Effects of Multivitamins on Overall Self-Assessed Health

by

Xin Lu

(Under the direction of Professor Scott Atkinson)

Abstract

This paper examines the relationship between vitamin usage and self-assessed health of United States residents and addresses the issue of endogeneity of vitamin use. First, preliminary OLS regressions are run, which ignore endogeneity. Then two instruments are introduced in two separate bivariate probit models. Finally, a difference in differences probit regression is run using data from years 1986 and 2006. The two data sets used in this paper are the 1986 National Health Interview Survey and the 2006 National Health and Nutrition Examination Survey. Vitamin usage was found to be significant in all regressions that do not ignore endogeneity.

Index words: econometrics, health, instrumental variables, vitamins

Effects of Multivitamins on Overall Self-Assessed Health

by

Xin Lu

A Thesis Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Masters in Arts

Athens, Georgia

2012

Effects of Multivitamins on Overall Self-Assessed Health

by

Xin Lu

Approved:

Major Professor: Scott Atkinson

Committee: Christopher Cornwell
David Mustard

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2012

Effects of Multivitamins on Overall Self-Assessed Health

Xin Lu

July 31, 2012

Acknowledgments

I would like to thank the university's Economics Department's BA/MA program for providing me with the chance to take graduate level classes in preparation for graduate school. I am especially grateful for the guidance provided by my thesis committee: Professors Scott Atkinson, Chris Cornwell, and David Mustard. I would also like to thank the Mathematics Department at the university for the instruction the faculty there has provided me.

Contents

List of Tables

Chapter 1

Introduction

Usage of multivitamins to supplement regular food intake has become increasingly common in the United States. Much of the observed increase has been a result of multiple studies published in the past two decades linking the taking of vitamins with decreased risk of many diseases. Gahche et al. (2011) found from the National Health and Nutritional Examination Surveys that while 30% of the population used multivitamin supplements in 1988, that figure has steadily increased to 39% of the total population by 2006. However, despite the amount of literature that exists on the subject, there is no general consensus in academia as to just how supplements affect individual health.

Finding the effect of multivitamins on health is difficult because of the endogenous nature of multivitamin use as an explanatory variable. Those who buy and use supplements on a regular basis are more likely to take other precautions to their health in choices such as diet and exercise for which we have no data. Because of this, estimates of any partial effects of vitamins are inconsistent unless an instrumental variable procedure is used. Although many people recognize this issue and know it needs to be corrected, in practice it is very difficult to do. The ideal instrument needs to be correlated with vitamin usage but somehow be exogenous, that is, uncorrelated with the error term.

Another difficulty of estimating the long-term causal effect of vitamins is that not much longitudinal data is available for these studies. It wasn't until the past two decades when this kind of information was collected through surveys and interviews (Rock 2007). Even now, the cost involved in collecting extensive data on vitamin usage makes estimation difficult. Also, of the data that have been gathered, many of them do not have the same sample population year after year. Time dummies can be used to capture general trends in health through time, but these regressions would not contain information on the effects of prolonged usage of multivitamins. Knowing whether someone has recently taken a supplement does not give further information about whether he or she takes them regularly over time.

Chapter 2

Literature Review

Some studies on vitamin use do not run into the issue of endogeneity becuse they are experimental. In a controlled sixteen-week experiment, Macpherson et al. (2011) concluded that using multivitamins and minerals to supplement diet could improve memory in elderly women. The experiment was double-blind and memory was assessed at the beginning and end of the trial with computer-assessed tests. The two groups performed roughly equally in most of the tests, with the non-placebo group displayed improved reaction time in the spatial memory test. The advantage of this study lies in the randomness of the distribution of vitamins, which takes away the endogenous nature of multivitamin usage as a variable. Harris et al. (2011) conducted a similar experiment using older men as subjects instead of older women. Instead of memory, this study looked at the effect that multivitamin supplements would have on general mood and stress. Participants were randomly assigned to take either a supplement or a placebo. General mood, health, depression and anxiety were assessed through questionnaires before and after the treatment. The group who took the vitamin displayed statistically significant improvement in self-assessed general health relative to the control group. However, no such improvement was visible in categories such as anxiety and stress. While the randomness of these studies effectively makes vitamin usage an exogenous

variable, most people take vitamins to prevent diseases, not to improve memory or alleviate stress. These studies are too narrow in scope to contain information about effects on overall health.

Much of the existing literature on multivitamin use has been focused on their effects on specific diseases. Using logistic regression models with data from the National Institutes of Health-American Association of Retired Persons Diet and Health Study, Song et al. (2011) found that multivitamin intake did not result in increased future risk of diabetes in older adults. In addition, those who took vitamin C and calcium were associated with a lowered risk of being diagnosed with the disease. The data had information on multivitamin intake for a population during the years from 1995 to 1996 and then from 2004 to 2006. This study mentioned the possibility of a diabetes diagnosis between the two study periods affecting the respondents choice of taking a multivitamin. To reduce the effect of an endogenous explanatory variable, Song et al. only looked at those in the sample who were diagnosed with diabetes after the year 2000. It is therefore defensible to say that having diabetes would not have determined whether one took supplements in the first few years of data. However, this does not preclude the possibility that a diabetes diagnosis could have been a causal factor in whether a respondent chose to take multivitamins in the follow-up years.

