

User Guide Tutorial #5-B

Intergenerational Correlations in Health Outcomes Ngina Chiteji

and Frank Stafford December, 2003

I. Introduction and Overview of the Panel Study of Income Dynamics

Many people are interested in correlations in health outcomes across generations. We often joke about living till 100 if we have an elderly relative who has lived a long time, or we speculate about whether a child will inherit mommy's strong constitution or daddy's height. Such interest often stems from the fact that ailments and physical attributes can be transmitted genetically, although this is not the only reason that health outcomes may be similar across the generations. Environmental factors also affect health, and in many instances parents and children share the same physical environment. Irrespective of the cause, it is

clear that many health outcomes exhibit some correlation across the generations. The purpose of this tutorial exercise is to demonstrate how one might look for such correlations across three generations of the same family. The tutorial's focus is on general health. It demonstrates how one might conduct the type of analysis that is necessary to determine whether children who are healthy tend to have mothers who are also generally healthy, and grandmothers in a similar situation. It will draw on data from the 1997 Child Development

Supplement (CDS) for the youngest generation, and data from the 1986 family files to obtain information about the two older generations. Upon completion, you should have a good idea of how the Panel Study of Income Dynamics (PSID) can be used to make extended intergenerational comparisons.

As noted in the companion tutorial (5-A), the PSID is a nationally representative, longitudinal dataset that contains a wide variety of economic and socio-demographic data about individuals and families. It has followed families and their offspring, and the families set up by these offspring, since 1968, and in 1997 and 2001 special supplements were devoted specifically to gathering information about dependent children in PSID families. Accordingly, the data can support a number of possible intergenerational analyses.

II. The Sample Exercise Examining Correlation in Health Outcomes

When analyzing the situation of the youngest generation (the children) low birthweight will be used as a proxy for poor health. Because low birthweight has been linked to medical problems later in life, it can be taken as an early guide to an individual's health situation (see FIFCFS, 2001, for example). For the adults, the PSID contains an actual question about general health. Once the exercise has been completed, you will be able to fill in the blank cells of the Table 1, to determine whether the likelihood of a child's having a low birthweight appears to be associated with having a mother or a grandmother whose health is also poor.

Table 1. Intergenerational Health Patterns

	All Children	Mother's Health	Grandmother's health
Birthweight	poor	not poor	poor
% low			

% not low					
-----------	--	--	--	--	--

III. Getting your data at the Data Center

To do our analysis we need to obtain three different customized datasets at the Data Center. The information from these datasets will be merged eventually, however when using Excel it is easiest to begin working with separate datasets for each generation. Accordingly, we start by obtaining a dataset that contains information about the children (the third generation). Then, we obtain a dataset that contains information about the health status of these children's moms (the second generation). And, finally, we have the Data Center output a dataset containing health data for grandmothers (generation one). To then use these three datasets you will need to carry out pairwise matching (or linking) based on identifier variables.

What variables will you need for your kids datasets?

In the dataset for generation three, you need six variables. If you have ever seen a birth announcement, you probably know that birthweight is reported as a pound-ounce combination, as 10 lbs and 11 ounces for example (for a rather large baby). Accordingly, one has to factor in information from two variables from the Child Development Supplement--Q1A8P (pounds) and Q1A8O (ounces)--in order to determine how much a given child weighed at birth. Because the CDS also contains a question about the child's general health at birth, we choose this variable (Q1A11) too, to go along with our birthweight information, for purposes of comparison.

Next we need some variables that allow us to link each child to his/her mom. When using CDS data one can establish this link by obtaining three variables that contain information about the child's primary caregiver (PCG). Two of these variables will be used simultaneously. They are the 1968 family identification number of the PCG (PCG68ID) and the PCG's person number (PCGPN). In the PSID, an individual's 1968 family id number and person number are variables which are critical because they are the variables that are used to uniquely identify each individual ever included in the survey. The crucial point to remember, when contemplating use of the 1968 family id and the 1968 person number, is that one can use these two variables to locate an individual throughout the different years of the PSID (and its supplements), and to determine the original sample family from which an individual has descended. (A refresher sample of post-1968 immigrants, or their adult children, was added to the PSID in 1997. They too have identifier variables.) In the CDS one is able to obtain these identification variables for each child, a n d for the child's primary caregiver. (One also can obtain them for each child's secondary caregiver--the OCG--but we will not be working with the OCG data). This facilitates the process of matching information about moms with information about children. Each PCG's unique identifier can be used to locate information about her from different years in the PSID, and this can then be matched to the information reported about the child directly in the CDS. Finally, we will use the variable recording the PCG's relationship to the child (the variable "PCG") to identify the cases in which the mom is the primary caregiver of the child. This allows us to be certain that when we use the PCG68ID and PCGPN to match information about adults' health status (taken from the PSID's individual files) to the records of children (taken from the CDS), we are doing such in a way that attaches information about the mother (rather than the father or some other relative) to the information about the birthweight of the child. (The PCG is not the mom by construction in the CDS, so there are some instances in which a child's primary caregiver is not the mom. In this tutorial we consider the cases where the PCG is the mom, including an adoptive mother, for ease of exposition. As noted in Tutorial 4, we do not lose too many cases by imposing this requirement.)

What variables will you need for the moms dataset?

The information about mothers' health will be obtained from a question about general health

conditions that was asked for different members of the family in the 1986 wave of the PSID. These are questions H15, H37, and H69e. While additional data on individuals' health status have been collected in more recent years, in 1997 for example, the tutorial focuses on the 1986 data. This will allow it to show how data from different years can be combined for analysis purposes.

In the dataset for generation two, you need a total of 6 variables. These variables are (1) ER32000 (sex of the individual), (2) ER30001 (1968 family identification number for the individual at hand), (3) ER30002 (1968 person number for this individual) (4) ER30527 (H1/37/69 Health Status in '86), (5) ER32009 (the 1968 family identification number of the mother of the individual at hand), and (6) ER32010 (the person number of the individual's mother).

