the end of block 3, and decreases in the middle of block 3. These random patterns of the median estimates of both estimators indicate that there is not a monotonic relationship between the median estimates from both estimators and the relevance control parameter δ at these low levels of correlations. These random patterns of the median estimates of both estimators also indicate that there is a high level of uncertainty involved due to the low levels of correlations.

Changes in the median absolute bias of the OLS estimator of β_1 exhibit pattern of slight increase and decrease, and then increase and decrease again. The median absolute bias of the IV estimator of β_1, overall, exhibits increasing pattern, except when the relevance control parameter δ declines from 0.014 to 0.01. The changes in the median absolute bias of the OLS estimator of β_2 display increasing pattern moving down block 3, except for the case, when the relevance control parameter δ declines from 0.014 to 0.01. The median absolute bias of the IV estimator of β_2 increases to the point, where the relevance control parameter δ equals to 0.07, and decreases thereafter. Once again, it should be pointed out that there is not a monotonic relationship between the median

absolute biases of the estimates from both estimators and the relevance control parameter δ at these low levels of correlations, and that there is a high level of uncertainty involved due to the low levels of correlations.

The partial correlation of R^2p_1 overall displays a decreasing pattern, except when the relevance control parameter δ decreases from 0.014 to 0.01. The partial correlation of R^2p_2 exhibits slight pattern of decrease at the beginning of block 3, and slight pattern of increase at the end of the block. As the relevance control parameter δ declines, so does the percentage of times the Hausman test rejects the null hypothesis. It starts out with 1, and then decreases becoming 0 moving down block 3 of table 3.1.

Block 3 - Averages

When summarizing block 3 using average values, it should be noted that in block 3 the OLS estimator of β_1 is overestimated, on average, by 0.985824, whereas, the IV estimator of β_1 does a slightly better job, because it is only overestimated, on average, by 0.728072. The situation changes when comparing the OLS estimator of β_2 to the IV estimator of β_2. The OLS estimator of β_2 is overestimated again, but only, on average, by 0.109694, whereas the IV estimator of β_2 is overestimated, on average, by 0.194336, meaning that in this case OLS does a slightly better job than IV.

When summarizing block 3 in terms of median absolute biases we see that the average of the median absolute bias of the OLS estimator of β_1 is 0.985824, whereas the average of the median absolute bias of the IV estimator of β_1 is 0.85341, meaning that IV is doing a slightly better job compared to OLS. The average of the median absolute bias of the OLS estimator of β_2 is 0.109694, which is relatively less than the average of the

median absolute bias of the IV estimator of β_2 of 0.557518, implying that in this case OLS does a better job than IV. These results for the median estimates and the median absolute biases are in accordance with the fact that the X_1 is more endogenous than X_2.

In block 3 of table 3.1 the average value of the in R^2p_1 is 0.004323978, and the average of R^2p_2 equals to 0.002764576. Also, in block 3 the average of the percentage of the times the Hausman test rejects the null hypothesis is equal to 0.4.

Across the Blocks Comparison

When conducting analysis across the blocks in table 3.1, it should be noted that the only parameter that is changing is the endogeneity control parameter γ, which keeps increasing moving down the blocks. Because the endogeneity control parameter γ governs the correlation between X's and the disturbance term, increase in the endogeneity control parameter γ raises the endogeneity of both X_1 and X_2, although the level of endogeneity for X_1 increases more than the level of endogeneity of X_2.

When comparing the OLS estimator of β_1 with the IV estimator of β_1 it becomes obvious that the IV estimator does a better job than the OLS, as expected. Although both median estimates keep increasing across the blocks, however, the OLS estimator of β_1 goes up by greater margin than the IV estimator of β_1. The opposite is true for the OLS estimator of β_2 and the IV estimator of β_2, indicating that the OLS outperforms the IV in terms of median parameter estimation, as expected, because the level of endogeneity associated with X_1 is relatively higher than the level of endogeneity associated with X_2. Overall, there is not a monotonic relationship between the OLS estimator of β_1, the OLS estimator of β_2, the IV estimator of β_2 and the endogeneity control parameter γ. This

implies that there is high level of risk involved these low levels of correlations. However, a monotonic relationship can be observed between the IV estimator of β_1 and the endogeneity control parameter γ.

When comparing the median absolute bias of the OLS estimator of β_1 with the median absolute bias of the IV estimator of β_1 it is clear that the IV performs better than the OLS, as expected. Although, the median absolute biases of both estimators go up moving down the blocks the amount of increase in the median absolute bias of the OLS estimator is slightly bigger than that of the median absolute bias of the IV estimator of β_1. However, the OLS estimator does a better job relative to the IV estimator in terms of the median absolute bias of β_2, as expected, because the median absolute bias of the OLS estimator of β_2 is greater than the median absolute bias of the IV estimator of β_2 for every level of the endogeneity control parameter γ. It needs to be pointed out that there is a non-monotonic relationship between the median absolute bias of the OLS estimator of β_1, the median absolute bias of the OLS estimator of β_2, the median absolute bias of the IV estimator of β_2 and the endogeneity control parameter γ, meaning that there is a high level of uncertainty involved because of the low levels of correlations. However, there is a monotonic relationship between the median absolute bias of the IV estimator of β_1 and the endogeneity control parameter γ.

The averages of both partial correlations decline moving down the blocks, which is expected, because the increasing endogeneity control parameter γ indirectly raises the endogeneity level of the instruments through their correlation with the regressors making them less relevant. Also, the average partial correlation associated with X_2 is slightly bigger as opposed to the average partial correlation associated with X_1 for every level of

the endogeneity control parameter γ, except for the case when the endogeneity control parameter γ equals to 0.9. The relatively low level of endogeneity of X_2 is the reason why the average R^2p_2 is slightly bigger than R^2p_1.

Even though the average of the percentage of times the Hausman test rejects the null hypothesis, overall, increases moving down the blocks, which is expected, due to the fact that the regressors become more and more endogenous, however, it slightly goes down by 0.02, moving from block 2 to block 3.

Summary of Experiment 1 (200 Observations per Row)

In summary, it should be pointed out that all the results presented in table 3.1 are consistent with the expectations discussed earlier. Particularly, the decreases in the relevance control parameter δ led to biased IV estimation (relevance problem), and the same result was observed when there were increases in the endogeneity control parameter γ (endogeneity problem). What is somewhat surprising is that there does not seem to be a monotonic relationship between the median parameter estimates from both estimators and the relevance control parameter δ. A non-monotonic relationship was also observed between the median absolute biases of the parameter estimates from both estimators and the relevance control parameter δ. In both cases a non-monotonic relationship suggests that there is a high level of uncertainty involved due to the low levels of correlations.

Also, a non-monotonic relationship is observed between the median parameter estimates from both estimators (except for the IV estimator of β_1) and the endogeneity control parameter γ, as well as between the median absolute biases of the parameter estimates from both estimators (except for the IV estimator of β_1) and the endogeneity

control parameter γ, implying that there is a high level of uncertainty involved because of the low levels of correlations.

The OLS estimator seems to be doing a better job in contrast to the IV estimator in terms of the parameter estimation and in terms of bias with respect to the independent variable of X_2. The explanation for that is the relatively low level of endogeneity of X_2, in which cases the OLS estimator is expected to perform better. However, the IV estimator outperforms the OLS estimator in terms of the parameter estimation and bias with respect to the independent variable of X_1, which is explained by the relatively high level of endogeneity associated with that variable.

The decreasing pattern of both partial correlations within each block, and across the block is consistent with the expectations. Also, it is noteworthy that the average value of the partial correlation between X_1 and Z_1 is lower compared to the average value of the partial correlation between X_2 and Z_2 when the endogeneity control parameter γ equals to 0.3 and 0.7. However, the average value of the partial correlation between X_1 and Z_1 is greater compared to the average value of the partial correlation between X_2 and Z_2 when the endogeneity control parameter γ equals to 0.9.

The declining pattern of the percentage of the times the Hausman test rejects the null hypothesis within each block is in accordance with the expectations. The decrease in the percentage of the times the Hausman test rejects the null hypothesis within each block is explained by the decrease in the value of the relevance control parameter δ. However, the percentage of the times the Hausman test rejects the null hypotheses, overall, exhibits an increasing pattern reading down the blocks. The amount of the increase in moving from block 1 to block 2 is equal to 0.16. The percentage of the times the Hausman test

rejects the null hypothesis then declines slightly (0.02) in moving from block 2 to block 3. This pattern may be explained by two factors. First, the value of the endogeneity control parameter γ increases by 0.4 when moving from block 1 to block 2, and by 0.2 when moving from block 2 to block 3. Second, these results again underscore the non-monotonic relationship between the endogeneity control parameter γ and the Hausman test statistic.

Experiment 2 (2000 Observations per Row)

The overall purpose of analyzing the data in table 3.1 was to get insights for Monte-Carlo experiment. Even though results in table 3.1 are useful, however, they are not representative of a real-world estimation problem, because they involved 10 samples for each true parameter setting. Yet, in real-world estimation problem there is only one sample drawn from dataset implicitly employing only one combination of both the relevance control parameter δ and the endogeneity control parameter γ, as opposed to 10 samples that were drawn when estimating results in table 3.1. Given the fact that the results in table 3.1 are not representative of a real-world estimation problem, data should be generated that are comparable to the real-world estimation problem. For this purpose, all of the 10 trials of 200 observations per row are going to be pooled together forming one trial of 2000 observations for each known combination of control parameters. As a result, 15 models are going to be estimated, one model for each combination of the relevance control parameter δ and the endogeneity control parameter γ. Point estimates are going to be reported, as opposed to before where 150 (15x10) models were estimated and median of the estimates were reported. Also, an important objective of this analysis is

to obtain and interpret the results that are going to be compared with the results obtained from using the method of the directed graphs using the same data. The results obtained from estimating 15 models of 2000 observations each per row are presented in table 3.2.

