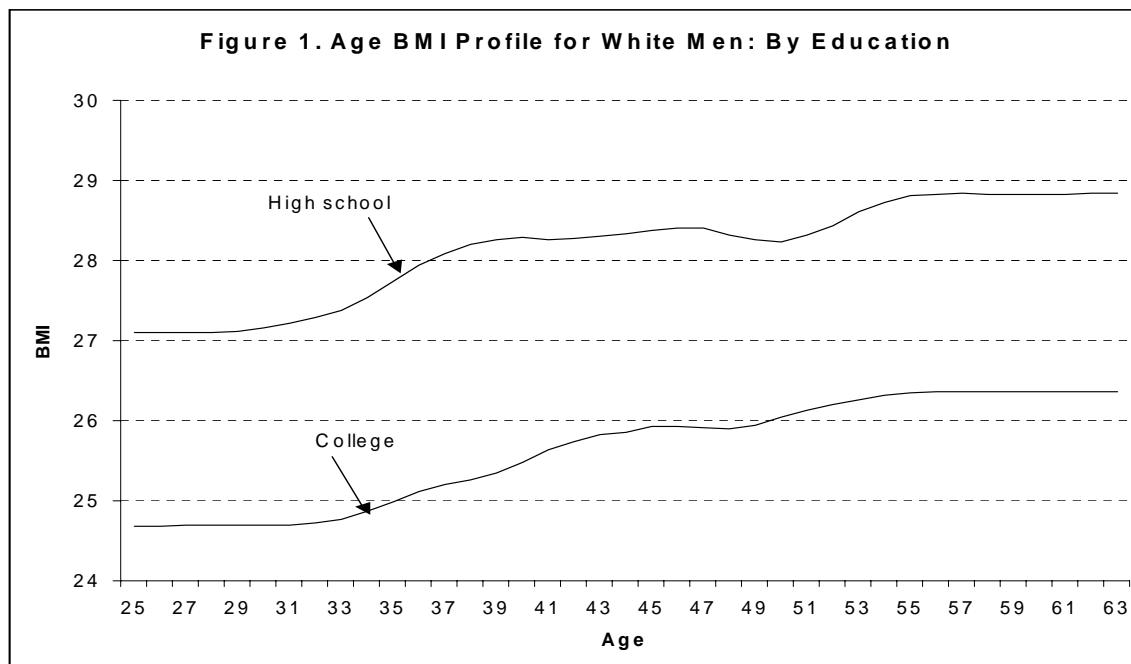


In this tutorial we demonstrate how one can use the earnings data contained in the PSID--which



represent an example of data that the PSID has collected every survey year--to estimate the extent to which earnings are correlated across generations. More generally, this case also serves to illustrate how researchers conduct intergenerational analyses, and it enables the tutorial to highlight some of the challenges that researchers encounter when attempting to study labor market, wealth or health outcomes in a multi-generational setting.

The tutorial is lengthy. So, here is a guide to help you understand how it is structured (in case you feel there are sections you want to skip or de-emphasize). And you can always skip past or back to a section or subsection by using the selection bar on the left of your screen. Or you can download the document and companion files in [PDF](#) or [Microsoft Word](#) format if that is more convenient. [Section II](#) provides an overview of the PSID. If you already have done some of the [other tutorials](#), you probably know a lot about the PSID already and you can breeze through Section II. [Section III](#) discusses the PSID's labor income data, and explains how one can navigate the data files to learn what data the survey has collected over the years and archived in the [Data Center](#). It concludes (in [Part III-D](#)) with a general discussion of the different research strategies and methodological issues that a researcher might contemplate when setting up an intergenerational study, and a discussion of the explicit strategy that the tutorial will adopt. [Section IV](#) --by far the longest section of the tutorial--presents two different examples of ways intergenerational analysis can be conducted. For each example the tutorial shows the steps you need to take in order to get the data that you need to test the hypothesis identified above. Upon completion of each example you will have an actual numerical estimate of the degree to which fathers' and sons' earnings are associated. The first example ([Example 1--presented in IV-B](#)) illustrates an approach to conducting intergenerational analysis that relies on a measure of earnings taken from a fixed calendar year for each generation. The second example ([Example 2--presented in IV-C](#)) shows how to construct measures of earnings that incorporate information from many years of a worker's life, instead. Before getting to these examples however, Section IV starts with a discussion of a new PSID tool that allows one to match fathers and sons to each other (Part

A of Section IV). Note that you cannot do either example without going through the steps needed to match fathers and sons, so make sure you definitely read Section IV-A.

II. Background about the PSID

Why use the PSID for intergenerational analyses? The PSID is a nationally representative sample of U.S. individuals (men, women, and children) and the family units in which they reside. (See the overview for more about the design of the PSID.) It began in 1968 and since that time it has continued to follow PSID families (and the individuals residing in them) over time--annually from 1968-1997 and biennially since 1997. The study also follows the offspring of the original sample families as they leave their natal units to form their own households. Because of this rather unique design, the dataset contains information about individuals covering a number of years, and during a variety of life stages; and it also contains information about multiple generations of a family. This allows researchers to use the PSID to compare the outcomes of parents and their adult children. (The dataset also can support analysis across three generations, as many of the "children" of the original sample families are now old enough to have had children of their own.)

The PSID's content is broad, including sociological and psychological measures in addition to a rich array of economic and demographic information about family income, family composition changes, receipt of public assistance, and employment. The PSID collects both family-level information, such as whether the family lives in a rental unit or owns its own home, and information about individuals. While some information is collected about all individuals in the family unit, the greatest level of detail is ascertained for the primary adults heading the family unit--heads and wives (or cohabitators for couples who are not legally married). Overall, the PSID has collected information about 7,000 families and almost 65,000 individuals, with information spanning as much as 36 years of their lives for the earliest members of the survey. Because maintaining the comparability of the data throughout time is crucial for a panel study, the general design and content of certain variables have remained largely unchanged over the years. This permits the PSID to offer time series for core content--income, employment, family composition changes, and demographic events. More information about the survey's core topics and about supplemental questions can be found in Tables 1, 2 and 3 at in the overview discussion.

III. Intergenerational Data: the PSID Data Center

To analyze the research hypothesis one needs information about earnings from labor-related activities and one needs information from two generations of any given family. The PSID has collected data about individuals' labor income since 1968. This information about labor income is available for both heads and wives, except for a few odd cases--1968, 1977 and 1978--where no labor income data for wives was collected. In what follows we discuss the variables that are relevant for heads only because the tutorial focuses on fathers and sons (to keep the task manageable), and as described in on-line Tutorial 1, the information about men will be captured in the head-related variables since the tradition in the PSID has been to code men as household heads in families in which a husband (or a male cohabitor) is present. The discussion of the earnings data (and how to find the relevant variables) is followed by a discussion of strategies that a researcher can take to test the tutorial's hypothesis (Part D below).

A. Labor income data available

As shown in Figure 2, the specific names for the labor income variable change from year to year. (Variable names can change across years in the PSID even when the data collected are the same.) Additionally, starting in 1994, the labor income data are presented in a disaggregated form--with

earnings that come in the form of wages (or salary, overtime, commissions, et cetera) separated from labor income that is derived from farm activity, and that derived from unincorporated business enterprises.

Figure 2. Labor income variables for different PSID years

year 2003	year 1993	year 1986	year 1979	year 1972
ER24116, ER24105, ER24109	V23323	V13624	V6767	V2498
year 2001	year 1992	year 1985	year 1978	year 1971
HDEARN01, FRMINC01, HDBUSY01	V21484	V12372	V6174	V1897
year 1999	year 1991	year 1984	year 1977	year 1970
HDEARN99, FRMINC99, HDBUSY99	V20178	V11023	V5627	V1196
year 1997	year 1990	year 1983	year 1976	year 1969
HDEARN97, FRMINC97, HDBUSY97	V18878	V9376	V5031	V514
year 1996	year 1989	year 1982	year 1975	year 1968
HDEARN96, FRMINC96, HDBUSY96	V17534	V8690	V3863	V74
year 1995	year 1988	year 1981	year 1974	
HDEARN95, FRMINC95, HDBUSY95	V16145	V8066	V3463	
year 1994	year 1987	year 1980	year 1973	
HDEARN94, FRMINC94, HDBUSY94	V14671	V7413	V3051	

Notes: Prior to 1994, there is just one labor income variable for the head. Beginning in 1994, head's labor income is reported in disaggregated form (in 3 separate variables--one for earnings from all activities except farming and unincorporated businesses, another for farm income, and a third for labor income generated in unincorporated business).

How does a researcher determine what variables needed to study a single phenomenon over several years? That is to say, how *did* we identify the variables listed in Figure 1? The PSID's "cross-year index" is one useful device for determining what type of coverage the PSID has for different variables throughout the history of the survey. This index allows one to look up a topic and to determine (a) the years that data about the topic was collected and (b) the exact variable names in each survey year. The "Variable Selection By Search" feature at the PSID Data Center offers a second way of locating variables over time.

B. Variable names using the cross-year index

"The new interactive cross-year index can be used. The index is essentially a grid listing specific topics covered in the PSID and the different years that the PSID has collected data on the topic. The cross-year index therefore allows one to identify years in which each particular topic was covered in the survey, and the specific types of information collected about the topic in any given year, along with the relevant variable names for each year. For example, under the "Family Data Index," if one scrolls down to "INCOME", one sees a subsection titled "Labor," and boxes to indicate years in which this information is available for heads and wives.

C. Using the search and browse features

A second way one can determine whether the PSID contains multi-year data for any given topic is to use the "Variable Selection By Search" feature at the Data Center. As shown below in Figure 3, the third bullet item on the Data Center's main page contains a clickable link titled "By Search" that allows for variable selection by searching. Clicking on this will take the user to a screen where a keyword can be entered to search the PSID website for information related to the keyword. Accordingly, one could identify all of the data that the PSID has collected about labor income over the years by searching for terms such

as "labor" or "labor income." Note that there are four different fields that one can use for a keyword search when using the PSID's search and browse features. The in "*question or explanation text*" option typically provides the broadest search results. If this is selected, the Data Center will look to determine whether the keywords are mentioned in the question that underlies a specific variable, or in the text explaining the variable name and content. To understand the importance of these fields, consider the following example: If one simply searches for "labor income" in the "variable label" field, the search only will yield variables for which the full phrase "labor income" is part of the variable label (which occurs in 1968 and in the years following 1993). However, in many years (such as 1969-1992), the PSID uses an abbreviated form of the phrase for the variable label, and it is only in the question text or the explanation of the variable that the full phrase "labor income" can be found. When using the search feature to try to identify variable names across a long span of time, it is helpful to search in a number of different fields or to truncate the phrase or name of the keyword that one is interested in (substituting "labor inc" for "labor income" for example, in order to pick up any years in which an abbreviated form of the phrase was used for the variable label). This is one reason that it is easiest to do an exhaustive search using the cross-year index. The search and browse feature is a handy second option, however.

Figure 3. Tools for searching and browsing in the PSID

D. Getting started

Conducting a multi-generational study can be a complicated exercise. A researcher has to make decisions about how to connect the different generations of the family tree, and about how to measure the labor market outcome of interest for each generation. This section of the tutorial walks the user through the kinds of decisions a researcher must make in the two-generation case. In the three generation case the data archive is going to have the most information on one set of grandparents and the most on *either* the father or *mother* in that set.