Why do you want these particular variables? The first variable will be used in a subsetting command to create our customized dataset for moms. The information about general health status is reported at the individual-level, and is contained in files containing information about all individuals in the PSID in 1986. Because our objective is to obtain health information for moms only, we know that we are only interested in obtaining this individual-level data for women. Accordingly, it behooves us to instruct the Data Center to restrict our output to cases in which the individual in question is female. These will correspond to cases in which the ER32000 variable takes on a code value of two. (You can confirm these code values by examining the documentation once you have selected your variables, as explained below.) The second and third variables will be used to construct the PSID's unique identifier for each woman whose record we obtain. These will then be matched to the identifiers for the PCGs that we have in our kids dataset when we put our three datasets together. We want the fifth and sixth variables because they are variables that can be used to construct an identification variable that uniquely identifies the mothers of the individuals for whom we are gathering information. This is critical because our three-generation analysis requires us to match information about kids, moms, and grandmothers, however we cannot link the grandmothers of the CDS kids together with those kids in a direct fashion because there is no variable designated as "grandmother's" 1968 family ID or person number. Instead, we have to link grandmothers (generation one) and grandchildren (generation three) indirectly--by using the children's moms as the bridge. As noted above, the CDS provides the 1968 family ID and person numbers for each child's PCG, who is the mom in most instances. In the step that we are discussing here we will be ensuring that we draw the 1968 identification variables for these moms' mothers, so that when we put our moms together with their CDS children, we can also use the identifiers found for moms' mothers to simultaneously attach the information about each child's grandmother to the child. So, just to make sure we are all thinking about our strategy in the same way and that we understand how the 4 identification variables that we are including in the dataset for generation two will be used, let's review a bit: We will use the moms' identification variables to link moms with the CDS kids, and then we will use moms' mothers' identification variables to add in information about the grandmothers.

Finally, remember that the entire point of this tutorial exercise is to compare the health of the members of the different generations. ER30527 is the general health status variable that we will use to characterize moms' health. It contains the responses to a query asking whether the health of the individual in question is in excellent, very good, good, fair or poor.

One last point to note: Note that for the type of exercise outlined in this tutorial we do not need to obtain any information from the PSID yearly family files, and that this allows us to forego worrying about the "relationship to head" variables that some users may be familiar with from other PSID tutorials. In some instances data about individuals is collected for the heads and wives of PSID families only, and recorded within the family file (rather than in the individual files). If one ever wants to work with any of this type of information ---race, age, or an individual's labor income, for example---one would have to restrict one's analysis to individuals who are heads (or wives), and one would use the relationship to head variable for the year in question to do such. While you might have expected us to have to take this route, i.e., to restrict ourselves to instances in which the mom was a head or wife in a 1986 family, for example, to make sure that she was in her own household and not a dependent individual in another person's household, it is not necessary to impose such a restriction here because the 1986 health data was collected for all individuals in

the family, not just the head and wife. (For 1999, 2001, and 2003, however, these data are collected for heads and wives only.)

What variables will you need for the grandmothers' dataset?

The information that we want for generation one is very similar to the information that we obtained for generation two. Our primary interest is in obtaining information about individuals' health status, and this information will come from the 1986 individual file. However, we also need identification variables so that we can link grandmothers with their offspring (the moms). Furthermore, to construct this particular dataset, we make use of the relationship-to-head information provided in the PSID. (This is largely to keep the size of the resultant Excel file manageable. It is not a variable that is central to the analysis.) In total then, we need the following six variables: (1) ER32000--the variable listing the individual's sex, (2) ER30527--the variable characterizing the individual's general health status, (3) ER30001--the 1968 family ID number of the individual, (4) ER30002--the 1968 person number for an individual, (5) ER30500--the variable characterizing an individual's relationship to the head of the 1986 family that he/she resided in, and (6) ER30499--a variable which is used in conjunction with the relationship-to-head variable in instances in which we want to make sure the individual listed as household head is currently in the family.

What is the general strategy for compiling the grandmothers dataset? As with the moms dataset, we draw information from two "data groups" in the Data Center. The gender information comes from the "Sampling Variables" data group, and the remainder of the variables come from the 1986 individual file. Additionally, we use the sex variable when the Data Center prompts us for subsetting criteria. We also will use the relationship-to-head variable and the sequence number at this juncture. As a consequence, instead of drawing all individuals who exist in the PSID in 1986, the Data Center will produce a customized dataset that only contains information for individuals who happen to be women, and for women who are old enough to have set up their own households only (rather than someone who is a six-year old daughter of someone else, for example). This is important because the PSID contains information on a large number of individuals (close to 70,000 people in 1997 for example), yet we know that we only want individuals who happen to be grandmothers in 1997, which means we know we will not need the health status of every individual in the PSID. Instead, we can go a long way toward eliminating individual records that do not stand a chance of making a match to our CDS kids right from the start (individuals who are male for example).

How (and where) do you get your data for the kids dataset?

The PSID Data Center can be found at <http://simba.isr.umich.edu>. To construct the dataset with the children's information, you want to choose "variable selection" and then check the box for 1997 Child Development Supplement (CDS) under "select data groups."

Screenshot #1 Select Data Group(s)

Select Data Group(s)	
PSID Data	
<input type="checkbox"/>	Sampling variables, birth & marital variables (subset), gender
<input type="checkbox"/>	PSID Individual
<input type="checkbox"/>	PSID Family
<input type="checkbox"/>	Income Plus
<input type="checkbox"/>	Work Hours and Wages
<input type="checkbox"/>	Family Wealth
<input type="checkbox"/>	Family Weights
<input type="checkbox"/>	Other Family Data
Child Development Supplement	
<input checked="" type="checkbox"/>	CDS
<input type="checkbox"/>	Time Diaries

If you hit continue you will be taken to a screen where you can select specific files from the 1997 CDS:

Screenshot #2 Select data categories and years

Select data categories and years	
CDS	
<input checked="" type="checkbox"/>	1997 Demographic File - including weights
<input checked="" type="checkbox"/>	1997 Primary Caregiver Child File
<input type="checkbox"/>	1997 Primary Caregiver Household File
<input type="checkbox"/>	1997 Child File
<input type="checkbox"/>	1997 Other Caregiver Child File
<input type="checkbox"/>	1997 Other Caregiver Household File
<input type="checkbox"/>	1997 Fathers Outside the Home Child File
<input type="checkbox"/>	1997 Fathers Outside the Home Household File
<input type="checkbox"/>	1997 Elementary/Middle School Teacher File
<input type="checkbox"/>	1997 Preschool/Daycare Teacher File
<input type="checkbox"/>	1997 Home-Based Care Provider File
<input type="checkbox"/>	1997 Elementary/Middle School Administrator File
<input type="checkbox"/>	1997 Preschool/Daycare Administrator File
<input type="checkbox"/>	1997 Food Security Status File
<input type="checkbox"/>	1997 TD Followup

