Senior Project
Department of Economics

"Dollar by the Pound: A Replication Study on How BMI Affects Wages"

Hannah Kretch

May 2013

Advisors: *Francesco Renna*

Abstract

I replicate Cawley (2004) to see if body mass index (BMI) negatively affects hourly wages for white females. Obesity is one of the variables of interest in this replication study, which will be a binary variable indicating if the respondent has a BMI equal greater than or equal to 30. I question whether the coefficient for dummy variable is negative in the wage equation. The study sample includes white females from the NLSY 1979 cohort. I use a 2-Stage Least-Squares model to address any endogeneity. I look at the sibling's BMI of the respondents as the instrumental variable in order to properly conduct the 2-Stage Least-Squares. By replicating Cawley (2004) with white females in 2000, I do get similar results in the OLS regressions, however lower results in the 2-Stage Least-Squares models.[1]

[1] I would like to thank Dr. Francesco Renna and Dr. Zheng Fang for their guidance in writing this research paper.

Table of Contents

I. Introduction

According to a survey conducted in 2009-2010 survey conducted by the National Health and Nutrition Examination survey, 66.8% Americans over the age of 20 are considered to be either overweight or obese. 35.7% of adult Americans over the age of 20 considered obese, given by the same source. People are more likely to be obese than smoke cigarettes (Baum and Ford 2004). A wage penalty will be examined to see the relationship between body mass index (BMI) and wages. Body mass index is a measure of a person's body shape that takes into account the person's height and weight. I calculate a person's body mass index (BMI) using weight in kilograms divided by height in meters squared. If BMI is greater than or equal to 30, the obesity indicator is equal to 1, otherwise it is zero. Cawley (2004) extends the literature to account for the endogeneity of BMI. The central research question is:

Does BMI cause pay differences in wages for white females in the workforce?

This research question will examine the relationship between these two variables, and control for typical covariates in the female wage equation such as marital status, region, work experience, ect. My main hypothesis created from Cawley's (2004) model is:

Using a 2SLS model, having a higher BMI will cause a decrease a white female's hourly wages.

I replicate Cawley (2004) by using data from the National Longitudinal Survey of Youth 1979 cohort (NLSY79), but will only be looking into the year 2000. The motivation of this study is to bring awareness to Americans on how one's aspect of health can positively or negatively affect wages, and to produce replicated findings by observing the effects that BMI has on wages for females using the model in Cawley (2004). The awareness on how BMI can affect the wages will encourage a healthier lifestyle for Americans who are motivated to succeed in their career. Weight has not been a topic that has commonly been brought up in such lectures; education and experience are the two main attributes people are taught to seek out when furthering their career. By looking at the findings from this study, it is questioned if an aspect of one's health can contribute to an increase in hourly wages.

The paper will proceed as follows: Section II will review the literature. Section III will present a theoretical and empirical model. Section IV will introduce the data. Section V will discuss the results. Section VI will discuss the limitations and conclude. Section VII will list references. Finally, Section VIII will list all of the tables in the Appendix.

II. Literature Review

Cawley (2004) presents a large data set from the NLSY79 to estimate the effects of the indicator variable *BMI* has on wages. By using an OLS estimation and a 2-Stage Least Squares, Cawley (2004) discovers that white obese females have lower wages. A weight increase of two standard deviations, or around 65 pounds, is associated with a 9% reduction in wages. The theoretical model utilized is the Mincer equation. Cawley utilizes an instrumental variable, *Sibling's BMI*, to account for endogeneity in the 2-Stage Least Squares model. The instrumental variable is related to the respondent's BMI, but not the respondent's wages.

Averett and Korenman (1996) conduct a similar study with a sample from the NLSY79 for the year 1988 with ages ranging from 23-31 years. They examine the relationship of BMI and wages in the labor market, marriage market, and for the market with the remaining respondents. In addition to hourly earnings, Averett and Korenman (1996) find differences by weight status in family income, marital status, and spouses earnings. Averett and Korenman (1996) use simultaneous relations between economic status and body weight in order to account for endogeneity. Lastly, Averett and Korenman (1996) compare same-sex sibling's weight to address bias from heterogeneity. Averett and Korenman (1996) raise awareness of the social and psychological pressures, and how it can be linked to eating disorders. Averett and Korenman (1996) estimate the relations between the economic status in 1988 and a BMI measure. The BMI measure is taken by computing the average of the BMIs in 1981 and 1982, when sampled members were between the ages of 16-25 years. The results show evidence of discrimination in the labor market against obese women, however there is no evidence of a difference in obese African American female's economic statuses compared to other African American women (Averett and Korenman 1996). Averett and Korenman (1996) also concluded that 50-85% of lower economic statuses were accounted from the differences in marriage probabilities and in spouse's earnings.

Brunello and D'Hombres (2007) conduct a study using the European Community Household Panel to observe the impact of weight on wages with nine European countries. From their study, a 10% increase in the average BMI will reduce the earnings for males by 3.27% and females by 1.86%. The negative effect of BMI on wages is statistically significant (Brunello and D'Hombres 2007).

Wada and Tekin (2009) used data on bioelectrical impedance analysis (BIA) to measure the body composition that is made up of body fat and fat-free mass. With information from NHANES III and samples from the NLSY79, Wada and Tekin (2009) use body composition to estimate the wage equation. Wada and Tekin (2009) retrieve parameter estimates from the NHANES III by having two separate regressions for fat-free mass and body fat. They estimate separate models for males and females to account for gender differences. The R-squared values are found to be high for both genders (Wada and Tekin 2009). The Mincer equation is utilized as the theoretical model within this study. Wada and Tekin (2009) also account for unobservable family-specific factors by using same-sex sibling. The results showed that the body-fat is associated with a decrease in wages for white males and white females.

Baum and Ford (2004) use data from NLSY79 cohort and exclude respondents that have not completed their education, are younger than 18 years of age, are self employed respondents, and are in the armed forces. Four separate approaches are used within this study. The first method estimates a wage model by using person-year observations. The second method estimates a fixed effect model for an individual level. Baum and Ford (2004) have each respondent report two wage observations, and they are averaged out for the fixed effect model. The third model assumes that heterogeneity at a given time is family specific. This model allows the unobserved heterogeneity to change over time (Baum and Ford 2004). The final approach is the combination of the second and third models, by comparing respondent's wages with their sibling's wages over time. The obesity wage penalty is found to be consistent over the first 20 years of a person's career (Baum ad Ford 2004). Job discrimination, health-related factors and/or obese workers' behavior patterns all contributed to the wage penalty.

III. Model

The theoretical model used behind Cawley (2004) and my replicated model is from the Mincer Equation. The Mincer equation is the human capital earnings function, which takes the natural log of *Earnings* as a function of years of *Work Experience*, years of *Work Experience* squared, and years of *Education.*

This paper adopts the empirical model in Cawley (2004):

$$(1)\ \ln W_i = B_i\beta + X_i\gamma + \varepsilon_i$$

The variable ε is the residual component, and the variable X represents a vector of variables that have an effect on wages (Cawley, 2004). The variable B represents the variable of interest for each respondent *i*. Depending on which model is being run, *B* can represent either *BMI, Obese, Overweight, Underweight,* or *Weight*. Because BMI can be affected by wages and personal characteristics, Cawley (2004) creates the equation for *B*:

$$(2)\ B_i = X_i\gamma + W_i\alpha + Z_i\phi + NG_i^B + \xi_i$$

Within Equation (2), *W* continues to represent wages, and *X* is the same vector of variables. *Z* is a vector of variables that do not directly affect wages but have an effect on BMI. G^W is the influence of genetics. NG^W is the non-genetic factors on BMI. Finally, ξ represents the residual BMI (Cawley, 2004). The equation below represents the components of ε (Cawley, 2004):

$$(3)\ \varepsilon_i = G_i^W + NG_i^W + v_i$$

The residual ε has three components: the genetic component G^W, the non-genetic component NG^W, and a residual v that is i.i.d. over individuals. The pitfalls of the OLS estimation are represented on the right side of Equation (3). First, if $\alpha \neq 0$, then current wages may affect BMI. Second, genetic factors that may influence BMI, G^B, and genetic factors that affect wages, G^W, could be correlated. Third, non-genetic factors that influence BMI, NG^B, may be correlated with non-genetic factors that can affect wages NG^W (Cawley, 2004). This model will be replicated by looking at separate models on how wages are affected by *BMI, Obese Overweight & Underweight,* and *Weight.*

IV. Data

Table 1 represents the variables' names, as well as the summary statistics. The dependent variable *Wages* is logged for the log of hourly wages received. A sample of the population from the NLSY79 is used for the replication study in the year 2000. *Ln(wages)* is the dependent variable. *HGC_Mother* indicates the highest grade completed by the respondent's mother.