For each trial the following are reported in table 3.2: the correlation between X_1 and Z_1, the correlation between X_2 and Z_2, the values of the estimates of the OLS and the IV estimators, the absolute values of the biases of the parameter estimates, partial correlations between X's and Z's, and the Hausman P values (probability of making Type I error, where the null hypotheses is rejected given it is true).

Overall, the results are expected to be similar to the patterns in table 3.1 for the same reasons. Because the relevance control parameter δ is the only parameter that affects the correlation between X's and Z's, the correlations between X's and Z's are expected to go down as the relevance control parameter δ decreases and the endogeneity control parameter γ is held constant. Also, when the relevance control parameter δ decreases and the endogeneity control parameter γ is held constant the OLS parameter estimator is expected to perform better than IV parameter estimator in terms of parameter estimation and in terms of bias. A monotonic relationship is expected between the estimates from both estimators and the relevance control parameter δ, as well as between the absolute biases of the parameter estimates from both estimators and the relevance control parameter δ. Partial correlations are expected to go down as the relevance control parameter δ keeps decreasing and the endogeneity control parameter γ is held constant. Low partial correlation has its implications in terms of the Hausman test. According to Park and Davis, when the partial correlations are low, (weak instruments) the Hausman test is not very reliable (Park and Davis, 2001, p. 846). As a result, when the relevance

control parameter δ decreases and the endogeneity control parameter γ is held constant we would expect the Hausman P value to increase, implying that the probability of making Type I error, where the null hypotheses of the Hausman test is that the OLS and IV estimates do not differ, will go up.

As in table 3.1, the endogeneity control parameter γ governs the correlations between X's and the disturbance term. As such, the level of the endogeneity of X_1 will increase by more than the level of the endogeneity of X_2 due to the increase in the value of the endogeneity control parameter γ, because of the coefficient of λ, which equals 0.9. Given this increase in the levels of endogeneity of both X_1 and X_2, it is expected that the OLS estimator would outperform the IV estimator in terms of parameter estimation and bias for X_2 for each level of the endogeneity control parameter γ. However, the opposite is true for the regressor of X_1. Also, it should be pointed out that a monotonic relationship is expected between the parameter estimates from both estimators and the endogeneity control parameter γ, as well as between the absolute biases of the parameter estimates from both estimators and the endogeneity control parameter γ. A declining pattern for both partial correlations is expected when the relevance control parameter δ is held constant and the endogeneity control parameter γ increases. Given the increase in the level of endogeneity by increasing γ, when the relevance control parameter δ is held constant, we would expect the Hausman P value to go down, meaning that the probability of making Type I error (reject the null hypotheses given it is true) will go down.

As in table 3.1 the discussion of table 3.2 presented below is broken into three blocks: block 1, block 2, and block 3. The value of the endogeneity control parameter γ is

held constant at a certain level within each block and the value of the relevance control parameter δ is decreasing reading down the block.

Block 1 ($\gamma = 0.3$)

In block 1 of table 3.2 the relevance control parameter δ decreases, whereas the endogeneity control parameter γ is held constant at 0.3 reading down the block. The correlation between X_1 and Z_1, as well as the correlation between X_2 and Z_2, is exhibiting a decreasing pattern, which is expected every time the parameter that governs those correlations is exhibits decreasing pattern itself.

Due to the design in the setup, the changes in the relevance control parameter δ are expected to directly affect IV estimates. However, the OLS estimates are not going to be directly influenced by the changes in the relevance control parameter δ. Thus, the OLS estimator of β_1 is, overall, displaying an increasing pattern, except when the relevance control parameter δ decreases from 0.07 to 0.014, where it slightly decreases. The IV estimator of β_1 is overall displaying increasing pattern, except when the relevance control parameter δ goes down from 0.07 to 0.014. The OLS estimator of β_2 is exhibiting the pattern of decrease and increase, and then decrease and increase again. The IV estimator of β_2 is exhibiting the pattern of increase and decrease, and then increase and decrease again. As a result, we can say that there is not a monotonic relationship between the estimates from both estimators and the relevance control parameter δ at these low levels of correlations. This is very important information, which implies that there is a high level of uncertainty involved because of the low levels of correlations.

Table 3.2. Results from Monte-Carlo Experiment with the Number of Observations 2000 per Row

	δ	γ	Corr X_1-Z_1	Corr X_2-Z_2	$\beta_1$ols	$\beta_2$ols	$\beta_1$iv	$\beta_2$iv	Abs bias $\beta_1$ols	Abs bias $\beta_2$ols	Abs bias $\beta_1$iv	Abs bias $\beta_2$iv	R^2p_1	R^2p_2	Hausman Pval
	.1	.3	.63592	.66161	.47458	2.05219	.913899	1.99727	.47458	.05219	.086101	.00273	.4046	.43791	0
	.07	.3	.53747	.51279	.50337	2.04081	1.05944	2.10242	.50337	.04081	.05944	.10242	.30307	.24588	0
	.014	.3	.15165	.1026	.41652	2.07897	.976064	1.39132	.41652	.07897	.023936	.60868	.022679	.0099051	.00248
	.01	.3	.12252	.097641	.49163	2.029	1.09329	1.99688	.49163	.029	.09329	.00312	.014614	.0079143	.12476
	.008	.3	.061823	.034533	.49806	2.0436	1.2911	1.39662	.49806	.0436	.2911	.60338	.0059923	.0016953	.43555
Average			**.3018766**	**.2818348**	**1.476832**	**2.048914**	**1.0667586**	**1.776902**	**.476832**	**.048914**	**.1107734**	**.264066**	**.15019106**	**.14066094**	**.112558**
	.1	.7	.25332	.28457	2.09392	2.12638	.785543	1.87916	1.09392	.12638	.214457	.12084	.061464	.076988	0
	.07	.7	.21742	.24025	2.08922	2.10856	.885987	1.90476	1.08922	.10856	.114013	.09524	.046993	.056793	0
	.014	.7	.074489	.031578	2.087	2.12452	1.57834	1.26266	1.087	.12452	.57834	.73734	.0021056	.00090531	0
	.01	.7	.022688	.005562	2.07521	2.11388	1.3922	3.00886	1.07521	.11388	.3922	1.00886	7.39160D-06	.000013282	.00037
	.008	.7	.00048031	.021648	2.07273	2.11558	2.99307	3.95046	1.07273	.11558	1.99307	1.95046	0.00067127	.00024717	.02265
Average			**.113679462**	**.1167216**	**2.083616**	**2.117784**	**1.527028**	**2.40118**	**1.083616**	**.117784**	**.658416**	**.782548**	**.027808468**	**.026989352**	**.004604**
	.1	.9	.099845	.087231	1.98375	2.1087	1.25172	1.90139	.98375	.1087	.25172	.09861	.0096372	.0086279	0
	.07	.9	.068958	.073823	1.98644	2.10904	.984148	2.0894	.98644	.10904	.015852	.0894	.0048015	.0055805	0
	.014	.9	.036921	.028783	1.98552	2.11275	1.2472	1.57878	.98552	.11275	.2472	.42122	.0004059	.00024432	0
	.01	.9	.0047634	.031371	1.98929	2.10851	-.450693	0.032327	.98929	.10851	1.450693	1.967673	.000033216	.000048135	0
	.008	.9	-.013369	.0058544	1.98634	2.11211	2.7906	1.66916	.98634	.11211	1.7906	.33084	.00036557	.000011704	0
Average			**.03942368**	**.04541248**	**1.986268**	**2.110222**	**1.164595**	**1.4542114**	**.986268**	**.110222**	**.751213**	**.5815486**	**.003048677**	**.002902512**	**0**

The absolute bias of the OLS estimator of β_1 shows no real pattern. The absolute bias of the IV estimator of β_1 decreases at the beginning of the block, but then goes up slightly at the end of the block. The absolute bias of the OLS estimator of β_2 shows no real pattern as well moving down block 1. The absolute bias of the IV estimator of β_2 shows no real pattern either. Overall, there is a non-monotonic relationship between the absolute biases of the estimates of both estimators and the relevance control parameter δ at these low levels of correlations. Thus, there is a high level of uncertainty involved due to the low levels of correlations.

Both partial correlations, R^2p_1 and R^2p_2, exhibit a decreasing pattern reading down the block, as expected. Finally, the Hausman P value clearly exhibits increasing pattern reading down block 1, which is consistent with expectations.

Block 1 - Averages

Block 1 is summarized using absolute values. The average correlation between X_1 and Z_1 is 0.3018766 and the average correlation between X_2 and Z_2 is 0.2818348.

As for OLS estimator, β_1 is overestimated, on average, by 0.476832. Alternatively, the IV estimator for β_1 does much better job compared to OLS, because, on average, it is only overestimated by 0.0667586. However, when we compare the OLS and IV estimators for β_2, we see that the OLS estimate is, on average, overestimated by 0.048914, while the IV estimate is underestimated, on average, by 0.223098.

In block 1 the absolute bias of OLS parameter estimate of β_1 has an average of 0.476832. The absolute bias of IV parameter estimate of β_1 has an average absolute bias of 0.1107734. The absolute bias of OLS parameter estimate of β_2 has an average of

0.048914. The average of the absolute bias of IV parameter estimate of β_2 is 0.264066. It should be noted that there is no uniform result in terms of the performance of the IV estimator as the relevance control parameter δ changes, which is somewhat expected. This just highlights the relevance problem associated with these parameter settings. The above discussions on parameter estimates and absolute biases are in agreement with the fact that X_1 is more endogenous than X_2.

The results for partial correlations are totally expected. As the relevance control parameter δ decreases, so do the partial correlations between X's and Z's. The average partial correlation of X_1 and Z_1 is 0.15019106, and the average partial correlation of X_2 and Z_2 is 0.14066094.