There is also some literature available on the effects of multivitamins on general health. Using the Iowa Womens Health Study (IWHS), Mursu et al. (2011) concluded that certain supplements actually increase mortality in older women. However, the endogenous nature of the usage of supplements as an explanatory variable was not dealt with using an instrumental variable. Also, the population of the study was older women living in the United States, a country that has very few people suffering from malnutrition. While it is plausible that multivitamin usage would increase the lifespan of someone suffering deficiencies in his or her diet, it would be unlikely that this effect would be present in a population without these deficiencies.

The study in this paper uses data from the 1986 National Health Interview Survey (NHIS) and the 2006 National Health and Nutrition Examination Survey (NHANES). Although the NHIS was collected in years other than 1986, vitamin usage was not a statistic that was included in each year. Appending these two data sets allows us to look at the time trend of self-assessed health status. Instead of information on mortality, these sets both have a ranking of self-assessed health. Before using an instrumental variable approach, I will look at results from an ordinary least squares regression. In the next chapter, I first use county size and then a generated interaction variable of the respondent as an instrumental variable for vitamin use. Finally, I will take a difference-in-differences approach to see how self-assessed health responded to the Dietary Supplement Health and Education Act of 1994. Although I was not able to obtain the data from the IWHS as it protected by Health Insurance Portability and Accountability Act, the data set I use in this paper supplies the necessary data on vitamin intake of a random sample of U.S. citizens.

Chapter 3

OLS and Ordered Probit Estimation Assuming Exogeneity

3.1 Data Description

The 1986 National Health Interview Survey gathered vitamin usage, self-assessed health ranking, and other demographic information from residents of the United States. Counties in the United States were analyzed and broken down by demographic information. From these demographics, counties with similar demographics were grouped together and a representative county was randomly chosen. The survey population is a random sampling of children and adults in the United States from these representative counties. This particular data set differs from the Iowa Womens Health Study not only in the demographics, but in variables collected. While the IWHS focuses on older women, ages in the NHIS range from 2 to 99; there is no focus on a particular age. The IWHS is a longitudinal data set with information on vitamin use across time periods and contains information on age of death, if respondents passed away at some point during the study. The NHIS is collected regularly, but does not provide a balanced panel, since the sample of individuals differs for each cross

section. Therefore, age of death is not available and so a general rating of respondents' overall health is used instead.

The nature of the available data made it necessary to create various dummy variables for categorical data. Because vitamin usage is the variable with which this paper is mostly concerned, 3754 observations that were missing data on whether the respondent took a vitamin in the past two weeks were dropped from the data set. A dummy variable $vitamin2wks$ was then created which had a value of 1 if the observation reported taking a multivitamin in the past two weeks and 0 otherwise. Similarly, the new dummy variable $female$ has a value of 1 if the respondent is female and 0 otherwise.

The original data also gives information about the racial background of the indivuals. Black was coded as *black*, Asian as *Asian*, and all others, excluding white as *raceother*. The white category was dropped from estimation.

The NHIS data contains information on whether the respondent has ever been married, and if so, whether he or she is divorced or widowed. If not, there is further data on whether the spouse lives in the same household as the individual. For the purposes of this study, the relevant variable will be whether the respondent has a spouse living in the same house. If marriage has any effect on the health status of an individual, we would only expect those effects to manifest if the individual is in regular contact with that spouse. In light of this, the dummy variable *married* was created to take on a value of 1 if the individual is married and the spouse is living in the same space and 0 otherwise. Those aged 14 and under are placed in the latter category.

It is also likely that being a veteran of the armed forces would have an impact on one's health status. The NHIS has collected information about whether the participants have participated in various wars such as WWI, WWII, the Korean War, and the Vietnam War. The *veteran* dummy created from this information will have a value of 1 without regard to which war the individual fought and 0 otherwise. If there are differences in health status

among the veterans of the different wars, these effects would be better dealt with by the *age* regressor. Although there can be arguments made as to the intensity and violence of the different wars, by the year 1986, there is no way to do so uniformly. Unfortunately, almost 30% of respondents provided no information on whether they had served or not. These observations will be dropped in regressions that use this *veteran* regressor.

Three dummies for different regions were created as well. Northeast, Midwest, and West will all be compared to the excluded region, South. As stated before, the health variable is a self-assessed categorical ranking of health. Category 1 denotes 'poor', 2 denotes 'fair', 3 denotes 'good', 4 denotes 'very good', and 5 denotes 'excellent'.

The data on household income was rescaled because of its nonlinear nature. Up to an annual amount of $20,000, the data on income was broken down by every thousand dollars. In the original NHIS file, a 0 in the data would translate to less than $1,000 in income and a 1 in the data would translate to an income of $1,000 to $1,999 dollars. Each increment of 1 unit would correspond to an increase of $1,000 up until the code equalled 20. From there, the discrete indicator would then increment with every additional $5,000 earned instead. This is the case until the indicator 26, which is the code for more than $50,000. Instead of converting these amounts into dummy variables, they were re-coded such that each unit increment represents a $10,000 increase in annual income. Now, a 0 indicates an annual income of less than $10,000 and a 1 indicates an annual income of less than $20,000. A value of 5 will indicate income that is higher than $50,000. This is truncation of the data, but the issue is minor for this data set. An income of $50,000 would be almost $90,000 in 2011 dollars in terms of buying power. One would expect any effects of income on health to be minor after this benchmark.