HGC_Father indicate*s* the highest grade completed by the respondent's father. *HGC* indicates the highest grade completed by the respondent. *AFQT* is a proxy for cognitive ability. *Enrolled* is an indicator variable of whether or not the respondent is in enrolled in school in the year 2000. *Sales, Administrative, Technical, Managerial, Service, Farming, Repair, Assemblers, Transportation,* and *Laborers* are all different classifications for the variable *Occupation. Full_Time* is an indicator variable takes value of 1 if the survey taker works full time (works more than 20 hours a week). *Tenure* represents job tenure in weeks. *Region* is broken up into different categories: *NE, NC, S, W,* and *NM. Marital,* is categorized into *Nottogehter* and *Married. Work_experience* is calculated from the respondent's age and years in school. The equation consists of $Work\_experience = Age - HGC - 6$. *Work_experience2* is the squared value of *Work_experience,* in order to align it with the Mincer Equation. *Weight* is the current weight in pounds of the respondent in the year 2000, and is a variable of interest while being controlled for height. *Height* is the current height in inches of the respondent recorded in the year 2000. *Height* and *Weight* are also used to compute *BMI. BMI* is a variable of interest and is calculated by using the equation $BMI = (Weight*703)/Height^2$. *Underweight*[2], *Overweight*[3], and *Obese*[4] are indicator variables created from *BMI,* and are variables of interest. Finally, *Sibling_bmi*[5] represents the BMI of the respondents' siblings in the year 2000. This variable will be the instrumental variable (IV) used for the 2-Stage Least Squares model.

[2] Underweight is classified as a BMI less than 18.5.

[3] Overweight is classified as a BMI greater than 25 but less than 30.

[4] Obese is classified as a BMI equal to or greater than 30.

[5] Sibling_bmi is calculated the same way as the respondent's BMI

V. Results

The OLS regression results with *BMI* as the variable of interest is shown in Table (2). The statistically significant variables are noted within Table (2), and economically significant variables are italicized.

Table (3) shows the results of the variables of interest for the two regressions of the 2-Stage Least Squares (2SLS), the three OLS regressions ran without using the entire sample of white females, and the three OLS regressions using only the IV sample. The IV sample represents the sample used in the 2SLS, without any more observations. The lines separating some of the variables of interest are separating the three regressions. The results from both the OLS models are shown for a regression looking at (A) *BMI*, a regression looking at (B) *Weight*, and a regression looking at (C) *Underweight, Overweight,* and *Obese.* There are two 2SLS models conducted. For Model (A)'s 2SLS, *BMI* is the endogenous variable. Model (B)'s 2SLS has *Weight* as the endogenous variable. *Siblings_BMI* is used as the IV for both Model (A) and Model (B). As stated before, the *Weight* variable is controlled for height. With the separate regressions and different variables of interest, results were found for the replication study.

Model (A) shows that if BMI is increased by 1 kg/m^2, the log of wages will decrease by 0.008, without controlling for endogeneity. When endogeneity is controlled by using the 2SLS, the log of hourly wages will decrease by 0.017as BMI increases by 1 kg/m^2. When the OLS regression was ran with only the IV sample, the log of wages will decrease by 0.008 for a BMI increase by 1 kg/m^2.

In Model (B), the results display that as a respondent's weight increases by one pound, the log of hourly wages will decrease by 0.0014, without controlling for endogeneity. By controlling for endogeneity, as the weight increases by one pound, the log of hourly wages will decrease by 0.0028. Running an OLS with the IV sample shows that the log of wages decreases by 1.39E-3 as the weight increases by 1 pound.

Looking at the third model (C), *Obesity* is interpreted as the log of wages will decrease by 0.119 for every respondent who is obese. However, the log of wages will decrease for people who are categorized as *Obese* by 0.069 using only the IV sample. If the respondent is classified as *Overweight*, then the log of wages will decrease by 0.045 with the entire sample of white females, and decrease by 0.005 with only the IV sample. Finally, if the respondent is classified as *Underweight*, then the log of wages will decrease by 0.01 dollars with the entire sample, but will decrease by 0.089 using the IV sample. It appears that in Model (C), respondents face the largest pay penalty when they are in the category of *Obese*, than any other classification.

Table (4) shows the results from Cawley (2004) in the same format as Table (3). Cawley's (2004) results are used to compare to the replicated results. From this, it appears that the parameter estimates from the OLS regressions were similar. However, the IV results that were replicated were much smaller than Cawley's (2004).

VI. Conclusions and Limitations

Cawley's (2004) OLS results with the entire sample appear to be half the value his IV results. This is not the case in the replication study. The differences in the IV results could be due to the differences in sample sizes. Cawley conducts his study with 10,800 person-year observations in his IV sample and 25,843 person-year observations in this entire sample of white females. My replication study had 1,446 observations for my entire sample of white females and only 388 in my IV sample. Cawley's sample does not indicate how many unique respondents were listed. Due to the number of years his study was conducted, respondents had the potential of being interviewed 17 times. Since my replication study is conducted for one year, each respondent is a unique white female. A unique respondent represents a respondent that has been interviewed for the study once. The differences in results can also be due to the fact that not all of Cawley (2004)'s regressors were accounted for within the replication model. The variables that are cannot be retrieved are the county's unemployment rate, white collar vs. blue collar, the number of children ever born, and the age of the youngest child. Finally, the variable *Work_experience* is not able to be directly retrieved, so I use the algebraic equation, $Work\_experience = Age - HGC - 6$, in order to calculate the potential work experience of the respondents. The potential work experience will take place of the actual work experience in my replication study.

Both results from Cawley (2004) and this study find that a pay penalty does exist for white females who are obese. Also, the pay penalty is associated with a higher BMI, a higher weight, and overweight.

V. References

Averett, Susan, and Sanders Korenman. *The Economic Reality of the Beauty Myth*. 2. 31. University of Wisconsin Press, 1996. 304-330. Print.

Baum II, Charles, and William Ford. "The Wage Effects of Obesity: a Longitudinal Study." *Health Economics*. 13 (2004): 885-899. Print.

Brunello, Giogrio, and Beatrice D'Hombres. "Does Body Weight Affect Wage Evidence from Europe." *Economics and Human Biology*. 5.1 (2007): 1-19. Print.

Cawley, John. *The Impact of Obesity on Wages*. 2. 39. University of Wisconsin Press, 2004. 451-474. Print.

Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of obesity in the United States, 2009 2010. NCHS data brief, no 82. Hyattsville, MD: National Center for Health Statistics. 2012. Print.

Wada, Roy, and Erdal Tekin. "Body Composition and Wages." *NBER Working Paper Series*. w13595. (2007): 1-19. Web. 11 Feb. 2013. <http://ezproxy.uakron.edu:2048/login?url=http://ezproxy.uakron.edu:2421/docview/56599171?accountid=14471>.

