



Panel Study of Income Dynamics

[PSID Guide](#) | [Data and Documentation](#) | [Publications and Conferences](#) | [Child Development Overview](#) | [User Guide](#) | [FAQ](#) | [News](#) | [Data Quality](#) | [Funding Opportunities](#) | [Tutorials](#) | [Contact](#)

User Guide Tutorial #5-A

Intergenerational Analysis Using the Panel Study of Income Dynamics (PSID) -- With an Application to Homeownership

Ngina Chiteji and Frank Stafford

December, 2003

I. Introduction

In recent years there has been growing interest in exploring connections among family members, particularly connections that span across the generations. In both research and policy circles there has been interest in a range of issues, from father-son correlations in occupation, educational attainment, and earnings (Chadwick and Solon, 2002; and Solon, 1992, for example); to mother-daughter correlations in fertility decisions and welfare receipt (An, Haveman and Wolfe, 1993 for example); to the influence that parent's income, wealth and socio-economic status has on their children's well-being (Conley, 1999 for example); to intra-family transfers and parent-child similarity in asset ownership and wealth levels (Charles and Hurst 2002, and Chiteji and Stafford, 1999, for example). Because of its unique design, the Panel Study of Income Dynamics (PSID) is one of the premier datasets for conducting intergenerational analyses. It contains extensive economic and socio-demographic information about families and their relatives; and, because it is a longitudinal survey that has followed families and their offspring since 1968, it contains information on partial and complete life histories for each family/individual, over different stages of their lives. In Tutorial 5 we present two exercises that illustrate ways to access data resources when conducting intergenerational research using the PSID. While not exhaustive (since the PSID includes a wide variety of data elements and therefore lends itself to a multitude of applications), the sample exercises do serve to demonstrate how PSID data can be used to answer intergenerational research questions in a comprehensive yet fairly easy fashion.

The two sample exercises presented focus on the use of PSID data to examine cross-generational correlations in homeownership, and similarities in health conditions. The first exercise, presented here in Tutorial 5-A, shows how one can use the PSID to put a two-generation dataset together to examine adults with their parents in prior years of the PSID. The second, presented as a separate tutorial (5-B), illustrates a three-generational analysis using PSID data.

II. Overview of the PSID

Why is the PSID an ideal dataset for exploring connections between generations? First, the PSID is nationally representative and contains a vast amount of information about the social and economic characteristics of U.S. families and the individuals within them. The focus of the data is economic and demographic, with substantial detail on income sources and amounts, employment, family composition changes and housing, but since the 1980s there also has been an extensive set of health and well-being measures which can be used to portray the full life course and intergenerational connections. Extensive information about asset-holdings and wealth also is available in several years (in the 1984, 1989, 1994, 1999 and 2001), for example, and additional

information about the quality of individuals' and families' lives, such as information about charitable giving and volunteer time is also available in several years.

The second reason that the PSID data are ideal for conducting cross-generational analyses is that the PSID follows individuals as they reside in ever-changing families over time, along with the new families that emerge as the children in the original families age and move on to set up their own households. This means that one is able to obtain a wealth of information about different generations of a given family tree in the same dataset. In fact, because the PSID has been following families and individuals since 1968, in some applications the files can be used not just to put information about children and their parents together, but to create a file of information across three generations of a family (the children, their parents and the grandparents).

The PSID currently obtains information on about 8,000 families and has collected information on about 65,000 individuals. It therefore supports possible analyses of the U.S. population as a whole, or of special sub-groups. One can decompose the population by race or gender for example, or across different generations, and often still have sample sizes that are large enough to draw reliable conclusions about families' circumstances and their behavior.

III. Sample Exercises Demonstrating the Way the PSID can be used in Intergenerational Analyses

How might one put together a dataset with information about parents and their children, and what type of question might one answer with such data? How one puts an intergenerational dataset together can depend upon the nature of the question that one is asking. Some common challenges emerge regardless of what type of analysis one is doing, however.

The first issue any researcher will face is the need to link parents with their offspring. When examining dependent children, this task is facilitated by the PSID's inclusion of information that uniquely identifies, for a given year, both the family that a child lives in and the head and, when relevant, the wife of that family within the files containing detailed information about children. Researchers interested in conducting extensive analyses of dependent children might choose to use the PSID's 1997 [Child Development Supplement](#) (CDS), for example. As noted in [Tutorial 4](#), the CDS contains a wealth of information about children and the different settings in which their lives unfold, such as their school environment and neighborhood context; and it also provides basic family-level information for the family in which the child resides, such as total family income and the employment status of the head and wife of that family, who are typically the child's parents, for the year that the survey was done. The CDS also lists an identification number for the family in which the child resides, along with the unique identifiers for the primary caregiver of the child (typically the mother) and for each child's second caregiver (often the father), and a researcher can use these identification variables to attach detailed information about the child's parents or its family, obtained from the PSID's annual family file, to each child in the CDS. The process of linking offspring and parents is therefore fairly straightforward in this case, since one is able to obtain information about children along with some information about their caregivers (again, typically the parents) in one location, and because the CDS gives the user the information that is needed to locate other information about the child's family and parents elsewhere in the PSID (in the yearly or "core" family and individual files for example). [Tutorial 4](#) offers an example of the way that CDS data can be used, and how one can match information about children with information about a parent. [Tutorial 5-B](#) uses the CDS data to conduct an analysis that also merges in information about children's grandparents.