The Hausman P values are expected to be inversely related to the relevance control parameter δ. This is exactly what happens, as the relevance control parameter δ decreases the probability of rejecting the null hypothesis increases. The average of the Hausman P value is 0.112558.

Block 2 ($\gamma = 0.7$)

In block 2 of table 3.2, the level of endogeneity is increased by increasing the endogeneity control parameter γ from 0.3 to 0.7. Even though the correlation between X_2 and Z_2 slightly increases in the last row of block 2 where the relevance control parameter δ equals to 0.008, overall, both correlations between X's and Z's exhibit decreasing pattern moving down the block.

The OLS estimator of β_1 exhibits no real pattern, as the relevance control parameter δ decreases. The IV estimator of β_1 displays an overall increasing pattern as the

relevance control parameter δ decreases. The OLS estimator of β_2 exhibits no real pattern as well. The IV estimator of β_2 displays no real pattern either. A conclusion can be drawn that there is not a monotonic relationship between the estimates from both estimators and the relevance control parameter δ at these low levels of correlations. One of the implications of the non-monotonic relationship between the estimates of the both estimators and the relevance control parameter δ is that there is a high level of risk is present because of the low levels of correlations.

The absolute bias of the OLS estimator of β_1 exhibits decreasing pattern moving down the block. The absolute bias of the IV estimator of β_1 shows no real pattern moving down the block. The absolute bias of the OLS estimator of β_2 also exhibits no real pattern moving down the block. The absolute bias of the IV estimator of β_2 shows no real pattern as well. It should be pointed out that there is not a monotonic relationship between the absolute biases of the estimates from both estimators and the relevance control parameter δ at these low levels of correlations. And, due to these low levels of correlations there is a high level of risk present.

Both partial correlations clearly display decreasing pattern moving down the block, in spite of the fairly insignificant increase in the value of R^2p_2 at the end of the block when the relevance control parameter δ goes down from 0.01 to 0.008. The Hausman P value also displays an increasing pattern moving down the block.

Block 2 - Averages

Average values of all the column headings are used summarizing the block. The average correlation between X_1 and Z_1 is 0.11367946, and the average correlation

between X_2 and Z_2 is 0.1167216. The reason for the decrease in the values of both correlations is the decrease in the value of the relevance control parameter δ.

In the second block we see that the OLS estimator of β_1 is overestimated, on average, by 1.083616, whereas the IV estimator of β_1 is only overestimated, on average, by 0.527028. This implies that IV parameter estimate of β_1, on average, performs better than OLS parameter estimate of β_1. However, when comparing the OLS parameter estimate of β_2 with the IV parameter estimate of β_2, it becomes clear that the OLS estimator does a better job. The OLS parameter estimate of β_2 is overestimated, on average, by 0.117784, yet the IV parameter estimate of β_2 is overestimated, on average, by 0.40118.

The average absolute bias of the OLS estimator of β_1 is 1.083616, and the average absolute bias of the IV estimator of β_1 is 0.658416. This supports the point that the IV estimator does a better job compared to the OLS estimator in terms of the absolute bias of the parameter estimate of β_1. The average absolute bias of the OLS estimator of β_2 is 0.117784, and the average absolute bias of the IV estimator of β_2 is the 0.782548. This implies that the OLS estimator performs better than the IV estimator in terms of the absolute bias of the parameter estimate of β_2. It should be pointed out that these discussions of parameter estimates and biases are in accordance with the expectations given the fact that X_1 is more endogenous than X_2.

The average R^2p_1 is equal to 0.027808468, and the average R^2p_2 is equal to 0.026989352. The average of the Hausman P value equals to 0.004604.

Block 3 ($\gamma = 0.9$)

In block 3 the level of the endogeneity control parameter γ is raised from 0.7 to 0.9. Both correlations between X's and Z's exhibit a declining pattern reading down the block, as expected. Even though the correlation between X_2 and Z_2 slightly increases when the relevance control parameter δ declines from 0.014 to 0.01, but the amount of the increase is insignificant and it does not distort an overall decreasing pattern of the correlation between X_2 and Z_2.

The OLS and IV estimators of β_1 both display no real pattern moving down the block. The OLS and IV estimators of β_2 also display no real pattern moving down the block. As such, there is not a monotonic relationship between the parameter estimates from both estimators and the relevance control parameter δ at these low levels of correlations. As a result there is high level of uncertainty involved, because of the low levels of correlations.

The absolute bias of the OLS estimator of β_1 displays no real pattern moving down the block. The absolute bias of the IV estimator of β_1 shows an overall increasing pattern moving down the block. The absolute bias of the OLS and IV estimators of β_2 shows no real pattern moving down the block. Again, there is not a monotonic relationship between the absolute biases of the estimates from both estimators and the relevance control parameter δ at these low levels of correlations. The direct implication of the non-monotonic relationship is that there is a high level of uncertainty involved, because of the low levels of correlations.

Both partial correlations clearly display decreasing pattern reading down block 3, in spite of the insignificant increase in the value of R^2p_1 at the end of the block when the

relevance control parameter δ goes down from 0.01 to 0.008. The Hausman P value is equal to 0 moving down block 3.

Block 3 - Averages

The absolute average correlation for X_1 and Z_1 is 0.04477128, and the average correlation between X_2 and Z_2 is 0.0454125. In terms of the parameters estimate of β_1, the IV estimator is doing relatively better job than the OLS estimator. Specifically, the OLS estimator of β_1 is overestimated, on average, by 0.986268, whereas the absolute average number of overestimation of IV estimator of β_1 is 0.3448722. In case of the parameter estimate of β_2 the OLS estimator is doing relatively better job than the IV estimator. The OLS estimator of β_2 is overestimated, on average, by 0.110222, whereas the IV estimator of β_2 is underestimated, on average, by 0.5457886.

The average absolute bias of the OLS estimator of β_1 is 0.986268, and the average absolute bias of the IV estimator of β_1 is 0.751213. This comparison implies the IV estimator does a better job compared to the OLS estimator in terms of the absolute bias of the parameter estimate of β_1. The average absolute bias of the OLS estimator of β_2 is 0.110222, and the average absolute bias of the IV estimator of β_2 is the 0.5815486. This means that the OLS estimator performs better than the IV estimator in terms of the absolute bias of the parameter estimate of β_2.

The average R^2p_1 is equal to 0.003048677 and the average R^2p_2 is equal to 0.002902512. The average P value of the Hausman test is equal to 0.

Across the Blocks Comparison

When summarizing the results in table 3.2 across the blocks, it should be pointed out that the only parameter that is changing is the endogeneity control parameter γ, which keeps increasing reading down the blocks. The correlations between X's and Z's exhibit the expected decreasing pattern reading down the blocks.

Comparing the OLS estimator of β_1 with the IV estimator of β_1 it should be pointed out that the IV estimator does a better job than the OLS, as expected. Although both parameter estimates display the same pattern of increase, and then decrease reading across the blocks, however, the OLS estimator of β_1 goes up by greater margin than the IV estimator of β_1 moving from bock 1 to block 2. Also, the OLS estimator of β_1 goes down by less than the IV estimator of β_1 moving from block 2 to block 3. However, the OLS outperforms the IV in terms of parameter estimation associated with X_2, as expected. Even though both estimates show the same pattern of increase when moving from block 1 to block 2, and then decrease when moving from block 2 to block 3, however, the OLS estimator of β_2 changes by relatively less than the IV estimator of β_2. What seems to be somewhat surprising is that there does not seem to be a monotonic relationship between the parameter estimates from both estimators and the endogeneity control parameter γ. This implies that there is a high level of uncertainty present due to the given low levels of correlations.

The same results are observed when summarizing the results in table 3.2 across the blocks in terms of absolute biases. Comparing the absolute bias of the OLS estimator of β_1 with the absolute bias of the IV estimator of β_1 it becomes clear that the IV estimator performs better than the OLS estimator, as expected, because the absolute bias

of the OLS estimator is bigger compared to the absolute bias of the IV estimator of β_1 for every level of the endogeneity control parameter γ. Yet, the OLS estimator does a better job relative to the IV estimator in terms of the absolute bias of β_2, as expected. That is true because the absolute bias of the OLS estimator of β_2 is greater than the absolute bias of the IV estimator of β_2 for every level of the endogeneity control parameter γ. A non-monotonic relationship can be observed between the absolute bias of the OLS estimator of β_1, the OLS estimator of β_2, the IV estimator of β_2 and the endogeneity control parameter γ, except in the case of the IV estimator of β_1. A non-monotonic relationship means that there is a high level of uncertainty involved due to the low levels of correlations.

The averages of both partial correlations decrease reading across the blocks, which are expected, because the increasing endogeneity control parameter γ indirectly raises the endogeneity level of the instruments, because of their correlation with the regressors, making them less relevant.

The average of the P value of the Hausman test exhibits the expected pattern by declining across the blocks in table 3.2. This is expected, because when there is an increase in the level of the endogeneity in the regressors the Hausman test should communicate the idea of using the instruments. And this is reflected in the decreasing pattern of the Hausman P value, which implies that the probability of making Type I error goes down.

Summary of Experiment 2 (2000 Observations per Row)

In summarizing table 3.2, it should be noted that all the results are consistent with the expectations. As expected, the values of the correlations between X's and Z's displayed declining pattern moving down the blocks.

The OLS tends to outperform the IV in terms of the parameter estimation and in terms of bias with respect to X_2. The relatively low level of endogeneity associated with X2 can explain this. Alternatively, the IV seems to be doing a better job in terms of parameter estimation and in terms of bias with respect to the X_1. The reason for that is the relatively high level of endogeneity of X_1 in which case IV tends to perform better than the OLS. Another important observation is that there is not a monotonic relationship between the parameter estimates and the relevance control parameter δ, as well as between the biases of the parameter estimates and the relevance control parameter δ. This means that there is a high level of risk present due to the low levels of correlations. Overall, there does not seem to be a monotonic relationship between the parameter estimates and the increasing endogeneity control parameter γ, as well as between the biases of the parameter estimates and the increasing endogeneity control parameter γ, indicating that there is a high level of risk present due to the low levels of correlations.