V. Appendix

Table 1: Summary Statistics

Variable Name	Description	Expected Sign	Mean	Std Dev	Min	Max
Sibling_BMI	Sibling's BMI	-	26.26	6.62	15.36	91.22
HGC_Mother	Highest Grade Completed by Mother in years	+	12.00	2.18	0.00	19.00
HGC_Father	Highest Grade Completed by Father in years	+	12.26	3.05	2.00	20.00
Region	Region of Residence	+/-	2.39	0.89	1.00	4.00
Marital	Marital Status of the Respondent	+	1.19	0.89	0.00	6.00
AFQT	Cognitive Ability score	+	58.35	26.82	0.82	100.00
Enrolled	Currently enrolled in school	+/-	0.04	0.19	0.00	1.00
Occupation	Industry of Occupation	+/-	284.19	217.47	5.00	889.00
Full_time	Full-time vs. Part-time	+	0.90	0.29	0.00	1.00
Tenure	Job Tenure in weeks	+	337.67	297.29	1.00	1151.00
Work_Experience	Work Experience in years	+	18.45	2.90	10.00	25.00
Work_Experience2	Work_Experience^2 in years	+	348.91	105.04	100.00	625.00
HGC	Highest Grade Completed in years	+	13.97	2.33	8.00	20.00
Wages	Hourly Wages in dollars-dependent variable	N/A	15.70	10.79	2.40	75.00
BMI	BMI of Respondent in kg/m^2	-	26.26	6.62	15.36	91.22
Overweight	Overweight vs. Not-Overweight	-	0.28	0.45	0.00	1.00
Underweight	Underweight vs. Not-Underweight	-	0.02	0.13	0.00	1.00
Obese	Obese vs. Not-Obese	-	0.21	0.41	0.00	1.00
Height	Height of Respondent in inches	+	64.71	2.51	59.00	72.00
Weight	Weight of Respondent in pounds	-	156.49	42.01	84	600

Table 2: OLS Results

Parameter Estimates		
Variable	Parameter Estimate	t Value
Intercept	2.196	4.760
HGC_Mother	0.006	0.820
HGC_Father**	0.011	2.030
AFQT***	0.002	2.800
Enrolled	-0.026	-0.360
Technical	-0.006	-0.080
Sales***	-0.200	-3.950
Administrative***	-0.232	-6.020
Service***	-0.462	-9.600
Farming*	-0.245	-1.790
Repair	-0.064	-0.740
Assemblers**	-0.190	-2.570
Transportation	-0.151	-1.350
Laborers*	-0.214	-1.840
Nottogether	0.169	1.080
NC**	-0.102	-2.450
S*	-0.075	-1.780
W	0.051	1.070
NM	-0.035	-0.730
Full_Time	-0.032	-0.710
Tenure***	4.1E-4	8.660
HGC***	0.050	4.950
Work_experience	-0.024	-0.620
Work_experience2	0.000	0.430
BMI***	-0.008	-3.790
***p<.01	**p<.05	*p<.10
99% Condifence Interval	95% Confidence Interval	90% Confidence Interval

Table 3: OLS and 2SLS Results from Replication

Model	Variable of Interest	OLS	IV	OLS Using IV Sample
(A)	BMI[6]	-0.008[7]	-8.20E-4	-0.008
		(-3.79)[8]	(-0.05)	(-2.13)
(B)	Weight in pounds[9]	-1.38E-3	-1.10E-04	-1.39E-3
		(-3.78)	(-0.05)	(-2.20)
(C)	Underweight[10]	0.029	N/A	-0.089
		(0.31)	N/A	(-0.45)
	Overweight	-0.056	N/A	-0.005
		(-1.71)	N/A	(-0.09)
	Obese	-0.109	N/A	-0.069
		(-3.06)	N/A	(-1.02)
Number of Respondents		1446	388	388

[6] First model looking at BMI as the variable of interest.

[7] Parameter Estimates

[8] t-statistics

[9] Second model looking at Weight in pounds as the variable of interest, controlling for height.

[10] Third model looking at Underweight, Overweight, Obese as the variables of interest.

Table 4: Cawley (2004)

Model	Variable of Interest	OLS	IV	OLS Using IV Sample
(A)	BMI[11]	-0.008[12]	-0.017	-0.010
		(-7.01)[13]	(-3.38)	(-6.10)
(B)	Weight in pounds[14]	-0.0014	-0.0028	-1.60E-3
		(-6.98)	(-3.40)	(-5.97)
(C)	Underweight[15]	-0.01	N/A	-0.01
		(-0.53)	N/A	(-0.53)
	Overweight	-0.045	N/A	-0.045
		(-3.52)	N/A	(-3.52)
	Obese	-0.119	N/A	-0.119
		(-6.76)	N/A	(-6.76)
Number of Respondents		25,843	10,800	10,800

[11] First model of Cawley (2004) looking at BMI as the variable of interest.

[12] Parameter estimate

[13] t-statistic

[14] Second model of Cawley (2004) looking at Weight in pounds as the variable of interest, controlling for height.

[15] Third model of Cawley (2004) looking at Underweight, Overweight, Obese as the variables of interest.

```
options nocenter validvarname=any nolabel;

*---Read in space-delimited ascii file;

data new_data;


infile "C:\Users\hnk6\Desktop\kretch\kretch.dat" lrecl=467 missover DSD DLM='
' print;
input
  R0000100
  R0000149
  R0006500
  R0007900
  R0214700
  R0214800
  R0481600
  R0618301
  R0779800
  R0998900
  R0999000
  R1773900
  R4125100
  R4125101
  R4125200
  R4125300
  R4125400
  R4125500
  R4125600
  R4125700
  R4125800
  R4125801
  R4125900
  R4126000
  R4126100
  R4126200
  R4126300
  R4126400
  R4126500
  R4126501
  R4126600
  R4126700
  R4126800
  R4126900
  R4127000
  R4127100
  R4127200
  R4127201
  R4127300
  R4127400
  R4127500
  R4127600
  R4127700
  R4127800
  R4127900
  R4127901
  R4128000
```

