How Dietary Behaviors Effect Income

By Noam Rosenberg

Abstract

This paper describes how Dietary Behaviors can effect the total yearly income of an individual. I find that while shopping for groceries, a person's inclination to check the label of the product prior to purchase can have a 5.68 % effect on Income. I attempt to rule out reverse causation by implementing the Hausman Statistic to test the Instrumental Variables against the Predictive/Explanatory Variable in my model.

Introduction

This paper seeks to investigate whether Dietary Behaviors can effect an individual's yearly Income. I start off by examining three possible candidates to be explanatory variables via a log linear model for which income is the dependent variable. 1. "TIMES ATE FOOD FROM A FAST FOOD RESTAURANT IN THE PAST 7 DAYS" 2. "DOES R READ NUTRITIONAL INFORMATION WHEN SHOPPING FOR FOOD" 3. "DOES R READ INGREDIENTS WHEN SHOPPING FOR FOOD".

So what might be the connection between Dietary Behaviors and Income? It could be that Healthy Behaviors in general are connected to Income. Vasilios D. Kosteas in his paper "The Effect of Exercise on Earnings: Evidence from the NLSY " finds that engaging in regular exercise yields a six to ten percent wage increase. I was unable to account for this effect in my model, meaning my results could be endogenous and this third variable which exists inside the error of my model could account for a portion of my results. Healthy Dietary Behaviors would likely be positively correlated with other healthy behaviors such as regular exercise which in turn we know has a positive effect on income. Not all hope is lost though. We know from the afore mentioned paper that the effect is between six to ten percent which in a log linear model means a relatively small coefficient of 0.06 – 0.1. This is a small effect, and multiplied by the coefficient that Dietary Behaviors would have when regressed on Regular Exercise which in itself should be quite small relieves us of some of our worries. If this situation proves to be accurate, I expect a minuscule positive bias which was likely already rounded off when presenting my results.

So How about the reverse effect, could then a higher Income then be causing people to behave in a healthier manner as per their diet? By examining milk products as well as soft drinks, where healthier, low calorie products are cheaper or the same price as those options considered less nutritious "An Exploration of the Relationship Between Income and Eating Behavior" uses a metric they call "energy density", which is amount of calories divided by amount of grams of product consumed in a two day window, as the dependent variable. The results show that people with a lower household income are consuming much more calories per gram of product. This is considered much less nutritious in the context of the US and similar western countries where obesity is considered the source of many medical

issues as well as a major health problem in itself. The study suggests that cost does not play an important role in why low income people choose not to eat nutritiously. Instead the study suggests that low income people are less willing to undergo the cost of giving up less tasty products today because future health gain may be less valuable to them.

However, this study fails to account for some important variables, one of which is “Health Limitations”. I find that Health Limitations are responsible for a 47% decrease in Income, when in a log linear model the coefficient appears as -0.63, this is a substantial effect. It also seems highly likely that “Health Limitations” have a strong positive effect on the dietary behaviors addressed in the study. This then would cause a strong positive bias when measuring the effect of Income on Dietary Behaviors, and would be the cause of a highly distorted coefficient. Moreover, this study fails to rule out reverse causation via statistical methods such as the Hausman coefficient which I use thus de legitimizing their results even more.