Click on the box next to "Demographic file," and the box beside "Primary Caregiver Child File." This will take you to a screen allowing you to select specific variables:



Screenshot #3 Variable Selection

1997 Demographic File - including weights: 3,563 observations	
SEQNUM97 - 1997 SEQUENCE NUMBER	▲
RELHD - RELATIONSHIP TO HEAD	☰
AGECORE - AGE OF CHILD AT 1997 PSID CORE INTERVIEW	
SEXCH - SEX OF CHILD	
CHRA6 - CHILD RACE (A6)	
BIRTHMON - CHILD'S BIRTH MONTH	
BIRTHYR - CHILD'S BIRTH YEAR	
CHIWMON - CHILD INTERVIEW- MONTH COMPLETED	
CHIWDAY - CHILD INTERVIEW- DAY COMPLETED	
CHIWYR - CHILD INTERVIEW- YEAR COMPLETED	▼

1997 Primary Caregiver Child File: 3,563 observations	
Q1A0 - EXACT TIME	▲
Q1A1 - ASSESSMENT	☰
Q1PC1 - PASSAGE COMPREHENSION	
Q1PC2 - PASSAGE COMPREHENSION	
Q1PC3 - PASSAGE COMPREHENSION	
Q1PC4 - PASSAGE COMPREHENSION	
Q1PC5 - PASSAGE COMPREHENSION	
Q1PC6 - PASSAGE COMPREHENSION	
Q1PC7 - PASSAGE COMPREHENSION	
Q1PC8 - PASSAGE COMPREHENSION	▼

Now select the following 7 variables. From the 1997 CDS Demographic File select (1) PCG (the relationship of the child's primary caregiver to the child), (2) PCGID68 (the 1968 family ID number of the primary caregiver), (3) PCGPN (the person number of the primary caregiver), and (4) CH97PRWT. (This last variable provides the sampling weights for the CDS children, who are age 0-12, as of 1997.) From the CDS Child File you want to select your last three variables: (5) Q1A8P (birthweight-pounds), (6) Q1A8O (birthweight-ounces), and (7) Q1A11 (health at birth). In case you are wondering about the "view" option beside each variable name, this is a feature that allows you to remind yourself what a particular variable is, and how it is coded (to make sure you really have the correct variable, for example). If you click on the word "view" the Data Center will respond by displaying a window containing codes and the actual question asked along with other pertinent documentation for the variable that you have inquired about. Note that the Data Center automatically adds two variables that you need to your list--these are the child's 1968 family identification number and the child's person number. (Note that the Data Center also automatically adds three other variables--ER33401, ER33402, and ER33403. You will not need these, but we will refrain from deleting them until later. They are often critical for analysis using PSID data, as seen in the companion tutorial 5-A for example, so the Data Center is designed to ensure that they are included by default, even though this is one of the rare cases in which they are not needed.)

Screenshot #4 PSID/CDS Data Cart Contents

PSID/CDS Data Cart Contents						
Total PSID Variables Selected: 10						
Year	Data File	Name	Label	Data Group	<input type="checkbox"/>	
1997	PSID Individual Data by Years	ER33401	1997 INTERVIEW NUMBER	PSID Individual	<input type="checkbox"/>	[View]
1997	PSID Individual Data by Years	ER33402	SEQUENCE NUMBER 97	PSID Individual	<input type="checkbox"/>	[View]
1997	PSID Individual Data by Years	ER33403	RELATION TO HEAD 97	PSID Individual	<input type="checkbox"/>	[View]
1997	Demographic File - including weights	PCG	relationship of PCG to child	CDS	<input type="checkbox"/>	[View]
1997	Demographic File - including weights	PCGID68	PSID 1968 ID of PCG	CDS	<input type="checkbox"/>	[View]
1997	Demographic File - including weights	PCGPN	PSID PN of PCG	CDS	<input type="checkbox"/>	[View]
1997	Demographic File - including weights	CH97PRWT	child level weight	CDS	<input type="checkbox"/>	[View]
1997	Primary Caregiver Child File	Q1A8P	BIRTH WEIGHT (POUNDS)	CDS	<input type="checkbox"/>	[View]
1997	Primary Caregiver Child File	Q1A8O	BIRTH WEIGHT (OUNCES)	CDS	<input type="checkbox"/>	[View]
1997	Primary Caregiver Child File	Q1A11	HEALTH AT BIRTH	CDS	<input type="checkbox"/>	[View]
Automatic Individual Identification Variables						
Name		Label				
ER30001 (ID68)		1968 INTERVIEW NUMBER		[View]		
ER30002 (PN)		PERSON NUMBER 68		[View]		

Next hit "Get Data and/or Codebook" and a screen depicting several output options will be displayed.

Screenshot #5

There are 5 important steps to take here. First, one of the options that you have surrounds the format of your codebook (and whether you even want one). The default is no codebook, however if you want a codebook to be delivered to you so that you can remember how the variables that you have selected are coded, and what questions were asked to get your data, you should specify a codebook type (html or pdf). The second option you have concerns the format of your data. Click on the circle next to Microsoft Excel Spreadsheet to follow this tutorial. Third, in the subsetting box, you want to enter the following command: PCG = 1 The number "one" is the code for an instance in which the child's primary caregiver is the mother, so this command instructs the Data Center to limit the output to these cases. (As noted in Tutorial 4, over 95 percent of children have primary caregivers who are mothers, so we do not lose too many cases by restricting the analysis in this fashion, while we gain much in terms of ease of exposition. If we were to include children whose mothers were not their PCGs we could not make use of the PCGID68 and PCGPN variables that are provided in the children's records from the CDS to locate these moms' health information elsewhere in the PSID. This would complicate the task of matching children's and moms' information, though it does not render the task impossible. The inquisitive user can think about how one might do such matching after he/she understands the direct route that we illustrate here in the tutorial.) The fourth step is to provide your e-mail address if you want the Data Center to e-mail you to alert you when your dataset is ready. The e-mail will include the location of the Excel file containing your dataset, labels, and a link to your codebook if you have asked for one. (If you do not want to provide an e-mail address, you have the option of simply waiting at the Data Center while it produces your dataset. A job completion notice, such as the one depicted below, will be displayed on the screen once the dataset has been created.) The fifth and last step for you to take at this point is to click on the "CDS kids only" box at the bottom of the screen. This tells the Data Center that you want information about children only at this point.