With respect to the partial correlations, it needs to be pointed out that both of them exhibit expected decreasing pattern both within each block and across the blocks. Also, it is noteworthy that the average value of the partial correlation between X_1 and Z_1 is greater compared to the average value of the partial correlation between X_2 and Z_2 when the endogeneity control parameter γ equals to 0.3, 0.7, and 0.9.

With respect to the P value of the Hausman test, it needs to be pointed out that the increasing pattern of the P value of the Hausman test within each block is caused by the relevance problem. However, the declining pattern of the P value of the Hausman test across the blocks is explained by the endogeneity problem.

Summary of OLS vs. IV

In general, we see that our expectations about the results in table 3.1 and table 3.2 being the same turned out to be plausible. We can clearly see that decreases in the relevance control parameter δ indeed brought up the relevance problem, and increase in the endogeneity control parameter γ brought up the endogeneity problem.

Specifically, in both tables the OLS estimator of β_1 tends to be overestimated by greater margin as opposed to the IV estimator of β_1 for a given level of the relevance control parameter δ and also for a given level of the endogeneity control parameter γ. However, the opposite is true for the β_2 parameter estimate of both estimators. That is, the OLS estimator of β_2 tends to perform better compared to the IV estimator of β_2 for a given level of the relevance control parameter δ and also for a given level of the endogeneity control parameter γ. Perhaps most importantly, a non-monotonic relationship between the parameter estimates from both estimators and the relevance control parameter δ, as well as between the parameter estimates from both estimators and the endogeneity control parameter γ, due to the low levels of the correlations was the case in table 3.1, as well as in table 3.2, pointing out to the fact that there is a high level of risk present due to the low levels of correlations.

Directed Graphs

When conducting the Hausman test it was implicitly assumed that X's were independent variables that cause the dependent variable Y, and that Z's were legitimate instruments for X's, meaning Z's cause X's. However, in a real-world setting, the researcher usually designates a priori what will be the explanatory variables and what will be the instruments. In other words, in the real-world setting the true structural causal relationship between variables is unknown to the researcher. A useful tool would be one that told the researcher the structural causal relationship between the variables. Under certain assumptions, the method of the directed graphs is such a tool. This method is expected to yield legitimate structural causal relationship between the variables. The following discussion on the theory of directed graphs is borrowed from the work of Bessler (*Introduction to Directed Graphs*, 2002) and Pearl (*Causality*, Cambridge University Press, 2000).

Description of the Directed Graphs

Even in the case when the "true" system is unknown to the researcher, the method of the directed graphs can be used to identify causal relationship between variables in a dataset. One needs to obtain the variance-covariance matrix from different variables for the directed graphs to work. The reliability of a directed graph algorithm for determining causality rests on three assumptions:

1. Causal sufficiency;
2. Markov condition;
3. Stability (faithfulness) condition.

According to the *causal sufficiency* condition a casually sufficient set of variables must be included in an observational dataset. However, one needs to make sure that no other variable that might possibly cause the other two variables is missing in the dataset. For example, if variable X causes both Y and Z and X is not considered, then an apparent causal flow from Y to Z (or from Z to Y) may be spurious because both Y and Z are caused by X. In econometric terminology, there can be no omitted variables.

According to the *Markov condition* if Z is caused by Y and Y is caused by X, the underlying joint probability distribution for the three variables can be factored as:

$$\Pr(X,Y,Z) = \Pr(X)\Pr(Y \mid X)\Pr(Z \mid Y) \quad (3.6)$$

To put it in words, the causal flows that we are trying to identify must respect a genealogy condition; a genealogy condition states that only direct causal variables (called parents) should be conditioning variables, if we want to fully capture the variable generating probability distribution. There is no need to condition on indirect variables, or parents of parents (i.e. grandparents).

Pearl gives the following definition about the *stability condition*:

> Let I (P) denote the set of all conditional independence relationships embodied in P (probability distribution). A causal model M = (D, Θ_d) generates a stable distribution if and only if P (D, Θ_d) contains no extraneous independences. The *stability condition* states that, as we vary the parameters from Θ to Θ', no independence in P can be destroyed; hence the name "stability" (Pearl, 2000, p. 48).

A graph contains a set **V** of non-empty vertices (representing variables), a set **M** of non-empty marks, which are symbols at the ends of undirected edges, and a set **E** of edges, or lines, representing a certain relationship. Two vertices (variables) are said to be

adjacent if a line connects them. When we have vertices or variables X and Y the following graphs can be observed:

1. An undirected graph consists only of undirected lines, and is denoted as X—Y. This means that a relationship exists between X and Y, however, the algorithm is unable to identify which way the causal flow goes. In other words, the algorithm can't determine whether X causes Y, or the other way around.
2. A directed graph consists only of directed edges, and is denoted as $X \rightarrow Y$. This means that X causes Y. In other words, the values of Y will be changed, every time we change the values of X.
3. A graph with confounders or unobserved common causes denoted as X ⇢ Y.

Directed graphs allow for directed cycles (e.g., $X \rightarrow Y$, $Y \rightarrow X$), but do not allow for self-loops (e.g., $X \rightarrow X$). A graph that does not include directed cycles is referred to as *acyclic*. Any graph that is both directed and acyclic is called a *directed acyclic graph* (DAG). Only a directed acyclic graph will be used in this thesis. Various relationships in a graph are denoted through the terminology of kinship (e.g., parents, children, descendants, ancestors, spouses). For example, consider the graph in Figure 3.1.

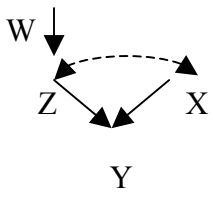

Figure 3.1. A graph including both directed edges and confounders

In the figure above Y has two parents, X and Z, three ancestors, X, Z, and W, and no children. X has no parents (hence no ancestors), one spouse, Z, and one child, Y. Y has one grandparent, W.

Consider the following underlying model:

$Y = \beta_0 + \beta_1 X + U$

$X = \lambda Z + (1- \lambda)U$

where U is disturbance to the equation. In the case when the variables Y, X, Z and U are all observed then the relationship in Figure 3.2 would be expected.

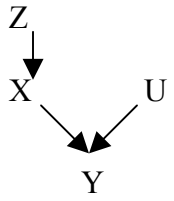

Figure 3.2. A graph showing causal flow between Z, X, Y, and U Y when each variable is observed

However, when U is not observed the directed graph would look as in Figure 3.3.

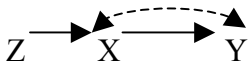

Figure 3.3. A confounding arc embracing causal flow $X \rightarrow Y$ **(U is not observed)**

In Figure 3.2 it is clear that both X and U cause Y. However, in Figure 3.3, when U is not observed we can't be sure which way the casual flow would go. That is, the total effect of X on Y cannot be identified because of a confounding arc (dashed) embracing a causal link between X and Y and represents the existence in the diagram of unobserved variable U. However, in case of a linear model, the total effect of X on Y can become identifiable due to the addition of an arc to a confounding arc pattern. That is the computation of the total effect of X on Y will be possible through instrumental variable formula, if we can find variable Z that is highly correlated with X, but not correlated with U (simply adding an arc $Z \rightarrow X$).

With this brief background on directed graphs, the same data used for estimating the results in table 3.2 are going to be analyzed using the method of the directed graphs. Given the underlying causal structure and assuming that the conditions for the legitimate instruments are satisfied, it is expected that both independent variables, X_1 and X_2, are going to cause Y. Also it is expected, that both instruments, Z_1 and Z_2, are going to cause X_1 and X_2, respectively. In terms of arrows, the following is expected: $Z_1 \rightarrow X_1 \rightarrow Y$, and $Z_2 \rightarrow X_2 \rightarrow Y$.

Analysis Using the Directed Graphs (0.05 Significance Level)

For the analysis of the directed graphs Tetrad II (Scheines, Spirtes, Glymour, and Meek 1994) software was used. Tetrad II employed the variance-covariance matrix from the set of variables, Y, X_1, X_2, Z_1, and Z_2. The results from the directed graphs are presented in two tables: table 3.3 and table 3.4. The results in table 3.3 and in table 3.4 will be presented at the 0.05 and 0.005 significance levels, respectively. Although the

standard significance level used is 0.05, however, in case of a large sample size, it is strongly suggested to decrease the significance level. Because in our case, the sample size is 2000 observations, it provides us with motivation to report results at the two significance levels. Each table is broken into three blocks: block 1, block 2, and block 3. In each block the endogeneity control parameter γ is held constant at a certain level and the relevance control parameter δ is decreasing moving across the block.

Block 1 (0.05 significance level)

In block 1 of table 3.3 the endogeneity control parameter γ is held constant at 0.3 and the relevance control parameter δ is decreasing moving across the block. When the relevance control parameter δ equals to 0.1 and 0.07, both X_1 and X_2 cause Y, Z_1 causes Y, and there are undirected edges from X_1 to Z_1, as well as from X_2 to Z_2, meaning the algorithm is not able to sort out the causal flow between X_1 and Z_1, as well as between X_2
and Z_2. Also, when the relevance control parameter δ equals to 0.07, there is an undirected edge from X_1 to Z_2, meaning the algorithm is unable to sort out causal flow between X_1 and Z_2.

When the relevance control parameter δ equals to 0.014 and 0.01, both X_1 and X_2 cause Y, and there are undirected edges from X_1 to Z_1, as well as from X_2 to Z_2, meaning the algorithm is not able to sort out the causal flow between X_1 and Z_1, as well as between X_2 and Z_2. Also, when the relevance control parameter δ equals to 0.014, Y causes Z_2.