R4128100
R4128200
R4128300
R4128400
R4128500
R4128600
R4128601
R4128700
R4128800
R4128900
R4129000
R4129100
R4129200
R4129300
R4129301
R4129400
R4129500
R4129600
R4129700
R4129800
R4129900
R4130000
R4130001
R4130100
R4130200
R4130300
R4130400
R4130500
R4130600
R4130700
R4130701
R4130800
R4130900
R4131000
R4131100
R4131200
R4131300
R4131400
R4131401
R4131500
R4131600
R4131700
R4131800
R4131900
R4132000
R4132100
R4132101
R4132200
R4132300
R4132400
R4132500
R4132600
R4132700
R4132800
R4132801
R4132900
R4133000

```
  R4133100
  R4133200
  R4133300
  R4133400
  R4133500
  R4133600
  R4133700
  R4133701
  R4133800
  R4133900
  R4134000
  R4134100
  R4134200
  R4134300
  R4134400
  R4134500
  R6540000
  R6592900
  R6659100
  R6888100
  R7005200
  R7005700
  R7006800
  R7007000
  R7007300
  R7007700
  T0897400
  T0897500
  T2053900
  T2054000
  T3024800
  T3024900
;
array nvarlist _numeric_;


*---Recode missing values to SAS custom system missing. See SAS
      documentation for use of MISSING option in procedures, e.g. PROC FREQ;

/*
do over nvarlist;
  if nvarlist = -1 then nvarlist = .R;
  if nvarlist = -2 then nvarlist = .D;
  if nvarlist = -3 then nvarlist = .I;
  if nvarlist = -4 then nvarlist = .V;
  if nvarlist = -5 then nvarlist = .N;
end;
*/

  label R0000100 = "ID# (1-12686) 79";
  label R0000149 = "HH ID # 79";
  label R0006500 = "HGC BY RS MOTHER 79";
  label R0007900 = "HGC BY RS FATHER 79";
  label R0214700 = "RACL/ETHNIC COHORT /SCRNR 79";
  label R0214800 = "SEX OF R 79";
  label R0481600 = "HEIGHT OF R 81";
  label R0618301 = "PROFILES AFQT PRCTILE 2006 (REV) 81";
label R0779800 = "HEIGHT OF R 82";
label R0998900 = "F HEIGHT OF R IN FEET 83";
label R0999000 = "F HEIGHT OF R IN INCHES 83";
label R1773900 = "HEIGHT OF R 85";
label R4125100 = "DOES R HAVE SIBLINGS? 93";
label R4125101 = "PUBID OF INT YTH SIBLING CORR TO 1ST SIB ON ROSTER";
label R4125200 = "1ST SIBLING OLDER/YNGER THAN R? 93";
label R4125300 = "# YR OLDR/YNGR THAN R IS 1ST SIB 93";
label R4125400 = "SEX OF 1ST SIBLING 93";
label R4125500 = "HIGHST GRADE EVR COMPLETE 1ST SIBLNG 93";
label R4125600 = "# OF CHILDREN 1ST SIBLING EVER HAD 93";
label R4125700 = "AGE OF 1ST SIBLING AT 1ST BIRTH 93";
label R4125800 = "DOES R HAVE A 2ND SIBLING? 93";
label R4125801 = "PUBID OF INT YTH SIBLING CORR TO 2ND SIB ON ROSTER";
label R4125900 = "2ND SIBLING OLDR/YNGR THAN R? 93";
label R4126000 = "# YR OLDR/YNGR THAN R IS 2ND SIB 93";
label R4126100 = "SEX OF 2ND SIBLING 93";
label R4126200 = "HIGHST GRADE EVR COMPLETE 2ND SIBLNG 93";
label R4126300 = "# OF CHILDREN 2ND SIBLING EVER HAD 93";
label R4126400 = "AGE OF 2ND SIBLING AT 1ST BIRTH 93";
label R4126500 = "DOES R HAVE A 3RD SIBLING? 93";
label R4126501 = "PUBID OF INT YTH SIBLING CORR TO 3RD SIB ON ROSTER";
label R4126600 = "3RD SIBLING OLDR/YNGR THAN R? 93";
label R4126700 = "# YR OLDR/YNGR THAN R IS 3RD SIB 93";
label R4126800 = "SEX OF 3RD SIBLING 93";
label R4126900 = "HIGHST GRADE EVR COMPLETE 3RD SIBLNG 93";
label R4127000 = "# OF CHILDREN 3RD SIBLING EVER HAD 93";
label R4127100 = "AGE OF 3RD SIBLING AT 1ST BIRTH 93";
label R4127200 = "DOES R HAVE A 4TH SIBLING? 93";
label R4127201 = "PUBID OF INT YTH SIBLING CORR TO 4TH SIB ON ROSTER";
label R4127300 = "4TH SIBLING OLDER/YNGER THAN R? 93";
label R4127400 = "# YR OLDR/YNGR THAN R IS 4TH SIB 93";
label R4127500 = "SEX OF 4TH SIBLING 93";
label R4127600 = "HIGHST GRADE EVR COMPLETE 4TH SIBLNG 93";
label R4127700 = "# OF CHILDREN 4TH SIBLING EVER HAD 93";
label R4127800 = "AGE OF 4TH SIBLING AT 1ST BIRTH 93";
label R4127900 = "DOES R HAVE A 5TH SIBLING? 93";
label R4127901 = "PUBID OF INT YTH SIBLING CORR TO 5TH SIB ON ROSTER";
label R4128000 = "5TH SIBLING OLDER/YNGER THAN R? 93";
label R4128100 = "# YR OLDR/YNGR THAN R IS 5TH SIB 93";
label R4128200 = "SEX OF 5TH SIBLING 93";
label R4128300 = "HIGHST GRADE EVR COMPLETE 5TH SIBLNG 93";
label R4128400 = "# OF CHILDREN 5TH SIBLING EVER HAD 93";
label R4128500 = "AGE OF 5TH SIBLING AT 1ST BIRTH 93";
label R4128600 = "DOES R HAVE A 6TH SIBLING? 93";
label R4128601 = "PUBID OF INT YTH SIBLING CORR TO 6TH SIB ON ROSTER";
label R4128700 = "6TH SIBLING OLDR/YNGR THAN R? 93";
label R4128800 = "# YR OLDR/YNGR THAN R IS 6TH SIB 93";
label R4128900 = "SEX OF 6TH SIBLING 93";
label R4129000 = "HIGHST GRADE EVR COMPLETE 6TH SIBLNG 93";
label R4129100 = "# OF CHILDREN 6TH SIBLING EVER HAD 93";
label R4129200 = "AGE OF 6TH SIBLING AT 1ST BIRTH 93";
label R4129300 = "DOES R HAVE A 7TH SIBLING? 93";
label R4129301 = "PUBID OF INT YTH SIBLING CORR TO 7TH SIB ON ROSTER";
label R4129400 = "7TH SIBLING OLDR/YNGR THAN R? 93";
label R4129500 = "# YR OLDR/YNGR THAN R IS 7TH SIB 93";
label R4129600 = "SEX OF 7TH SIBLING 93";
label R4129700 = "HIGHST GRADE EVR COMPLETE 7TH SIBLNG 93";
label R4129800 = "# OF CHILDREN 7TH SIBLING EVER HAD 93";
label R4129900 = "AGE OF 7TH SIBLING AT 1ST BIRTH 93";
label R4130000 = "DOES R HAVE AN 8TH SIBLING? 93";
label R4130001 = "PUBID OF INT YTH SIBLING CORR TO 8TH SIB ON ROSTER";
label R4130100 = "8TH SIBLING OLDER/YNGER THAN R? 93";
label R4130200 = "# YR OLDR/YNGR THAN R IS 8TH SIB 93";
label R4130300 = "SEX OF 8TH SIBLING 93";
label R4130400 = "HIGHST GRADE EVR COMPLETE 8TH SIBLNG 93";
label R4130500 = "# OF CHILDREN 8TH SIBLING EVER HAD 93";
label R4130600 = "AGE OF 8TH SIBLING AT 1ST BIRTH 93";
label R4130700 = "DOES R HAVE A 9TH SIBLING? 93";
label R4130701 = "PUBID OF INT YTH SIBLING CORR TO 9TH SIB ON ROSTER";
label R4130800 = "9TH SIBLING OLDR/YNGR THAN R? 93";
label R4130900 = "# YR OLDR/YNGR THAN R IS 9TH SIB 93";
label R4131000 = "SEX OF 9TH SIBLING 93";
label R4131100 = "HIGHST GRADE EVR COMPLETE 9TH SIBLNG 93";
label R4131200 = "# OF CHILDREN 9TH SIBLING EVER HAD 93";
label R4131300 = "AGE OF 9TH SIBLING AT 1ST BIRTH 93";
label R4131400 = "DOES R HAVE A 10TH SIBLING? 93";
label R4131401 = "PUBID OF INT YTH SIBLING COR TO 10TH SIB ON ROSTER";
label R4131500 = "10TH SIBLING OLDR/YNGR THAN R? 93";