Now you can hit "submit" and the Data Center will get to work to put together your customized dataset containing information about children. At this point, you can simply wait for the Data Center to finish, or you can close this window, and begin a new internet session so you can get to work compiling your moms' dataset.

Screenshot #6 Job completion notice

Select Data Group(s)	
PSID Data	
<input type="checkbox"/>	Sampling variables, birth & marital variables (subset), gender
<input type="checkbox"/>	PSID Individual
<input type="checkbox"/>	PSID Family
<input type="checkbox"/>	Income Plus
<input type="checkbox"/>	Work Hours and Wages
<input type="checkbox"/>	Family Wealth
<input type="checkbox"/>	Family Weights
<input type="checkbox"/>	Other Family Data
Child Development Supplement	
<input checked="" type="checkbox"/>	CDS
<input type="checkbox"/>	Time Diaries

How to construct the moms dataset

To obtain the data for your dataset for generation two, begin a new session at the Data Center, and select "PSID individual" as your data group, along with the "Sampling variables" data group option.

Screenshot #7 Select Data Group(s)

PSID Data	
<input checked="" type="checkbox"/>	Sampling variables, birth & marital variables (subset), gender
<input checked="" type="checkbox"/>	PSID Individual
<input type="checkbox"/>	PSID Family
<input type="checkbox"/>	Income Plus
<input type="checkbox"/>	Work Hours and Wages
<input type="checkbox"/>	Family Wealth
<input type="checkbox"/>	Family Weights
<input type="checkbox"/>	Other Family Data
Child Development Supplement	
<input type="checkbox"/>	CDS
<input type="checkbox"/>	Time Diaries

Next, under the list of year options specified for the individual-level data (the PSID individual files), check the box next to 1986. Additionally, you want to check the box beside "All years" under the "sampling variables" heading.

Screenshot #8 Select data categories and years

Select data categories and years									
Sampling variables, birth & marital variables (subset), gender									
<input checked="" type="checkbox"/> All Years									
PSID Individual Data by Years									
<input type="checkbox"/> 2001	<input type="checkbox"/> 1999	<input type="checkbox"/> 1997	<input type="checkbox"/> 1996	<input type="checkbox"/> 1995	<input type="checkbox"/> 1994	<input type="checkbox"/> 1993	<input type="checkbox"/> 1992	<input type="checkbox"/> 1991	<input type="checkbox"/> 1990
<input type="checkbox"/> 1989	<input type="checkbox"/> 1988	<input type="checkbox"/> 1987	<input checked="" type="checkbox"/> 1986	<input type="checkbox"/> 1985	<input type="checkbox"/> 1984	<input type="checkbox"/> 1983	<input type="checkbox"/> 1982	<input type="checkbox"/> 1981	<input type="checkbox"/> 1980
<input type="checkbox"/> 1979	<input type="checkbox"/> 1978	<input type="checkbox"/> 1977	<input type="checkbox"/> 1976	<input type="checkbox"/> 1975	<input type="checkbox"/> 1974	<input type="checkbox"/> 1973	<input type="checkbox"/> 1972	<input type="checkbox"/> 1971	<input type="checkbox"/> 1970
<input type="checkbox"/> 1969	<input type="checkbox"/> 1968								

Next you need to select three variables from the "All years--Sampling variables" list. These variables are (1) ER32000 (sex of individual), (2) ER32009 (1968 family ID of mother), and (3) ER32010 (person number of mother). Furthermore, you need to select one variable from the 1986 individual file. That variable is ER30527 (H1/37/69 Health Status in 86). After you have selected these variables hit "continue" and a datacart depicting the variables in your dataset will appear. Note that the Data Center adds the last two variables that we want, ER30001 and ER30002, automatically. You do not need to select them separately. (Also, note that the Data Center also adds three other variables automatically. They are ER30498, ER30499 and ER30500. As noted earlier, these three are sometimes critical for PSID analyses so the Data Center assumes that you might need them, even though this tutorial presents one of the rare instances in which you will not. We will delete them later in the exercise.)

Now hit "Get Data" and the Data Center will display a screen listing several output options for you to choose from.

Screenshot #9

Output Options

Select codebook type:

HTML (no frames)
 HTML (frames)
 PDF
 XML
 No thanks, I don't need a codebook

Codebook only (no data will be output)

Select data output type:

ASCII Data File Data Definition Statements: SAS
 SPSS
 Stata

Microsoft Excel Spreadsheet

dBase Data File (DBF) / Spreadsheet

SAS Transport Data File

Subsetting Criteria: Enter optional subsetting criteria. Example: 1 <= ER30001 <= 2930 [\[More Examples\]](#)
Please note that the variables used in subsetting criteria must be in your [variable selection](#).

ER32000 = 2

Email Notification: (Optional) [What's this?](#)

Email Address: Project Name (Optional):

Make my cart available to the public: [What's this?](#) Yes No

Data File Options
For PSID Individual Level Data - [What's this?](#)

PSID core individuals	CDS intergenerational individuals
<input type="radio"/> All individuals <input type="radio"/> Only heads current in *any* years selected	<input checked="" type="radio"/> PCG <input type="radio"/> OCG <input type="radio"/> Children

Once again, there are 5 important steps to take at this stage. First, you want to indicate whether you would like a codebook and what format you would like it to be in. Second, to follow this tutorial, you want to instruct the Data Center to deliver your dataset in Excel format. Third, you want to enter a subsetting command. That command is, ER32000 =2 This instructs the Data Center to only output information for individuals who are female. Fourth you want to enter your e-mail address if you are having your dataset, labels, and codebook e-mailed to you. Finally, in the CDS intergenerational individuals section in the bottom right hand corner of the screen, you want to fill in the circle next to the words "PCG." This last step instructs the Data Center to impose a further restriction on your dataset. It will not only restrict the output to cases in which the individuals are women, it also will limit the output to women who happen to be PCGs. (This means you won't be troubled with records for individuals who happen to be women but not to have had any children age 0-12 as of 1997. We do not want these individuals because we know a priori that their unique individual identifiers will not match any of the child records that are contained in our dataset for generation three.)