Finally, the 1986 NHIS data contains a number of continuous variables. These include age measured in years, years of completed education, height measured in inches, weight measured in pounds, and the number of people living in the household.

3.2 Model Specification

Before using any instruments to account for endogeneity of multivitamin usage, I first look at the results of an ordinary least squares regression since typically non-economists would ignore this. The specification of the model is the most basic model where each covariate enters the regression linearly:

$$\begin{aligned} health = \beta_0 + \beta_1 vitamin2wks + \beta_2 educationyears + \beta_3 familysize + \beta_4 familyincome \\ + \beta_5 height + \beta_6 weight + \beta_7 female + \beta_8 asian + \beta_9 black \\ + \beta_{10} raceother + \beta_{11} married + \beta_{12} veteran \\ + \beta_{13} northeast + \beta_{14} midwest + \beta_{15} west + \beta_{16} age + \epsilon \end{aligned} \quad (3.1)$$

I define this set of explanatory variables, including a vector of ones, as X. A single observation or row of this matrix will be denoted $\mathbf{x}_i$. For this model to be identified, we need X to have full rank. Because dummy variables were created in such a way to avoid perfect multicollinearity, this condition is easily satisfied. In order for this model to be consistent, we also need to assume that $\mathbf{E}(\mathbf{x}_i^T \epsilon) = 0$, or that the error is uncorrelated with each regressor. This is the reason why gauging any causal effects of vitamins on health is difficult to do. People who take vitamins tend to be the ones who are most concerned about their health, have pre-existing health conditions, or know from blood tests that they are deficient in certain basic vitamins and are taking other precautions to live a healthier life. However, none of this information is available and therefore remains in the error term. As a result, we have an endogenous variable in vitamin usage and we cannot make consistent inferences about estimated coefficients. However, that does not make an ordinary least squares regression completely useless. We can still specify a regression in such a way that makes the regressors

make error term as close to null as possible. By placing as many sensible covariates in the model as possible and having a very high R^2, the condition $\mathbf{E}(\mathbf{x}_i^T \epsilon) = 0$ may still be satisfied. In addition, such a regression mirrors what non-economists (who ignore endogeneity) would produce. It is not necessary to assume homoskedasticity of the errors in specification (3.1). The equation will be estimated using robust standard errors.

To test the robustness of specification (3.1), I then drop sets of explanatory variables and rerun the OLS regression. The first set of variables I drop are the *married* (married-spouse in household) and geography covariates. *Age*, *height*, *weight* and the racial background dummies are all physical traits. It is unlikely that these variables would not affect one's valuation of his or her own health. Years of education and family income are unlikely to be irrelevant as well. People with more education are more likely to know how to maintain their health and those with higher income would be more able to afford better quality food and health care. Since income is believed to be a very important factor in determining health, it would make sense that having a larger family results in a lower health rating due to there being less money available for each individual. Being married and living with a spouse could possibly lead to a change in health, but that is not certain to be the case. It may be the case that being married allows people to split the amount of work to be done, both inside and outside the home. This could theoretically have some positive impact on health. On the other hand, there is little indication of how strong each marriage is. The only information we have as to the strength of the marriage is that the couple is living together. It is more likely that just being married is not enough to affect health, but being in a particularly strong or weak one is. The geography variables were dropped because they encompass such large and diverse regions that on average, people in each of the four regions should have similar rankings of health. In each section of the U.S., there are large cities, towns, and rural areas. Unless variations in health can be attributed to differences in the climate in each of the areas, these dummy variables may end up irrelvant.

In the next regression, I drop the *married* and *veteran* covariates. Intuitively, being a war veteran would cause one to have a lower health ranking. I then run the regression again but with veteran and the geographic location dummies dropped. Finally, because it is the variable with which this paper is most concerned, I run a regression dropping the *vitamin2wks* indicator. Also, since the variable of interest, *healthrank*, is given by ordered categories, we will also look at the results from an ordered probit estimation.

3.3 Results

The results of the five OLS regressions are reproduced in Table 3.1. T-statistics are included underneath estimated coefficients and significant estimates are marked with stars. Results of the ordered probit model follow.

From the first regression, specified by (3.1), we see that many of the covariates included are statistically significant even at the 1% level. The only variables not statistically significant at any level are *asian*, *raceother*, *married*, *vitamin2wks*. Veteran status had significant estimated coefficients in every specification in which it was included, all of them with a negative sign. This suggests that there are long term health concerns for all veterans, even those who did not fight in the front lines. Each of the geographic region dummies were significant and positive as well, implying that the south is less healthy than the rest of the country. With the exception of being African American, it does not appear that there are many differences in health among U.S. citizens of different races. Every other variable has a significant estimated coefficient with the expected sign. People who are younger, skinnier, more educated, with more income and smaller familes all feel healthier relative to their peers.