Table 3.3. Results from the Directed Graphs Using Artificially Created Data with the Number of Observations 2000 for Each Combination of δ and γ (0.05 Significance Level)

Block 1				
δ=.1; γ=.3	**δ=.07; γ=.3**	**δ=.014; γ=.3**	**δ=.01; γ=.3**	**δ=.008; γ= .3**
x1 -> y	x1 -> y	x1 -> y	x1 -> y	x1 -> y
x2 -> y	x2 -> y	x2 -> y	x2 -> y	x2 -> y
z1 -> y	z1 -> y	y -> z2	x1 -- z1	x1 -- z1
x1 -- z1	x1 -- z1	x1 -- z1	x2 -- z2	z2
x2 -- z2	x1 -- z2	x2 -- z2		
	x2 -- z2			
Block 2				
δ=.1; γ=.7	**δ=.07; γ=.7**	**δ=.014; γ=.7**	**δ=.01; γ=.7**	**δ=.008; γ=.7**
x1 -> y	x1 -> y	x1 -> y	x1 -> y	x1 -> y
x2 -> y	x2 -> y	x2 -> y	x2 -> y	x2 -> y
z1 -> y	z1 -> y	z1 -> x1	z1	x1 -- z2
x1 – z1	x1 -- z1	z1 -- z2	z2	z1
x2 – z2	x2 -- z2			
Block 3				
δ=.1; γ=.9	**δ=.07; γ=.9**	**δ=.014; γ=.9**	**δ=.01; γ=.9**	**δ=.008; γ= .9**
x1 -> y	y -- x1	x1 -> y	x1 -> y	x1 -> y
x2 -> y	y -- x2	x2 -> y	x2 -> y	x2 -> y
x1 -- z1	x1 -- x2	z1	z1	z1
x2 -- z2	x1 -- z1	z2	z2	z2
	x2 -- z2			
	z1 -- z2			

When the relevance control parameter δ equals to 0.008, both X_1 and X_2 cause Y, and there is an undirected edge from X_1 to Z_1, meaning the algorithm is unable to sort out causal flow between X_1 and Z_1. Also, Z_2 is independent, meaning it does not cause any other variables.

Block 2 (0.05 significance level)

In block 2 of table 3.3 the level of endogeneity is increased to 0.7 and the relevance control parameter δ is decreasing moving across the block. When the relevance control parameter δ equals to 0.1 and 0.07, both X_1 and X_2 cause Y, Z_1 causes Y, and there are undirected edges from X_1 to Z_1, as well as from X_2 to Z_2, meaning the algorithm is not able to sort out the causal flow between X_1 and Z_1, as well as between X_2 and Z_2.

When the relevance control parameter δ equals to .014, both X_1 and X_2 cause Y, Z_1 causes X_1, and there is an undirected edge from Z_1 to Z_2, meaning the algorithm is unable to sort out causal flow between Z_1 and Z_2.

When the relevance control parameter δ equals to .01, both X_1 and X_2 cause Y, and both Z_1 and Z_2 are independent, meaning they do not cause any other variables. When the relevance control parameter δ equals to .008, both X_1 and X_2 cause Y, and there is an undirected edge from X_1 to Z_2, meaning the algorithm is unable to sort out the causal flow between X_1 and Z_2. Also, Z_1 is independent, meaning it does not cause any other variables.

Block 3 (0.05 significance level)

In block 3 of table 3.3 the level of endogeneity is raised up to 0.9 and the relevance control parameter δ is decreasing moving across the block. When the relevance control parameter δ equals to 0.1, both X_1 and X_2 cause Y, and there are undirected edges from X_1 to Z_1, as well as from X_2 to Z_2, meaning the algorithm is not able to sort out the causal flow between X_1 and Z_1, as well as between X_2 and Z_2.

When the relevance control parameter δ equals to .07 there are undirected edges from Y to X_1, from Y to X_2, from X_1 to X_2, from X_1 to Z_1, from X_2 to Z_2, from Z_1 to Z_2, meaning the algorithm is unable to sort out the causal flows between Y and X_1, Y and X_2, X_1 and X_2, X_1 and Z_1, X_2 and Z_2, Z_1 and Z_2.

When the relevance control parameter δ equals to .014, .01, and .008, both X_1 and X_2 cause Y, and both Z_1 and Z_2 are independent, meaning they do not cause any other variables.

Summary of the Directed Graphs (0.05 significance level)

The results presented in table 3.3 are inconsistent with our expectations in every block. Although almost in every block both X's caused Y, as expected, however, at these low levels of correlations there is not a clear-cut pattern of causal flows between Z's and X's. Thus, a conclusion can be drawn, that at the 0.05 significance levels the method of the directed graphs does not do a good job in terms of identifying the instruments.

Analysis Using the Directed Graphs (0.005 Significance Level)

As it was mentioned above, the results from the directed graphs are also going to be reported at the 0.005 significance level, because of the large sample size of 2000 observations.

Block 1 (0.005 significance level)

In block 1 of table 3.4 the endogeneity control parameter γ constant at the level of 0.3, whereas the relevance control parameter δ is decreasing reading across the block.

When the relevance control parameter δ equals to 0.1 and 0.07, X_1, X_2, and Z_1 cause Y, and there are undirected edges from X_1 to Z_1, as well as from X_2 to Z_2, meaning the algorithm is not able to sort out the causal flow between X_1 and Z_1, as well as between X_2 and Z_2.

When the relevance control parameter δ equals to 0.014 and 0.01 both X_1 and X_2 cause Y, and there are undirected edges from X_1 to Z_1, as well as from X_2 to Z_2, meaning the algorithm is not able to sort out the causal flow between X_1 and Z_1, as well as between X_2 and Z_2. When the relevance control parameter δ equals to 0.008, both X_1 and X_2 cause Y, and both Z_1 and Z_2 are independent, meaning that they do not cause any other variables.

Block 2 (0.005 significance level)

In block 2 of table 3.4 the level of endogeneity is raised to 0.7 and the relevance control parameter δ is decreasing reading across the block. When the relevance control parameter δ equals to 0.1 and 0.07 both X_1 and X_2 cause Y, Z_1 causes X_1, and there is an undirected edge from X_2 to Z_2, meaning the algorithm is unable to sort out causal flow between X_2 and Z_2.

When the relevance control parameter δ equals to 0.014, 0.01, and 0.008, both X_1 and X_2 cause Y, and both Z_1 and Z_2 are independent, meaning that they do not cause any other variables.

Table 3.4 Results from the Directed Graphs Using Artificially Created Data with the Number of Observations 2000 for Each Combination of δ and γ (0.005 Significance Level)

Block 1				
δ=.1; γ=.3	**δ=.07; γ=.3**	**δ=.014; γ=.3**	**δ=.01; γ=.3**	**δ=.008; γ=.3**
x1 -> y	x1 -> y	x1 -> y	x1 -> y	x1 -> y
x2 -> y	x2 -> y	x2 -> y	x2 -> y	x2 -> y
z1 -> y	z1 -> y	x1 -- z1	x1 -- z1	z1
x1 -- z1	x1 -- z1	x2 -- z2	x2 -- z2	z2
x2 -- z2	x2 -- z2			
Block 2				
δ=.1; γ=.7	**δ=.07; γ=.7**	**δ=.014; γ=.7**	**δ=.01; γ=.7**	**δ=.008; γ=.7**
x1 -> y	x1 -> y	x1 -> y	x1 -> y	x1 -> y
x2 -> y	x2 -> y	x2 -> y	x2 -> y	x2 -> y
z1 -> x1	z1 -> x1	z1	z1	z1
x2 -- z2	x2 -- z2	z2	z2	z2
Block 3				
δ=.1; γ=.9	**δ=.07; γ=.9**	**δ=.014; γ=.9**	**δ=.01; γ=.9**	**δ=.008; γ=.9**
x1 -> y	x1 -> y	x1 -> y	x1 -> y	x1 -> y
x2 -> y	x2 -> y	x2 -> y	x2 -> y	x2 -> y
z1 -> x1	x1 -- z1	z1	z1	z1
x2 -- z2	z2	z2	z2	z2

Block 3 (0.005 significance level)

In block 3 of table 3.4 the level of endogeneity is increased to 0.9 and the relevance control parameter δ is decreasing reading across the block. When the relevance control parameter δ equals to 0.1, both X_1 and X_2 cause Y, Z_1 causes X_1, and there is an undirected edge from X_2 to Z_2, meaning the algorithm is not able to sort out the causal flow between X_2 and Z_2.

When the relevance control parameter δ equals to 0.07, both X_1 and X_2 cause Y, and there is an undirected edge from X_1 to Z_1, meaning the algorithm is unable to sort out the causal flow between X_1 and Z_1. Also, Z_2 is independent, meaning it does not cause any other variables.

When the relevance control parameter δ equals to 0.014, 0.01, and 0.008, both X_1 and X_2 cause Y, and both Z_1 and Z_2 are independent, meaning that they do not cause any other variables.

Summary of the Directed Graphs (0.005 significance level)

The results from the directed graphs at the 0.005 significance level show that even though in most of the cases both X_1 and X_2 do cause Y, however, this method is unclear on assigning causal flows between X's and Z's, meaning that directed graphs still does not perform well in terms of identifying the instruments given the low levels of correlations between variables.

Summary of the Directed Graphs Across the Significance Levels

The method of directed graphs, overall, did not perform well in terms of identifying the instruments. Even though, in both tables the algorithm was able to identify causal flows between X's and Y, however, it failed to sort out the causal flows between X's and Z's, thus failing to determine whether Z's can be used as legitimate instruments for X's. For some combinations of the relevance control parameter δ and the endogeneity control parameter γ the algorithm was able to assign causal flow between Z_1 and X_1, but it did not do that with respect to Z_2 and X_2.