Methodology

Be it because of omitted variable bias or because of reverse causality, endogeneity has proven itself to be of some concern in my model. An omitted explanatory variable would cause my variable to be correlated with the error of the model, creating a biased estimator of how Dietary Behaviors affect Income. A reverse relationship where Income effects Dietary Behaviors would also create a correlation with the error again rendering our estimator biased. At first I use an Ordinary Least Square Regression model, but in an attempt to deal with the risk endogenous variables and biased estimators I later employ Instrumental Variables via the 2 Stage Least Squared Regression method. Instrumental Variables must not be directly independently correlated with the dependent variable, but they must be sufficiently correlated with the suspected endogenous explanatory variable. For this I begin by using a dummy variable which describes if the person in question has ever been diagnosed with Diabetes or high blood sugar, my reasoning being that people who have been diagnosed would take extra care of what they eat, creating a strong correlation with variables associated with Dietary Behaviors, yet should be uncorrelated to income directly, at least not to any significance. Next I employ a dummy variable that measures adverse or allergic reaction to serums drugs or medicine. I would have preferred a variable measuring allergic reactions to specific foods for obvious reasons, but I settle for this. Allergic reactions should be uncorrelated to Income while being somewhat correlated with Dietary Behaviors, surprisingly while later testing this I find the correlation to be somewhat significant for reasons I can’t quite explain. I start out by inserting each instrumental variable into an Ordinary Least Squared Regression predictive model of Income. I do this to check for a direct correlation with Income and make sure my Instrumental variable has no place in my original model. Once I make sure that my Instrumental Variable is not directly and independently correlated with Income I later run the first stage of the 2 Stage Least Squared Regression to measure the level of correlation with my suspect endogenous variable and to test its significance. Once I’m satisfied that the variable fits the conditions of being an Instrumental Variable I finally run a 2SLS Regression. Once this is done I should be able to get a sense of whether my Dietary Behaviors variable is in fact exogenous to Income by hypothesis testing it for endogeneity against my Instrumental Variables via the Hausman statistic.

Data

This paper makes use of the data from the 2012 Wave of the National Longitudinal Surveys of Youth 1979 dataset (NLSY79). The NLSY79 conducted surveys every year from 1979 to 1994, from then on the survey was conducted only in even numbered years up to 2012. The survey began following Dietary Behaviors in 2002, however the omitted variable bias mentioned in the intro of this paper of physical activity is present only in year 1998 and 2000. The initial sample size is 12686 people. The people are excluded if they were not present or did not answer the survey for reasons which seem to be invalid to the author of this paper, and the author had no way to replace or estimate those missing values. This was evaluated for each survey question individually. Other factors including link function necessities and variable values that were deemed incompatible left the final sample size as 4960.

People with valid reasons for not being present or giving an answer had the values adjusted accordingly and as seen fit. When asked about the "NUMBER OF WEEKS WORKED SINCE LAST INTERVIEW" those individuals who were not even asked the question because it did not apply to them, were regarded by the author as people who are currently not in the job market https://www.nlsinfo.org/content/cohorts/nlsy97/using-and-understanding-the-data/nlsy97-documentation). These values were adjusted from -4 to 0 for the purpose of later being summed up to a single variable expressing total work experience.

When comparing two similar variables, one labeled "AGE OF R AT INTERVIEW DATE" the other "AGE OF RESPONDENT", the latter is ruled out for containing less information and for having valid skips were the first contains viable information.

When summing up the total work experience of an individual I had to account for missing information for reasons not deemed valid, to account for this I created a yearly average of weeks worked, for every individual user, assigning a weight twice as large for those years only surveyed every other year, then I replaced that value with the invalid one mentioned above. This allowed me to create a model for income with a variable I call work experience given as number of weeks worked in lifetime.

I changed weight and Height measured by Pounds and Feet to Metric Units. "SAMPLE RACE" was changed from a categorical variable to two dummy variables "Black" and "Hispanic". "Eye Color" was transformed into a single dummy variable "Light Eyes", all those with light colors were given the value 1, and all others 0.

Five people had valid reasons for not being asked about "Health Limitations relating to work" - ((([Would your health keep you from working now?]==1) OR ([Limited in kind of work due to accident or injury?]==1) OR ([Limited in amount of work due to accident or injury?]==1)), other == 0). The author assigned these five people with a 0 value.

A BMI variable was constructed using data from "Height" and "Weight". Using visualization techniques "Age" and "Experience" were found to have parabolic connection with "Income" and "Log Income", and as such "Age squared" and "Experience (weeks) squared" variables were constructed. The variable "Income" was log transformed into "Log Income". This also meant I had to drop all people with no income last year, people who made 0 income.