Locating fathers and sons

As noted at the outset of this tutorial, we would like to know whether sons' earnings are correlated with fathers', i.e. whether fathers with high labor income tend to have sons who also have high incomes, and whether low earning dads have sons who tend to be low earners. And we may want to know covariates predicting son's earnings other than the dad's earnings. For example, both sons and dads earnings may be related via labor market hours. For our immediate purpose, thinking of the paths through which the generational earnings relate, though intriguing, is just a distraction! To do this analysis, we need data covering sons and their fathers. More specifically, one wants a flat file with observations on a number of different men, each of whose record includes information about the individual's dad. Fortunately, the PSID website has a tool that makes it easy to match individuals with parents--the "Family Identification Mapping System" (FIMS). FIMS is readily available. It is possible to create a dataset containing fathers and sons without FIMS, but FIMS greatly simplifies the process because it can construct data files of the unique identifier for PSID individuals, their 1968 ID and Person Number. Referred to in terms of relational data files this 1968 ID (ER30001) and Person Number (ER30002) pair is their 'primary key'. The 1968 ID is the family they were a part of in 1968 or are a lineal descendent from that family who has split off to form their own family. From this information records are kept of each individual's relationship to others in the genealogical family, whether they are co-residential or are living on their own or as part of another family – commonly through marriage. The files are organized with pre-specified relationships (such as individuals and their siblings; or individuals and their parents; or individuals, parents and grandparents). Accordingly, this tutorial will show the user how to use FIMS to search for and create files which consist of pre-specified matching of family members. Here sons and dads. But one can get files of siblings, or children and their grandparents....

Temporal issues

Beyond identifying relationships between individuals, one will need to identify the proper analysis variables for the exercise. The tutorial will illustrate two ways that one can examine the correlation between fathers' and sons' (the id mapping from FIMS) and earnings of sons and dads (the variables from the Data Center). As a first approach, we will examine the labor income of each generation in fixed calendar years, taking a recent period (2001) for sons, and a calendar year further back in time (1973) to observe fathers' labor income. The logic employed here is to choose a time period for the fathers that could be expected to correspond to a period in which they would be working (not retired) and raising their families. Suppose, for example, that we would like to study male Baby Boomers (men born between 1946 and 1964) and their dads. This would give us a sample of individuals age 37 to 55 in 2001, for whom we could use the variable HDEARN01 (which is the earnings of the head variable found in the PSID in 2001) to ascertain current earnings. However, since we want to use 1973 to observe the dads' income, we will need data for two different variables (HDEARN01 and V3051) to do our intergenerational analysis, since the PSID changes the name of the labor income variable over time. The main point is that the first approach we call calendar year (or years)-based based matching. As a second approach, the tutorial will demonstrate how one can examine the correlation across generations while picking a specific age or age range in the lifecycle to observe earnings (rather than arbitrarily selecting a calendar year in which to observe the younger generation and a different one in which to observe the older generation). Matching fathers and sons at similar points in the life-cycle (say in their 40s) is beneficial because it means one should be observing fathers and sons at the same phase in their work lives. This approach will require us to take information from a number of different survey years (for each generation), as different individuals reach their 40s during different calendar years, to create what we will refer to as life stage matching. Often this matching will boil down to an age range. But what that age range might be would depend on the research question. Presumably the relation of BMI across generations would be better in some form of life stage matching rather than picking arbitrary years. One form of matching generations would be to match individuals in a

given calendar year! But that would be a case where and life course and panel aspects of the analysis would disappear, possibly with siblings of the 'same generation'. FIMS would still be helpful.

Some general challenges

The discussion of the two possible approaches that one can take to measuring the intergenerational correlation in earnings illustrates some general challenges that researchers face when making comparisons across generations. As alluded to above, one challenge that intergenerational researchers face is determining how best to measure parental outcomes. There are several ways to think about determining whether fathers and sons have similar outcomes in some domain. As note, one could examine fathers' and sons' outcomes during the same time period, i.e., during the same calendar year (using HDEARN01 from 2001 to attempt to measure earnings for both generations, for example). The obvious drawback of this approach is that by 2001 many of the fathers of baby boomers will not be earning any labor income, so this is not the best way to measure fathers' earnings. (Many dads would be expected to have missing data for the HDEARN01 variable since it asks about earnings from one's current work, and deceased and non-working individuals will not have been routed through the question sequence.)

Another approach is to pick a calendar year in which one knows that most of the baby boomers would have been dependents, and to seek out information about fathers in that particular survey year. For example, in 1973 individuals born between 1946 and 1965 would have been between the ages of 8 and 27. Accordingly, if a researcher maps the baby boomers back to the families they resided in 1973, and draws data on the fathers' labor income during that year, the researcher will be able to obtain information about most of the fathers of the baby boomers in his sample. One limitation of this approach, however, is that the researcher is likely to be unable to obtain information for the all the older baby boomers, since individuals 18 or older may already have left their parents' households, meaning that it will not be possible to match them with their natal units (parents) in 1973 if one relies on co-residence to do the matching.

A third issue that researchers conducting intergenerational analysis face, also alluded to in previous discussion, is determining whether it is important to observe parents and their children at the same stage in the lifecycle. When analyzing father-son income correlations, we might want to compare the incomes of the two generations at the same point in the life-stage (meaning when the fathers and sons were roughly the same age), because individuals face rising earnings profiles. This fact implies that if one were to measure sons' incomes when the sons were in their 20s and to compare this to the fathers' incomes at a different age, say when the dads were 50, one would not be making a valid comparison between the incomes of the two generations. The dads already would have reached the point in the lifecycle when individual earnings peak, while some sons would be early in their work in careers in which earnings tend to be much lower than average lifetime earnings. Instead, when possible, it can make sense to pick an age that is considered to be a peak or otherwise 'representative' earning age, depending on one's theory, for individuals and to obtain one's data for all fathers and sons at this pre-specified age. An even more complex type of life stage approach would be to attempt to match the discounted earnings of sons and dads by observing both generations over, say, as large a portion as possible of the years that correspond to age 20-70 and calculating present values. (Or for BMI possibly some average over a substantial part or multiple years of the life course (Davis and Gebremariam, 2005)).

So, one must also decide whether to use a single year or multiple years to measure the phenomenon of interest for each generation, whether going the *calendar year* based or the *life course* based route. One attractive feature of multiple years from the panel, in the context of labor income, is that, like BMI (Kim McGonagle and Stafford, 2001) income can fluctuate from year to year. Accordingly, a measure that

averages income over several years can paint a better picture of an individual's true earnings (his long-run earnings capacity) than a single year's observation does.

Now that we have contemplated the key issues that intergenerational a researcher faces when attempting to decide how to set up his or her study, we are ready to move ahead and to actually do some analysis. We start by getting our data from the PSID's [Data Center](#).

IV. IG Analysis: Illustrating Two Approaches

This section discusses the steps that you need to take in order to obtain the data for your dataset and to assemble your dataset so that you can analyze it. The tutorial provides its discussion using applications of SAS, but the PSID's Data Center allows a user to retrieve data in SPSS, Stata, and Excel format, so it is possible to use other software programs to do one's analysis.

To organize your thoughts as you maneuver through this section (Section IV), remember that this tutorial will provide two examples of ways that a researcher might conduct an intergenerational study of earnings. Example 1 shows the user how to compare the earnings of two different generations using a fixed calendar year for each generation, meaning that the researcher observes the earnings of each generation only once. Example 2 walks the user through a more complicated approach. In it, we show the user how to examine earnings at particular age synchronized points in the lifecycle and using more than one year's worth of information to gauge an individual's earnings. Because both examples require the use of FIMS, this section begins with a discussion of how to get the FIMS dataset. After it, sections B and C proceed with the discussion of the two examples.

A. Getting the FIMS data

In order to get started, we need to use FIMS to construct a generation map. To use the FIMS data extraction tool to do this, go to [FIMS](#).

Step 1: File with sons' and dads' identification numbers

FIMS is very flexible and allows many choices of ID matches. We have provided ID maps with the Child Development Supplement (CDS) file as noted in [Tutorial 4](#). These 'mapfiles' have been used by many CDS researchers and are produced automatically with CDS data files. There the persons of interest to CDS users almost always include the Primary Caregiver (PCG) and sometimes include the Other Caregiver (OCG) or the grandparents. FIMS can also be used to produce these automatically generated CDS ID map files, but allows a very wide variety of user choices.

Because our analysis requires information about fathers and sons, we want an inter-generational map (not an intra-generational one). Under the "Parent" option highlight the bar for "biological and adoptive." Under the "Generation map" option, choose "individuals to parents" since we want to link individuals with their parents. Under "Map type" select "balanced map." (A balanced map is a generation map that does not contain missing 1968ID-PN data in any of the generations selected. Missing data is most likely to be a concern when mapping to grandparents or higher generations, so in those instances a researcher might opt for an unbalanced map.) The idea of balanced and unbalanced generational id *maps* is parallel to the idea of balanced and unbalanced panel *data* discussed in [Tutorial 3](#). Under "Output options" you want to select "wide file" for the "file format." This will give you a file that is organized so that each row represents a given individual in the PSID, with his parents' identification numbers (1968ID-PN) appearing in different columns

of that row. Figure 4 shows what you should see at the FIMS portion of the PSID website after following these instructions.

Figure 4. Screenshot of FIMS data extraction tool

Note that the screenshot displayed in Figure 4 shows that we have chosen to have the FIMS datafile delivered as ASCII data that can be read using SAS. One other thing to notice about the screenshot (Figure 4) is that FIMS has several user help screens. These actually help, we think! You can get to them by selecting on any of the items listed under "User Help." This list includes specific instructions about how to merge FIMS files with data taken elsewhere from the PSID. For additional information about FIMS, consult the [FIMS documentation](#).

Step 2 Downloading your FIMS dataset

What will happen after you select "submit"? Once your datafile of ID's has been created, a link to both the ASCII data and the SAS commands will appear just below the submit button. You can download these files by clicking on each. What kind of information is contained in the datafile that you receive? Figure 5 shows what the file that will be sent to you looks like. (Of course, you will have to do a bit of SAS programming to read the data into a SAS file before you can display it in this fashion, but more about that later!!)

Figure 5. Subset of the ID observations from the SAS file sent by FIMS

The first two columns in the datafile list the two variables that can be used to uniquely identify each individual in the PSID. ER30001 is the "1968 interview number" (the id number for the original sample family interviewed back at the start of the PSID from which the individual descends). ER30002 is the individual's "person number." As noted earlier, remember that each individual in the PSID has a unique 1968 family id number and a person number (1968ID-PN). Taken together, this pair of variables uniquely identifies each individual in the PSID. They can be combined to create a single unique identifier variable. (More specifically, many researchers create a new variable by multiplying ER30001 by 1,000 and then adding ER30002 to this.) To further assist you in thinking about how one can view these individual identifier variables, one might think about the '68 id number as being akin to a last name (a family name) like "Clinton," while the person number is akin to an individual's first name (perhaps "Bill"). Used in combination, the two variables then yield an identifier ("Bill Clinton"), which can be used to distinguish "Bill Clinton" from other members (and descendants) of the Clinton family (such as Hillary or Chelsea).