CHAPTER IV

AN INVESTIGATION OF ACTUAL FAT INTAKE

In the previous chapter the Hausman test and the method of the directed graphs were applied to the artificially generated dataset. However, in this chapter the Hausman test and the method of the directed graphs are going to be applied to the dataset on actual fat intake.

Data Description

The data consist of 1778 observations from the 1994-1996 from the Continuing Survey of Food Intakes by individuals (CSFII) and Diet and Health Knowledge Survey (DHKS) conducted by the Human Nutrition Information Service of the U.S. Department of Agriculture. These data have been used in the paper by Park and Davis on studying the endogeneity of health information (Park and Davis 2001). The CSFII data contains information on socioeconomic variables and nutrient intake for members of a representative sample of US household over a time period of two nonconsecutive days. The CSFII was followed with the DHKS in such a manner so that information from it could be linked to information from the CSFII. Twenty years and older individuals who participated in the CSFII were contacted approximately three weeks after they responded to the CSFII and asked a series of questions about their diet-health knowledge and attitudes. The DHKS was conducted in such a way so that only one respondent per household took part in the survey.

The dependent variable in the analysis is *total fat* intake for the individual and is calculated as a simple average of fat intake in grams during two-day time period. The explanatory variables are hypothesized to be as follows. *Total Fat/Disease Knowledge* is measured based on the responses to several questions on respondents' awareness on the relationship between fat intake and diseases. Each respondent is graded from 0 to 100 based on a special answer key (see Park and Davis for discussion, p. 844). Total fat/Disease Knowledge is considered endogenous as in Park and Davis. *Household Size* is represented by the number of members of household. *Age* is represented by the actual age of a main meal planner in years. *Income* is represented by total household income in $1,000. *TV* is represented by the average hours of watching TV on daily basis. *Body Mass Index* (*BMI*) is represented by the ratio of body weight (kilograms) to squared height (meters), and, for obvious reasons, it is also considered endogenous. *College* is awarded 1 if a main meal planner has more than 12 years education, otherwise, 0. *Smoker* is awarded 1 if a main meal planner is a smoker, otherwise, 0. *Special Diet* is awarded 1 if a main meal planner is on a special diet, otherwise, 0.

The summary statistics are given in table 4.1. In table 4.1 the standard deviation (38.20501) of Total Fat is almost half the size of the mean (72.83802), meaning a wide range of intake across the sample. The average of the total fat/disease knowledge for Total Fat is 50.04687, while the standard deviation is 19.84786. The average household size is almost three, and the average age of main meal planner is almost fifty. The average annual income is $33.04641. The average hours of TV watching per day is almost three, and the average BMI is approximately twenty-six, while the standard deviation is just above six. Because the answer associated with such variables as college,

smoker, and special diet was either “yes” or “no”, hence their respective means indicate the percentage of respondents that gave positive answers.

Table 4.1. Variables, Definitions, and Summary Statistics for 1994

Variable	Definitions and Units	Mean and Standard Deviation
Total fat	Two day average intake in grams	72.83802 (38.20501)
Total fat/disease knowledge	Grade on Total Fat/Disease questions relative to nutritionist	50.04687 (19.84786)
Household size	Number of members of household	2.64961 (1.48841)
Age	Age of main-meal planner in years	49.52868 (17.25673)
Income	Total household income in $1,000	33.04641 (25.34951)
TV	Average hours of TV watching per day	2.7635 (2.22715)
Body mass index	Ratio of body weight (kilograms) to Squared (height meters)	25.89314 (6.33006)
College	1 if attending school beyond 12th grade; zero otherwise	0.4252 (0.49451)
Smoker	1 if smoker; zero otherwise	0.25759 (0.43743)
Special diet	1 if on special diet; zero otherwise	0.18504 (0.38844)

It is expected that total fat/disease knowledge will have a negative effect on total fat intake, because as consumers become more and more knowledgeable about the possible diseases associated with fat consumption they are going to reduce their consumption of fat.

It is unclear what the sign on the household size will be. Household size could have a positive or negative impact on Total Fat intake. It is hard to hypothesize about the effect of age, because it strongly depends upon cultural backgrounds and dietary habits (Variyam, Blaylock, and Smallwood 1998).

Income could have a positive or negative impact on Total Fat intake. An increase in a household's income will allow household members to buy normal or luxurious goods, thus cutting down on the consumption of relatively cheaper goods that contain fat. However, individuals might view products containing fat as their major source of protein, and thus keep consuming products containing fat even if their income goes down.

TV watching is expected to be negatively related to fat consumption. Because television can be considered as one of the major source for consumers, extensive TV watching will allow consumers to become more knowledgeable about potential diseases linked to fat consumption and eventually reduce their intake of fat. However, TV watching might contribute to increased fat intake, but that would be a direct result of a sedentary lifestyle.

The effect of the body mass index is unclear because it might be the case when the people with higher BMI might cut down their further consumption of fat trying to stay in a good shape. However, it can also be the case when people with higher BMI

might get more of calories from foods rich in fat because they are larger (Variyam, Blaylock, and Smallwood 1998).

College is expected to have a negative effect on fat intake. It is expected that an additional year in college will allow improving of individuals' information processing ability. Smoker is predicted to have positive effect on fat consumption given the indifferent attitude towards the health issues on smokers' part.

Special diet is expected to be negatively related with fat consumption because people on diet tend to pay a great deal of attention to their consumption patterns, which is mostly fat-free food oriented, because excessive fat consumption creates conducive conditions for gaining weight.

Estimation Procedure: OLS and IV

Before presenting the results, it should be pointed out that the following statistics are going to be reported at the 5% significance level: the OLS and IV parameter estimates for the abovementioned variables, R-squared from the OLS and IV estimation, Hausman P-value, and partial correlations R^2p_1 and R^2p_2 just as in Monte-Carlo experiment. The R-squared statistic – sometimes referred to as the coefficient of determination – is the percent of the variation that can be explained by the regression equation. The Hausman P-value is the probability of rejecting the null hypothesis, given it is true. The null hypothesis of the Hausman test is that the OSL and IV estimates are the same. R^2p_1 and R^2p_2 are measuring instrument relevance between Total fat/Disease Knowledge and the rest of the variables, and BMI and the rest of the variables, respectively. The instruments were calculated according to the method developed by Lewbel (1997), using second and

third moments of variables as instruments. When Park and Davis (2001) investigated the endogeneity of the health information, they also used Lewbel's method. According to Lewbel, if x_i is an element of the X matrix, then $r_1 = (y_1 - \overline{y_1})(y_2 - \overline{y_2})$, and $r_i = (x_i - \overline{x_i})(y_2 - \overline{y_2})$ are all legitimate instruments, in addition to the x_i variables, and the IV estimator is consistent. $\overline{x_i}$ and $\overline{y_i}$ indicate averages of the variables. In the present discussion, continuous variables, such as total fat/disease knowledge, age, household size, income, TV, are used to form instruments for Body Mass Index (BMI) and total fat/disease knowledge (y1). The IV generation process looks the following way:

z1 = (y1-72.8380)*(h1-50.046) (4.1)

z2 = (age-49.528)*(h1-50.046) (4.2)

z3 = (hhsz-2.649)*(h1-50.046) (4.3)

z4 = (inc-33.046)*(h1-50.046) (4.4)

z5 = (tv-2.763)*(h1-50.046) (4.5)

z6 = (y1-72.8380)*(bmi-25.8931) (4.6)

z7 = (age-49.528)*(bmi-25.8931) (4.7)

z8 = (hhsz-2.649)*(bmi-25.8931) (4.8)

z9 = (inc-33.046)*(bmi-25.8931) (4.9)

z10 = (tv-2.763)*(bmi-25.8931) (4.10)

Below is the representation of the real-world model in mathematical form so that it could be compared to the actually estimated model:

$$y1=\beta_0+\beta_1 h1+\beta_2 BMI+\beta_3 hhsz+\beta_4 age+\beta_5 inc+\beta_6 TV+\beta_7 college+\beta_8 smoker+\beta_9 sdiet+\varepsilon_1 \quad (4.11)$$

$$h1 = \gamma_0+\gamma_1 z1+\gamma_2 z2+\gamma_3 z3+\gamma_4 z4+\gamma_5 z5+\gamma_6 hhsz+\gamma_7 age+\gamma_8 inc+\gamma_9 TV+\gamma_{10} college+ \gamma_{11} smoker+\gamma_{12} sdiet+\varepsilon_2 \quad (4.12)$$

$$BMI = \delta_0+\delta_1 z6+\delta_2 z7+\delta_3 z8+\delta_4 z9+\delta_5 z10+\delta_6 hhsz+\delta_7 age+\delta_8 inc+\delta_9 TV+ \delta_{10} college+\delta_{11} smoker+\delta_{12} sdiet+\varepsilon_3 \quad (4.13)$$

Equation (4.11) is the structural equation of interest. Equation (4.12) and (4.13) define total fat/disease knowledge and BMI, respectively, in terms of instruments. The parameters ε_1, ε_2, and ε_3 are disturbances to the equations (4.11), (4.12), and (4.13), respectively.

The values of the abovementioned statistics are presented in table 4.2. The results in table 4.2 show that the OLS parameter estimate for total fat/disease knowledge is a negative -.082927 and is significant at the 6% level, while the IV parameter estimate is also negative, which is consistent with our expectations, but is about 19 times as large (-1.64178) and is significant at the 5% level. The OLS parameter estimate for BMI is a positive .611346 and significant at the 5% level, while the IV parameter estimate is a negative -1.29042, but is about twice as large and significant at the 5% level. These positive and negative effects of BMI on fat consumptions are possible, but troubling that a different estimation procedure can cause the sign to change.