label R4131600 = "# YR OLDR/YNGR THAN R IS 10TH SIB 93";
label R4131700 = "SEX OF 10TH SIBLING 93";
label R4131800 = "HIGHST GRADE EVR COMPLETE 10TH SIBLNG 93";
label R4131900 = "# OF CHILDREN 10TH SIBLING EVER HAD 93";
label R4132000 = "AGE OF 10TH SIBLING AT 1ST BIRTH 93";
label R4132100 = "DOES R HAVE AN 11TH SIBLING? 93";
label R4132101 = "PUBID OF INT YTH SIBLING COR TO 11TH SIB ON ROSTER";
label R4132200 = "11TH SIBLING OLDER/YNGER THAN R? 93";
label R4132300 = "# YR OLDR/YNGR THAN R IS 11TH SIB 93";
label R4132400 = "SEX OF 11TH SIBLING 93";
label R4132500 = "HIGHST GRADE EVR COMPLETE 11TH SIBLNG 93";
label R4132600 = "# OF CHILDREN 11TH SIBLING EVER HAD 93";
label R4132700 = "AGE OF 11TH SIBLING AT 1ST BIRTH 93";
label R4132800 = "DOES R HAVE A 12TH SIBLING? 93";
label R4132801 = "PUBID OF INT YTH SIBLING COR TO 12TH SIB ON ROSTER";
label R4132900 = "12TH SIBLING OLDR/YNGR THAN R? 93";
label R4133000 = "# YR OLDR/YNGR THAN R IS 12TH SIB 93";
label R4133100 = "SEX OF 12TH SIBLING 93";
label R4133200 = "HIGHST GRADE EVR COMPLETE 12TH SIBLNG 93";
label R4133300 = "# OF CHILDREN 12TH SIBLING EVER HAD 93";
label R4133400 = "AGE OF 12TH SIBLING AT 1ST BIRTH 93";
label R4133500 = "DID R REPORT AT LEAST 12 SIBLINGS? 93";
label R4133600 = "R'S YOUNGEST SIB LISTED? (RE) 93";
label R4133700 = "IS R'S YOUNGEST SIBLING LISTED? 93";
label R4133701 = "PUBID OF INT YTH SIBLING COR TO 13TH SIB ON ROSTER";
label R4133800 = "YOUNGEST SIBLING OLDR/YNGR THAN R? 93";
label R4133900 = "# YR OLDR/YNGR THAN R YNGST SIB 93";
label R4134000 = "SEX OF YOUNGEST SIBING 93";
label R4134100 = "HIGHST GRADE EVR COMPLETE YNGST SIBLNG93";
label R4134200 = "# OF CHILDREN YNGST SIBLING EVER HAD 93";
label R4134300 = "AGE OF YOUNGEST SIBLING AT 1ST BIRTH 93";
label R4134400 = "R HAVE ANY OTHR SIBLNGS NOT MENTIONED 93";
label R4134500 = "TOTAL # OF SIBLINGS R HAS 93";
label R6540000 = "R CURRENTLY ATTENDING REG SCHOOL? 2000";
label R6592900 = "CPS OCCUPATION 1980 CODE L1 2000";
```

```
   label R6659100 = "TTL HRS/WK AT JOB, HOME > 20 L1 2000";
   label R6888100 = "HOW MUCH DOES R WEIGH 2000";
   label R7005200 = "TOTAL TENURE JOB #01 2000";
   label R7005700 = "HRLY RATE OF PAY JOB #01 2000";
   label R7006800 = "REGION OF RESIDENCE 2000";
   label R7007000 = "MARITAL STATUS 2000";
   label R7007300 = "HIGHEST GRADE COMPLTD (REV) 2000";
   label R7007700 = "WKS WRKD IN PAST CAL YR 2000";
   label T0897400 = "R HEIGHT IN FEET 2006";
   label T0897500 = "R HEIGHT IN INCHES 2006";
   label T2053900 = "R HEIGHT IN FEET 2008";
   label T2054000 = "R HEIGHT IN INCHES 2008";
   label T3024800 = "R HEIGHT IN FEET 2010";
   label T3024900 = "R HEIGHT IN INCHES 2010";

/*---------------------------------------------------------------------*
 *  Crosswalk for Reference number & Question name                     *
 *---------------------------------------------------------------------*
 * Uncomment and edit this RENAME statement to rename variables
 * for ease of use.  You may need to use  name literal strings
 * e.g.  'variable-name'n   to create valid SAS variable names, or
 * alter variables similarly named across years.
 * This command does not guarentee uniqueness

 * See SAS documentation for use of name literals and use of the
 * VALIDVARNAME=ANY option.
 *---------------------------------------------------------------------*/
  /* *start* */

 RENAME
  R0000100 = 'CASEID'n
  R0000149 = 'HHID'n
  R0006500 = 'HGC_Mother'n
  R0007900 = 'HGC_Father'n
  R0214700 = 'Race'n
  R0214800 = 'Gender'n
  R0481600 = 'Height_1981'n
  R0618301 = 'AFQT'n
  R0779800 = 'HEIGHT_1982'n
  R0998900 = 'FFER51A_1983'n
  R0999000 = 'FFER51B_1983'n
  R1773900 = 'Height'n
  R4125100 = 'SIBROS_ANY_SIBS_1993'n
  R4125101 = 'sibid_01'n
  R4125200 = 'SIBROS_OLDER_YOUNGER01_1993'n
  R4125300 = 'SIBROS_YRS_OLDER_YOUNGER01_1993'n
  R4125400 = 'SIBROS_GENDER01_1993'n
  R4125500 = 'SIBROS_HGC01_1993'n
  R4125600 = 'SIBROS_NUMKIDS01_1993'n
  R4125700 = 'SIBROS_AGEAT1STBIRTH01_1993'n
  R4125800 = 'SIBROS_ANY_MORE_SIBS01_1993'n
  R4125801 = 'sibid_02'n
  R4125900 = 'SIBROS_OLDER_YOUNGER02_1993'n
  R4126000 = 'SIBROS_YRS_OLDER_YOUNGER02_1993'n
  R4126100 = 'SIBROS_GENDER.02_1993'n
  R4126200 = 'SIBROS_HGC02_1993'n
  R4126300 = 'SIBROS_NUMKIDS02_1993'n
```

```
R4126400 = 'SIBROS_AGEAT1STBIRTH02_1993'n
R4126500 = 'SIBROS_ANY_MORE_SIBS02_1993'n
R4126501 = 'sibid_03'n
R4126600 = 'SIBROS_OLDER_YOUNGER03_1993'n
R4126700 = 'SIBROS_YRS_OLDER_YOUNGER03_1993'n
R4126800 = 'SIBROS_GENDER03_1993'n
R4126900 = 'SIBROS_HGC03_1993'n
R4127000 = 'SIBROS_NUMKIDS03_1993'n
R4127100 = 'SIBROS_AGEAT1STBIRTH03_1993'n
R4127200 = 'SIBROS_ANY_MORE_SIBS03_1993'n
R4127201 = 'sibid_04'n
R4127300 = 'SIBROS_OLDER_YOUNGER04_1993'n
R4127400 = 'SIBROS_YRS_OLDER_YOUNGER04_1993'n
R4127500 = 'SIBROS_GENDER04_1993'n
R4127600 = 'SIBROS_HGC04_1993'n
R4127700 = 'SIBROS_NUMKIDS04_1993'n
R4127800 = 'SIBROS_AGEAT1STBIRTH04_1993'n
R4127900 = 'SIBROS_ANY_MORE_SIBS04_1993'n
R4127901 = 'sibid_05'n
R4128000 = 'SIBROS_OLDER_YOUNGER05_1993'n
R4128100 = 'SIBROS_YRS_OLDER_YOUNGER.05_1993'n
R4128200 = 'SIBROS_GENDER05_1993'n
R4128300 = 'SIBROS_HGC05_1993'n
R4128400 = 'SIBROS_NUMKIDS05_1993'n
R4128500 = 'SIBROS_AGEAT1STBIRTH.05_1993'n
R4128600 = 'SIBROS_ANY_MORE_SIBS05_1993'n
R4128601 = 'sibid_06'n
R4128700 = 'SIBROS_OLDER_YOUNGER06_1993'n
R4128800 = 'SIBROS_YRS_OLDER_YNGR06_1993'n
R4128900 = 'SIBROS_GENDER06_1993'n
R4129000 = 'SIBROS_HGC06_1993'n
R4129100 = 'SIBROS_NUMKIDS06_1993'n
R4129200 = 'SIBROS_AGEAT1STBIRTH06_1993'n
R4129300 = 'SIBROS_ANY_MORE_SIBS06_1993'n
R4129301 = 'sibid_07'n
R4129400 = 'SIBROS_OLDER_YOUNGER07_1993'n
R4129500 = 'SIBROS_YRS_OLDER_YOUNGER07_1993'n
R4129600 = 'SIBROS_GENDER07_1993'n
R4129700 = 'SIBROS_HGC07_1993'n
R4129800 = 'SIBROS_NUMKIDS07_1993'n
R4129900 = 'SIBROS_AGEAT1STBIRTH07_1993'n
R4130000 = 'SIBROS_ANY_MORE_SIBS07_1993'n
R4130001 = 'sibid_08'n
R4130100 = 'SIBROS_OLDER_YOUNGER08_1993'n