The results appear to be extremely robust. In each of the regressions where at least one variable was dropped, each statisically significant estimated coefficient remained significant and no coefficient became significant after another was dropped. In addition, the signifi-

Table 3.1: OLS results

	(1)	(2)	(3)	(4)	(5)
	healthrank	healthrank	healthrank	healthrank	healthrank
vitamin2wks	-0.0290	-0.0258	-0.0279	-0.0249	
	(-1.18)	(-1.05)	(-1.14)	(-1.02)	
age	-0.0154***	-0.0154***	-0.0158***	-0.0158***	-0.0154***
	(-20.45)	(-20.79)	(-21.83)	(-21.50)	(-20.55)
educationyears	0.0600***	0.0600***	0.0595***	0.0596***	0.0593***
	(13.51)	(13.55)	(13.46)	(13.46)	(13.46)
familysize	-0.0348***	-0.0356***	-0.0363***	-0.0349***	-0.0338***
	(-3.57)	(-4.05)	(-4.13)	(-3.58)	(-3.48)
income	0.131***	0.131***	0.129***	0.130***	0.130***
	(14.57)	(15.21)	(14.99)	(14.56)	(14.52)
height	0.0157***	0.0156***	0.0152***	0.0152***	0.0155***
	(3.65)	(3.65)	(3.56)	(3.56)	(3.62)
weight	-0.00268***	-0.00269***	-0.00271***	-0.00271***	-0.00267***
	(-9.17)	(-9.22)	(-9.28)	(-9.29)	(-9.13)
female	-0.133***	-0.133***	-0.107**	-0.108**	-0.137***
	(-3.81)	(-3.81)	(-3.24)	(-3.27)	(-3.94)
asian	-0.0655	-0.0508	-0.0586	-0.0444	-0.0638
	(-0.79)	(-0.62)	(-0.71)	(-0.54)	(-0.77)
black	-0.226***	-0.235***	-0.227***	-0.238***	-0.224***
	(-6.58)	(-6.99)	(-6.69)	(-7.01)	(-6.52)
raceother	-0.106	-0.0920	-0.106	-0.0933	-0.104
	(-1.41)	(-1.23)	(-1.41)	(-1.25)	(-1.38)
married	-0.00490			-0.00990	-0.00514
	(-0.17)			(-0.34)	(-0.18)
veteran	-0.0883*	-0.0864*			-0.0876*
	(-2.37)	(-2.33)			(-2.35)
northeast	0.0256		0.0273		0.0258
	(0.80)		(0.85)		(0.81)
midwest	0.0281		0.0293		0.0278
	(0.87)		(0.91)		(0.87)
west	0.0691*		0.0669*		0.0666*
	(2.07)		(2.00)		(2.00)
_cons	3.114***	3.142***	3.145***	3.170***	3.122***
	(10.41)	(10.54)	(10.53)	(10.65)	(10.44)
N	7767	7767	7767	7767	7767
R^2	0.207	0.207	0.207	0.207	0.207
adj. R^2	0.206	0.205	0.206	0.205	0.205

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

cant coefficients were significant at the same levels in every specficiation. The results from these regressions imply that the overall effect that taking vitamins has on health is neglible. However, the R^2 in each case remained between 0.217 and 0.218. This suggests that there is still a lot of ϵ that was left out of the model. We still cannot make the assumption that $\mathbf{E}(\mathbf{x}_i^T \epsilon) = 0$. Therefore, these estimates are possibly inconsistent and we make use of instrumental variables in the next chapter instead.

Estimates from the ordered probit regression in Table 3.2 display similar results. Using a multivitamin does not appear to have a significant effect on self-assessed health. With the exception of the *west* geographic variable, every significant covariate from the first OLS regression remained statistically significant in ordered probit. Table 3.3 contains the cutoff points for this ordered probit regression.

Table 3.2: Ordered Probit Estimated Coefficients

	(1) healthrank
vitamin2wks	-0.0263
	(-1.01)
age	-0.0151***
	(-18.88)
educationyears	0.0575***
	(12.21)
familysize	-0.0347***
	(-3.33)
income	0.147***
	(15.11)
height	0.0144**
	(3.18)
weight	-0.00260***
	(-8.56)
female	-0.154***
	(-4.14)
asian	-0.0740
	(-0.84)
black	-0.224***
	(-6.22)
raceother	-0.129
	(-1.63)
married	-0.0346
	(-1.13)
veteran	-0.133***
	(-3.37)
northeast	0.0251
	(0.74)
midwest	0.0197
	(0.58)
west	0.0686
	(1.92)

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3.3: Ordered Probit Cutoff Points

cut1	
_cons	-1.437***
	(-4.53)
cut2	
_cons	-0.653*
	(-2.06)
cut3	
_cons	0.284
	(0.90)
cut4	
_cons	1.086***
	(3.43)
N	7767

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Chapter 4

IV Estimation

4.1 County Size as Instrumental Variable

The premise of using a consumer's county size as an instrument lies primarily in the disparity in prices of consumer goods among the different areas. In this situation, the best instrument to use would be the price of multivitamins faced by each consumer in the data set. Although prices in a free market are determined by both the supply and demand of a good, each consumer takes the market price as given because he or she is unable to individually cause price fluctuations. It is for this reason that prices would be the ideal instrument; it is completely exogenous. However, data this specific about prices of an individual good is not available and so a proxy must be used. For this paper, a smaller county will refer to a county with a smaller population.