Constructing a dataset on parents and their adult offspring is slightly more complicated. In this instance a researcher is required to match parents and their kids together by tracing the adult children back to their families of origin. This step is facilitated by the fact that the individual

files at the PSID **Data Center** give the identification numbers for the families in which an individual resided in ever year that an individual appears in the PSID along with information about the individual's relationship to the head of this family (in the annual individual files), and by the fact that the Data Center also lists two key identification variables for the mother and father of each individual that appears in the dataset (under the "sampling variables, birth and marital variables" data group listed at the Data Center). These two variables are the 1968 family identification number and the 1968 person number--key variables in the PSID because they allow one to uniquely identify each individual in the study. One can use the two to create a (composite) identification variable unique to each individual, and this variable can then be used to identify the proper records of information for each individual in the PSID from different years.

Below, we demonstrate how one can subset the PSID data to obtain a customized dataset with information about parents and adult children from different decades. The dataset will then be used to explore the following research questions: Is there a connection between parents' homeownership behavior and that of their children's? That is to say, do individuals who grew up in families in which the parents owned their home have a greater tendency to go on to own their own homes than individuals who were not similarly "exposed" to home-ownership when they were dependent children? We then provide a discussion of the steps one might take to create a customized dataset with information about parents and adult children in the same year. And, in tutorial 5-B, a companion piece to this tutorial, an exercise is presented to show how the PSID data can be used to analyze connections across three generations of the family tree. This exercise focuses on health outcomes.

IV. Analyzing Homeownership Across the Generations

The objective of the following exercise is to obtain the data necessary to compute the rates of homeownership for young adults with different family background characteristics. As shown in Table 1, we will be interested in determining if a greater fraction of individuals who grew up in families with parents who owned their own home go on to own homes during adulthood than those whose parents were renters. As part of this we will calculate the percentage of young adults who come from a background in which their parents did not own their home but go on to own their home nonetheless. Filling in Table 1 allows the user to determine whether those whose parents exposed them to homeownership appear to be different from those who did not. (While economists typically are interested in explaining economic phenomenon in addition to describing patterns, we will leave the analysis needed to explain any gaps in homeownership rates that we find for the interested and enterprising user to do on his or her own.)

Table 1. Comparison of homeownership rates by whether one's parents owned their home or rented

	<i>All young families</i>	<i>Young families of individuals whose parents owned their home during their childhood</i>	<i>Young families of individuals whose parents did not own their home during their childhood</i>
Percent owning their home	?	?	?

What data will we need?

We will need to obtain data about the housing situation of individuals "now" and housing of their parents in "prior years." A series of housing questions has been asked each year in the PSID, 1968 to the present. For example, respondents are asked whether they own their

residence or whether they rent it, how much rent they pay, whether they live in public housing, and, for homeowners, the value of their home in terms of what it could be sold for. To view sample housing questions see the housing section (Section A) of the PSID questionnaire at http://psidonline.isr.umich.edu/Data/Documentation/cai_doc/2001_interview_year/Section_A____Housing or, see <ftp://ftp.isr.umich.edu/pub/src/psid/questionnaires/q71.pdf> for 1971. In this tutorial we will use the responses to questions A19 in 2001 and C2 from 1971 (whether the individual owns or rents). As you will observe from examining the questionnaire, however, there is much other additional information about housing for researchers to use.

How and where will we get the data?

We will restrict our analysis to young adult families. We will define these families as those in which the head or the spouse is an individual in the 30 to 38 age range as of 2001. We put an age restriction on our sample for a few reasons. First, because we want individuals who are old enough to be out on their own, i.e. to be living independently or in families in which they are either heads or wives, and individuals who also are mature enough to be in the portion of their life cycle in which they would be "settled" enough for homeownership to be a reasonable option for them. If we think about the universe of people usually considered to be adults in U.S. society, we have anyone age 18 or older. However, when one thinks carefully about the situation of some of these individuals, it is clear that not everyone above the age of 17 would be expected to be in circumstances in which homeownership would be sensible. College students, for example, are not expected to be homeowners since many of them live in dorms and since many are only temporarily residing in the locations in which they are attending college. Similarly, it typically takes college graduates a while to become settled in their careers and to enter into a life situation in which they are likely to want to own a home.