For Dietary Behaviors I looked at three different specific dietary behaviors. “TIMES ATE FOOD FROM A FAST FOOD RESTAURANT IN THE PAST 7 DAYS” has values ranging from 0 to 14 and 15+, this seems to be an accurate border as only 4 people reported eating fast food 15 times or more.

The next variable is categorical “R READ INGREDIENTS WHEN SHOPPING FOR FOOD?” And has 5 values Always, Often, Sometimes, Rarely, Never and Don’t buy food. Thankfully only 47 people reported not buying food at all, I dropped all individuals reporting “Don’t buy food”, and that goes for the next variable as well “R READ NUTRITIONAL INFORMATION WHEN SHOPPING FOR FOOD?”, where 50 individuals reported “Don’t buy food”. These two variables were converted to both Ordinal Variables ranging from 1 to 5, 1 being the lowest and 5 the highest, and to dummy variables, were “Always” and “Often” categories are transformed to 1 and all the rest to 0.

The “DOCTOR EVER DIAGNOSED DIABETES OR HIGH BLOOD SUGAR?” and “ADVERSE/ALLERGIC REACTION TO SERUM, DRUG OR MEDICINE” are both Dummy Variables with 1 indicating Yes and 0 indicating No. They are both Instrumental Variables I will use to search for signs of endogeneity in my model. First I will use them separately the again together.

Results

In Table 1 you will see the results of me running an OLS Log Linear Regression Model. Interestingly enough only the dummy version of “R READ NUTRITIONAL INFORMATION WHEN SHOPPING FOR FOOD?” had significant results in this regression model. Other variables I collected which were insignificant in the model were also siphoned out. The results show that reading nutritional information off the label whilst shopping for food with a frequency of “Often” or more will increase income by approximately 5.7%. This is only one specific kind of Dietary Behavior, adding up different types of Dietary Behaviors could show the effect of Dietary Behaviors on Income to be quite meaningful.

Table two shows the most significant results the author could find for the first of the 2 Stage Least Squared Regression. When checking the significance of the two variables “DOCTOR EVER DIAGNOSED DIABETES OR HIGH BLOOD SUGAR?” and “ADVERSE/ALLERGIC REACTION TO SERUM, DRUG OR MEDICINE” via a F test, the significance to the first stage model is 0.063 which would be considered significant only if we’re being generous and rejecting the null hypothesis that those two variables don’t belong in the first stage model at a 10% significance level. However, I choose to include these results anyway because checking for endogeneity via the Hausman statistic can’t hurt, since at this point I’m looking for any sign of endogeneity. Getting a significant result via the Hausman test, would allow me to reject the null hypothesis that “R READ NUTRITIONAL INFORMATION WHEN SHOPPING FOR FOOD?” is exogenous to “Log Income” and as such assume endogeneity. However, In this case preforming the Hausman test leaves me with extremely insignificant results. Meaning the author of this paper was unable to find any statistical reason to believe that Dietary Behaviors suffer from reverse causality or omitted variable bias in the original OLS model. This decision must now be made solely on a theoretical basis.

Conclusions

Unable to find any sign of endogeneity in my model I must admit that I only did this out of a realization that I must show at least the level of skepticism as any of the readers of this paper. Truthfully I do not believe there to truly be a theoretical basis for a reverse relationship between Dietary Behaviors and Income. Even the research paper, "An Exploration of the Relationship Between Income and Eating Behavior", fails, in my opinion to offer a convincing theoretical explanation for why Income effects Dietary Behaviors. Simply stating that low income people place less value on their future selves health, offers no underlying explanation of this effect. Why does a person's Income level cause them to place less value on their future state of health? The study does not attempt to answer this question, and also fails to take into account explanatory variable which have strong theoretical basis for being in the model such as BMI, current health state as well as different health limitations, all of which could render the results insignificant. The author of this paper racked his brain for an answer to the above question as well as the more general question of how Income may effect dietary behaviors, but without success. Further thought should be given to this matter though before being ruled out.

Variables not accounted for in my model should only generate a very miniscule Omitted Variable Bias due to what should be relatively small effects of independent on dependent variables and small correlations between explanatory variables. However, data should be collected if we wish to truly rule out the effects of Omitted Variable Bias.