In theory, a place that is relatively rural will not have as many selection in terms of suppliers. Although each consumer in a smaller county may have a similar demand curve for multivitamins to people who live in more populated areas, suppliers are not as likely to ship their goods there. There are fewer modes of transportation available and the population demand is likely not as high as the population demand of a larger county. This would lead

to a lower overall price of multivitamins in rural areas. On the other hand, it is possible that because the demand for vitamins is lower, the equilibrium price ends up being lower as well. The population demand of a larger county will be comparatively high so suppliers are able to charge higher prices for multivitamins. This is not to say that individual suppliers are actively able to control the market price, only that the equilibrium price will be higher. Analogous to what was said before, it is also possible that there are more competitors on the supply side in the city, leading to an overall lower cost of vitamins.

The data set used in this section is the same one used in the previous chapter. There is an additional ordered variable, county size. Counties with less than 100,000 people are coded 1 and counties with 100,000 to 249,000 people are coded 2. Those with populations greater than 250,000 but less than 1,000,000 are coded 3 and counties with populations of more than 1,000,000 are coded 4. The county size codes are not scaled linearly, so it is best to group them into two categories and create a new dummy variable as the instrument.

From the data, we see that being in the highest and lowest populated counties was positively correlated with vitamin usage and being in the two medium populated counties was negatively correlated ($\rho_1 = 0.0135$, $\rho_2 = -0.0233$, $\rho_3 = -0.087$, $\rho_4 = 0.0164$). We generate the new dummy variable, *mediumcounty* to be 1 if the original county code was 2 or 3, and 0 otherwise.

4.2 Model Specification

Like the previous model, we are interested in the effects that vitamins have on an individual's health. However, we also now have a second reduced form equation for vitamin usage. Although ordinary least squares was used as a preliminary approach in the previous chapter, the measure of health given in the data is not necessarily linear. It would be incorrect to intuit that someone who ranked his or her health a 4 feels twice as healthy as another who

responded with a 2. The ranking system is not linear, but it is ordered. Therefore, an ordered probit model will be more appropriate here. Let $\mathbf{x}$ denote the column vector of covariates as defined before and $\mathbf{z}$ denote the column vector of exogenous covariates and county size. The model to be estimated the takes the form

$$health^* = \mathbf{x}^T\beta + \epsilon \tag{4.1}$$

$$vitamin2wks^* = \mathbf{z}^T\delta + \mu \tag{4.2}$$

where $health^*$ and $vitamin2wks^*$ are latent variables. Equation (4.1) is the structural model which expresses the dependent variable, $health$ in terms of the explanatory variables. Equation (4.2) is the reduced form which expresses the endogenous variable $vitamin2wks$ in terms of only the exogenous covariates from (4.1) and the instrumental variable. Note that while we observe the variable $vitamin2wks$, what is estimated by equation (4.2) is not the same thing. The same is true of $health$ and $health*$. Models of the form given by (4.1) and (4.2) are usually estimated with a two step regression. First the reduced form (4.2) is estimated and then its fitted values are substituted in for the endogenous variable in the structural model (4.1). Standard errors are then bootstrapped to correct the estimated standard errors. However, because vitamin usage is a binary variable, it also is not linear. Predicted values from a probit procedure on (4.2) will not be either 1 or 0. We use the equation

$$\hat{vitamin2wks} = 1[\mathbf{\Phi}(vitamin2wks^* > 0)] \tag{4.3}$$

where $\mathbf{\Phi}(\cdot)$ is the cumulative distribution function of the standard normal distribution to obtain predictions on vitamin use. Let $\mathbf{x}_1$ be the column vector of only exogenous covariates such that

$$health^* = \mathbf{x}_1^T\beta_1 + vitamin2wks\beta_k + \epsilon \tag{4.4}$$

with $\beta = [\beta_1^T \ \beta_k]^T$. We can define analogous variables δ_1^T such that $\delta = [\delta_1^T \ \delta_k]^T$ and $\mathbf{z}_1^T$ such that

$$vitamin2wks^* = \mathbf{z}_1^T\delta_1 + vitamin2wks\delta_k + \mu \tag{4.5}$$

but it should be noted that $\mathbf{x}_1$ and $\mathbf{z}_1$ are the same vector. In order for it to make sense for us to plug the values from (4.3) into (4.1), we need for

$$\begin{aligned}\mathbf{P}(health = 0|\mathbf{z}) &= \mathbf{P}(\mathbf{z}_1^T\beta_1 + vitamin2wks\beta_k + \epsilon < \alpha_1|\mathbf{z}) \\ &= \mathbf{\Phi}(\alpha_1 - \mathbf{z}_1^T\beta_1 - \beta_k\mathbf{\Phi}(\mathbf{z}^T\delta))\end{aligned} \tag{4.6}$$

to hold for some estimated cutoff value α_1. Equation (4.5) implies that

$$vitamin2wks = \mathbf{\Phi}(\mathbf{z}^T\delta) \tag{4.7}$$

but we know from (4.3) that this is untrue because $\hat{vitamin2wks} = 1[\mathbf{\Phi}(vitamin2wks^* > 0)]$. Since a two-stage regression is not possible with this data, we will use an automated full information maximum likelihood procedure instead. The *bioprobit* command is available to use for models where both dependent variables of both equations are discrete and ordered. Using maximum likelihood instead of a two stage regression also obviates the need to correct standard error estimates by bootstrapping.