We also want to keep our sample from getting too large to be easily manipulated in Excel, since the tutorial will be using this program for data analysis. Importantly, we also need to ensure that our individuals are young enough to be offspring of the original PSID sample families, so that information about their parents will be available. This is the rationale for putting an upper bound on the age range. (Of course, for researchers who are experienced users of Excel, or for those who plan to use SAS, SPSS, or Stata to analyze their data, there is not necessarily a need to arbitrarily limit the age range and size of the dataset as we are doing in this tutorial.)

What kinds of information do we want about our young adults? For our analysis we need information about their homeownership status, and information about their parents' homeownership status during the young people's childhoods. In addition, we will include the family income of the parents in 1971 and family income of the adult children in 2001. This will allow a check of whether any observed home-ownership pattern across the generations may be simply reflecting an income correlation across the generations. A full analysis would be more involved and allow for the fact that income, particularly normal or permanent income of families, is a strong predictor of home ownership and housing consumption. We will get the information about our young adult families from the 2001 files. To obtain the information about parents in a simple and straightforward fashion, we will examine the parent's housing situation 30 years prior to the adult children's (a time at which the younger generation would have been ages 0 to 8 and therefore living as dependents in their parents homes). This means we need data for the younger generation (the adult children) from 2001 and that we will need to connect it to data about the older generation (the parents) from 1971. To obtain this information we will need to draw from two types of PSID "files" or "data groups." We will use individual-level data for 2001 and 1971 and we will want to combine this with family-level data from these same years. (See [Tutorial 1](#) for a more in-depth discussion of the data groups listed at the Data Center.) The individual data are necessary because they allow us to obtain identification information for the families that our young individuals resided in during the years 2001 and 1971. The family-level information is necessary because the housing information is actually collected at the family-level and stored as information

in the PSID "family files." This means that once we know the individual's family of residence in each year, we can find out whether that family of residence owned or rented. For 1971, the family that an individual lives in will correspond to the family that he/she grew up in. In 2001, the family will be a household in which the individual of interest is either a head or wife, and we will be able to exclude the thirty-somethings who have not left home and set up independent households.

How do we match parents with their adult offspring?

In order to use the Data Center to obtain information about the two generations simultaneously, we will draw a few additional variables beyond our analysis variables. As discussed earlier, when the PSID collects data about individuals, it records the individual's "status" in the family in which he/she resides. The PSID does this by noting each individual's relationship to the head of the family in which they are living. Accordingly, a spouse of the head is listed as the "wife" and the relationship to head variable is coded to reflect this status. (By convention, when a family is headed by two adults, the adult male is usually treated as the head. See [Tutorial 1](#) for additional discussion of this matter.) Similarly, children receive codes indicating their relationship to the head. In early years this was "child of head," while there are separate codes to indicate a son or daughter in later years. An extensive list of the codes for this variable can be viewed at the Data Center when one selects the relation to head variable. Or, one can go to <http://psidonline.isr.umich.edu/data/Documentation> to view a complete PSID codebook. We want to select the relationship to head variable for both 2001 and 1971 when we choose variables via the Data Center. We will use these data to make sure that our young adults are heads or wives in 2001 (meaning they have set up independent households of their own), and to make sure that the family identification numbers that we obtain for 1971 represent families in which the individuals were dependent children during 1971.

V. Getting the Data for the Homeownership Exercise

To construct the dataset that you need to do the analysis of homeownership, go to the Data Center at <http://simba.isr.umich.edu> and click on "variable selection." Next, when presented with the list of data groups under "PSID Data," you want to check the box for PSID individual data, the box for PSID family data, and the box for Income Plus data.

Screenshot #1 Select Data Group(s)

Select Data Group(s)	
PSID Data	
<input type="checkbox"/>	Sampling variables, birth & marital variables (subset), gender
<input checked="" type="checkbox"/>	PSID Individual
<input checked="" type="checkbox"/>	PSID Family
<input checked="" 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

This will take you to a screen with a range of years for individual data and a range of years for family data, each spanning back to 1968 when the PSID first began.



Screenshot #2 Select data categories and years

Select data categories and years									
PSID Individual Data by Years									
<input checked="" 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 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 checked="" type="checkbox"/> 1971	<input type="checkbox"/> 1970
<input type="checkbox"/> 1969	<input type="checkbox"/> 1968								
PSID Core Family Data									
<input checked="" 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 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 checked="" type="checkbox"/> 1971	<input type="checkbox"/> 1970
<input type="checkbox"/> 1969	<input type="checkbox"/> 1968								
Income Plus									
<input checked="" type="checkbox"/> 2001	<input type="checkbox"/> 1999	<input type="checkbox"/> 1997	<input type="checkbox"/> 1996	<input type="checkbox"/> 1995	<input type="checkbox"/> 1994				