The analogy used here is not quite correct. While Clinton works nicely because it is a relatively uncommon name in practice, it is important to note here that if we had two different 1968 families with the same surname (like "Jones"), they would each have a different '68 family id number. (It is also important to

note that if Chelsea Clinton were to marry and change her last name, the PSID would still link her back to her original '68 family with her original '68 family id number.) Thinking about this analogy is illuminating however. It allows one to understand why there are several rows with "4" in the ER30001 column. Each of these rows represents an individual that is a member of one extended family. You will notice that each row with a "4" in the ER30001 column has a distinct number in the ER30002 column. Again, this is because the PSID is using the person number to identify each member of the '68 family (and extended family) that was assigned a "4" for the family id number.

The two columns with the individual's identifier variables are followed by a series of columns that contain information about the individual's parents. Notice that each column begins with a prefix of either "ER30001" or "ER30002." This is because the PSID aims to give you each parent's unique identifier. However, in this instance the variable names also include extensions--that is to indicate whether the particular parent is a father (for which "F" appears as the last letter of the variable name), and to indicate whether it is an adoptive father (where "A" appears in the second to last letter of the variable name) or a birth parent (in which case the "F" is preceded by a "_" and not a letter). So, for example, the individual identified in the third row of the datafile that FIMS sent you, has a father whose family id number is "4" and whose person number is "1". You will notice that this individual has no entries in either the "ER30001_P_AF" column or the "ER30002_P_AF" columns. Similarly for the mom, the person's mother has the family id number "4" but her person number is "2" (which is different from the father's). This information is found in the ER3001_P_M and ER30002_P_M columns (where "P" stands for parent and "M" denotes mom). There are no data in the "ER30001_P_AM" or "ER30002_P_AM" columns. That is because the individual is not an adopted child.

So what is significant about this datafile that FIMS created for you? It lists the individuals who appear in the PSID along with identification numbers for their parents. This will allow us to use this file to create a subset or sample that contains male baby boomers, with their fathers' ids already attached to their records. We will use the fathers' id numbers to merge in information about the fathers' earnings, and we will use the individuals' (sons') id numbers to attach information about their earnings in order to examine the correlation between the two.

3. FYI (only): Interesting sample composition issues

Skip and proceed to [Part B](#) if you are eager to get started analyzing data.

When constructing a cross-generational dataset, selection issues that might affect the characteristics of the sample can emerge. In a sense it is possible to view the PSID as following a bloodline over time rather than a person per se, although following a bloodline requires one to track individuals over time. Additionally, a researcher conducting intergenerational analysis can be viewed as being interested in family dynasties. Recognizing this fact becomes important because different dynasties may have different generational spans, which could have implications for a researcher's analysis. Two examples come to mind.

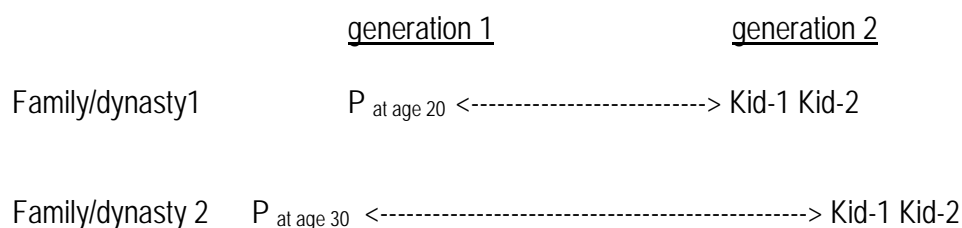
First, consider the case of two different families, each with two children, in a situation in which the two sets of parents began childbearing at different ages. Suppose the first family began having children when the parents were 20 years old, while the second did not become parents until they were 30. This means that the span between generations will be different in the two families. As shown in Figure 6, in the first family there will be a 20-year gap between the age of the parents and the age of their adult children. For

the second set of parents, the gap will be at least 30 years. So suppose we wanted to draw a sample of adult kids and their parents in a given year (say 2005). The PSID would easily contain all four adult children (on the assumption that all were grown and living independently). However, we would expect to be less likely to observe parents who had their kids later in life than the ones who had their kids while in their 20s, because the former would be older than the latter, and subsequently might be more likely to be deceased (*ceteris paribus*) during the year in which we are conducting our study. In this case, our sample of adult children and their parents might contain usable information for the adults from family dynasties with short generation spans only.

If people who have kids at young ages are substantially different from those with children later in life, particularly in terms of some unobserved characteristic that might be associated with the analysis variables, a researcher may need to be concerned about sample selection. For example, if it is the less educated who tend to begin childbirth early, and one is analyzing parent-child correlations in earnings, then the researcher might need to worry that the parent-child pairs in her sample are likely to be disproportionately composed of individuals whose parents were not highly educated (which would have implications for their earnings). Note, however, that this potential bias from observing adult children-parent pairs from short generation family dynasties should be reduced if the panel-intergenerational observation period is very long. Then information about parents from both long and short generations will be included.

Figure 6.

Generation span comparisons across dynasties--where members of generation 2 represent the unit or focal point of analysis ("P" denotes parent)



Second, consider the case of two sets of parents who began childbearing at the same age, but whose fertility decisions differ in terms of the number of children borne. Let the first family have 5 kids and the second 1. Here there is little difference in generation span across dynasties, but there is a difference in family size (as depicted in Figure 7). This has implications for the likelihood that one will be able to observe all of the offspring of the first family in any given year, since it is possible that some of the children will not be adults living independently yet, which means there will be limited information available for them (because the PSID collects most of its information--particularly that surrounding labor market activity--for individual when they serve in the roles of household heads and their wives or partners). If the missed children are, *de facto*, observationally non-distinct from the included ones however, the failure to observe some children should not affect one's analysis.

Figure 7. Differences in Family size



Family/dynasty 2 P at age 20 <-----> Kid-1

B. Intergenerational analysis: fixed calendar years

In this subsection the tutorial shows you how to obtain data that one can use to compare fathers' and son's earnings, when the data for each generation come from different calendar years. Because there has been a substantial amount of interest in baby boomers in recent years, we make them the focal point of this analysis. We will take the boomers' earnings from 2001 and compare it to their dads' earnings in 1973. This is a point in time that is almost 30 years earlier than the time being used for the boomers, so the boomers would have been relatively young during it, and the fathers should have been in a phase of life in which they were young enough to be raising families and somewhere in the midst of their work lives rather than being old enough to be in the retirement phase.

We organize the discussion of the fixed calendar year example in the following way. First, we discuss the variables you will need to construct a dataset containing all necessary information about sons; and follow this with a discussion of explicit steps to be taken to obtain these data from the PSID's Data Center. Second, the tutorial describes the variables one needs for dads, and shows one how to obtain these data. Third, the tutorial provides instructions for merging the son data and the dad data and the FIMS data. Finally, we describe steps to take to do your analysis on your combined dataset.

Step 1. Selecting the analysis variables for sons

(a) What variables do you need for sons? Son's earnings will come from the calendar year 2001, so-- as noted in Figure 1--HDEARN01. For the purpose of the tutorial, we will ignore the other two labor income-related measures (HDBUSY01 which is labor income generated from unincorporated enterprises, and FRMINC01, which compiles asset and labor income from farm activity). We ignore these for computational ease in the tutorial and because very few baby boomers have these sources of labor income. We also need a few PSID variables that provide useful descriptive information about individuals: (i) the variable indicating an individual's sex (ER32000) to identify males in the PSID, (ii) a variable indicating the age of the household head in 2001 (ER17013) to identify baby boomers, and (iii) two additional variables (ER33602 and ER33603), which will allow us to make sure we assign the family-level earnings data to the appropriate individual in the family (i.e., the one who earned it). Because there are a number of "positions" that any given individual in the PSID could hold in his or her family--head, wife, dependent child, for example--we need those last two variables noted (ER33602 and ER33603) to help ensure that the individuals we select from the PSID Data Center happen to be the males who are the heads of their households in 2001. This way it is legitimate to use the head's earnings data to characterize their earnings.

We also will need the two variables that allow us to create a unique identifier variable for each individual, and a statistical weight. The id variables were introduced to you back in the section discussing FIMS. As you will recall, the PSID assigns a 1968 family id variable or "ID68" (ER30001) to each individual in the PSID, along with a person number "PN" (ER30002). Together, these two variables can be used to create a single variable that uniquely identifies any individual in the PSID. We need a weight variable because the PSID data is nationally representative when weights are used for analysis purposes. The 2001 individual weight variable is ER33637. Finally, while at the Data Center you will be shown how to obtain data about work hours (HDTOT01). This will be useful if you decide to do follow-up analysis examining fathers' and sons' work hour choices. And this way the selection of more variables for other types of intergenerational analysis can be illustrated.

(b) How do you get these data? Go to the PSID Data Center, which should take you to a screen like the following (Figure 8) and then click "By File."

(Note: as a shortcut to selecting variables separately as in figures 9-12, we have saved a public "data cart" in the Data Center where you can easily get these variables with one click. To get this cart, login to the Data Center and be sure you have no variables in your cart. Then goto the "Variable Selection By Cart" page <http://simba.isr.umich.edu/VS/c.aspx> and enter the email datacenterhelp@isr.umich.edu and click the [Submit] button. Select the cart titled "Tutorial IG: analysis variables for sons." These variables will now be in your cart. Click "Checkout" and you should see a page that looks like the one in *Figure 13*.)

Figure 8. Screenshot of the PSID Data Center

▪ Now you should see a screen showing the different PSID sub-files (or "data groups") in which the PSID data are stored. We will be using variables that come from 3 different data groups in order to obtain that data that we need for sons. Expand the file tree so that the following three nodes are showing:

- 1) PSID Family-Level Data -> PSID Main Family Data
- 2) PSID Individual-Level Data -> PSID Individual Data by Years
- 3) PSID Individual-Level Data -> Summary Variables (Sampling Variables, Family History Variables, and Sex)

Figure 9. Screenshot of the Data Groups options

- Next you need to indicate the years for which you want data. Expand the three nodes above so that the years of available data are displayed. This list notes every year in which the PSID collected data for the given file. For our exercise one needs data from "2001" from the "PSID Family-Level Data -> PSID Main Family Data" and "PSID Individual-Level Data -> PSID Individual Data by Years" nodes, and "all years" (the only selection available) under the "Sampling variables" node. Click the [+] sign next to these yearly indicators on each of the three nodes.

Figure 10. Screenshot showing how to select data categories by year

▪Now you are ready to select the actual variables that you need to analyze sons. There were a total of 10 variables discussed under Step 1(a) above. You will be selecting 7 of them here (the last two will be given to you automatically by the Data Center).

- 1) Under the first node, "**PSID Family-Level Data -> PSID Main Family Data -> 2001**" select the age of head variable (ER17013), then scroll down towards the bottom in the same selection box and select the labor income variable (HDEARN01) and the work hours variable (HDTOT01).
- 2) Under the second node, "**PSID Individual-Level Data -> PSID Individual Data by Years -> 2001**" select the variables ER33602, ER33603, and ER33637.
- 3) Finally, under the third node, "**PSID Individual-Level Data -> Summary Variables (Sampling Variables, Family History Variables, and Sex)**", select the sex of individual variable (ER32000).