Table 4.2. OLS vs. IV Results

	OLS		IV	
Variable	Estimated Coefficient	P-value	Estimated Coefficient	P-value
C	74.6934	.000	202.277	0
H1	-.082927	.058	-1.64178	0.01
BMI	.611346	.000	-1.29042	0.026
HHSZ	-.707586	.270	-0.511417	0.558
AGE	-.307067	.000	-0.262223	0.001
INC	.132610	.000	0.113823	0.024
TV	.289013	.459	-0.13106	0.821
COLL	-1.36992	.479	-4.2308	0.121
SMOKER	9.44550	.000	7.36382	0.009
SDIET	-18.8229	.000	-16.2921	0
R-squared OLS	0.090777			
R-squared IV	0.010781			
Hausman P-value	0			
R^2p_1	0.0086461			
R^2p_2	0.10461			

The OLS parameter estimate for household size is a negative -.707586 and not significant at the 5% level, while the IV parameter estimate is negative -.511417 and not significant at the 5% level. The OLS parameter estimate for age is a negative -.307067 and significant at the 5% level, while the IV parameter estimate is a negative -.262223 and significant at the 5% level. The negative effect of age on fat consumption might be due to the fact that older people are less active and require fewer calories, hence consuming less and less fat. The OLS parameter estimate for income is a positive .132610 and significant at the 5% level, while the IV parameter estimate is positive .113823 and significant at the 5% level. Although both OLS and IV parameter estimates for income contradict to our expectations, however, the positive effect to income might be explained by the fact that consumers tend to buy energy dense food rich in fat as they enjoy increase in their income. The OLS parameter estimate for TV is a positive .289013 and not significant at the 5% level, while the IV parameter estimate is a negative -.13106 and not significant at the 5% level. The OLS parameter estimate for college is a negative -1.36992 and not significant at the 5% level, while the IV parameter estimate is a negative -4.2308 and not significant at the 5% level. The OLS parameter estimate for smoker is a positive 9.44550 and significant at the 5% level, while the IV parameter estimate is a positive 7.36382 and significant at the 5% level. The positive effect of smoker on fat consumption is consistent with our expectations. The OLS parameter estimate for special diet is a negative -18.8229 and significant at the 5% level, while the IV parameter estimate is a negative -16.2921 and significant at the 5% level. The negative effect of special on fat consumption is in agreement with our expectations.

The R-squared statistic from the OLS estimation is a positive 0.090777, meaning that independent variables can account for just above 9 percent of the variation in consumption of total fat. The R-squared statistic from the IV estimation is a positive 0.010781, meaning that independent variables can account for just above 1 percent of the variation in consumption of total fat.

The Hausman P-value equals to 0, which means we can reject the null hypotheses without fearing of making Type I error, according to which the null hypotheses is rejected given that it is true. Considering the Hausman P-value conclusion is drawn that the OLS parameter estimate is significantly different from the IV parameter estimate.

Both R^2p_1 and R^2p_2 are small and equal to .0086461 and .10461, respectively. These small partial correlations mean that the IV estimator won't be reliable, and the Hausman test might be misleading.

Summarizing the results in table 4.2, it needs to be pointed out that both OLS and IV parameter estimates for total fat/disease knowledge are negative, however, the IV parameter estimate is about 19 times greater than the OLS parameter estimate. Also, it is noteworthy that the OLS parameter estimate for total fat/disease knowledge is significant at the 6% level, while the IV parameter estimate is significant at the 5% level. The OLS parameter estimate for BMI is positive and significant at the 5% level, while the IV parameter estimate is negative, but is about twice as large and significant at the 5% level. Even though these positive and negative effects of BMI on total fat intake are possible, however, it is somewhat troubling that the sign can be changed due to the different estimation procedure. Reading down the table we see that the coefficient estimates from both estimators have the same signs, except for coefficient estimates associated with the

hours of TV watching per day. Specifically, the OLS parameter estimate for TV is a positive and not significant at the 5% level, while the IV parameter estimate is negative and not significant at the 5% level.

Both coefficients of determination from the OLS and IV estimation are positive, although fairly small too. However, R-squared associated with the OLS regression is slightly larger than that of the IV estimation, meaning that the OLS performs better than the IV in terms of explaining the amount of variation in the dependent variable.

The direct conclusion reached from the Hausman P-value is that the IV estimator should be preferred over the OLS estimator. But, the message that we get from both partial correlations is that the result of the Hausman test may be misleading due to the fact that both partial correlation are fairly small.

Directed Graphs

Next, the method of directed graphs is implemented to the same real-world dataset. It is expected for the total fat intake to be caused by total fat/disease knowledge, BMI, and the rest of the exogenous variables. Also, according to the IV generation process it is expected for total fat/disease knowledge to be caused by z1, z2, z3, z4, and z5 and other exogenous variables, as well as for BMI to be caused by z6, z7, z8, z9, and z10, and other exogenous variables, if our priors on instruments are correct. The results from the method of directed graphs are presented in table 4.3 at two significance levels, .005 and .05. In table 4.3 y1 represents total fat intake, sdiet represents special diet, inc represents income, hhsz represents household size, coll represents college, h1 represents

total fat/disease knowledge, and z1, z2, z3, z4, z5, z6, z7, z8, z9, and z10 represent instruments.

Table 4.3. Directed Graphs Results

S.L.=0.05			S.L.=0.0050		
age	->	y1	age	->	y1
inc	->	y1	y1	->	smoker
bmi	->	y1	y1	->	sdiet
smoker	->	y1	z1	->	y1
sdiet	->	y1	y1	->	z6
y1	->	z1	age	->	hhsz
z6	->	y1	inc	->	hhsz
y1	->	z9	hhsz	->	z3
z1	->	h1	hhsz	->	z8
age	->	hhsz	coll	->	age
inc	->	hhsz	age	->	smoker
hhsz	->	sdiet	age	->	sdiet
hhsz	->	z3	age	->	z7
hhsz	->	z8	coll	->	inc
coll	->	age	inc	->	smoker
age	->	smoker	tv	--	z5
age	->	sdiet	coll	->	bmi
age	->	z7	bmi	->	sdiet
coll	->	inc	z6	->	bmi
inc	->	z6	z7	->	bmi
z5	->	tv	bmi	->	z8
z10	->	tv	bmi	->	z9
coll	->	bmi	coll	->	smoker
sdiet	->	bmi	z2	->	z1
z6	->	bmi	z1	->	z4
z7	->	bmi	z3	->	z2
z8	->	bmi	z3	->	z4
bmi	->	z9	z7	->	z8
z1	->	z2	z7	->	z9
z1	->	z4	h1		
z1	->	z5	z10		
z2	->	z3			
z3	->	z4			
z8	->	z3			
z8	->	z7			
z7	->	z9			
z10	->	z7			

According to the results in table 4.3 at the 0.05 significance level age causes total fat intake, household size, smoker, special diet, and z7, income causes total fat intake, household size, and z9, BMI causes total fat intake, and z9, smoker causes total fat intake, special diet causes total fat intake, and BMI, total fat intake causes z1, and z9, household size causes special diet, z3, and z8, college causes age, income, and BMI, z1 causes total fat/disease knowledge, z2, z4, and z5, z2 causes z3, z3 causes z4, z5 causes TV, z6 causes total fat intake, and BMI, z7 causes BMI, and z9, z8 causes BMI, z3, and z7, z10 causes TV, and z7.

In mathematical form, the TETRAD output at the 0.05 significance level can be represented similar to the equations (4.11) – (4.13):

$$y1= \beta_0+\beta_1 age+\beta_2 inc+\beta_3 bmi+\beta_4 smoker+\beta_5 sdiet+\beta_6 z6 \quad (4.14)$$

$$h1= \gamma_0+\gamma_1 z1 \quad (4.15)$$

$$BMI= \delta_0+ \delta_1 coll+\delta_2 sdiet+\delta_3 z6+\delta_4 z7+\delta_5 z8 \quad (4.16)$$

Unlike the equation (4.11), the equation (4.14) does not include such variables as total fat/disease knowledge, household size, college and TV, meaning that these variables do not cause the total fat intake. Equation (4.15) does not contain such instruments as z2, z3, z4, z5 and other exogenous variables in contrast to the equation (4.12). Equation (4.16) includes only such instruments as z6, z7 and z8 unlike equation (4.13) and it only includes college and special diet of all the other endogenous variables.

According to the results in table 4.3 at the .005 significance level age causes total fat intake, household size, smoker, special diet, and z7, total fat intake causes smoker, special diet, and z6, income causes household size, and smoker, household size causes z3, and z8, college causes age, income, BMI, and smoker, BMI causes special diet, z8, and

z9, z1 causes total fat intake, and z4, z2 causes z1, z3 causes z2, and z4, z6 causes BMI, z7 causes BMI, z8, and z9. The algorithm is unable to sort out the causal flow between TV and z5. Also, z10 and total fat/disease knowledge are independent, meaning that they do not cause any other variables.

In mathematical form the TETRAD output at the 0.005 significance level can be represented similar to the equations (4.11) and (4.13):

$$y1= \beta_0+\beta_1 age+\beta_2 z1 \quad (4.17)$$

$$BMI= \delta_0+ \delta_1 coll+\delta_2 z6+ \delta_3 z7 \quad (4.18)$$

When comparing equation (4.17) with equation (4.11), we see that such variables as total fat/knowledge disease, household size, income, TV, BMI, college, smoker, and special diet are left out of the equation (4.17). The equation that would possibly define total fat/knowledge disease does not include any instruments or any other exogenous variables. The equation (4.18) includes only such instruments as z6 and z7, missing the rest of the instruments, such as z8, z9, and z10. It also does not include only of the other exogenous variables, except college.

The results presented in table 4.3 are not consistent with our expectations. Specifically, at the .05 significance level the results show that total fat intake is caused by exogenous variables, such as age, income, BMI, smoker, special diet. However, such exogenous variables as college and household size do not cause total fat intake. Furthermore, the only instrument that causes total fat/disease knowledge is z1. Also, z6, z7, and z8 are the only instruments that cause BMI. The same results can be observed from equations represented in mathematical form (i.e. (4.14), (4.15), and (4.16)).