On this screen you want to select 2001 and 1971 for both the individual-level data and the family-level data, and for 2001 you will need "income plus" file data, since from 1994 the income data as a subset of family-level data constitute their own separate data group. (Just to make sure we're all on the same page, that's five boxes to check overall.) Hit "continue" and you will be taken to a screen where you can select your variables. From the 2001 individual data you want to highlight

the following 4 variables: ER33601 (2001 interview number—the family id number for the year 2001), ER33602 (the sequence number for 2001), ER33603 (the relationship to head variable), and ER33604 (age in 2001). To select non-contiguous variables you simply need to hold the control key as you use the mouse to click on and highlight the variables that you want to select. From the 1971 individual data variable list you want to select 2 variables: ER30067 (the 1971 interview number—the family id number in the year 1971) and ER30069 (the relationship to head variable for 1971). From the 2001 core family data, you also want to select 3 variables: ER17002 (the 2001 family interview [ID] number), ER17043 (the variable indicating whether the family owned its home or rented), and ER20394 (the family weight). From the 2001 income plus file you want: FAMINC01 (total family income). From the 1971 family file you need to select 2 variables. They are V1967 (the variable indicating whether the family owned its home or rented in 1971) and V2226 (total family income in 1971). Once you have selected all these variables and checked 'continue', your screen should depict a datacart that looks as follows:

Screenshot #3 PSID/CDS Data Cart Contents

PSID/CDS Data Cart Contents						
Total PSID Variables Selected: 13						
Year	Data File	Name	Label	Data Group	<input type="checkbox"/>	
1971	PSID Individual Data by Years	ER30067	1971 INTERVIEW NUMBER	PSID Individual	<input type="checkbox"/>	[View]
1971	PSID Individual Data by Years	ER30068	SEQUENCE NUMBER 71	PSID Individual	<input type="checkbox"/>	[View]
1971	PSID Individual Data by Years	ER30069	RELATIONSHIP TO HEAD 71	PSID Individual	<input type="checkbox"/>	[View]
1971	PSID Core Family Data	V1967	OWN OR RENT? 39:11	PSID Family	<input type="checkbox"/>	[View]
1971	PSID Core Family Data	V2226	TOT FU MON INC 19426	PSID Family	<input type="checkbox"/>	[View]
2001	PSID Individual Data by Years	ER33601	2001 INTERVIEW NUMBER	PSID Individual	<input type="checkbox"/>	[View]
2001	PSID Individual Data by Years	ER33602	SEQUENCE NUMBER 01	PSID Individual	<input type="checkbox"/>	[View]
2001	PSID Individual Data by Years	ER33603	RELATION TO HEAD 01	PSID Individual	<input type="checkbox"/>	[View]
2001	PSID Individual Data by Years	ER33604	AGE OF INDIVIDUAL 01	PSID Individual	<input type="checkbox"/>	[View]
2001	PSID Core Family Data	ER17002	2001 FAMILY INTERVIEW (ID) NUMBER	PSID Family	<input type="checkbox"/>	[View]
2001	PSID Core Family Data	ER17043	A19 OWN/RENT OR WHAT	PSID Family	<input type="checkbox"/>	[View]
2001	PSID Core Family Data	ER20394	CORE/IMMIGRANT FAMILY WEIGHT NUMBER 1	PSID Family	<input type="checkbox"/>	[View]
2001	Income Plus	FAMINC01	TOTAL FAMILY INCOME 2000	Income Plus	<input type="checkbox"/>	[View]
Automatic Individual Identification Variables						
Name		Label				
ER30001 (ID68)		1968 INTERVIEW NUMBER		[View]		
ER30002 (PN)		PERSON NUMBER 68		[View]		

Note that you can view documentation for each variable by clicking on the “view” statement below the book icon in the variable list. This is helpful if you are unsure how a variable is coded or if you want to see the actual question text that was asked (when applicable or explanation text for generated variables).

Finally, note that the Data Center automatically adds a few variables to your cart. Because the 1968 family interview number (ER30001) and the 1968 person number (ER30002) are key identifier variables in the PSID, the Data Center will give you these variables automatically. After you have selected all your variables and viewed your variable list to confirm that you have all the variables that you need, you want to click “Get Data and/or Codebook.” This will take you to a screen where the Data Center allows you to specify the form that you want your dataset to take.

What to do if you do not know the exact names of the variables that you want: As a quick aside you should note that if you did not know the names of the variables that you wanted, you could still construct your variable list at the Data Center. The Data Center has a feature that

allows you to type in a keyword(s) in order to identify variables. For example, suppose we did not know that ER17043 was the variable name for the variable indicating whether a family owned or rented its home in 2001. If this were the case we could simply select “search and browse” instead of clicking on “variable list” at the Data Center’s opening page. This would take you to a screen where you can search for variables using keywords or browse the variable list by category. If you choose “variable search” you are taken to a screen where you can type in keywords to conduct a variable search. If you type the words own and rent in the search box, and then select “and” from the “search type” box you can instruct the Data Center to search through the list of variable names in the family files to identify ones that contain the words “own” and “rent.” If you choose the word "own" you would have many variables since people are asked about owning other things and because the word "own" also appears as an adjective in several places. (Remember to also click beside the box for PSID family while at this screen.) After you submit this search, the Data Center would give you a list of variables with these two words in the name. The first one listed comes from the 2001 family file (as indicated in the year column to the left of the variable names). This is our familiar ER17043. (You could confirm that this corresponds to question A19, the own or rent question, by clicking on the view command to the right of the variable name and label. This will open a documentation box that tells you a bit about the variable.) Upon determining that this is the variable that you wanted, you could simply add it to your data cart by clicking the box under the green plus sign; this instructs the Data Center to add this variable to your list of desired variables.