R4130200 = 'SIBROS_YRS_OLDER_YOUNGER08_1993'n
R4130300 = 'SIBROS_GENDER08_1993'n
R4130400 = 'SIBROS_HGC08_1993'n
R4130500 = 'SIBROS_NUMKIDS08_1993'n
R4130600 = 'SIBROS_AGEAT1STBIRTH08_1993'n
R4130700 = 'SIBROS_ANY_MORE_SIBS08_1993'n
R4130701 = 'sibid_09'n
R4130800 = 'SIBROS_OLDER_YOUNGER09_1993'n
R4130900 = 'SIBROS_YRS_OLDER_YNGR09_1993'n
R4131000 = 'SIBROS_GENDER09_1993'n
R4131100 = 'SIBROS_HGC09_1993'n
R4131200 = 'SIBROS_NUMKIDS09_1993'n
R4131300 = 'SIBROS_AGEAT1STBIRTH09_1993'n
```

```
  R4131400 = 'SIBROS_ANY_MORE_SIBS09_1993'n
  R4131401 = 'sibid_10'n
  R4131500 = 'SIBROS_OLDER_YNGR10_1993'n
  R4131600 = 'SIBROS_YRS_OLDER_YNGR10_1993'n
  R4131700 = 'SIBROS_GENDER10_1993'n
  R4131800 = 'SIBROS_HGC10_1993'n
  R4131900 = 'SIBROS_NUMKIDS10_1993'n
  R4132000 = 'SIBROS_AGEAT1STBIRTH10_1993'n
  R4132100 = 'SIBROS_ANY_MORE_SIBS10_1993'n
  R4132101 = 'sibid_11'n
  R4132200 = 'SIBROS_OLDER_YOUNGER11_1993'n
  R4132300 = 'SIBROS_YRS_OLDER_YOUNGER11_1993'n
  R4132400 = 'SIBROS_GENDER11_1993'n
  R4132500 = 'SIBROS_HGC11_1993'n
  R4132600 = 'SIBROS_NUMKIDS11_1993'n
  R4132700 = 'SIBROS_AGEAT1STBIRTH11_1993'n
  R4132800 = 'SIBROS_ANY_MORE_SIBS11_1993'n
  R4132801 = 'sibid_12'n
  R4132900 = 'SIBROS_OLDER_YNGR12_1993'n
  R4133000 = 'SIBROS_YRS_OLDER_YNGR12_1993'n
  R4133100 = 'SIBROS_GENDER12_1993'n
  R4133200 = 'SIBROS_HGC12_1993'n
  R4133300 = 'SIBROS_NUMKIDS12_1993'n
  R4133400 = 'SIBROS_AGEAT1STBIRTH12_1993'n
  R4133500 = 'SIBROS_12_SIBS_REPORTED_1993'n
  R4133600 = 'SIBROS_YOUNGEST_SIB_1993'n
  R4133700 = 'SIBROS_YOUNGEST_SIBREP_1993'n
  R4133701 = 'sibid_13'n
  R4133800 = 'SIBROS_OLDER_YOUNGER13_1993'n
  R4133900 = 'SIBROS_YRS_OLDER_YOUNGER13_1993'n
  R4134000 = 'SIBROS_GENDER13_1993'n
  R4134100 = 'SIBROS_HGC13_1993'n
  R4134200 = 'SIBROS_NUMKIDS13_1993'n
  R4134300 = 'SIBROS_AGEAT1STBIRTH13_1993'n
  R4134400 = 'SIBROS_ANY_SIBS_NOT_LISTED_1993'n
  R4134500 = 'SIBROS_TOTAL_NUM_SIBS_1993'n
  R6540000 = 'Enrolled'n
  R6592900 = 'Occupation'n
  R6659100 = 'Full_Time'n
  R6888100 = 'Weight'n
  R7005200 = 'TENURE'n
  R7005700 = 'Wages'n
  R7006800 = 'Region'n
  R7007000 = 'Marital'n
  R7007300 = 'HGC'n
  R7007700 = 'Weeks_worked'n
  T0897400 = 'Q1110_A_2006'n
  T0897500 = 'Q1110_B_2006'n
  T2053900 = 'Q1110_A_2008'n
  T2054000 = 'Q1110_B_2008'n
  T3024800 = 'Q1110_A_2010'n
  T3024900 = 'Q1110_B_2010'n
;
  /* *finish* */

run;
```

```
proc means data=new_data n mean min max ;
      var hhid caseid height weight ;
run;

*---Read in space-delimited ascii file;

data new_data2;


infile 'default.dat' lrecl=11 missover DSD DLM=' ' print;
input
  R0000100
  R7007500
;
array nvarlist _numeric_;


*---Recode missing values to SAS custom system missing. See SAS
      documentation for use of MISSING option in procedures, e.g. PROC FREQ;

do over nvarlist;
  if nvarlist = -1 then nvarlist = .R;  /* Refused */
  if nvarlist = -2 then nvarlist = .D;  /* Dont know */
  if nvarlist = -3 then nvarlist = .I;  /* Invalid missing */
  if nvarlist = -4 then nvarlist = .V;  /* Valid missing */
  if nvarlist = -5 then nvarlist = .N;  /* Non-interview */
end;

  label R0000100 = "CASEID";
  label R7007500 = "AGE AT INTERVIEW DATE 2000";
 rename
      R0000100 = 'CASEID'n
      R7007500 = 'Age'n
;
      run;

proc means data=new_data2 n mean min max;
run;


/*---------------------------------------------------------------------*
 *  FORMATTED TABULATIONS                                              *
 *---------------------------------------------------------------------*
 * You can uncomment and edit the PROC FORMAT and PROC FREQ statements
 * provided below to obtain formatted tabulations. The tabulations
 * should reflect codebook values.
 *
 * Please edit the formats below reflect any renaming of the variables
 * you may have done in the first data step.
 *---------------------------------------------------------------------*/

/*
proc format;
value vx1f
  35='35'
  36='36'
  37='37'
  38='38'
  39='39'
  40='40'
  41='41'
  42='42'
  43='43'
  44='44'
;
*/

/*
 *--- Tabulations using reference number variables;
proc freq data=new_data;
tables _ALL_ /MISSING;
  format R7007500 vx1f.;
run;
*/

/*
*--- Tabulations using default named variables;
proc freq data=new_data;
tables _ALL_ /MISSING;
  format AGEATINT_2000 vx1f.;
run;
*/

options nocenter validvarname=any;
data data0;
merge new_data new_data2/*(caseid))*/;
by caseid;
run;



data datatotal;
set data0;

wages=(wages/100);
if wages < 1 then delete;
if wages > 500 then DELETE;
lnwages=Log(wages);
AFQT=AFQT/1000;
if height <=0 then delete;
if weight<=0 then delete;
male = 0;
female = 0;
if gender = 1 then male = 1;
if male = 1 then delete;
if gender = 2 then female = 1;
if gender < 0 then delete;
if HGC_Mother < 0 then delete;
if HGC_Father < 0 then delete;
if AFQT < 0 then delete;

* We have a question about how to treat the -4 (valid skip) on enrollment. We
comment out this section, just in case we need to convert -4 of enrolled to 0
(0 means not enrolled). ;
```

```
* if Enrolled = -4 then Enrolled = 0 -4=valid skip, reason to believe not
enrolled
* if Enrolled = -5 then Enrolled =. -5=was not interviewed;

* One experiment is to code negative values of Enrolled to missing (.) ;
* we could see how this treatment limit our sample size ;
if Enrolled=-4 then Enrolled=0;
if Enrolled =-5 then Enrolled=.;
if occupation < 0 then delete;
if full_time < 0 then delete;
if tenure < 0 then delete;
if region < 0 then delete;
if marital < 0 then delete;
if HGC < 0 then delete;
if weeks_worked < 0 then delete;
BMI=((weight*703)/(height*height));
obese=0;
if (BMI>=30) then obese=1;
else obese =0;
if (BMI<18.5) then underweight=1;
else underweight = 0;
if (25<BMI<30) then overweight =1;
else overweight=0;
if race = 1 then hsp = 1;
else hsp=0;
if race = 2 then blk = 1;
else blk=0;
if race = 3 then white = 1;
else white=0;
if white=0 then delete;


if region = 1 then NE =1;
else NE = 0;
if region= 2 then NC=1;