Figure 11 (below) provides a partial shot of the screen that you should be facing at the moment. As you can see, one selects the variables one wants by highlighting them. You do this by clicking on the variable name. And, note that when selecting more than one variable for a list, you need to hold down the "Ctrl" key while clicking on the variables of interest. When you have finished selecting your variables, click the "Add To Cart" button. (Note that the screenshot below depicts one, not all, of the variables you need to select. Because some of the variables--such as HDEARN01 and HDTOT01, and variables in the remaining two nodes--are found only after scrolling far down in a particular data group, or in the tree, it is not always possible to display all of the highlighted variables at once.

Figure 11. Variable selection screen

▪Now you should be faced with a screen that lists all the variables that you have chosen (or, in PSID jargon, the contents of your "data cart"). The screen should look like the one below.

Figure 12. Data cart contents screen

Note that while you only selected 7 variables, the Data Center automatically adds three. Two of these are the variables you need to construct the unique identifier for individuals. The third is the 2001 family interview number, which tracks the family unit in which each PSID individual resided in during 2001. This variable often is needed for analysis, so the Data Center always assumes that you may want it, even if you do not specify such.

If you cannot remember the specifics of a particular selected variable, you can double-click on the variable name or label and a codebook box will appear reminding you about the nature of the variable that you selected (what it measures, for example, or what question was asked to obtain it). If you are missing some data, this is the point at which to stop to go back and get it. If you use the "back" button on your internet browser, or the "Variable Selection" option in "Data Center" menu at the top of the page, you can select whatever variable(s) you may have forgotten. The PSID Data Center will automatically add the new selections to your existing list, so that when you return to the data cart contents screen your data cart will show the original selections and the new additions. If you accidentally selected a variable that you do not need, just click on the empty check box to the left of the variable and then hit "Delete Checked Nodes" at the top of the screen.

- If you have all the variables that you need, you are ready to hit "Checkout" at the top of the screen! If you are not currently logged in to the Data Center, you will be promoted to do so.
- Once you've hit "Checkout" you will be on the "Checkout" page and should see the "Output Options" screen (shown below). This is the point at which you will tell the Data Center how you would like your data delivered to you, and whether you want a codebook. Since this tutorial will do its demonstration using SAS, click on the circles next to "ASCII data file" and "SAS." Also click on the circle next to PDF if you want your codebook in PDF format (which is the easiest to read). Then, check "Send links to my Email" if you want your dataset delivered by e-mail. *The big advantage of checking this option is that it means you do not have to wait around at your computer while your dataset is being created. Instead, you can just check your e-mail at your convenience and get your data that way. Hmm...have you had dinner yet?* NOTE ALSO that we have entered text in the subsetting box. Since the tutorial focuses on sons (rather than all adult children), it makes sense to instruct the Data Center to send information on individuals who are male only. We do this by entering SAS code for restricting the ER32000 variable to take on a value of 1 (which indicates that we have an individual who is male). Additionally, since we know that our information about earnings will come from the variable denoting the earnings of the head of the household, we need to make sure that the individuals we select are heads of household. We do this using a variable that tells us what relationship a given individual has to the head of the household (ER33603). We want it to take on the value of 10, which indicates that the person was actually the head of the household in 2001. Finally, we use the sequence number variable (ER33602) to ensure that we have heads who are physically present in their households in the chosen year. Restrict this variable to take on a value of 1. (Those of you who are savvy with SAS know that you can easily do this subsetting in SAS after you get your dataset. However, since the Data Center allows us to do it upfront, we do it here.)

Figure 13. Output options screen

▪Now hit "submit" and the Data Center will get to work. You should get a dataset along with a SAS starter program that will allow you to read the ASCII data into SAS in no time.

Step 2. Selecting the variables for dads

While you are now undoubtedly excited because after all this you've actually gotten some data, you are not, however, done. The data you chose Step 1 (above) only represents some of the data you will need for your analysis. You now need to revisit the Data Center to create a dataset that will contain information about *dads' and their earnings*. You will need to take similar actions as you did in Step 1, but you will be choosing slightly different variables because the information for dads will come from the 1973 wave of the PSID.

(a) What variables do you need for dads?

We need heads' annual earnings (V3051) to measure the earnings of dads. We need V3096 (sex of the household head) so that we can restrict the dataset to household heads who are males. We also need the relationship to head variable and sequence number variable (both for 1973), for the reasons that we selected these variables when obtaining the data for sons (noted in Step 1). This means we need to select ER30118 and ER30119. As always, we will need the two PSID variables that allow us to create a unique identifier for each individual in the PSID--ER30001 and ER30002--as discussed in Step 1. Finally, we want a measure of heads work hours (V3027).

(b) How do you get these data?

- To obtain your variables you will need to follow procedures similar to the ones outlined in Step 1. These involved (i) going to the PSID Data Center, (ii) selecting particular data groups from which your variables will be selected, (iii) choosing the relevant year for which you will want data, (iv) selecting the actual variables, (v) viewing your variable list, and (vi) choosing your desired output options. Figures 14 through 16 depict the screens you will face as you navigate through the Data Center.

(Note: as a shortcut to selecting variables separately as in figures 14-15, we have saved a public "data cart" in the Data Center where you can easily get these variables with one click. To get this cart, login to the Data Center and be sure you have no variables in your cart. Then goto the "Variable Selection By Cart" page <http://simba.isr.umich.edu/VS/c.aspx> and enter the email datacenterhelp@isr.umich.edu and click the [Submit] button. Select the cart titled "Tutorial IG: analysis variables for dads." These variables will now be in your cart. Click "Checkout" and you should see a page that looks like the one in *Figure 16*.)

- The three main differences between what you will do here and what you did in Step 1 are that: (a) we will only work with the PSID individual and PSID family sub-files here, (b) we do not need to select an age variable as we did before, and (c) we do not need to select a weight variable, since we will be able to use the one we already selected (in Step 1) for our statistical analysis.

- Finally, note that when you get to the screen where you select your output preferences (shown in Figure 16)--in addition to selecting your codebook preference, ASCII data, and SAS definition statements--you will want to enter a subsetting command that is slightly different from the one we used in Step 1. Though similar in spirit, because dads' data come from 1973, the variables that we need to work with here are different. The subsetting command you want to enter is $ER30119 = 1$ and $ER30118 = 1$ and $V3096 = 1$. The first equation tells the Data Center to restrict the dataset to individuals who are heads of their households (the code for head in 1973 is "1," which is different for the value used in 2001). The second equation restricts the cases to those in which the head was present in the household. The third equation ensures that we only obtain data for men.

- Remember to check "Send links to my email" if you want your dataset e-mailed to you.

Figures 14 through 16 Screenshots for the Dad's data selection part of the exercise

Step 3. Merging the files created in Steps 1 and 2 with your FIMS dataset

At this stage you are ready to take the files that the Data Center created for you, and to construct 3 separate SAS datasets from them, which you can then merge together to create one SAS dataset to use for your analysis. Most Windows versions of SAS are quite user friendly these days. After you have unzipped the files that the Data Center has sent you, it should be easy to work with them by simply clicking on the SAS program sent by the Data Center. This will cause your computer to automatically open a SAS session.

Manipulating your FIMS dataset to ready it for the merge

Let's begin with the FIMS dataset. You should have two "files" for this--one with a SAS extension (the SAS starter program) and one with a "txt" extension (the ASCII data to be read into SAS). If you click on the file with the SAS extension you will open up a session of SAS, and a program should appear on your screen. It should look like the following:

Figure 17. SAS starter program associated with the FIMS dataset

As you can see, with this program SAS is almost ready to start inputting the data that you asked for from the FIMS Data Center. You simply need to add a libname statement at the beginning of the program and to specify a "path name" so that SAS knows where to look for the ASCII text data. If you stored this datafile on your c: drive after downloading it at FIMS, then substitute the code "c:" for the current phrase "[path]". If you stored your data somewhere else, just tell SAS where that happens to be. Appendix A contains some sample code to show you what a transformed program might look like.

Note that the second row of text in the screenshot tells you how many rows your dataset has. This is equivalent to the number of observations in the dataset, i.e., the number of individual records that you have

been given. You may be thinking, "Why are there only about 32,000 observations if there are over 60,000 individuals in the PSID?" While FIMS is designed to produce records that allow one to link any given individual to his or her parents or offspring (descendants and ancestors more generally for families for which four generations have participated in the study), linkage information is not available for all PSID individuals in FIMS. There are some individuals in the PSID for whom no parent identification information is available. This can occur for several reasons. For example, if an individual who marries into the PSID by marrying a respondent who is the adult child of an original 1968 family, the PSID will contain parent information for the child of the 1968 family, but there would be no records for the parents of the spouse who married this respondent (because the spouse was unknown to the PSID until he married someone in the sample). For a more detailed discussion of the difference between the number of cases in FIMS and the number of individuals in the PSID as a whole, consult the FIMS documentation.

- The most important thing for you to do at this stage is to create unique identifiers for the individual whose records are being presented to you in the FIMS dataset, and a variable to uniquely identify each individual's dad. (Note that at this point we cannot really talk about id variables for "sons" per se, because the file that FIMS sent us contains records for men and women, so the individuals may be sons or daughters. However, once we merge the FIMS dataset with our dataset created in Step 1, we will be able to narrow this dataset down to sons. For this reason, we will start labeling our id variables as if they correspond to sons. We will rename the ER30001 and ER30002 variables and then create a single, unique identifier from them.)

- To rename the ER30001 and ER30002 variables, you can use a SAS rename statement in your datastep like,

```
(Rename = (ER30001 = son_68id ER30002 = son_pn))
```

or you can add a line of code to your program like,

```
son_id68 = ER30001;
son_pn = ER30002;
```

- To create an individual identification (id) variable for each person in FIMS, you simply instruct SAS to create a new variable by multiplying ER30001 by 1000 and then adding ER30002 to this product.

```
son_id = ER30001*1000 + ER30002;
```

- To create an identification variable for dads, recall that FIMS sends you information about moms and dads, and about adopted parents and natural parents. The ERxxxx variables with suffices attach to them correspond to ER30001 and ER30002 variables for parents (with an "A" indicating an adoptive parent and "M" indicating mother while "F" indicates father). Since we are not interested in mothers, we only need to create a variable to store a unique identifier for the dads. We do such with language such as the following:

```
if (ER30001_P_F = . and ER30001_P_AF = .) then do; dad_id = .; end;

else do;
if ER30001_P_F ^= . then dad_id = ER30001_P_F*1000 + ER30002_P_F;
else dad_id = ER30001_P_AF*1000 + ER30002_P_AF;
end;
```

You should recognize this language as a standard SAS "do loop" with "if-then" statements embedded in it. We are instructing SAS to create the unique identifier for each dad by taking the dad's '68 id number and multiplying it by 1,000 and then adding the dad's person number to this product. We also are telling SAS to use the ER30001xxx and ER30002xxx variables that do not have an "A" in them as long as the natural dad's information is not absent. To the contrary, if there is missing data in the columns representing the natural dad, we have told SAS to look to the columns containing information about adoptive dads in order to create the dads id variable. Furthermore, if there is an individual for whom data is missing for both adoptive and natural dads, SAS assigns a missing value marker [the dot symbol "."] to the newly created dad_id variable to denote this.