Choosing appropriate output options

Now that you have reached the screen where the Data Center asks for the format you want your data delivered in, you should select “Microsoft Excel Spreadsheet” for the purposes of this tutorial, and indicate whether you would like the Data Center to create a codebook to go with your customized dataset. (The default option for the codebook is no.) You also will want to instruct the Data Center to subset your data at this stage. This means we need to type the following command in the subsetting criteria box:

ER30069 = 3 and ((ER33603 = 10 and ER33602 = 1) or ER33603 = 20 or ER33603 = 22) and (ER33604 > 29 and ER33604 < 39)

What is the point of this subsetting command? It instructs the Data Center to restrict your dataset to cases in which an individual was a dependent child in a family in 1971 and a head or wife of a 2001 family. This means you are choosing individuals who had formed their own households by 2001, but who were dependents back in 1971. The last command in the string of subsetting criteria limits the age of the heads and wives that we are drawing to the 30 to 38 age bracket (since we said earlier that this is the age range that we wanted to work with). We also add an additional restriction when working with the 2001 relationship to head criteria: In instances in which the individual is the head of the 2001 family we want to make sure they are a current head, which is why we add the restriction that ER33602 be equal to 1 for the head.

Two final matters. First, if you want the Data Center to e-mail you to tell you when your dataset is ready you should enter your e-mail address in the box provided. Second, doublecheck to make sure that you have “All individuals” selected under the Data File Options section of the page. (This is the default, so the circle beside “all individuals” should be filled in automatically.) NOW hit submit and let the Data Center do its work. When your dataset has been created, you will receive an e-mail indicating that it is ready, and allowing you to download it. If you did not provide an e-mail address you should receive a job completion notice on the screen before you.

VI. Using Excel on Your Output Subset

You should have a dataset with 13 variables (13 columns,) and 1085 observations. When working with this dataset there are a few Excel commands that will come in handy. Table 2 lists these commands and offers a brief explanation of each one. You may want to use the first one so that you can scroll down through all the records of your dataset without losing sight of the variable names. Looking through the dataset is a good way to verify that you subsetted correctly when you were at the Data Center. For example, as you look down the column for ER33603 all of your cells should contain a "3." This is because we asked the Data Center to restrict our dataset to cases in which we have an adult who was a dependent child (or step-child) back in 1971. You may also want to scroll through the dataset to make sure that the variable for the 2001 age (ER33604) only takes on values ranging from 30 to 38.

Table 2. Some useful Excel commands

Command	Explanation and instructions for use
"Freeze pane"	This command allows you to retain the image of your first row of variable names as you scroll down through the dataset. To freeze the list of variable names in this fashion, position your cursor in the second cell of the first column, and then click on "windows" in the command bar at the top of the Excel file, and then select "freeze pane."
"Fill down "	This allows you to quickly copy a formula (or entry) from one cell of a column to the remaining cells of that column. To implement the command, drag the cursor to the bottom right corner of the cell whose entry you wish to copy. (A plus sign should appear at this point.) Then, simply left click on your mouse and drag the cursor down to the last row of the column. When you reach the end of the column and release your grip the entire column will be filled in.
"Copy" and "Paste Special"	Using these two commands in succession allows you to convert any formulas that you have entered in a cell into the actual values that are calculated by that formula. To do this you first highlight the cells containing the formula entries, then select "edit" from the command menu, and "copy" from the resulting list. Next, return to the "edit" entry on the command menu and select "Paste Special". Then click on the box next to "values" under the paste options.

Steps to take to perform your desired calculations

1. First, before you begin to make any alterations to your dataset you probably want to copy your entire dataset to a second worksheet, so that you can use this one for calculations. (This way if you mess up you will still have your original dataset handy.) To do this click on "edit" from the command menu, and then choose "move or copy sheet." In the box that appears tab down to "move to end" to highlight that option, and then click on the "create a copy" option. This produces a new worksheet with a second copy of your entire dataset.
2. Delete all columns except the ones containing the following variables: ER30067 (the 1971 interview number), V1967 (the home own/rent variable for 1971), V2226 (the 1971 family income variable), ER33601 (the 2001 interview number for the family in which our individual resided in the year 2001), ER17043 (the home own/rent variable for 2001), ER20394 (the family weight variable), and FAMINC01 (the 2001 family income variable). This leaves you with 7 columns.