Now you can hit the submit command and the Data Center will get to work to create your customized dataset with information about moms. Again, at this point you have the option of simply closing the window and opening a new internet session so that you can get to work selecting the variables for your third and final dataset without having to wait for the Data Center to finish your moms dataset first. (And, again, if you entered an e-mail address you will know when your moms dataset is ready because you will receive it via e-mail. Or, as noted under the guidelines for the kids dataset, if you did not provide an e-mail address you will see a screen with a job completion notice once the Data Center has finished.)

How to construct the grandmothers dataset

To obtain the data for the generation-one dataset, you will again need to begin a new session at the Data Center. And, once again, you select "PSID Individual" and "Sampling Variables" as your two data groups. Under the list of options specified after you hit continue, you want to check the box next to 1986 as you did for the moms dataset, and to check the box beside the phrase "All years."

Next choose the following four variables from the list of options presented on the screen: (1) ER32000 from the "All years" category, (2) ER30499 from the 1986 Individual file, (3) ER30500 from the 1986 Individual file, and (4) ER30527 from the 1986 Individual file. Your last two desired variables, ER30001 and ER30002, will be added automatically by the Data Center. (Note that the Data Center also will add ER30498 to your variable list, under the assumption that you might also need it, because it is typically a critical one for PSID analyses. We will delete it later in the exercise.) At this point you will see a screen of the sort presented above in Screenshot # 8.

Now hit "Get Data" and the screen listing the output options will materialize. At this point, you probably can guess what you need to do next. First, you need to tell the Data Center whether you want a codebook created, that you want an Excel file created, and if/where to e-mail your dataset, labels and codebook. Additionally, you want to enter a subsetting command. That command is as follows:

```
ER32000 = 2 and ((ER30500 = 10 and ER30499 = 1) or ER30500 = 20 or  
ER30500 = 22)
```

This subsetting command instructs the Data Center to limit the output to individual data for women, and to restrict output to records for females who were living independently (as heads or wives rather than dependent children) in 1986. The subsetting commands need to be typed carefully and there is help for subsetting formats in the Data Center. (Again, this is done largely to restrict our file size to one that is manageable in Excel, yet it should not dramatically affect the ultimate results.) Finally, at the bottom of the screen you can select the box for "all individuals."

Screenshot #10

Output Options

Select codebook type:

HTML (no frames)
 HTML (frames)
 PDF
 XML
 No thanks, I don't need a codebook

Codebook only (no data will be output)

Select data output type:

ASCII Data File Data Definition Statements: SAS
 SPSS
 Stata

Microsoft Excel Spreadsheet

dBase Data File (DBF) / Spreadsheet

SAS Transport Data File

Subsetting Criteria: Enter optional subsetting criteria. Example: 1 <= ER30001 <= 2930 [\[More Examples\]](#)
Please note that the variables used in subsetting criteria must be in your [variable selection](#).

ER32000 = 2 and |(ER30500 = 10 and ER30499 = 1) or ER30500 = 20
or ER30500 = 22

Email Notification: (Optional) [What's this?](#)

Email Address: Project Name (Optional):

Make my cart available to the public: [What's this?](#) Yes No

Data File Options
For PSID Individual Level Data - [What's this?](#)

PSID core individuals	CDS intergenerational individuals
<input checked="" type="radio"/> All individuals <input type="radio"/> Only heads current in *any* years selected	<input type="radio"/> PCG <input type="radio"/> OCG <input type="radio"/> Children

Next hit "submit" and the Data Center will begin creating your dataset with information for generation-one (the grandmothers).

IV. Using Excel on your output subsets

In working toward the answer to your research question, you will need to perform some operations on each dataset separately, before you merge your information from the three generations together. The steps that you need to take are outlined below.

A. Working with your CDS kids dataset

1. First, after opening your Excel file containing your first dataset, you want to make a duplicate copy of the worksheet to use for calculations. This way, if you make a mistake, you will still have your original dataset to go back to. To make a copy of your worksheet choose "edit" from the command menu at the top of the screen, then select "move or copy sheet," then highlight the phrase "move to end," and click beside "create a copy."
2. Next, you may want to use the "freeze pane" command to permanently lock your column labels into place so that they remain visible if you scroll down through the dataset. To do this, position your cursor in cell A2, then choose "window" from the command menu, and then select "freeze panes". A line should appear separating row one from the remaining rows of the dataset, and you will see that if you move downward in your dataset the row of column headings remains visible, even if you go all the way down to the bottom of the

dataset. ADDITIONALLY, at this point you can delete the three variables that you will not need, ER33401, ER33402 and ER33403, which were automatically added by the Data Center. You can delete them by selecting highlighting their variable names, and then choosing "edit" from the command menu, and then "delete" and then "entire column(s)".

3. Now let's really get to work. Create a variable that combines the pound and ounces data on birthweight. To do this, we want to create a new column at I to hold the composite birthweight measure. The new column can be created by positioning the cursor in column I and then choosing "insert" from the command menu, and then column. This will create a new column to the left of the one initially containing the cursor. Label this column "total birthweight," and enter the following formula in cell I2:

$$=G2*16 + H2$$

Then, copy this formula into the remaining cells. One easy way to do this is by using Excel's "fill down" feature. If you are not familiar with this feature from work with previous tutorials, all you need to do is to position your cursor at the bottom right corner of cell I2. Then, press down on your mouse and drag the cursor down to the end of the column. As you move downward you should see that Excel is highlighting cells as you go. Once you reach the last cell in the column you can release your grip on the mouse and Excel will automatically fill in all the cells you have highlighted with the formula. (And, it automatically adjusts the formula so that it makes the appropriate row references in each cell to give you the correct total birthweight formula each time.) Now use Excel's "copy" and "paste special" feature to convert the formulas into values. (You do this by highlighting the cells you want to copy--the entire column that is--and then choosing "edit" from the command menu at the top of your screen, and then choosing "copy". Then you reposition your cursor in the first cell of column I, and then return to the "edit" option on the command menu. Select "paste special" and Excel will display a box with options. Choose "values." This instructs Excel to copy the values computed by the formula into each cell.)

4. Now take a moment to glance down at some of the values you are getting for your new, total birthweight measure. Some of them are outrageous--as high as 996004 ounces. This is not reasonable. You're getting such numbers because some of the responses for the birthweight (pounds) variable and the birthweight (ounces) variable given by the Data Center have values like 998 and 999. These are not real birthweights. The PSID simply uses these high values to indicate instances in which the child's birthweight was not known or the individual did not respond to the question, and this is explained for another application in Tutorial #1 (see <http://psidonline.isr.umich.edu/Guide/tutorials/default.aspx>). Accordingly, you now need to eliminate the cases with such values. To do this, you have to sort the dataset by birthweight (pounds) and delete the rows containing either 998 or 999. Then you want to re-sort the dataset by birthweight (ounces) and delete the rows containing either 998 or 999. After you do this, you'll see that you have eliminated such extreme and unreal values from your "total birthweight" column.