- These newly created unique id variables will be important as we move forward with the tutorial exercise. They will be used to merge that datasets obtained in Steps 1 and 2.
- You probably will want to save your new dataset as a permanent SAS dataset so that you can retrieve it later. Try a simple name like "FIMSdata" or "FIMS_ids" to help you remember that this is a dataset based on FIMS that contains the unique identifier variables that you created for sons and dads.

Manipulating your sons dataset (from Step 1) to make it ready for the merge

Your key focus at this stage will be on working with the dataset that you selected at the Data Center in Step1 to (a) parse the dataset down so that it only contains baby boomers (and not all men in the PSID), and (b) to create a unique identifier variable for these men so that you can merge this dataset in with your FIMS dataset (created above).

We start as we started above--by clicking on the "file" containing the simple starter program that the Data Center sent you when you requested the sons dataset (the file with the "sas" extension at the end). A SAS session should open and you should see the beginnings of a SAS program--something like the following:

Figure 18. Sons dataset SAS starter program

Once again, you will need to specify the location of a drive (in place of the current phrase "[path]" in the first line of the program), so that SAS knows where to look for the ASCII data (the file with the ".txt" extension that the Data Center sent you). Moreover, you need to add a libname statement to the program.

What other modifications do you want to make to the starter program? *Three* additions are required.
(a) You need to add language to instruct SAS to limit the SAS dataset that it will create to individuals who are baby boomers. Having been born between 1946 and 1964 baby boomers would be between the ages of 37 and 55 in 2001. Accordingly, we want to instruct SAS to create a new dataset that drops any individuals who do not lie in this age range. We can do this with a simple output command:

if 36 < ER17013 < 56 then output;

(b) You also want to add a new variable to the dataset, so that we can use it as a unique identifier for these baby boomers, who are sons of someone to whom we will seek to match them later. Let's call this variable "son_id." We need the following language to create it:

```
son_id = ER30001*1000 + ER30002;
```

(c) Finally, we will rename the labor income variable for sons. Let's call it "s_labory" where the "s" stands for sons and the "y" stands for income. (If you've taken any economics courses, you're undoubtedly familiar with the use of "y" to symbolize income.) You can rename using the following language;

```
s_labory = HDEARN01
```

*LASTLY, make sure you create a permanent SAS dataset (as you did for your FIMS data) so that you will have a dataset that reflects the changes you made. This way it will be available for later use, and you will not have to re-do all the steps when you want to merge this data with your FIMS-based dataset.

Appendix B contains an example of a modified version of the sons starter program that accomplishes the three tasks outlined above.

Manipulating the dads dataset (from Step 2) to prepare it for the merge

By now you probably get the system: you will need to work with the SAS starter program that the Data Center sent you in order to convert the ASCII data it sent into a SAS dataset. Click on the starter program, and then add a libname statement to it, and then specify the location of the ASCII dataset.

What type of language do you need to add to this SAS program?

(a) We need to add a bit of code to convert the ER30001 and ER30002 variables into one single variable that uniquely identifies the individuals in this dataset (who are prospective dads of the baby boomers). To do this, you need language like the following,

```
dad_id = ER30001*1000 + ER30002;
```

(b) You also may find it helpful to rename the 1973 labor income and work hours variables so that you will remember that these are the variables containing information about fathers' annual labor income and fathers' annual hours of work respectively. You can do this by creating two new variables. Let's name them "d_labory" for dad's labor income (to correspond with the label we used for sons) and "d_hours" for dads' hours of work. To rename variables you can use SAS' rename command:

```
(rename = (V3027 = d_labory   V3051 = d_hours) )
```

(c) Finally, do not forget to create a permanent SAS dataset that stores your new, modified dataset. This will enable you to recall the dataset for later use.

Appendix C contains an example of a modified version of the dads starter program that incorporates all of this new language.

Step 4. How to analyze your data

Now you are ready to merge your 3 datasets together. We will do this by first merging the FIMS-based dataset onto the sons dataset, and then by merging the dads dataset in. Note that none of your files contain the same number of observations. This means you will have to *be careful* when you merge. We recommend using a simple merge process and then keeping track of your process with an "in" statement. Also remember that you must sort your datasets by the variable you want to merge on before you can do a merge. Finally, remember that you have to instruct SAS to read the data from the permanent datasets that you created, which means you need to use the libname when you instruct SAS to use a particular dataset.

For example, here is some sample code that one might use to merge the sons data with the FIMS-based dataset in a case where you named your FIMS-based dataset "FIMS_ids" and your sons dataset "son_vars".

```
libname today "c:";
run;

proc sort data= today.FIMS_ids out=FIMS;
by son_id;

proc sort data=today.son_vars out = sons;
by son_id;

data merged1;
  merge sons (in = s)
        FIMS;
  by son_id;
  if s=1 then output;
run;
```

And, here is some sample code that you might use to merge in the dads dataset (which we have denoted as "dad_vars" in the following example)

```
proc sort data=merged1 out=sorted;
by dad_id;

proc sort data= today.dad_vars out=dads;
by dad_id;

data merged2;
  merge sorted (in = st)
        dads;
  by dad_id;
  if st=1 then output;
run;
```

With this you should have a dataset named "merged2" that contains all of the sons information combined with the dads information. This is the dataset that can be used to analyze the relationship

between adult sons' their dads' earnings. The most commonly used method of computing the association between sons' and dads' earnings in economics is to use the elasticity of sons' earnings with respect to fathers'. To get this, regress son's earnings on dad's earnings. However, we need to take logs of the labor income variables first, and to use the logged labor income variables in the regression. The following code accomplishes this:

```
data analysis;
  set merged2;
  if s_labory > 0 then log_s_y = log(s_labory);
  else log_s_y = .;

  if d_labory > 0 then log_d_y = log(d_labory);
  else log_d_y = .;
run;

proc reg data=analysis;
  model log_s_y = log_d_y;
  weight ER33637;
run;
```

Note that when we instruct SAS to run the regression we tell SAS to incorporate the statistical weights (the ER33637) variable in the analysis. Note also, that the sample code listed above has language designed to handle the fact that it is not possible to take the log of zero. This could be quite problematic: fortunately only a few individuals may have zero labor income in a given year. Yet, we had to add language to tell SAS to assign a missing value to the newly created log variables anytime the original variables (s_labory and d_labory) took on values that could not be logged.

Some concluding remarks about this fixed calendar year example

At this point you should have some output to interpret. The coefficient on the "log_d_y" variable in your regression gives you the estimate of the elasticity of son's earnings with respect to dad's. You should have an estimated elasticity of about 0.25. You have accomplished your task of analyzing the connection between son's and dad's earnings at this point! This is cause for celebration! Before celebrating too much, it is worth noting some limitations of the approach that we used, however.

As noted earlier, observing earnings in just one year for an individual has two key drawbacks. It can lead us to mismeasure individual earnings, since in any given year an individual may be experiencing unusually low earnings (if it is a recession year for example and the individual has been laid off) or unusually high earnings. For this reason, as shown by Solon (1992), when possible it is better to have a measure of individual earnings that average a number of years' worth of data. (Solon finds that averaging over five years provides a good way to resolve this measurement error problem.) Moreover, as shown in Mincer (1974), workers' earnings often rise until around the age of 40, so one can argue that earnings measure close to this peak gets a good handle on an individual's earnings capacity. In the next exercise--which you can do now if you are not *life stage matching* tired--we show right here and now how to put together a dataset that will allow you to compute an earnings elasticity estimate that overcomes these drawbacks.

C. Intergenerational analysis: life course matching

This example requires some time to get through it, yet it is exciting because it demonstrates some of the techniques that most experts use when attempting to conduct intergenerational research for scholarly journals and books. The goal of the example is to use information from a number of years to characterize sons' and dads' earnings. Moreover, a primary emphasis will be *the life course approach* which here will be based on using years from the time period in which different sons and dads were in their 40s, because of the life cycle earnings effects reviewed by Mincer (1974).

Step 1 Selecting your variables

(a) What variables will you need?

(i) *Labor income variables*: We want information about individual earnings, but here we want earnings observed in a number of years for each individual. Additionally, because different individuals will be in their 40s during different calendar years, we cannot simply take the earnings variables associated with a limited set of calendar years. Instead, we need the data collected for head's labor income *throughout the history of the PSID--from 1968 through 2003!*

(ii) *Age variables*: Because we want to look at men sometime during their 40s, we will need to know the age of every man in our dataset. We will use the age of household head variables (from the yearly PSID family files) for this. Again, we will need this variable for every year of the PSID--from 1968 through 2003.

(iii) *Sex*: Because fathers and sons are both male, as was the case for Example 1, Example 2 requires us to obtain information about an individual's sex. This means we will need ER32000 (from the Sampling Variables data group). It is the variable we used earlier to determine whether an individual was male.

(iv) *Work hours variables*: Because we know you may get ambitious later, and be inspired to attempt to examine intergenerational correlations in other aspects of labor market activity beyond earnings, the tutorial will show you how to obtain the variables detailing the number of hours the household head spent working in any given year.

(v) *Relationship to head and sequence number variables*: In Example 1, we saw how the relationship to head variable (and sequence number) could be used to identify individuals who were heads of households in a given year. For Example 2, we need this team of variables for every wave of the PSID, since we cannot know, a priori, what calendar years we will be turning to in order to retrieve information about any given son or dad's earnings. (As you will recall from Example 1, we need this information about relationship to head status to properly assign the labor income data--which comes in the form of "head's labor income"--to the correct individual in the family. These data come from the yearly individual files.)

(vi) *The individual weight variable for 2001 (from the individual file)*. This variable allows us to make sure that our results are nationally representative. (We discussed the use of statistical weights in Example 1.)

(b) How do you get these data?

You will probably not be surprised to learn that you need to return to the PSID Data Center to get your data. Before you head there, however, you need to know what the variable names are for the 7 different types of data that are listed above. Figure 2 listed the variables one needs to obtain head's labor income in every single year of the PSID, and you could easily figure out what names the PSID has given to all the other variables in different survey years by using the cross-year index that was discussed earlier in this tutorial (in Section IIIB). We provide a list below however, to save you some time.