3. Insert a new, blank column beside your ER17043 column. How do you do this? Position your cursor at the column immediately to the right of the ER17043 column (in column F that is). Then select "insert" from the Excel command menu, and choose "column" from the list of options that appears. Doing this will create a new column right next to the ER17043 column. Why are we doing this? As you will recall, ER17043 is the variable indicating whether the younger generation of families that we are interested in rented or owned their home (in the year in which we are observing them--2001). In the PSID this variable can take on a range of values. A code of "1" indicates that the family owned its home, while the PSID assigns a code of "5" for renters, and the value "8" for any family that says that it neither owns nor rents. We want to transform this data to make it easier to work with, because it will be easier to do the calculations if we use the number 0 to represent instances in which the family is a renter. (We will still use the number 1 to indicate instances in which the family owns its home.) Accordingly, we want a new column beside the original and we want to label it "young families own/rent 0/1" by typing this phrase in the first row of the new column. What do you want to do with this new column? In cell F2, enter the following formula:

```
=if(E2=1,E2,if(E2=5,0,"."))
```

This instructs Excel to enter a one in the cell if ER17043 takes on a value of 1, or to enter a zero in column F if ER17043 takes on a value of 5, and to enter a dot (a ".") if ER17043 takes on a value of 8. Next you want to copy this formula into every row of column F. To do this you can use the "fill down" command discussed earlier. Then, you may want to use the "copy" and "paste special" commands to instruct Excel to save the actual values in the column instead of the formula. (Remember, Table 2 tells you how to use Excel's copy/paste special feature.) Finally, as a last step, you want to sort your entire dataset by our new variable and to delete any observations for which the "young families own/rent 0/1" variable takes on a non-numerical value (i.e., the instances in which we have a "." in a row of this column). To sort the dataset by this variable, position your cursor in the top row of this column, and then select "data" from the command menu. Choose "sort" from the options list that appears, and you will see that Excel asks you what variable you want to sort the dataset by. (It should have the young families variable listed--if not, click on the triangular icon next to whatever variable name is listed and you will see a list of the names of all the variables in your dataset and you can then simply scroll down until you find the variable that you want.) Why are you sorting the dataset by this variable? Doing such reorganizes your dataset so that all the observations with a 0 for the newly created variable "young families own/rent 0/1" are grouped together, followed by the observations where that variable takes on a value of 1, followed by the records for which the variable has a "." reported for it. You can then eliminate these latter cases by simply highlighting the rows containing this symbol and then choosing "edit" from the command menu, then clicking "delete" from the options presented, and "entire row" when prompted further.

4. Now we want to make sure that we do not have any observations in which parents neither own nor rent their homes. As you will recall, we are using 1971 to obtain our background information about the adults leading the young families that we are interested in (to gauge their exposure to homeownership as a child, that is). The responses for the variable V1967 ("own/rent" for 1971) therefore correspond to information about the parents of the young families in our dataset. Again we want to make sure that we eliminate records in which the response was "neither own nor rent" (i.e., instances in which the variable takes on a value of 8). We therefore want to sort the dataset by this variable, so that all observations with the value 8 are grouped at the end of the dataset. To sort the dataset in this fashion, position your cursor in the first row of the column containing the V1967 variable (this should be column B), and then select "data" from the command menu, and choose "sort" from the list of options that appears. As was the case earlier, a box will appear asking what variable you wish to sort by (and V1967 should be highlighted as the default option). After the dataset has been sorted, you want to scroll down to the end to locate any records where this variable takes on a value of 8. You can then delete these observations by

highlighting the associated rows, and then selecting "edit" from the command menu, and then "delete" and then "rows" and then "entire row" as you did before in step 3.

5. Next we want to create a variable that isolates instances in which the parents of the adults leading our young families (the older generation) owned their home. To do this we will create a new column that stores such information. Position your cursor in column G and select "insert" from the Excel command menu, then choose "column." This will create a new, blank column at G, and you want to label it "parents own 0/1." In cell G2, insert the following formula: $=if(b2=1,b2,0)$ Then copy this formula to the remaining cells of the column. (Remember, you can use the "fill down" command to do this; AND you can use the "copy/paste special" feature to convert the formulas to values after you have copied the formula to all the rows of column G.) What is the purpose of this step? With this step you are re-coding your parental ownership variable (the 1971 homeownership variable) so that it takes on a value of 1 for any parents who own their home and a value of zero otherwise. Having the responses coded in this fashion will make it easier to compute the homeownership rate for parents with the data.

6. Now we want to create a variable that isolates instances in which the parents of the adults leading our young families were renters back when our young adults of interest were growing up. To do this we create another new column beside the "parents own" column. To do this, position your cursor in column H. Select "insert" from the command menu, and then choose "column" from the options list. This will create a new, blank column at column H. You probably want to label this column "parents rent" to serve as a reminder that this will be a column that tells us whether the parents were renters. In cell H2 enter the following formula: $=if(b2=5,1,0)$ Then copy this formula to the remaining cells of the column, and use the copy/paste special feature to transform your formulas into values when you are done. What are you doing with this step? You are creating a variable that directly indicates whether the parents were renters (with a 1 for yes and a zero for no).