else NC=0;
if region=3 then S=1;
else S=0;
if region=4 then W=1;
else W=0;
if marital = 0 then NM=1;
else NM=0;
if marital = 1 then M=1;
else M=0;
if marital = 2 then nottogether=1;
if marital = 3 then nottogether=1;
if marital = 6 then nottogether=1;
else nottogether=0;
if 3<=occupation<=199 then managerial=1;
else managerial= 0;
if 203<=occupation<=235 then technical=1;
else technical=0;
if 243<=occupation<=285 then sales=1;
else sales=0;
if 303<=occupation<=389 then administrative=1;
else administrative=0;
if 403<=occupation<=469 then service=1;
```

```
else service=0;
if 473<=occupation<=499 then farming=1;
else farming=0;
if 503<=occupation<=699 then repair=1;
else repair=0;
if 703<=occupation<=799 then assemblers=1;
else assemblers=0;
if 803<=occupation<=859 then transportation=1;
else transportation=0;
if 863<=occupation<=889 then laborers=1;
else laborers=0;
if age<1 then delete;


  if Height_1981<0 then Height_1981=.;
  if HEIGHT_1982<0 then HEIGHT_1982=.;
  if FFER51A_1983<0 then FFER51A_1983=.;
  if FFER51B_1983<0 then FFER51B_1983=.;
if SIBROS_ANY_SIBS_1993<0 then SIBROS_ANY_SIBS_1993=.;
  if sibid_01<0 then sibid_01=.;
  if SIBROS_OLDER_YOUNGER01_1993<0 then SIBROS_OLDER_YOUNGER01_1993=.;
  if SIBROS_YRS_OLDER_YOUNGER01_1993<0 then
SIBROS_YRS_OLDER_YOUNGER01_1993=.;
  if SIBROS_GENDER01_1993<0 then SIBROS_GENDER01_1993=.;
  if SIBROS_HGC01_1993<0 then SIBROS_HGC01_1993=.;
  if SIBROS_NUMKIDS01_1993<0 then SIBROS_NUMKIDS01_1993=.;
  if SIBROS_AGEAT1STBIRTH01_1993<0 then SIBROS_AGEAT1STBIRTH01_1993=.;
  if SIBROS_ANY_MORE_SIBS01_1993<0 then SIBROS_ANY_MORE_SIBS01_1993=.;
  if sibid_02<0 then sibid_02=.;
  if SIBROS_OLDER_YOUNGER02_1993<0 then SIBROS_OLDER_YOUNGER02_1993=.;
  if SIBROS_YRS_OLDER_YOUNGER02_1993<0 then
SIBROS_YRS_OLDER_YOUNGER02_1993=.;
  if SIBROS_GENDER02_1993<0 then SIBROS_GENDER02_1993=.;
  if SIBROS_HGC02_1993<0 then SIBROS_HGC02_1993=.;
  if SIBROS_NUMKIDS02_1993<0 then SIBROS_NUMKIDS02_1993=.;
  if SIBROS_AGEAT1STBIRTH02_1993<0 then SIBROS_AGEAT1STBIRTH02_1993=.;
  if SIBROS_ANY_MORE_SIBS02_1993<0 then SIBROS_AGEAT1STBIRTH02_1993=.;
  if sibid_03<0 then sibid_03=.;
  if SIBROS_OLDER_YOUNGER03_1993<0 then SIBROS_OLDER_YOUNGER03_1993=.;
  if SIBROS_YRS_OLDER_YOUNGER03_1993<0 then
SIBROS_YRS_OLDER_YOUNGER03_1993=.;
  if SIBROS_GENDER03_1993<0 then SIBROS_GENDER03_1993=.;
  if SIBROS_HGC03_1993<0 then SIBROS_HGC03_1993=.;
  if SIBROS_NUMKIDS03_1993<0 then SIBROS_NUMKIDS03_1993=.;
  if SIBROS_AGEAT1STBIRTH03_1993<0 then SIBROS_AGEAT1STBIRTH03_1993=.;
  if SIBROS_ANY_MORE_SIBS03_1993<0 then SIBROS_ANY_MORE_SIBS03_1993=.;
  if sibid_04<0 then sibid_04=.;
  if SIBROS_OLDER_YOUNGER04_1993<0 then SIBROS_OLDER_YOUNGER04_1993=.;
  if SIBROS_YRS_OLDER_YOUNGER04_1993<0 then
SIBROS_YRS_OLDER_YOUNGER04_1993=.;
  if SIBROS_GENDER04_1993<0 then SIBROS_GENDER04_1993=.;
  if SIBROS_HGC04_1993<0 then SIBROS_HGC04_1993=.;
  if SIBROS_NUMKIDS04_1993<0 then SIBROS_NUMKIDS04_1993=.;
  if SIBROS_AGEAT1STBIRTH04_1993<0 then SIBROS_AGEAT1STBIRTH04_1993=.;
  if SIBROS_ANY_MORE_SIBS04_1993<0 then SIBROS_ANY_MORE_SIBS04_1993=.;
  if sibid_05<0 then sibid_05=.;
  if SIBROS_OLDER_YOUNGER05_1993<0 then SIBROS_OLDER_YOUNGER05_1993=.;
  if SIBROS_YRS_OLDER_YOUNGER05_1993<0 then
SIBROS_YRS_OLDER_YOUNGER05_1993=.;
  if SIBROS_GENDER05_1993<0 then SIBROS_GENDER05_1993=.;
  if SIBROS_HGC05_1993<0 then SIBROS_HGC05_1993=.;
  if SIBROS_NUMKIDS05_1993<0 then SIBROS_NUMKIDS05_1993=.;
  if SIBROS_AGEAT1STBIRTH05_1993<0 then SIBROS_AGEAT1STBIRTH05_1993=.;
  if SIBROS_ANY_MORE_SIBS05_1993<0 then SIBROS_ANY_MORE_SIBS05_1993=.;
  if sibid_06<0 then sibid_06=.;
  if SIBROS_OLDER_YOUNGER06_1993<0 then SIBROS_OLDER_YOUNGER06_1993=.;
  if SIBROS_YRS_OLDER_YNGR06_1993<0 then SIBROS_YRS_OLDER_YNGR06_1993=.;
  if SIBROS_GENDER06_1993<0 then SIBROS_GENDER06_1993=.;
  if SIBROS_HGC06_1993<0 then SIBROS_HGC06_1993=.;
  if SIBROS_NUMKIDS06_1993<0 then SIBROS_NUMKIDS06_1993=.;
  if SIBROS_AGEAT1STBIRTH06_1993<0 then SIBROS_AGEAT1STBIRTH06_1993=.;
  if SIBROS_ANY_MORE_SIBS06_1993<0 then SIBROS_ANY_MORE_SIBS06_1993=.;
  if sibid_07<0 then sibid_07=.;
  if SIBROS_OLDER_YOUNGER07_1993<0 then SIBROS_OLDER_YOUNGER07_1993=.;
  if SIBROS_YRS_OLDER_YOUNGER07_1993<0 then
SIBROS_YRS_OLDER_YOUNGER07_1993=.;
  if SIBROS_GENDER07_1993<0 then SIBROS_GENDER07_1993=.;
  if SIBROS_HGC07_1993<0 then SIBROS_HGC07_1993=.;
  if SIBROS_NUMKIDS07_1993<0 then SIBROS_NUMKIDS07_1993=.;
  if SIBROS_AGEAT1STBIRTH07_1993<0 then SIBROS_AGEAT1STBIRTH07_1993=.;
  if SIBROS_ANY_MORE_SIBS07_1993<0 then SIBROS_ANY_MORE_SIBS07_1993=.;
  if sibid_08<0 then sibid_08=.;
  if SIBROS_OLDER_YOUNGER08_1993<0 then SIBROS_OLDER_YOUNGER08_1993=.;
  if SIBROS_YRS_OLDER_YOUNGER08_1993<0 then
SIBROS_YRS_OLDER_YOUNGER08_1993=.;
  if SIBROS_GENDER08_1993<0 then SIBROS_GENDER08_1993=.;
  if SIBROS_HGC08_1993<0 then SIBROS_HGC08_1993=.;
  if SIBROS_NUMKIDS08_1993<0 then SIBROS_NUMKIDS08_1993=.;
  if SIBROS_AGEAT1STBIRTH08_1993<0 then SIBROS_AGEAT1STBIRTH08_1993=.;
  if SIBROS_ANY_MORE_SIBS08_1993<0 then SIBROS_ANY_MORE_SIBS08_1993=.;
  if sibid_09<0 then sibid_09=.;
  if SIBROS_OLDER_YOUNGER09_1993<0 then SIBROS_OLDER_YOUNGER09_1993=.;
  if SIBROS_YRS_OLDER_YNGR09_1993<0 then SIBROS_YRS_OLDER_YNGR09_1993=.;
  if SIBROS_GENDER09_1993<0 then SIBROS_GENDER09_1993=.;
  if SIBROS_HGC09_1993<0 then SIBROS_HGC09_1993=.;
  if SIBROS_NUMKIDS09_1993<0 then SIBROS_NUMKIDS09_1993=.;
  if SIBROS_AGEAT1STBIRTH09_1993<0 then SIBROS_AGEAT1STBIRTH09_1993=.;
  if SIBROS_ANY_MORE_SIBS09_1993<0 then SIBROS_ANY_MORE_SIBS09_1993=.;
  if sibid_10<0 then sibid_10=.;
  if SIBROS_OLDER_YNGR10_1993<0 then SIBROS_OLDER_YNGR10_1993=.;
  if SIBROS_YRS_OLDER_YNGR10_1993<0 then SIBROS_YRS_OLDER_YNGR10_1993=.;
  if SIBROS_GENDER10_1993<0 then SIBROS_GENDER10_1993=.;
  if SIBROS_HGC10_1993<0 then SIBROS_HGC10_1993=.;
  if SIBROS_NUMKIDS10_1993<0 then SIBROS_NUMKIDS10_1993=.;
  if SIBROS_AGEAT1STBIRTH10_1993<0 then SIBROS_AGEAT1STBIRTH10_1993=.;