(Note: as a shortcut to selecting the variables below, we have saved a public “data cart” in the Data Center where you can easily get these variables with one click. To get this cart, login to the Data Center and be sure you have no variables in your cart. Then goto the “Variable Selection By Cart” page <http://simba.isr.umich.edu/VS/c.aspx> and enter the email datacenterhelp@isr.umich.edu and click the [Submit] button. Select the cart titled “Tutorial IG: Life course matching variables.” These variables will now be in your cart. Click “Checkout” and you should see a page that looks like the one in *Figure 21.*)

Figure 19. List of variables the user needs for Example 2

year	labor income of head	total hours worked— head	Relation ship to head	Sequenc e number	Age of head	Sex	2001 Individual weight
2003	ER24116	ER24080	ER33703	ER33702	ER21017	ER32000	-
2001	HDEARN01	HDTOT01	ER33603	ER33602	ER17013		ER33637
1999	HDEARN99	HDTOT99	ER33503	ER33502	ER13010		-
1997	HDEARN97	HDTOT97	ER33403	ER33402	ER10009		-
1996	HDEARN96	HDTOT96	ER33303	ER33302	ER7006		-
1995	HDEARN95	HDTOT95	ER33203	ER33202	ER5006		-
1994	HDEARN94	HDTOT94	ER33103	ER33102	ER2007		-
1993	V23323	V21634	ER30808	ER30807	V22406		-
1992	V21484	V20344	ER30735	ER30734	V20651		-
1991	V20178	V19044	ER30691	ER30690	V19349		-
1990	V18878	V17744	ER30644	ER30643	V18049		-
1989	V17534	V16335	ER30608	ER30607	V16631		-
1988	V16145	V14835	ER30572	ER30571	V15130		-
1987	V14671	V13745	ER30537	ER30536	V14114		-
1986	V13624	V12545	ER30500	ER30499	V13011		-
1985	V12372	V11146	ER30465	ER30464	V11606		-
1984	V11023	V10037	ER30431	ER30430	V10419		-
1983	V9376	V8830	ER30401	ER30400	V8961		-
1982	V8690	V8228	ER30375	ER30374	V8352		-
1981	V8066	V7530	ER30345	ER30344	V7658		-
1980	V7413	V6934	ER30315	ER30314	V7067		-
1979	V6767	V6336	ER30285	ER30284	V6462		-
1978	V6174	V5731	ER30248	ER30247	V5850		-
1977	V5627	V5232	ER30219	ER30218	V5350		-
1976	V5031	V4332	ER30190	ER30189	V4436		-
1975	V3863	V3823	ER30162	ER30161	V3921		-
1974	V3463	V3423	ER30140	ER30139	V3508		-
1973	V3051	V3027	ER30119	ER30118	V3095		-
1972	V2498	V2439	ER30093	ER30092	V2542		-
1971	V1897	V1839	ER30069	ER30068	V1942		-
1970	V1196	V1138	ER30045	ER30044	V1239		-
1969	V514	V465	ER30022	ER30021	V1008		-
1968	V74	V47	ER30003	na	V117		-

Because you are now an expert at navigating your way through the Data Center, we will not bother with a lengthy discussion of what to do when you get there here. Instead, we present the sequence of screenshots that you should see as you make your way through selecting your different data groups, years,

variables, and--ultimately--your output options. The biggest difference to note between what you do for Example 2 and what you did for Example 1, is that you will not be creating two separate datasets at the Data Center--a distinct one for sons and a distinct one for dads--as we did in Example 1. Instead, we can use the same dataset to retrieve information about sons' and dads' earnings. (We will simply need to make two copies of it after we get it from the Data Center!) Finally, remember to subset when you get to the output options screen (you want to set ER32000 = 1 so that you only receive data for men), and remember to enter your e-mail address if you want your files delivered to you via e-mail.

Figures 20 - 21. Screenshots illustrating the screens one sees when retrieving data for Example 2

Step 2. Preparing summary variables

If you instructed the Data Center to e-mail you your dataset, you should check your e-mail and then download the files sent to you so that they are stored somewhere on your hard drive (or on a cd, zipdisk or external drive). When you are ready to get to work on the analysis, you can click on the SAS starter program that the Data Center gave you (the file with the "sas" extension). This should open a session of SAS, and you will see a short SAS program designed to help you read the ASCII data into SAS. As noted earlier, before you can run this program you must specify a libname and the location of the ASCII data. Then you will be ready to get to work. Figure 22 shows what the SAS program sent by the Data Center should look like (although it has been amended to add a libname statement and the name of the directory in which the ASCII data is located).

Figure 22. Sample starter program sent by the Data Center

▪Now you need to create some new variables:

- (a) You will create an age variable that assigns information from the PSID's "age of the head" variable if the individual in question was the head of the household in the relevant year. (You'll need one of these new age variables for each year of the PSID.)
- (b) You will need to create similar variables for labor income and work hours.

We have listed sample code below to help you think about how to do this. You can see that it contains a statement that specifies the full list of variables that we would like to create (new variables for age for

every year of the PSID, new variables for labor income in each PSID year--noted as "labory"-- and new variables for work hours.

```

data test;
  set job58068;
  length default = 8   age1968 age1969 age1970 age1971 age1972 age1973
                        age1974 age1975 age1976 age1977 age1978 age1979
                        age1980 age1981 age1982 age1983 age1984 age1985
                        age1986 age1987 age1988 age1989 age1990 age1991
                        age1992 age1993 age1994 age1995 age1996 age1997
                        age1999 age2001 age2003

                        laby1968 laby1969 laby1970 laby1971 laby1972 laby1973
                        laby1974 laby1975 laby1976 laby1977 laby1978 laby1979
                        laby1980 laby1981 laby1982 laby1983 laby1984 laby1985
                        laby1986 laby1987 laby1988 laby1989 laby1990 laby1991
                        laby1992 laby1993 laby1994 laby1995 laby1996 laby1997
                        laby1999 laby2001 laby2003

                        wkhr1968 wkhr1969 wkhr1970 wkhr1971 wkhr1972 wkhr1973
                        wkhr1974 wkhr1975 wkhr1976 wkhr1977 wkhr1978 wkhr1979
                        wkhr1980 wkhr1981 wkhr1982 wkhr1983 wkhr1984 wkhr1985
                        wkhr1986 wkhr1987 wkhr1988 wkhr1989 wkhr1990 wkhr1991
                        wkhr1992 wkhr1993 wkhr1994 wkhr1995 wkhr1996 wkhr1997
                        wkhr1999 wkhr2001 wkhr2003 8;

  if ER30003 = 1 then do; age1968 = V117; laby1968 =V74; wkhr1968= V47; end;
  else do; age1968 = .; laby1968 = .; wkhr1968 = .; end;

  if (ER30022 = 1 and ER30021 = 1) then do; age1969 = V1008; laby1969 = V514;
  wkhr1969 = V465; end;
  else do; age1969 = .; laby1969 = .; wkhr1969 = .; end;

```

Toward the bottom of the program one sees language that tells SAS how to assign values to the new variables we are creating. These commands use a combination of the relationship to head variable and the sequence number for each year. For example, the line that states "if (ER30022 = 1 and ER30021 = 1) then do; age1969 = V1008" tells SAS that if the individual in question has a relationship to head value of 1 during the year 1969, which stands for head, and a sequence number of 1 during that year, which indicates that the head is present in the household in the given year, one should assign the value given by the PSID's age of household head variable in 1969 (V1008) to the newly created "age1969" variable. In the event that the individual in question is not the head of the household in 1969, SAS will instead assign a missing value (a dot or ".") for age. (We do this because, if the individual in question is not the household head, we cannot use V1008 to assign an age to him, since V1008 represents the age of the head.) Two key points to note as you are adding language to the SAS program: (1) For 1968 there is no sequence number aspect to this command, and (2) Beginning in 1983, the PSID switched to a two digit code for head of the household, so you need to set the relationship to head variable equal to 10 instead of 1. With this information in hand, you should be able to write the appropriate code to allow SAS to create your new variables all years from 1970 through 2003.

- Now we recommend reducing the size of your dataset to eliminate variables that you will not need to analyze father-son earnings. We need all the newly created variables, and we will need the statistical weight variable (ER33637), and we will need the individual identifier variables (ER30001 and ER30002);

however, we no longer need any of the other variables in the dataset. Accordingly, at this point it makes sense to either drop all the other variables or to simply create a new (a 2nd) dataset that keeps only the variables we need. As you know, SAS has "drop" and "keep" statements that one can use to do this.

Creating a labor income variable that incorporates information about an individual's earnings in a number of years

Now we want to construct age specific labor income measures (so that we have variables that indicate what an individual earned when he was age 40, 41, 42 et cetera, until the age of 50); and we also want to construct a labor income measure that averages an individual's earnings over this age range. To accomplish this task, we will need a command that involves the age variables we created and the labor income variables. What steps will we need to take in the process?

- (a) We have to recode the age variable to convert values of 998 and 999 to missing values. (We do not want our SAS program to think that nine-hundred and ninety-eight is someone's actual age. It isn't. Ditto for nine-hundred and ninety-nine.)
- (b) We need to adjust the yearly labor income data to account for the inflation that occurs over time. To do this you want to convert all past values into 2003 dollars using CPI data from the Bureau of Labor Statistics.
- (c) We write 11 lines of code to tell SAS how to determine what an individual earned when he was 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, and 50. (Note that there are 11 ages there; hence the 11 lines of code.)
- (d) We will create an average labor income variable for each individual.
- (e) You can repeat this process for work hours (no CPI adjustment required!) if you want to analyze the correlation between father's and son's hours of work.

How can you do the above 5 tasks in SAS?

- A simple array statement will allow you to easily recode missing values on age for each year:

```
array AGE(*) AGE1968 - AGE2003;
do i=1 to dim(AGE);
if AGE[i] =998 or AGE[i] =999 then AGE[i] = .;
end;
```

- To convert the labor income data to constant dollars (with 2003) as the base year, you can either download consumer price index (CPI-U or a deflator of your choice) data from the BLS and input that data into your dataset, or you can use the CPI calculator at the site to determine what the appropriate conversion factor is for each year and then instruct SAS to multiply the labor income variable for the given year by that value. For example, the CPI calculator (available at www.bls.gov) tells me that \$1 in 1980 is equivalent to \$2.23 in the year 2003, so I could convert the labor income variable corresponding to the year 1981 to dollars from the year 2003 by multiplying it by 2.23, with the following SAS code:

```
labory1981 = labory1981*2.23
```

so that SAS readjusts the labor income variable for the year 1981 to factor in the effects of inflation. This way we can make legitimate comparisons between earnings from different calendar years. (If you checked your codebook and looked at the documentation for the income data you know that the income data in any given survey year is actually the household's income for the prior year--so the labor income values reported in the 1981 PSID survey actually tell us what the head earned in 1980.) The BLS website is located at www.bls.gov

▪How do we get SAS to construct age specific labor income variables? You can use a series of array statements and some computations involving them.

Start with the following three statements to create arrays for labor income, work hours and age:

```
array LABOR_INC(*) LABY1968 -LABY2003;
array WORK_HR(*) WKHR1968 -WKHR2003;
array AGE(*) AGE1968 -AGE2003;
```

Next add code to tell SAS to perform operations involving the arrays:

```
do YEAR=1968 to 2003;

    if AGE[YEAR-1967]-1=40 then do; LABOR40=LABOR_INC[YEAR-1967]; end;
    .....
    .....
end;
```

*Note that the second line essentially tells SAS to look at the individual's age variable, and--if that variable takes on a value of 40 after subtracting 1 -- SAS then creates a variable representing labor income at age 40 using the labor income variable from the corresponding survey year. This is the intuition of the line that begins with "if AGE...." The reason the code says to subtract 1 is because the age recorded in a given PSID survey year is the age in that year; however, as noted earlier, the income data that individuals report is income for the previous calendar year. So, for example, a head who is age 41 in 1999 reports information about his 1998 income when the PSID surveys him in 1999. This means that the HDEARN99 variable is actually capturing the labor income of the head at a time in which the head was 40 (not 41).

If you wonder what's going on with the bracketed expression "[YEAR - 1967]," the reason we have this is to get SAS to set up an array. Normally arrays use an indexing scheme that begins with 1, for example "Age[1], Age[2], et cetera." For our purposes it is useful to exploit the fact that the PSID began in 1968 to mimic this type of array sequencing. This way we can use the number "1967" as an anchor, and if we subtract any PSID survey year we get an array sequencing that mimics a sequence that uses 1, 2,.... (Note for example that if you plug in 1968, you get "Age[1968-1967]," which is the same thing as saying "Age[1]" since 1968 minus 1967 is equal to one.