7. With steps 5 and 6 we have separated out the members of the younger generation whose parents were owners from those whose parents were renters. We have one more step to take before we can do any calculations, however. Because we need to use weights in order for the PSID to be nationally representative, we now want to create a column that combines the weights information with the parental-ownership information, and another column that combines the weights information with the parental-renting information. To do this, we will create two new columns to the right of the "parents own" column (i.e., to the right column G). To do this we need to position the cursor at column H. Next choose "insert" from the command menu, and select "column" from the options list. Then repeat your actions. You should now have two new, blank columns at H and I. Label the first one "weighted p-ownership" (column H) and label the second one (column I) "weighted p-renting." Now do the following two things: (a) First, enter the following formula in cell H2: $=(g2*k2)$ and then copy this formula into the remaining rows of column H, and then use the "copy/paste special" feature to convert the formulas into values. (b) Second, enter the following formula in cell I2: $=(j2*k2)$ and then copy it to the remaining cells of column I, and then convert the formula entries into values using the "copy/paste special" commands.

8. Pheew...now we're finally ready to do some number crunching. Find some blank space in the far right section of your worksheet (somewhere AFTER your columns containing data). To calculate the proportion of young families that own their homes enter the following formula in an empty cell: $=sumproduct(f2:f967,k2:k967)/sum(k2:k967)$

To calculate the homeownership rate among young families whose parents also owned their own homes enter the following formula in a blank cell: $=sumproduct(f2:f967,h2:h967)/sum(h2:h967)$

To calculate the percent of young families coming from a background of renting (as a child) who have since gone on to own their own home during adulthood enter the following formula in a blank cell:

= sumproduct(f2:f967,i2:i967)/sum(i2:i967)

You can now fill in Table 1!!! The first value that you calculated gives you the rate of homeownership among young families. The last two numbers that you computed break this population up into two different categories based on their family background (whether they grew up in households where their parents owned their home or rented). This allows you to determine whether homeownership rates differ depending upon whether one's parents owned their home or not. As you can see, the homeownership rate is higher for young families that come from a background of parental ownership than it is for young families whose parents were renters. The gap is pretty large.

9. In the previous step we saw that the homeownership rate is higher among those whose parents owned their home than it is for those whose parents did not. While this suggests the presence of an association between homeownership and exposure to homeownership as a child, it is reasonable to ask whether an alternative reason for the association exists. For example, someone might ask whether the apparent association between parental homeownership and homeownership among adult children simply reflects a correlation between parental and child incomes. The argument would be that if high income individuals are more likely to own homes than low income individuals, and if incomes are correlated across generations, then our finding that homeownership is more likely among individuals whose parents were homeowners might simply be an artifact of the cross-generational income correlation. In practice, research within economics does suggest that there is intergenerational correlation in incomes (Solon, 1992 for example). Accordingly, one might want to supplement the bivariate analysis in step 8 with an investigation that takes parental and child incomes into account. To do this we will use Excel's regression analysis tool. It can be found by clicking on "tools" from the command menu, and then selecting "data analysis," and then highlighting "regression" and clicking "OK". This will prompt Excel to display a box in which you need to enter some additional information. Excel will ask you for an "input-Y" range and an "input-X" range. The column of unweighted 2001 homeownership data will be your "input-Y" range. Before starting however, you may want to copy your dataset into a new worksheet so that you can use it specifically for the regression analysis. (Remember, to create a new copy of a worksheet you choose "edit" from the command menu, and then "move or copy sheet." Then, you highlight "move to end" and click on the box beside the phrase "create a copy.") Now you are almost ready to run your regression. Before specifying your "input-X" range, however, you want to make sure that you have the variables that will be serving as independent variables (the regressors) in adjoining columns. If the (1971) parents own variable, and the 2001 family income variable (FAMINC01), and 1971 family income variable (V2226) are not presently in adjacent columns, reorganize your dataset so that they are. (You can do this easily by inserting blank columns beside one of the variables and then copying the other two variables into them.) Finally, be sure to click in the box under the word "label" since your Excel file contains a row of labels.

10. **Now run your regression and view your regression results!** When looking at the table of regression output, you should find that the coefficient for the parental ownership variable is positive and that its t-statistic indicates that parental ownership has a statistically significant effect even when we take into account the influence of parents' income and adult kids' income on adult kids' homeownership. (For those who need a quick statistics refresher, as a rule of thumb, a t-statistic larger than two denotes statistical significance at a generally acceptable level.)