How do you sort your datasets? To sort a dataset one has to select "data" from the command menu and then "sort." Excel should highlight your entire dataset in response, and display a box allowing you to indicate which variable (or column) you want to use as the sort variable. If you have your cursor positioned in the column containing the variable that you want to sort by before you select "data" from the command menu, Excel will list this particular variable as the default sort option. (Otherwise, you have to choose from the list of variables given.) Also, make sure to indicate that your first row contains labels--rather than data--by checking on the box next to header row.

How do you delete rows? You do this by highlighting the rows you wish to delete and then choosing "edit" from the command menu at the top of your screen, and then "delete," and then "entire row" from the small box displayed in the middle of the screen.

5. Next you want to create a composite variable that combines the PCG's 1968 identification variable with

her person number, in order to create the special identifier variable that can be used to uniquely identify each individual in the PSID. Create a new, blank column beside the PCGPN column, at F, using the technique you learned in the previous step. Label this column "mom s unique ID." In cell F2 enter the following formula:

$$=(D2*1000) + E2$$

Now copy this formula into the remaining cells of the column by using the "fill down" feature. Then make sure you convert your formulas into values.

6. Before we move on to actual calculations, we need to create a unique identifier for each child whose record appears in our dataset. The formula and procedure for doing such is similar to the one that we used to create the unique identifier for PCGs. First create a new column at C. (By now you should be an expert at creating new columns.) Then enter the following formula in cell C2:

$$=(A2*1000) + B2$$

Then copy the formula in to the remaining cells of column C by using Excel s "fill down" feature. Remember to transform the formulas into actual values using the copy/paste feature when you have finished.

7. Now delete some columns to eliminate variables that you no longer need. Here is a list of the variables to keep: (1) child s unique id, (2) mom s unique id, (3) CH97PRWT (4) total birthweight, and (5) Q1A11 (the health at birth variable). Not much has been said about the CH97PRWT variable at this point. It is a sampling or statistical weight. (Do not confuse it with the birthweight variable, which tells us something about each individual s health condition. As discussed in previous tutorials and in the PSID User Guide (found at <http://psidonline.isr.umich.edu/Guide/UG/tabcont.html>) the statistical weight is a mathematical tool that is employed when we get to the calculations phase of the analysis to ensure that the results are nationally representative.)

8. Next we want to create an index to indicate whether a child s birthweight was low, as defined by the Center for Disease Control. Insert a blank column beside your total-birthweight column, at column E. Label it "low birthweight" and enter the following formula in cell E2:

$$=if(D2 < 88, 1, 0)$$

What is the point of this formula? Medical standards define a child as having a low birthweight if his/her birthweight is below 5.5 pounds. This is equivalent to 88 ounces, and since the total birthweight measure that appears in column D tracks birthweights in ounces, we want to compare the figure in that column to 88. The formula tells Excel to assign a value of 1 for any child whose birthweight is low, and to assign a value of 0 otherwise for our index. (We can think of this as having a "1" to indicate "yes, the birthweight is low," and a zero to indicate "no." Those of you who have some experience working with data will no doubt recognize that the low birthweight variable essentially represents an example of data that is qualitative--since a child s birthweight is either "low" or "not low." However, we want to use numbers to represent these qualitative responses, because it is easier for Excel to work with data entries that are numerical, rather than verbal.) OK, now copy the formula into the remaining cells of the column by using the "fill down" procedure. Then remember to transform your formulas into values using Excel s "copy/paste" special feature.

9. Next use the health status at birth variable to create one more index that summarizes information about children s early health. Use the blank column beside Q1A11, at column G, and label it "bad health." Enter the following formula in cell G2:

$$=(F2 < 3, 0, if(F2 > 3, ".", 1))$$

What does this formula do? It instructs Excel to assign a value of zero if the health at birth was said to be the same as other babies (denoted with a code of 1 for the Q1A11 variable), or better (denoted by a 2 for Q1A11). In the instances in which there was missing data for Q1A11 (when the variable takes on a value of either 8 or 9 because the respondent either did not know the answer to the question or refused to answer it), the "bad-health" index variable will be assigned the universal symbol for missing values--a period ("."). Finally, this formula assigns a value of one to the index variable if the health at birth was coded as either 3, which means worse than other babies.

What else? Make sure you copy the formula to the remaining cells in the column. And, make sure you convert your formulas to values after that.

10. Next, to get ready for the three-way merging that will eventually be done to put all three generations together, we want to label the current worksheet "gen3 merge." To label a worksheet you simply position your cursor at the bottom of the screen where each worksheet name appears and double-click. This allows you to type a name of your choice.

11. Now as a final step we want to resort the dataset by mom's unique id (or whatever name you assigned to the unique identifier for PCGs that we created back in step 5). To do this, position your cursor in column B. Select "data" from the commands menu, and then choose "sort." This should prompt Excel to ask if you want to sort the entire dataset by "moms unique id." Hit "OK." (If you see alphabetical column heading options instead, you need to take a moment to tell Excel that you have a header row--by checking the circle next to this question. FURTHERMORE, note that you want to sort the entire dataset by this variable--not just the column. If you only sort the column, you will be erroneously adjusting the mom ids for each of the children in your dataset.)

B. Working with the moms dataset

1. After opening your moms dataset, the first thing to do is to delete the three variables that we will not be using. As noted above, for this tutorial exercise ER33401, ER33402, and ER33403 are not needed (although they frequently are needed for analyses using the PSID). Accordingly, you can delete the columns containing these variables. (Remember, to delete a column you position your cursor in the column you wish to delete, or highlight the variable name, and then you choose "edit" from the command menu, then "delete" then "entire column.")

2. Now we are ready to prepare the dataset for generation two for use in the analysis. The first thing one wants to do is to create the unique identifier for moms by combining the information contained in the 1968 family ID number (ID 68) and the person number (PN) for each individual whose record appears in the dataset. The relevant variables here are ER30001 (ID 68) and ER30002 (PN). Remember, to create the unique identifier you multiply the 68 ID by one thousand and then add the person number to that value. Accordingly, you need to insert a new blank column at C, and in cell C2 you want to enter the following formula:

$$=A2*1000 + B2$$

Label this column "mom s unique id" and then copy the formula to the remaining rows of the column by using the "fill down" procedure, and then convert your formulas into values using the "copy/paste special" feature.