You now need to write 10 more lines of similar code to have SAS create labor income variables for ages 41 through 50. Use the existing line of code as your model, and just modify it to change the 40s into 41, 42 et cetera, where appropriate.

Note also, that if you envision wanting to analyze the correlation in work hours between fathers and sons, this is a good place to write code to tell SAS to create age-specific work hours variables. You'd do this using language that is similar to what you used for labor income. Add statements like the following next to your labor income statement (and before the end statement) in each of your 11 lines of code.

```
WK_HOURS40=WORK_HR[YEAR-1967];
```

▪How do you construct a measure of average earnings for each individual? This has to be done in steps. First, we have to sum up an individual's earnings over the period in which he or she is 40 to 50. Second, we instruct SAS to take notice of the number of years for which there was non-missing earnings information

for the individual. Third, we tell SAS to divide the sum by the number of years of non-missing data. You can use a series of statements such as the following to do this.

```
SUM_LABOR40_50=sum(of LABOR40-LABOR50);
N_LABOR40_50=N(of LABOR40-LABOR50);
AVR_LABOR40_50=SUM_LABOR40_50/N_LABOR40_50;
```

Moreover, if you want an average work hours variable to analyze later, you can write similar code for the age-specific work hours variables that you created in the previous step.

▪At this point, it probably makes sense to store the dataset with all these newly created variables as a permanent SAS dataset. Let's call it "many_yrs" since it is a dataset that contains many years worth of labor earnings information.

Step3. Working with a Sons-specific dataset and a Dads-specific dataset

You will now need to use the dataset you created above, to essentially make two copies of it--one that stores information for sons and one that stores information for dads.

A. Making the dataset with sons' information

1. You want to keep only some of the variables in the dataset "many_yrs," and you will want to rename others so that you will remember that they apply to sons' information (and so that SAS will not write over them when we get to the point of merging sons' information with dads'). We also recommend adding a statement to rename the statistical weight variable in order to make it easier to recognize. You can use language like the following to accomplish this:

```
libname sunshine "c:\...";
run;
```

```
data DATA_SON(keep=ID68_SON PN_SON AGE2000_SON indwgt LABY_2001_SON
AVR_LABOR40_50_SON);
  set sunshine.many_yrs;
```

```
ID68_SON=ER30001;
PN_SON=ER30002;
AGE2000_SON=AGE2001-1;
indwgt = ER33637;
LABY_2001_SON = LABY2001;
AVR_LABOR40_50_SON=AVR_LABOR40_50;
```

2. Next you want to create a unique identifier variable for the sons, so you can use it to merge with your FIMS dataset:

```
son_id = ID68_SON*1000 + PN_SON;
```

*If you decide to do this in the same dataset as the one renaming your variables (above) remember to add the name of this variable to the keep statement.

3. Now you are ready to run your SAS program. You probably want to store your results as a permanent SAS dataset.

B. Making the dataset with dads' information

Here you just want to replicate the steps that you took above to create labor income variables and the unique identifier. Just make sure you give the variables names to indicate that they contain dads' information. Code such as the following should work.

```
libname sunshine "c:\...";
run;

data DATA_DAD(keep= DAD_ID AVR_LABOR40_50_DAD);
  set sunshine.many_yrs;

DAD_ID=1000*id68+pn;

AVR_LABOR40_50_DAD=AVR_LABOR40_50;
run;
```

C. Merging the sons, dads and FIMS datasets

At this stage, you want to merge your FIMS dataset containing son identification variables and dad identification variables onto the sons dataset you created above (in Part A). Then you will want to merge the dad dataset created in Part B in. You should merge using your son_id variable as the "by" variable when you do the first merge; and you want to use your dad_id variable as the "by" variable for the second merge. (Remember to make sure your datasets are properly sorted before you do your merge.) Because we provided sample language showing how to do a merge for Example 1, we will not provide additional language here.) You may want to give your merged dataset a name like "son_fims_dad" and to save it as a permanent SAS dataset.

D. Analyzing the data

Now you are ready to estimate some intergenerational elasticities to portray the relation between fathers' and sons' earnings. Recall that we want to run the regression using logged income variables so that we can interpret the coefficient on father's earnings as giving us the elasticity of son's earnings with respect to fathers. This means you'll need to create log versions of your income variables before you run your regressions.

A regression with dads' income averaged over a number of years

Because we are particularly interested in baby boomers, our first regression will include language to instruct SAS to run the regression using information for baby boomers only. Below, we provide sample language to allow you to perform a regression that relates son's income in 2001 to a variable that averages dads' income over a number of years. Why average dads' income over a number of years? Remember that Solon (1992) reveals that a single year measure of earnings is likely to provide a poor measure of an individual's earnings, and that error in measurement creates estimation problems when the independent variable in a regression (father's labor income in our case) is poorly measured. If an independent variable is subject to measurement error, we know that its absolute value will be biased downward. We can correct for this by getting a better measure of father's earnings, and Solon's work tells us that a measure that averages father's income over a number of years should do the trick.

```
data SON_FIMS_DAD_2_A;
  set SON_FIMS_DAD;
  if 36<=AGE2000_SON<=54;
```

```

if LABY_2001_SON >0;
if AVR_LABOR40_50_DAD>0;
LOG_LABY_2001_SON = LOG(LABY_2001_SON);
LOG_AVR_LABOR40_50_DAD = LOG(AVR_LABOR40_50_DAD);

run;

proc reg data=SON_FIMS_DAD_2_A;
model LOG_LABY_2001_SON =LOG_AVR_LABOR40_50_DAD;
weight indwgt;
run;

```

A regression comparing sons and fathers at peak life-cycle points

Next you might note that the previous regression had the attractive feature of using observations on dads' earnings that come from the dads' peak earnings years. One might ask what would happen if we compared this to sons' earnings when sons were in their peak earning years--during the life cycle point shown to be where an individual's earnings profile will normally have reached a point where earnings better represent 'lifetime' earnings. The following language allows you to run this type of regression.

```

data SON_FIMS_DAD_3_A;
set SON_FIMS_DAD;

if AVR_LABOR40_50_SON>0;
if AVR_LABOR40_50_DAD>0;

LOG_AVR_LABOR40_50_SON = LOG(AVR_LABOR40_50_SON);
LOG_AVR_LABOR40_50_DAD = LOG(AVR_LABOR40_50_DAD);

run;

proc reg data=SON_FIMS_DAD_3_A;
model LOG_AVR_LABOR40_50_SON =LOG_AVR_LABOR40_50_DAD;
weight indwgt;
run;

```

Interpreting your results

What conclusions can we draw? If your (and our!) analysis was done correctly, your estimate for the elasticity of earnings for the first regression should be about 0.382. This is close to the 0.4 intergenerational elasticity found by Solon (1992), which is often taken as the standard in intergenerational work. Moreover, it illustrates Solon's point that the estimated elasticity rises as one incorporates earnings measures that contain more than a single year's worth of information. Your estimate from the second regression should be about 0.367. Is this *the* intergenerational elasticity of labor earnings between the Baby Boomers and their dads? As noted earlier since the PSID began in 1968, some of the older Boomers, born in 1946 or 1947, for example, could have already left their parents' family (split off) by 1968 and so would not be included. Are all these sons Baby Boomers? Not quite. In our analysis sample of 763, the number of Baby Boomers is 753 or approximately 99 percent of the sample. How did this happen? This can occur since some of the sons were born in before 1946 (and were older children living at home as of 1968). These sons would not be Baby Boomers. (They would be War Babies – born 1940-1945 but not as old as to be Children of the Depression Era or "CODA's" – or, since they are sons, perhaps, if present, they could be called "SODA's.") So we have observations on ten War Babies for whom we have observed their labor

income in the age range of 40-50 as well as for their the dads age 40-50. None of are cases are, or even could be, Post-Boomers since someone born in 1964 would only be 40 as of 2004.

Thinking about possible extensions

You could go back and create measures of work hours that are comparable to the ones you created for earnings, so that you could run a series of regressions to determine whether sons' and fathers' work hours are correlated. Moreover, if you are interested in exploring reasons that labor income might be correlated across generations, you might want to go back and put work hours in as a control variable in your earnings regressions. We leave these interesting extensions to the user. Now that you are armed with the tools that you need to do intergenerational analysis, you should be able to analyze work hours on your own. (Yeah!! You are now an expert!!!)

Note that as in Solon's analysis we restrict ourselves to the SRC sample of sons and dads, we observe an elasticity of .437. And for the SEO sample alone we observe an elasticity of .215. The result of truncating the SRC sample by income of the dads (to capture what may be operating in the SEO sample) can be illustrated by limiting the dad son pairs to dads whose average earnings were above (or below) the median. In each of these cases the estimated elasticity is about .20.

V. References

- Blanden, Jo and Stephen Machin (2003). "Cross-Generation Correlations in Union Status for Young People in Britain," *British Journal of Industrial Relations*, September 41(3): 391-415.
- Davis, Matthew M. and Achamyelah Gebremariam (2005) "Factors Associated with Five Year Trends in Childhood Weight Status: Panel Study of Income Dynamics, Child Development Supplement Waves I and II, Paper presented at the Institute for Social Research Symposium on the PSID-CDS,
- Gouskova, Elena; Frank Stafford and Ngina Chiteji (2006). "Financial Market Participation and Pension Holdings Over the Life Course," *Wealth Accumulation and Communities of Color in the United States: Current Issues*. (Jessica Gordon-Nembhard and Ngina S. Chiteji, eds), 2006.
- Gustman, Alan; Olivia Mitchell and Thomas Steinmeier (1994). "The Role of Pensions in the Labor Market: A Survey of the Literature," *Industrial and Labor Relations Review*, Vol. 47(3): 417-438.
- Lazear, Edward (1986). "Retirement from the Labor Force," in Ashenfelter and Layard (eds). *Handbook of Labor Economics*, Vol. 1, pp. 305-355, NY: Elsevier Science Publishers.
- Lee, Chul-In and Gary Solon (2006). "Trends in Intergenerational Income Mobility." *NBER Working Paper Series*, No. 12007. Available at <www.nber.org> Accessed on June 24, 2006.
- Mincer, Jacob (1974). *Schooling Experience and Earnings*, NY: National Bureau of Economic Research and Columbia University Press.
- Solon, Gary (1992). "Intergenerational Income Mobility in the United States," *American Economic Review*, June 82(3): 3-44.
- Solon, Gary (1999). "Intergenerational Mobility in the Labor market," in Ashenfelter and Card (eds). *Handbook of Labor Economics*, Volume 3A, pp. 1761-1800. Amsterdam: North Holland.
- Tabb, Rebecca (2004). "Intergenerational Labor Supply." Unpublished Manuscript, Stanford University Economics Department. June 5, 2004. Available at <www-econ.stanford.edu/academics/Honors_Theses/Theses_2004/Tabb.pdf>
- Treiman, Donald and Robert Robinson (1981). "Introduction: Stratification Theory and Research," in Trieman and Robinson (eds). *Research in Social Stratification and Mobility*, Volume 1. Greenwich, CG: JAI Press.