  if SIBROS_ANY_MORE_SIBS10_1993<0 then SIBROS_ANY_MORE_SIBS10_1993=.;
  if sibid_11<0 then sibid_11=.;
  if SIBROS_OLDER_YOUNGER11_1993<0 then SIBROS_OLDER_YOUNGER11_1993=.;
  if SIBROS_YRS_OLDER_YOUNGER11_1993<0 then
SIBROS_YRS_OLDER_YOUNGER11_1993=.;
  if SIBROS_GENDER11_1993<0 then SIBROS_GENDER11_1993=.;
  if SIBROS_HGC11_1993<0 then SIBROS_HGC11_1993=.;
  if SIBROS_NUMKIDS11_1993<0 then SIBROS_NUMKIDS11_1993=.;
  if SIBROS_AGEAT1STBIRTH11_1993<0 then SIBROS_AGEAT1STBIRTH11_1993=.;
  if SIBROS_ANY_MORE_SIBS11_1993<0 then SIBROS_ANY_MORE_SIBS11_1993=.;
  if sibid_12<0 then sibid_12=.;
  if SIBROS_OLDER_YNGR12_1993<0 then SIBROS_OLDER_YNGR12_1993=.;
  if SIBROS_YRS_OLDER_YNGR12_1993<0 then SIBROS_YRS_OLDER_YNGR12_1993=.;
  if SIBROS_GENDER12_1993<0 then SIBROS_GENDER12_1993=.;
  if SIBROS_HGC12_1993<0 then SIBROS_HGC12_1993=.;
  if SIBROS_NUMKIDS12_1993<0 then SIBROS_NUMKIDS12_1993=.;
  if SIBROS_AGEAT1STBIRTH12_1993<0 then SIBROS_AGEAT1STBIRTH12_1993=.;
  if SIBROS_12_SIBS_REPORTED_1993<0 then SIBROS_12_SIBS_REPORTED_1993=.;
  if SIBROS_YOUNGEST_SIB_1993<0 then SIBROS_YOUNGEST_SIB_1993=.;
  if SIBROS_YOUNGEST_SIBREP_1993<0 then SIBROS_YOUNGEST_SIBREP_1993=.;
  if sibid_13<0 then sibid_13=.;
  if SIBROS_OLDER_YOUNGER13_1993<0 then SIBROS_OLDER_YOUNGER13_1993=.;
  if SIBROS_YRS_OLDER_YOUNGER13_1993<0 then
SIBROS_YRS_OLDER_YOUNGER13_1993=.;
  if SIBROS_GENDER13_1993<0 then SIBROS_GENDER13_1993=.;
  if SIBROS_HGC13_1993<0 then SIBROS_HGC13_1993=.;
  if SIBROS_NUMKIDS13_1993<0 then SIBROS_NUMKIDS13_1993=.;
  if SIBROS_AGEAT1STBIRTH13_1993<0 then SIBROS_AGEAT1STBIRTH13_1993=.;
  if SIBROS_ANY_SIBS_NOT_LISTED_1993<0 then
SIBROS_ANY_SIBS_NOT_LISTED_1993=.;
  if SIBROS_TOTAL_NUM_SIBS_1993<0 then SIBROS_TOTAL_NUM_SIBS_1993=.;
  if Q1110_A_2006<0 then Q1110_A_2006=.;
  if Q1110_B_2006<0 then Q1110_B_2006=.;
  if Q1110_A_2008<0 then Q1110_A_2008=.;
  if Q1110_B_2008<0 then Q1110_B_2008=.;
  if Q1110_A_2010<0 then Q1110_A_2010=.;
  if Q1110_B_2010<0 then Q1110_B_2010=.;
```

```
work_experience = ((age - hgc) -6);
work_experience2 = work_experience*work_experience;
run;

proc reg;
model lnwages = HGC_Mother HGC_Father AFQT enrolled technical sales
administrative service farming repair assemblers transportation laborers
nottogether NC S W NM full_time tenure HGC Work_experience work_experience2
underweight overweight obese;
run;
proc reg;
model lnwages = HGC_Mother HGC_Father AFQT enrolled technical sales
administrative service farming repair assemblers transportation laborers
nottogether NC S W NM full_time tenure HGC Work_experience work_experience2
bmi;
run;
proc reg;
model lnwages = HGC_Mother HGC_Father AFQT enrolled technical sales
administrative service farming repair assemblers transportation laborers
```

```
nottogether NC S W NM full_time tenure HGC Work_experience work_experience2
weight height;
run;


/* This is the first method that uses proc sql, based on siblings are likely
to share the same household id (hhid) in 1979.
The potential problem with this method is , they could actually be cousins,
or other non sibling relations.
So we would like to use a second data source to collaborate. */


data ds1;
      set datatotal;
run;

data ds2;
      set datatotal;
run;


proc sql;
      create table ds3 as
      select A.*, B.bmi as sibling_bmi, B.caseid as sibling_caseid
      from ds1 as A, ds2 as B
      where  A.hhid= B.hhid and A.caseid ~= B.caseid and
( A.sibid_01 = B.caseid or A.sibid_02 = B.caseid or A.sibid_03 = B.caseid or
A.sibid_04 = B.caseid or A.sibid_05 = B.caseid or A.sibid_06 = B.caseid or
A.sibid_07 = B.caseid or A.sibid_08 = B.caseid or A.sibid_09 = B.caseid or
A.sibid_10 = B.caseid or A.sibid_11 = B.caseid or A.sibid_12 = B.caseid or
A.sibid_13 = B.caseid or
 B.sibid_01 = A.caseid or B.sibid_02 = A.caseid or B.sibid_03 = A.caseid or
B.sibid_04 = A.caseid or B.sibid_05 = A.caseid or B.sibid_06 = A.caseid or
B.sibid_07 = A.caseid or B.sibid_08 = A.caseid or B.sibid_09 = A.caseid or
B.sibid_10 = A.caseid or B.sibid_11 = A.caseid or B.sibid_12 = A.caseid or
B.sibid_13 = A.caseid )
      ;
      quit ;

proc print data=ds3 (obs=20);
run;

proc syslin 2sls;
endogenous bmi;
instruments sibling_bmi HGC_Mother HGC_Father Region Marital AFQT Enrolled
Occupation Full_time Tenure work_experience work_experience2 HGC;
model lnwages= HGC_father HGC_Mother AFQT Enrolled Occupation Full_Time
Tenure Region Marital HGC work_experience work_experience2 bmi;
run;
proc syslin 2sls;
endogenous weight;
instruments sibling_bmi HGC_Mother HGC_Father Region Marital AFQT Enrolled
Occupation Full_time Tenure work_experience work_experience2 HGC;
model lnwages= HGC_father HGC_Mother AFQT Enrolled Occupation Full_Time
Tenure Region Marital HGC work_experience work_experience2 weight height;
run;
proc means;
```

```
var sibling_bmi HGC_Mother HGC_Father Region Marital AFQT Enrolled Occupation
Full_time Tenure Work_experience work_experience2 HGC wages bmi overweight
underweight obese height weight;
run;
proc reg;
model lnwages = HGC_Mother HGC_Father AFQT enrolled technical sales
administrative service farming repair assemblers transportation laborers
nottogether NC S W NM full_time tenure HGC Work_experience work_experience2
underweight overweight obese;
run;
proc reg;
model lnwages = HGC_Mother HGC_Father AFQT enrolled technical sales
administrative service farming repair assemblers transportation laborers
nottogether NC S W NM full_time tenure HGC Work_experience work_experience2
bmi;
run;
proc reg;
model lnwages = HGC_Mother HGC_Father AFQT enrolled technical sales
administrative service farming repair assemblers transportation laborers
nottogether NC S W NM full_time tenure HGC Work_experience work_experience2
weight height;
run;


/* The second data sourse, is in 1993, the respondent identified the siblings
that are also in NLSY79.
sibid_01 to sibid_01 identifies the siblings' caseid.


*/
```