This study has come to an end, but the author would like to encourage any readers to further ponder and explore this issue. Also, if there is more relevant data out there it should be explored. The author was unsuccessful when looking for statistical reasons to reject the hypothesis that Dietary Behaviors are exogenous to Income, but more data could potentially supply us with different results, even if we have not yet found the theoretical basis to explain them.

Refrences

"The Effect of Exercise on Earnings: Evidence from the NLSY" - Vasilios D. Kosteas Cleveland State University - http://academic.csuohio.edu/kosteas_b/Exercise%20and%20Earnings.pdf

An Exploration of the Relationship Between Income and Eating Behavior - Susan E. Chen, Jing Liu, and James K. Binkley - http://ageconsearch.umn.edu/bitstream/123315/2/liu%20-%20current.pdf

To further expore my analysis please check out my portfolio - https://github.com/NoamRosenberg/Portfolio/blob/master/Dietary%20Behaviours.ipynb

Table 1:

OLS Regression

Dep. Variable:	Log Income	**R-squared:**	0.3288
	coef	**std err**	**P>|t|**
const	28.9285	7.092	0.000
Black	-0.1634	0.029	0.000
Male	0.3834	0.030	0.000
Highest Grade	0.1250	0.005	0.000
Health Limitations	-0.6347	0.042	0.000

Height	0.2 125	0.12 3	0. 0 8 4
Reads Nutri	0.0 553	0.02 6	0. 0 3 3
Marie d-ish	0.0 832	0.02 6	0. 0 0 2
Age	- 0.8 060	0.27 6	0. 0 0 3
Age square d	0.0 073	0.00 3	0. 0 0 7
Numb er of Jobs	- 0.0 220	0.00 2	0. 0 0 0
Experi ence (weeks)	0.0 028	0.00 0	0. 0 0 0

Experience (weeks) squared	-1.07e-06	1.08e-07	0.000
BMI	0.0043	0.0002	0.015

Table 2:

First Stage, “Allergy” & “Diabetes” regressed on “Reads Nutri”

Dep. Variable:	Reads Nutri	R-squared:	0.090
	coef	std err	P>|t|
const	6.3677	3.954	0.107
Black	-0.0440	0.016	0.006

Male	-0.1659	0.016	0.000
Highest Grade	0.0412	0.003	0.000
Health Limitations	0.0703	0.023	0.002
Height	0.0684	0.067	0.310
Maried-ish	0.0064	0.014	0.657
Age	-0.2650	0.154	0.085
Age squared	0.026	0.001	0.079

Number of Jobs	0.0033	0.001	0.001
Experience (weeks)	0.004	0.000	0.002
Experience (weeks) squared	-1.316e-07	5.92e-08	0.026
BMI	0.002	0.001	0.873
Allergy	0.803	0.055	0.147
Diabetes	0.833	0.045	0.066

Allergy = Diabetes = 0

F test: F=array([[2.76153202]]), p=0.0632923384849274, df_denom=4945, df_num=2>

Table 3:

2SLS Regression of “Reads Nutri” on “Log Income”

Instrument: “Allergy”, “Diabetes”

Instrumented: “Reads Nutri”

Dep. Variable:	Log Income	R-squared:	0.267
	coef	stderr	P>|t|
const	25.4162	9.0333	0.005

Black	-0.1397	0.0466	0.003
Male	0.4752	0.139	0.001
Highest Grade	0.1023	0.034	0.002
Health Limitations	-0.6741	0.073	0.000
Height	0.1722	0.141	0.223

Reads Nutri	0.6055	0.8100	0.4555
Maried-ish	0.0795	0.0028	0.0005
Age	-0.6592	0.3600	0.0067
Age squared	0.0058	0.0004	0.0099
Number of Jobs	-0.0239	0.0003	0.0000

Experience (weeks)	0.0026	0.0000	0.0000
Experience (weeks) squared	-9.3466e-07	1.55e-07	0.0000
BMI	0.0042	0.0002	0.027