4.3 Generated Interaction Variable as Instrument

Wooldridge (2010) suggested creating new interaction variables from known exogenous ones and using those as instruments to solve an endogeneity problem. These new variables have the advantage of being known to not be endogenous, but their usage is not necessarily justified

by economic theory. The covariates that are most obviously exogenous in the combined data set are age, $female$, and $height$. Creating three multiplicative interaction variables from these choices yield the conclustion that $age * female$ has the highest correlation coefficient with vitamin usage at 0.1431. The model then looks the same as before, with $age * female$ in place of $mediumcounty$. I will now switch over to the data set that includes information on both the NHIS in 1986 and NHANES in 2006.

4.4 Combined Data Set Description

Before going further, it may be helpful to discuss what data transformation was necessary in order to append one data set to the other. The NHANES did not include information on geographic location or county size, so it would have been impossible to use the appended set in the previous estimation. However, because the NHANES data was collected in a different time period than the NHIS data, appending the two sets provides information about the time trend of self-assessed health. Both data sets give categorical income information but the category ranges used are different. In addition, the ranges are defined in a way that makes it difficult to scale one in terms of the other. I apply the ceiling function to each of the observations and recode income in a way that it takes on the value of the highest possible value in its category. For example, someone who had an income of $10,000 to $14,999 is marked with a 15,000. Then I scale everything down by 1000 for easier interpretation of results. To distinguish observations in the different time periods, the dummy $year06$ was generated to have a value of 1 if the data was from 2006 and 0 otherwise.

4.5 Results

Table 4.1 give the estimated coefficients for each structural form estimation. The results here are much different than what was suggested by ordinary least squares. Both *bioprobit* regressions imply that vitamins actually do impact someone's health, or at least their own assessment of it. However, estimates of the other covariates are mostly aligned with those of the OLS regressions. Most of the coefficients that were statistically significant before have remained so here. Being a veteran is no longer significant, however. In terms of coefficient signs and significance levels, the two *bioprobit* procedures also performed very similarly to each other. Each variable that was statistically significant in the estimation with *mediumcounty* as an instrument was also statistically significant when $age * female$ was used instead. For the ones that are significant, the direction of the effect is the same in each case. Interestingly, the second regression suggests that there has been an overall decline in health in the United States. The estimated response probabilities of vitamin usage on health ranking are given as well in Table 4.2. Taking multivitamins is associated with increased probabilities in assessing health to be good, very good, or excellent and decreased probabilities in assessing health to be poor or fair. Using the interaction variable $age * female$ produces more extreme results than using *mediumcounty*. The estimated increase in probability of ranking one's health as excellent is 14.67% when $female * age$ is used which is almost twice the estimated increase when *mediumcounty* is used.

Coefficients for the reduced forms are given in Table 4.3 and cutoff points are shown in Table 4.4. Because reduced forms specify an endogenous variable from the structural model in terms of only known exogenous factors, they are by definition free from their own endogeneity issue. Because the estimated coefficients for the instrument in each reduced form is small, there may be cause to be wary that the instruments are weak. Bound et al. (1995) showed how this issue leads to inconsistent estimators in their critique of Angriest

and Krugers paper on using quarter of birth as an instrument for education. While this is something to keep in mind, it is still important to acknowledge that some sort of instrument is necessary when dealing with an endogenous explanatory variable. Any conclusions that are drawn from this study in regards to the effect of vitamins on health should be done keeping possible weakness of the instrument in mind.

Table 4.1: Structural Model Estimates

	(1) Estimates of structural form using mediumcounty as instrument		(2) Estimates of structural form using age*female as instrument
education	0.0690***	education	0.116***
	(8.97)		(17.53)
familysize	-0.110***	familysize	-0.129***
	(-8.61)		(-13.51)
income	0.110***	income	0.00996***
	(7.40)		(13.10)
height	0.0203***	height	0.0269***
	(3.43)		(5.85)
weight	-0.00183***	weight	-0.00440***
	(-4.22)		(-10.59)
female	0.369***	female	0.333***
	(7.69)		(8.98)
asian	-0.0878	veteran	0.0209
	(-0.82)		(0.56)
black	-0.353***	year06	-0.0939*
	(-7.30)		(-2.08)
raceother	-0.196	married	0.153***
	(-1.93)		(5.91)
veteran	-0.0559	vitamin2wks	0.433***
	(-1.09)		(3.51)
married	0.00395		
	(0.10)		
vitamin2wks	0.211*		
	(2.57)		
N	7767	*N*	14627

t statistics in parentheses
$^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

Table 4.2: Estimated Partial Effects of Taking Vitamins on Health Ranking

	Estimated partial response probabilities of vitamin2wks on self-assessed health ranking IV: MediumCounty	Estimated partial response probabilities of vitamin2wks on self-assessed health ranking IV: Female * Age
$\frac{\partial P(\text{health} = 1)}{\partial \text{vitamin2wks}}$	-.0822001	-.151929
$\frac{\partial P(\text{health} = 2)}{\partial \text{vitamin2wks}}$	-0.005213	-0.0507543
$\frac{\partial P(\text{health} = 3)}{\partial \text{vitamin2wks}}$	.0008089	0.0047076
$\frac{\partial P(\text{health} = 4)}{\partial \text{vitamin2wks}}$	0.0023811	0.0512754
$\frac{\partial P(\text{health} = 5)}{\partial \text{vitamin2wks}}$	0.0830091	0.1467002