3. Next, create a unique identifier for each mom s mother (i.e., our grandmothers) by using the variables ER320009 and ER32010. You will need a new, blank column to hold this new variable, so insert a blank column at G, and label it "mom s mother s id." Enter the following formula in cell G2:

$$= E2*1000 + F2$$

Next copy the formula into the remaining cells of the column and then convert the formulas into values.

4. Finally, we want to use the information contained in the 86 Health Status variable (ER30527) to create an index that reveals whether an individual is in relatively good or poor health. If you glance at the documentation for the variable (ER30527) you will see that the ranges of responses for this categorical variable are 1 to 5. Values from 1 to 3 indicate good or better health, while instances in which the response is coded as 4 or 5 correspond to situations in which the individual's health is fair or poor. If you scroll down your column, you notice that the 86 Health Status variable also can take on a value of 8 or 9. These correspond to cases in which the health status was unknown or where the question about health status was not applicable. Both are examples of missing data and when we create our index variable it will be coded to indicate such. One final note about the 86 Health Status variable: For some records this variable has been assigned a value of 0. This denotes a case of absent data--individuals who were not asked the health question (for example, because they were not in the PSID in 1986). As noted in Tutorial #4 (found at <http://psidonline.isr.umich.edu/Guide/tutorials/Default.aspx>) while this does not constitute missing data in the traditional sense, it will be treated as such when we create our new index variable.

OK, before creating the new variable, we need a blank column to put it in. Accordingly, use the blank column at H and label it "bad health--moms." Now enter the following formula in cell H2:

```
=if(and(G2 > 0, G2 < 4), 0, if(and(G2 > 3, G2 < 6), 1, "."))
```

Remember to copy this formula into the remaining cells of the column and to convert the formulas into values.

5. As a final step, we want to eliminate the individuals for which we do not have useful health information. Accordingly, we want to sort the dataset by the new bad-health variable, and then to delete all rows with a "." entered for this variable. (Once you sort the dataset in this fashion the records with missing values should all be grouped toward the end.)

6. Now create a new copy of this worksheet. Label it "gen2 merge." (Remember that to create a copy of a worksheet, you simply select "edit" from the command menu, and then "move or copy sheet," and then you instruct Excel to "create a copy" and to place this copy at the end of the workbook (by selecting "move to end" from the list of options that appears just above the words "create a copy"). Then delete all the columns in this worksheet except the ones containing the following variables: mom's id, mom's mother's id, and bad-health--moms.

7. The final step to take is to re-sort the dataset contained in this gen2-merge worksheet by

"mom's-mothers-id," so that the data will be ready when we get to the step where we merge this generation's data with data from the other two generations.

C. Working with the grandmothers dataset

1. After opening the dataset containing the information for grandmothers (the dataset for generation 1), the first thing to do is to delete the column containing the variable ER30498.

2. Next, with this grandmothers dataset one wants to take many of the steps that we took for the moms dataset. First, we need to create a unique identifier for each individual in the dataset. Call it "g-mom's id" and use the procedure that we used to create the unique identifier in the previous cases (step B1 for example). Second, we want to create a variable that indicates whether the grandmother had bad health. For this, we again turn to the 86 Health Status variable (ER30527). We need to transform it as we did above, for moms, in order to generate a new measure that applies to grandmothers. (You can label this "g-moms bad

health.") Use the same procedures that you used to construct the bad health--moms measure to create the new variable for grandmothers, and be sure to sort the dataset afterwards so that you can eliminate the rows containing missing data--as indicated by the symbol ".".

3. Next create a copy of this worksheet, and label it "gen1 merge." (Remember that to create a copy of a worksheet, you simply select "edit" from the command menu, and then "move or copy sheet," and then you instruct Excel to "create a copy" and to place this copy at the end of the workbook--by selecting "move to end" from the list of options that appears just above the words "create a copy"). In this new worksheet u o u Want to delete all columns except the following: g-moms id, and g-moms bad health.

4. Finally, you want to re-sort the dataset contained in the gen1-merge worksheet by the g-mom s id variable.

D. Merging the data from the three different dataset to create a 3-generational dataset

1. For this process we want to open a new, blank workbook in Excel. You probably want to give it a name. (How about "all_3_generations"?)

2. Next, open the moms dataset so that you can copy information from it into your new workbook. Go to the gen2-merge worksheet, highlight it and then select copy from the command menu. Then return to your empty workbook (the "all_3_generations" file) and paste this new information from generation 2 into it. (You may want to hit "save" at this point so that this modification is permanently captured in your file.)

3. Now open your grandmoms dataset and go to the gen3-merge worksheet in that file. Highlight all of the information contained in it and select "copy" from the command menu. Then return to your new file (the "all_3_generations" workbook) and paste the information from the grandmoms dataset into it. Make sure you copy this in formation into a BLANK space towards the right of the dataset--beginning at column E for example--so that you do not overwrite any existing information. If you do this correctly, you should now be looking at a file that displays your information about the moms on the left hand side of the dataset and your information about grandmothers on the right-hand side.

4. Now, if you do not already have one, insert a blank column in between the moms information and the information that came from the grandmothers dataset. This should be at column D. Label this blank column "assign g-mom health." In cell 2 of that column enter the following formula:

```
=if(lookup(B2, $E$2:$E$1423)=B2,lookup($E$2:$F$1423),".")
```

And, copy this formula into the remaining cells by using the fill down procedure. Then convert your formulas to values with the copy/paste special feature.

What is the purpose of this formula? It instructs Excel to look at the first entry in your column containing the moms mothers id numbers (column B), and then to look up and down the column containing entries for g-mom s ids (column E, which should have 1423 rows) to search for a match. Additionally, it tells Excel that when a match is found, Excel should assign the corresponding value from column F (which contains the index of g-mom bad-health index) in the blank column. This means Excel is attaching the grandmother s health information to each mom s record. If there is no match, the formula instructs Excel to assign the missing value symbol (to indicate that there is no health mom available for the grandmother).