(Some notes for the super-curious: It is not possible to run weighted regressions in existing versions of Excel, and Excel only offers the possibility of running ordinary least squares

regressions. There has been some debate in the social sciences literature about whether it is necessary to include weights in regressions [DuMouchel and Duncan, 1983, for example]. While it is conventional to use weights in regressions for many analyses of PSID data, particularly analyses that use income measures as the dependent variable, for the purposes of this tutorial the unweighted regression suffice to show how one might account for the influences of other phenomenon, such as parent-child correlations in income, in order to determine whether the association between parent and child homeownership can really be interpreted as suggesting that growing up in a family that owns its home has positive effects on homeownership because of the exposure that it provides. If you want to analyze homeownership in even more detail, you may want to add additional regressors, and to use a different estimation technique such as probit or logistic regressions, because there are shortcomings associated with the use of using ordinary least squares estimation when the dependent variable is a discrete variable such as our homeownership variable, which only takes on values of zero or one. You probably already know this from your statistics or econometrics course. However, if you want a refresher, see Greene, 2002 for a discussion of the issue.)

VII. Other Possible Types of Intergenerational Analyses--Food for Thought

Two interesting questions that researchers often encounter in doing intergenerational analysis is (1) whether the researcher will require repeated observations on each generation and (2) whether the analysis will involve time measures that are asynchronous. While we do not illustrate all of these issues with this simple tutorial, one of the advantages of using PSID data to do intergenerational research is that each can be dealt with in a straightforward fashion. For example, suppose one wanted to examine the correlation in income between our young families above and their parents in 1997, if one wanted to know whether families that are well-to-do tend to come from extended families that are similarly situated? This is an example of analysis using synchronous time measures, rather than observing each generation at a different point in time as we did in the exercise presented above. In this hypothetical exercise, the best way to match adult children and their parents would be through use of the 1968 family identification number and 1968 person number that the PSID provides for each individual in the survey. These two variables can be used to create a unique identifier variable for each individual in the PSID. (This is done by multiplying the 1968 family ID number by 1000 and then adding on the value of the person number.) To put parents and adult children together in this situation, you would simply instruct the Data Center to select individuals who were heads or wives in 1997, restricting them to lie between the ages 30 and 38 as we did above, and to output their 1997 family interview number and their 1997 family income. You also would instruct the Data Center to provide the 1968 family ID and person number of the head's (and wife's) mother, AND the 1968 family ID and person number of the head's (and wife's) father. This parental identification information is available for each individual in the PSID from the "Sampling Variables" data group, and it would allow you to connect the individuals who are heads or wives of 1997 families to their parents, and to merge the information of the parents' families (such as the 1997 income of the parent families) onto the records of the young families (those headed by individuals age 30-38). The merge is accomplished by noting that once you obtain a dataset with information for young families that includes the unique identifiers for the parents of these young families, you can then obtain a second dataset with information about PSID families and use the unique identifiers for the heads and wives of the second set of families to see if they coincide with the unique parental IDs for the first set of families. In instances in which they match you will have identified the parents of the first set of families. (Actually, the companion tutorial, 5-B, shows how one can use the 1968 family identification number and the 1968 person number to create a unique identifier for each individual in the PSID, and how one can merge information from one generation onto the records of another generation using the unique ID. Its analysis does not involve synchronous time however.)

A seasoned researcher will realize that one possible limitation of such the above approach

to analyzing correlations in family income across the generation is that a family's can fluctuate from year to year, so if one truly is interested in exploring the correlation of parents' economic status and their offspring's, one might rather have a measure of income that is indicative of each individual's "normal" income or a longer-run measure of income. (For those who have taken macroeconomics, this is akin to the difference between transitory income and permanent income that Milton Friedman stressed.) In such a case, one would want to measure family income over a range of years, rather than taking a one-year snapshot. This can be done with PSID data. Because the study is longitudinal, the information about each of the families in the study covers a wide range of years. This allows one to put together a balanced panel that covers more than one year. (For more information about how to assemble a balanced panel using PSID data see Tutorial #3.)

References

An, Chong-Bum; Robert Haveman and Barbara Wolfe (1993). "Teen Out-of-Wedlock Births and Welfare Receipt: The Role of Childhood Events and Economic Circumstances," *Review of Economics and Statistics*, May, Volume LXXV, No. 2

Chadwick, Laura and Gary Solon (2002). "Intergenerational Income Mobility among Daughters," *American Economic Review*, Volume 92:1, March, pp. 335-344.

Charles, Kerwin and Eric Hurst (2002). "The Correlation of Wealth Across Generations," forthcoming in *Journal of Political Economy*.

Chiteji, Ngina and Frank Stafford (1999). "Portfolio Choices....." our AER Papers & Proceedings Paper.

Conley, Dalton (1999). *Being Black, Living in the Red: Race, Wealth and Social Policy in America*, University of California Press.

Greene, William H. (2002). *Econometric Analysis*. Prentice Hall.

DuMouchel, William and Greg Duncan (1983). "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples," *Journal of the American Statistical Society*, 78 (383):535-543.

Solon, Gary (1992). "Intergenerational Income Mobility in the United States," *American Economic Review*, Volume 82:3, June, pp. 393-408.