Table 4.3: Reduced Form Estimates

	(1) Estimates of structural form using mediumcounty as instrument		(2) Estimates of structural form using age*female as instrument
education	0.0690***	education	0.116***
	(8.97)		(17.53)
familysize	-0.110***	familysize	-0.129***
	(-8.61)		(-13.51)
income	0.110***	income	0.00996***
	(7.40)		(13.10)
height	0.0203***	height	0.0269***
	(3.43)		(5.85)
weight	-0.00183***	weight	-0.00440***
	(-4.22)		(-10.59)
female	0.369***	female	0.333***
	(7.69)		(8.98)
asian	-0.0878	veteran	0.0209
	(-0.82)		(0.56)
black	-0.353***	year06	-0.0939*
	(-7.30)		(-2.08)
raceother	-0.196	married	0.153***
	(-1.93)		(5.91)
veteran	-0.0559	vitamin2wks	0.433***
	(-1.09)		(3.51)
married	0.00395		
	(0.10)		
vitamin2wks	0.211*		
	(2.57)		
N	7767	*N*	14627

t statistics in parentheses
$^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

Data Set: NHIS

t statistics in parentheses
$^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

Data Set: NHIS NHANES appended

Table 4.4: Cutoff Value Estimates

	(1) Estimates of probit cutoff values using mediumcounty as instrument		(2) Estimates of probit cutoff values using age*female as instrument
cutoff for reduced form	2.024^{***}	cutoff for reduced form	2.450^{***}
	(5.34)		(8.72)
cutoff1 for structural form	-0.415^{**}	cutoff1 for structural form	-0.712^{***}
	(-2.94)		(-4.67)
cutoff2 for structural form	-0.221^{*}	cutoff1 for structural form	-0.228
	(-2.10)		(-1.65)
cutoff3 for structural form	0.0113	cutoff1 for structural form	0.310^{*}
	(0.13)		(2.25)
cutoff4 for structural form	0.210^{*}	cutoff1 for structural form	0.762^{***}
	(2.12)		(5.12)
N	7767	*N*	14627

t statistics in parentheses
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

t statistics in parentheses
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Chapter 5

Difference in Differences Estimation

5.1 Dietary Supplement and Health Education Act

Data shows that vitamin usage became much more common in U.S. citizens from 1986 to the early 2000s. In 1994, the Dietary Supplement and Health Education Act (DSHEA) was passed through Congress. This act was responsible for much of this increase (Newhouser 2003). It made it much easier for companies to put their vitamin supplements on the market and keep them there. Although companies still had to get their multivitamins tested to prove that they were safe for consumption before they were placed on the shelves, the Food and Drug Administration would take responsibility thereafter. So as long as a company could get initial approval by the FDA, it would be free from the burden of proof afterwards. DSHEA also broadened the definition of a vitamin supplement, making it easier to market herbs and other amino acids as such. This change in the environment makes a difference in differences approach appropriate. The data set used will be the same as the combined 1986 NHIS and 2006 NHANES set used in the second *bioprobit* procedure. We will consider the effect of being in the treatment group in the time period after the treatment is in effect, assuming that income levels will also cause differences in health rankings. The difference-in-differences

estimation will be implemented as an ordered probit.

5.2 Model Specification

Since the price of vitamins was reduced for everyone in the country, it is difficult to pick a control group and a treatment group. However, it would make sense to believe that those who are less wealthy are the ones who benefited most from the DSHEA. After a certain level of income, anyone who wants to use supplements can buy them. A price decrease would not cause someone who is already using vitamins to buy more. I generate a new dummy variable *notwealthy*, which is equal to 1 if an observation's *income* is in the top income bracket of the time period and 0 otherwise. This will be our treatment group and being wealthy will be our control group. Then the model to be estimated is

$$healthrank = \gamma_0 + \gamma_1 yr2006 + \gamma_2 notwealthy + \gamma_3 yr2006 * notwealthy + \mathbf{w}^T\gamma + \rho \quad (5.1)$$

and γ_3 is our coefficient of interest and $\mathbf{w}$ is the vector of all other covariates.

5.3 Results

Table 5.1 displays the results of this analysis. The time dummy is now a proxy for the environmental change caused by DSHEA and this result suggests that the Act was significant in increasing self-assessed health ranking. It may be of interest to note that these results also suggest that health really has declined in the United States in the past two decades. Data is limited in this situation, but a more ideal difference in differences procedure would look at data on health and income level in the years 1994 and 1995. Even though the DSHEA was undoubtedly responsible for some of the increase in supplement usage, the 20 year gap

Table 5.1: Difference-in-Differences Estimates

	(1) healthrank
yr2006*notwealthy	0.158**
	(2.92)
age	-0.0133***
	(-23.78)
education	0.0971***
	(28.07)
familysize	-0.0394***
	(-5.71)
height	0.0232***
	(6.84)
weight	-0.00436***
	(-16.56)
married	0.153***
	(7.70)
veteran	-0.0359
	(-1.25)
female	-0.122***
	(-4.63)
notwealthy	-0.354***
	(-8.84)
yr2006	-0.607***
	(-12.11)
cut1	
_cons	-1.206***
	(-5.09)
cut2	
_cons	-0.294
	(-1.24)
cut3	
_cons	0.720**
	(3.04)
cut4	
_cons	1.574***
	(6.65)
N	14627

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

between the time periods is possibly too large for interpretation to be reliable; it is uncertain how much of the increase in popularity of vitamins can be attributed to the Act.