Neither are the results in accordance with our expectations at the 0.005 significance level. Specifically, age is the only exogenous variable that causes total fat intake. Furthermore, total fat/disease knowledge is independent, meaning that it is not caused by any of its instruments. Also, z6 and z7 are the only instruments causing BMI. The same points are supported by the equations (4.17) and (4.18).

The results obtained from analyzing the real-world data implementing the method of the directed graphs lead us to believe that this particular method does not perform well in identifying the instruments. There are three possible reasons why the method of the directed graphs may not perform well. The first reason is that, the method of the directed graph is correct and prior on model structure is wrong. The second reason is that, the method of the directed graph is unreliable and the priors are correct. And, finally, the third reason is some combination of the first two reasons mentioned above.

CHAPTER V

SUMMARY AND CONCLUSIONS

Summary

The main purpose of this thesis has been the evaluation of the performance of the Hausman test as opposed to the method of the directed graphs in terms of identifying the endogeneity of the health information given weak instruments.

For that purpose a Monte-Carlo experiment was carried out, wherein the underlying model structure was known. Overall, two Monte-Carlo experiments were carried out. The first experiment utilized fifteen datasets, each consisting of ten separate datasets. The median values of parameters were reported in table 3.1. The overall picture obtained from that experiment was that the OLS estimator performed better than the IV estimator in terms of the parameter estimation and absolute bias of the parameter estimates associated with the relatively less endogenous explanatory variable. Another important finding was that a there does not seem to be a monotonic relationship between the median parameter estimates from both estimators and the relevance of the instrument at levels of correlation. A non-monotonic relationship was also observed between the median parameter estimates from both estimators (except for the IV estimator of β_1) and the degree of endogeneity of the variable. In both cases a non-monotonic relationship suggests that there is a high level of risk involved because of the low levels of correlations. The percentage of the times the Hausman test rejects the null hypothesis exhibited declining pattern as instrument relevance decreased, and the percentage of the times the Hausman test rejects the null hypothesis overall exhibited increasing pattern as the degree of endogeneity increased. In both cases the Hausman statistic communicated

the right idea, however, it cannot be considered as a reliable method, because both the partial correlations, which measure instrument relevance, were fairly small (Park and Davis 2001).

The second Monte-Carlo experiment used fifteen datasets, with 2000 observations each. As a result, 15 models were estimated, one model for each degree of instrument relevance and the degree of endogeneity. Again, the OLS estimator proved to be doing a better job than the IV estimator in terms of parameter estimation and in terms of bias with respect to the relatively less endogenous explanatory variable. Another significant finding was that there did not appear to be a monotonic relationship between the parameter estimates from both estimators and the instrument relevance, as well as the parameter estimates and the degree of endogeneity. In both cases it means, that there is a high level of uncertainty present, due to the low levels of correlations. The decreasing pattern of the Hausman P-value within each block, and its increasing pattern across the blocks communicated the right idea, however, because the partial correlations that were used to measure instrument relevance were fairly small, we cannot consider the results of the Hausman test as reliable.

The method of the directed graphs was applied to the second Monte-Carlo experiment and did not do any better. Even though in the majority of cases it identified the dependent variable correctly, the method of the directed graphs failed to sort out the causal flow between the endogenous variables and their instruments. In other words, the directed graphs did not do a good job in identifying the instruments.

When analyzing the real-world data it was assumed that total fat/disease knowledge and BMI variables were endogenous for obvious reasons. The results of the

analysis using the real-world data presented in table 4.2 show that even though both OLS and IV parameter estimates for total fat/disease knowledge are negative, the IV parameter estimate (-1.64178) is about 19 times greater than that of the OLS (-.082927). Also, the IV parameter estimate is significant at the 5% level, while the OLS parameter estimate is significant at the 6% level. The OLS parameter estimate for BMI is positive and significant at the 5% level (.611346), while the IV parameter estimate for BMI is negative (-1.29042). Also, the IV parameter estimate for BMI is about twice larger than that of the OLS and significant at the 5% level. The Hausman P-value clearly suggests using the IV estimator as opposed to the OLS estimator, however, considering fairly small values of partial correlations, a conclusion is drawn that the result of the Hausman test might be misleading.

The method of the directed graphs did not do a good job in identifying the instruments, which implies that either the results obtained from the method of the directed graph are correct and our prior on model structure is wrong, or the method of the directed graph is unreliable and the priors on model structure are correct. But it might also be the case that the combination of these two possibilities may account for not considering the method of the directed graphs as a reliable method when dealing with low levels of correlations between variables.

Conclusions

The overall conclusion reached from conducting the Hausman test and the directed graphs using cross-sectional dataset is that neither of them performs well when correlations are low. As such, the advice from Nakamura and Nakamura should be

followed, wherein they call for reporting results from both OLS and IV estimations. Under certain assumptions the performance of the Hausman test was concluded as unreliable, and the performance of the method of the directed graphs also did not do a good job given low levels of correlations. These conclusions still hold for both artificially generated and real-world dataset. The major conclusion of this research is significant to those who are further willing to do a further study on finding a method, or a procedure that would possibly allow to better handle the endogeneity issue.

Assuming the objective of public policy is to decrease health care cost associated with consuming total fat by conducting health education programs, it is unclear whether the OLS or IV estimates are to be preferred. The OLS estimates indicate a smaller effect on total fat intake of health information than the IV estimates. To err on the side of caution would indicate that the OLS estimates would be preferred. Because these estimates are smaller, if they are underestimated, then the target level would be reached with less health information and less money than expected.

Research Scope and Extensions

Extensive research has been done so far to investigate the impact of the health information on the demand for food and different nutrients. In most of the papers the endogeneity issue was handled through making choices in favor of specific estimation procedures, while in the others it was handled through the improvement of the method of the data collection. However, the concept of the weak instruments was not mentioned in any of them (see literature review). This thesis not only shows the importance of the instrument relevance, but it also shows the unreliability of the Hausman test and the

directed graphs when the instruments are weak. Nevertheless, further research should be conducted using the dataset that is more representative. That is, besides main meal planners, the rest of the household members should be included in the survey. As a result more representative results might be obtained.

REFERENCES

Bessler, D. A. "Introduction to Directed Graphs," Lecture notes from AGEC 607, Research Methodology. Department of Agricultural Economics, Texas A&M University, 2002.

Bound, J., D.A. Jaeger, and R.M. Baker. "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak." *Journal of American Statistics Association* 90(1995):443-50.

Brown, D.J., and L.F Schrader. "Cholesterol Information and Shell Egg Consumption." *American Journal of Agricultural Economics* 73 (August 1990):548-555.

Buse, A. "The Bias of Instrumental Variables Estimators." *Econometrica* 60(1992):173-80.

Capps, O., Jr., and J.D. Schmitz. "A Recognition of Health and Nutrition Factors in Food Demand Analysis." *Western Journal of Agricultural Economics* 16(July1991):21-35.

Carlson, K.A., and B.W. Gould. "The Role of Health Information in Determining Dietary Fat Intake." *Review of Agricultural Economics* 16(1994):373-86.

Fries, J.F., C.E. Koop, and C.E. Beadle. "Reducing Health Care Costs by Reducing the Need and Demand for Medical Services." *New England Journal of Medicine* 329(1993):321-25.

Godfrey, L.G. "Instrument Relevance in Multivariate Linear Models." *Review of Economics and Statistics* 81(1999):550-52.

Griffiths, W. E., G.G. Judge, and R. Carter Hill. *Learning and Practicing Econometrics.* New York: John Wiley & Sons Inc, 1992.

Gujarati, D. N. *Basic Econometrics.* 4th ed., New York: McGraw-Hill, 2002.

Hausman, J. A. "Specification Tests in Econometrics." *Econometrica* 46(1978):1251-1271.

Kennedy, P. *A Guide to Econometrics.* 4th ed., Cambridge, MA: MIT Press, 1998.

Lewbel, A. "Constructing Instruments for Regressions with Measurement Error When No Additional Data are Available, with an Application to Patents and R&D." *Econometrica* 65(1997):1201-14.

Nakamura A., and M. Nakamura. "Model Specification and Endogeneity." *Journal of Econometrics* 83(1998):213-37.

Park, J., and Davis C. G. "The Theory and Econometrics of Health Information in Cross-Sectional Nutrient Demand Analysis." *American Journal of Agricultural Economics* 83(November 2001):840-851.

Pearl, J. *Causality.* Cambridge, MA: Cambridge University Press, 2000.

Scheines, R., P. Spirtes, C. Glymour, and C. Meek. *TETRAD II: User's Manual and Software*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 1994.

Shea, J. "Instrument Relevance in Multivariate Linear Models. A Simple Measure." *Review of Economics and Statistics* 79(1997):348-52.

Staiger, D., and J.H. Stock. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65(May1997):557-586.

U.S. Department of Agriculture, Economic Research Service. *ERS Information.* June 2002. Available online at http://www.ers.usda.gov/News/June2002Media.pdf.

Variyam, J.N., J.R. Blaylock, and D. Smallwood. "Information Effects of Nutrient Intake Determinants on Cholesterol Consumption." *Journal of Agricultural Resource Economics* 23(1998):110-25.

Variyam, J.N., J.R. Blaylock, and D. Smallwood. "A Probit Latent Variable Model of Nutrition Information and Dietary Fiber Intake." *American Journal of Agricultural Economics* 78(1996):628-39.

VITA

Name:	Rafael G. Bakhtavoryan
Permanent Address:	4 Saryan Street, Apt # 4 Yerevan, Armenia
Education:	MS., Agricultural Economics Texas A&M University, 2004
	B.S., Accounting and Audit Armenian Agricultural Academy
Work Experience:	Graduate Research Assistant, Agricultural Economics, Texas A&M University, College Station, Texas. (2001-2003)