NOTE: If you did not remember to sort both sets of information--the information taken from your moms dataset and the information taken from your grandmothers dataset--by the respective grandmother s unique identifier variables before you copied it your matching will not proceed correctly. If you get funny results at this stage, this may be the reason. If so, start again, and remember to sort the information from the gen2-

merge and gen1-merge worksheets prior to copying it. (As specified in steps B7 and C4, both sets were supposed to have been sorted, by "mom s mother s id" and "g-mom s id" respectively.)

5. Now you can eliminate all of the columns to the right of the "assign g-mom health" variable. Since you have now attached each grandmother s health information to their offspring (to each mom that is), you no longer need the gen1-merge portion of your dataset (the portion that was previously storing all your information about grandmothers for you).

6. Finally, you want to resort this dataset--this time using the "mom s id" variable. This readies us for matching these records with records from our child dataset.

7. Now insert 7 blank columns in the leftmost portion of your worksheet, i.e., at column A.. You are going to copy the information about the children (generation three) in this spot.

8. Now open up the child dataset that we were working with earlier (in part A). Go to the gen1-merge worksheet of the child dataset and copy its information into the blank columns that you created above in step 7. (Remember, you accomplish this by highlighting all of the data in the "gen1-merge" worksheet and then selecting "copy" from the command menu. After this, you close the window and return to your "all_3_generations" workbook. Then you select "paste" from the command menu. Make sure you do not overwrite any existing data . If you have followed the instructions exactly, 7 columns should be enough to hold your child-related information without spilling over to the pre-existing information about moms. However, if you have more than 6 columns in your "gen1-merge worksheet" you will need additional blank columns to copy your information into.)

At this point, you should have a worksheet before you that contains information about the first generation in columns A through G, followed by information about the moms (with grandmother s health information merged on) in columns I through L.

9. Now insert three blank columns in between the two different sets of information, at columns I, J, and K. Label the first blank column "attach mom s bad health" and label the second "attach g-mom s bad health." In the first column enter the following formula:

```
=if(lookup(B2,$L$2:$L$1423)=B2,lookup(B2,$L$2:$N$1423),".")
```

Copy this formula to the remaining cells of your column (column I presumably) by using the fill down procedure, and then don t forget to convert your formulas to values.

What does this formula do? It instructs Excel to take each PCG id number (contained in column B) and to use it to search for a match in the range of values for "mom s id" (in column L). Then, if a match is found, Excel knows to assign the corresponding value from column N in column I. This means that Excel is taking the value for the moms-bad-health variable and placing it in column I. In doing such, it is attaching the measure of mom s health to the record for each child that we have.

10. Now, in cell 2 of column J you want to enter the following formula, which will allow you to attach the grandmother s health information to each child record:

```
=if(lookup(B2,$L$2:$L$1423)=B2(lookup(B2,$L$2:$O$1423),".")
```

This formula instructs Excel to take each PCG id number and to search for it in column H and to attach the corresponding information from the grandmom s bad health variable (from column N) when a match is found. This amounts to assigning the value for the measure of grandmother s health to each child s record.

Again, you want to copy your formula to the remaining cells of the column (using the fill down procedure), and to convert the formulas into values (using Excel's copy/paste special feature).

11. Now you are done merging your 3 generations worth of information together. All you need to do now is to do your calculations so you can fill in Table 1.

12. First, to determine what percentage of children have a low birthweight, you need to find a blank cell in your worksheet and to enter the following formula into it:

$$=sumproduct(C2:C3296,E2:E3296)/sum(C2:C3296)$$

This computes the percent of children who were born with a low birthweight. The formula does such by combining the information from the CH97PRWT column (the statistical weights in column C) with the information about whether a child had a low birthweight that is contained in column E.

13. Next, to compute the percentage of children with low birthweight for moms who have poor health and for grandmothers whose health is poor, it is easiest to create 2 separate new worksheets in which to do the calculations for each. Accordingly, you want to use the "edit" then "move/copy worksheet" feature to make a duplicate copy of the worksheet (and a triplicate copy) after the current one (at the end).

Once you have created your two new copies of the worksheet, name the worksheet containing the first copy "kid-mom connections" and the second one "kid-grandmom connections." Now, using the kid-mom connections worksheet, resort your entire dataset by the "attach mom bad health" variable. This will instruct Excel to order the information so that the cases where this variable takes on a value of 0 appear first, and the cases in which it takes on a value of 1 appear afterwards, and the cases in which there is missing data (the "." cases) come at the end. You want to delete all the rows with the "." symbol. Then, in a blank cell somewhere (preferably towards the right end of the dataset) you want to enter the following formula:

$$=sumproduct(C2:C1935,E2:E1935)/sum(C2:C1935)$$

This computes the incidence of low birthweight among children whose moms have good health.

In another empty cell enter the following formula:

$$=sumproduct(C1936:C2026,E1936:E2026)/sum(C1936:C2026)$$

This computes the proportion of kids who had a low birthweight among kids whose mothers' health is poor.

14. Now go to the worksheet that you labeled "kids grandmoms connections." Sort this worksheet by the attach-grandmom-bad-health variable. Then delete all rows with missing data for this variable (the ones with "." entered instead of numerical values). Afterwards, find an empty cell and enter the following formula into it:

$$=sumproduct(C2:C1052,E2:E1052)/sum(C2:C1052)$$

This computes the proportion of kids with low birthweight among those whose grandmothers are healthy.

In another blank cell enter the following formula:

$$=sumproduct(C1053:C1425,E1053:E1425)/sum(C1053:C1425)$$

This calculation yields the percent of kids who had a low birthweight who have grandmothers who have poor health.

15. What do you make of your calculations? It looks like the likelihood of having a low birthweight (our poor health measure for the youngest generation) is higher among kids whose grandmothers are not very healthy either. For moms, however, the difference between children's health situations by whether or not the mom

16. If you are really interested in this issue, you can repeat these calculations using the "worse health" measure instead of the low birthweight measure as your measure of each child's health.

17. How about comparing mothers and grandmothers? To compute the percent of moms with poor health among those whose own mothers also had poor health enter the following formula in a blank cell:

References

Federal Interagency Forum on Child and Family Statistics (FIFCFS) 2001. America's Children: Key National Indicators of Well-being 2001. Federal Interagency Forum on Child and Family Statistics, Washington, DC: U.S. Government Printing Office. (This publication can be found on-line at the following URL: <http://www.childstats.gov/ac2001/ac01.asp>)