Chapter 6

Discussion

Whether vitamins really have a positive effect on a person's health and not just their perception of it is another question to be addressed. Unfortunately, there does not exist a completely objective measurement of a person's overall health. The closest approximation may be a ranking similar to the ones used in the NHIS and NHANES but given by a doctor. Even this does not completely take away bias, however. It would also be difficult to standardize the rankings given by each doctor. Even mortality is not necessarily a good measure of overall health because it doesn't take the quality of life into account. As it remains, people buy vitamins because they believe it is beneficial. Whether they really do feel a difference in their quality of life can conceivably be examined through a self-assessment of health.

There is no reason to doubt that vitamins and supplements are necessary when an individual is lacking nutrients. But in a country such as the United States, it is unlikely that someone is so malnourished that their health has been negatively impacted. In the same line of thinking, in order to accurately assess the impact that multivitamins have on health, we need to know more about the individuals' own baseline blood levels. Ideally, the data set would provide information about the nutrients already inside the body. Certainly an iron supplement would have different effects on someone who is severly lacking the mineral in his

blood and someone else who has a perfectly healthy concentration of it.

Effects of individual vitamins and minerals would be difficult to assess with any type of data set. Most supplements are taken in the form of a multivitamin and multicollinearity would make the individual coefficients hard to identify. Then another possible issue is that some vitamins and minerals have different effects on the body depending on whether they are taken alone or with other nutrients.

In addition to the information on nutrients already in the blood, it would be best to use a data set from the years 1994 to 1995 to more accurately capture the effects of DHSEA. This would make a difference in differences approach much more reliable. Although I used one set from before the act was passed and another one from afterwards, there is still a twenty year difference in the time periods. DHSEA cannot have been the only event to have impacted vitamin usage for the past two decades. As stated before, the ideal instrument to use for vitamin usage is the price faced by each individual, but that was not available.

Chapter 7

Conclusion

It is still not possible to say with certainty that vitamins are beneficial to a person's health. Two instrumental variable procedures provided very similar results and suggest that there are some positive effects from taking a multivitamin. The difference-in-differences estimation suggested a similar result, but the two time periods used are possibly too far apart for a reliable conclusion to be draw. One, however, has a relatively low correlation coefficient with vitamin usage and the other one is not based in economic theory. While the results are not conclusive, they should not be dismissed entirely either. No other non-experimental study on this topic has addressed the issue of endogeneity by using an instrumental variable. More importantly, the methodological issues raised in this paper must be considered by future researchers. I did not have access to data that are available to other institutions so there was some restriction in the instruments I could choose. However, it is enough for me to conclude that there is possible merit to taking multivitamin supplements.

Bibliography

[1] Bound, John, David Jaeger, Regina Baker "Problems with Instrumental Variables when the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak." *Journal of the American Statistical Association* 1995.

[2] Gahche, Jaime. "Dietary Supplement Use Among U.S. Adults Has Increased Since NHANES III (19881994)." *NCHS Data Brief 61.* 2011.

[3] Godfrey, Jody R. "Toward Optimal Health: Meir Stampfer, M.D., Dr.P.H., Discusses Multivitamin and Mineral Supplementation for Women." *Journal of Womens Health (15409996) 16, no. 7: 959-962.* 2007.

[4] Macpherson, Helen, Kathryn Ellis, Avni Sali, and Andrew Pipingas. "Memory improvements in elderly women following 16 weeks treatment with a combined multivitamin, mineral and herbal supplement." *Psychopharmacology 220, no. 2: 351-365.* 2012.

[5] Mursu J, Robien K, Harnack LJ, Park K, Jacobs DR. "Dietary Supplements and Mortality Rate in Older Women: The Iowa Womens Health Study. " *Archives of Internal Medicine* 2011.

[6] Neuhouser, Marian L. "Dietary Supplement Use by American Women: Challenges in Assessing Patterns of Use, Motives and Costs." *The Journal of Nutrition 133.6* 2003.

[7] Rock, Marian L. "Multivitamin-multimineral supplements: who uses them?" *The American Journal of Clinical Nutrition vol. 85 no.1* 2007.

[8] Sajaia, Zurab. "Maximum likelihood estimation of a bivariate ordered probit model: implementation and Monte Carlo simulations" *The Stata Journal vv, Number ii, pp. 118.*

[9] Song, Yiqing, Qun Xu, Yikyung Park, Albert Hollenbeck, Arthur Schatzkin, and Honglei Chen. "Multivitamins, individual vitamin and mineral supplements, and risk of diabetes among older U.S. adults. "*Diabetes care 34, (1): 108-14.* 2011.

[10] Wooldridge, Jeffrey M. "*Econometric Analysis of Cross Section and Panel Data* MIT Press, 2010. Print.