Appendices

Appendix A---FIMS starter program modified to illustrate the creation of a unique identifier variable for sons and a unique identifier variable for dads

```

libname test "c:\"; run;
/*****
Label          : fiml054_gid_BA_2_BAL_wide
Rows           : 32358
Columns        : 10
ASCII File Date : June 30, 2006
*****/
DATA GID_MAP_wide ;
ATTRIB
  ER30001          FORMAT=F4. LABEL="1968 INTERVIEW NUMBER"
  ER30002          FORMAT=F3. LABEL="PERSON NUMBER 68"
  ER30001_P_AF     FORMAT=F4. LABEL="1968 INTERVIEW NUMBER /PARENT /ADOPTIVE FATHER"
  ER30002_P_AF     FORMAT=F3. LABEL="PERSON NUMBER 68 /PARENT /ADOPTIVE FATHER"
  ER30001_P_AM     FORMAT=F4. LABEL="1968 INTERVIEW NUMBER /PARENT /ADOPTIVE MOTHER"
  ER30002_P_AM     FORMAT=F3. LABEL="PERSON NUMBER 68 /PARENT /ADOPTIVE MOTHER"
  ER30001_P_F      FORMAT=F4. LABEL="1968 INTERVIEW NUMBER /PARENT /FATHER"
  ER30002_P_F      FORMAT=F3. LABEL="PERSON NUMBER 68 /PARENT /FATHER"
  ER30001_P_M      FORMAT=F4. LABEL="1968 INTERVIEW NUMBER /PARENT /MOTHER"
  ER30002_P_M      FORMAT=F3. LABEL="PERSON NUMBER 68 /PARENT /MOTHER"
;
INFILE 'C:\fiml054_gid_BA_2_BAL_wide.txt' LRECL = 35 missover ;
INPUT
  ER30001          1 - 4
  ER30002          5 - 7
  ER30001_P_AF     8 - 11
  ER30002_P_AF     12 - 14
  ER30001_P_AM     15 - 18
  ER30002_P_AM     19 - 21
  ER30001_P_F      22 - 25
  ER30002_P_F      26 - 28
  ER30001_P_M      29 - 32
  ER30002_P_M      33 - 35
;
RUN ;

/* the following language allows one to create a variable that can be used to
uniquely identify the individuals in FIMS and another variable to construct a
unique identifier for their dads*/

data unique;
  set GID_MAP_wide;

  length son_68id 8;
  son_68id = ER30001;
  label son_68id = "son 68id";

  length son_pn 8;
  son_pn = ER30002;
  label son_pn = "sons person number";

  length son_id 8;
  son_id = ER30001*1000 + ER30002;
  label son_id = "son unique id variable";

  length dad_id 8;
  if (ER30001_P_F = . and ER30001_P_AF = .) then do; dad_id = .; end;

```

```
    else do;
      if ER30001_P_F ^= . then dad_id = ER30001_P_F*1000 + ER30002_P_F;
      else dad_id = ER30001_P_AF*1000 + ER30002_P_AF;
    end;
  run;

/*the following code creates a permanent SAS dataset containing the two newly
created unique ids */

data test.FIMS_ids (keep = son_68id son_pn son_id dad_id);
  set unique;
  run;
```

Appendix B---modified version of the SAS starter program sent by the Data Center when we selected the variables for sons

```

/* PSID DATA CENTER *****
JOBID          : 57857
DATA_DOMAIN    : PSID
USER_WHERE     : (ER32000 = 1) and (ER33603 = 10 an
FILE_TYPE      : Current Year Heads Individual Data
OUTPUT_DATA_TYPE : ASCII
STATEMENTS     : SAS Statements
CODEBOOK_TYPE  : PDF
N_OF_VARIABLES : 10
N_OF_OBSERVATIONS : 5245
MAX_REC_LENGTH : 38
DATE & TIME    : October 6, 2006 @ 11:06:36
*****/
libname sunshine "c:"; run;
FILENAME J57857 "[path]\J57857.txt" ;

DATA J57857 ;
ATTRIB
    ER30001  FORMAT=F4.    LABEL="1968 INTERVIEW NUMBER"
    ER30002  FORMAT=F3.    LABEL="PERSON NUMBER"                68"
    ER32000  FORMAT=F1.    LABEL="SEX OF INDIVIDUAL"
    ER17013  FORMAT=F3.    LABEL="AGE OF HEAD"
    HDEARN01 FORMAT=F7.    LABEL="LABOR INCOME OF THE HEAD 2000"
    HDTOT01  FORMAT=F6.1   LABEL="TOTAL HOURS OF WORK IN 2000 (HEAD)"
    ER33601  FORMAT=F4.    LABEL="2001 INTERVIEW NUMBER"
    ER33602  FORMAT=F2.    LABEL="SEQUENCE NUMBER"                01"
    ER33603  FORMAT=F2.    LABEL="RELATION TO HEAD"                01"
    ER33637  FORMAT=F6.3   LABEL="INDIVIDUAL WEIGHT NUMBER 1"     01"
;
INFILE J57857 LRECL = 38 ;
INPUT
    ER30001      1 - 4      ER30002      5 - 7      ER32000      8 - 8
    ER17013      9 - 11    HDEARN01     12 - 18    HDTOT01      19 - 24
    ER33601     25 - 28    ER33602     29 - 30    ER33603     31 - 32
    ER33637     33 - 38
;
run ;

data sunshine.son_vars;
set job54844;
son_id = ER30001*1000 + ER30002;
s_labory = HDEARN01;
if 36 < er17013 < 56 then output;
run;

```

Appendix C--modified version the SAS starter program that comes with the dataset that contains the variables for 1973 (the dataset for dads)

```

/* PSID DATA CENTER *****
JOBID          : 57845
DATA_DOMAIN    : PSID
USER_WHERE     : ER30119 = 1 and ER30118 = 1 and V3
FILE_TYPE      : Current Year Heads Individual Data
OUTPUT_DATA_TYPE : ASCII
STATEMENTS     : SAS Statements
CODEBOOK_TYPE  : PDF
N_OF_VARIABLES : 8
N_OF_OBSERVATIONS: 3776
MAX_REC_LENGTH  : 24
DATE & TIME    : October 6, 2006 @ 10:38:02
*****/
libname today "c:"; run;

FILENAME J57845 "c:\J57845.txt" ;

DATA J57845 ;
ATTRIB
  ER30001  FORMAT=F4.   LABEL="1968 INTERVIEW NUMBER"
  ER30002  FORMAT=F3.   LABEL="PERSON NUMBER"                                68"
  V3027    FORMAT=F4.   LABEL="HDS ANN WORK HRS60:10-13"
  V3051    FORMAT=F5.   LABEL="HDS TOT LABOR Y 61:21-25"
  V3096    FORMAT=F1.   LABEL="SEX OF HEAD"                                63:49"
  ER30117  FORMAT=F4.   LABEL="1973 INTERVIEW NUMBER"
  ER30118  FORMAT=F2.   LABEL="SEQUENCE NUMBER"                                73"
  ER30119  FORMAT=F1.   LABEL="RELATIONSHIP TO HEAD"                                73"
;
INFILE J57845 LRECL = 24 ;
INPUT
  ER30001      1 - 4      ER30002      5 - 7      V3027      8 - 11
  V3051      12 - 16      V3096      17 - 17      ER30117     18 - 21
  ER30118     22 - 23      ER30119     24 - 24
;
run ;

/*the code above reads the ASCII data into SAS */
/*the code below creates a new dataset with the new dad id variable and the
renamed work and labor income variables */

data dad_vars (rename = (V3027 = dad_hrs V3095 = dad_laby));
  set job54301;

  length dad_id 8;
  dad_id = ER30001*1000 + ER30002;
run;

/*the following language makes the above dataset a permanent SAS dataset */

data today.dad_vars;
  set dad_vars;
run;

```

Appendix D. Program to facilitate analysis that involves multi-year labor income measures

/*Note that this program instructs SAS to work with the dataset created in Step 2 of Part IV-C where the user constructed year specific measures for age, labor income, and work hours with the assumption here being that that dataset was called "igsweep" */

```
libname myproj 'K:\Elena_frank_ngi_IG_project';

data sweep(rename=(er30001=id68 er30002=pn er33637=indwgt));
set myproj.igsweep3;
run;

data FIMS(rename=(son_68id=id68_son son_pn=pn_son));
set myproj.fims_tut;
run;
#####
*RECODING MISSING AGE;
data sweep;
set sweep;
array AGE(*) AGE1968 -AGE2003;
do i=1 to dim(AGE);
if AGE[i]=998 or AGE[i]=999 then AGE[i]=.;
end;
run;

#####
*INFLATION ADJUSTMENT;

/*CPIU link:          ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt*/

data sweep;
set sweep;
LABY1968=LABY1968* 184/34.8 ;
LABY1969=LABY1969* 184/36.7 ;
LABY1970=LABY1970* 184/38.8 ;
LABY1971=LABY1971* 184/40.5 ;
LABY1972=LABY1972* 184/41.8 ;
LABY1973=LABY1973* 184/44.4 ;
LABY1974=LABY1974* 184/49.3 ;
LABY1975=LABY1975* 184/53.8 ;
LABY1976=LABY1976* 184/56.9 ;
LABY1977=LABY1977* 184/60.6 ;
LABY1978=LABY1978* 184/65.2 ;
LABY1979=LABY1979* 184/72.6 ;
LABY1980=LABY1980* 184/82.4 ;
LABY1981=LABY1981* 184/90.9 ;
LABY1982=LABY1982* 184/96.5 ;
LABY1983=LABY1983* 184/99.6 ;
LABY1984=LABY1984* 184/103.9;
LABY1985=LABY1985* 184/107.6;
LABY1986=LABY1986* 184/109.6;
```

```

LABY1987=LABY1987* 184/113.6;
LABY1988=LABY1988* 184/118.3;
LABY1989=LABY1989* 184/124 ;
LABY1990=LABY1990* 184/130.7;
LABY1991=LABY1991* 184/136.2;
LABY1992=LABY1992* 184/140.3;
LABY1993=LABY1993* 184/144.5;
LABY1994=LABY1994* 184/148.2;
LABY1995=LABY1995* 184/152.4;
LABY1996=LABY1996* 184/156.9;
LABY1997=LABY1997* 184/160.5;
LABY1999=LABY1999* 184/166.6;
LABY2001=LABY2001* 184/177.1;
LABY2003=LABY2003* 184/184 ;

```

```
run;
```

```

#####
#####;
/*Sweeping across the data to get AGE specific information*/

```

```

data SWEEP;
set SWEEP;

```

```

array LABOR_INC(*) LABY1968 -LABY2003;
array AGE(*) AGE1968 -AGE2003;

```

```
do YEAR=1968 to 2003;
```

```

if AGE[YEAR-1967]-1=40 then do; LABOR40=LABOR_INC[YEAR-1967];end;
if AGE[YEAR-1967]-1=41 then do; LABOR41=LABOR_INC[YEAR-1967];end;
if AGE[YEAR-1967]-1=42 then do; LABOR42=LABOR_INC[YEAR-1967];end;
if AGE[YEAR-1967]-1=43 then do; LABOR43=LABOR_INC[YEAR-1967];end;
if AGE[YEAR-1967]-1=44 then do; LABOR44=LABOR_INC[YEAR-1967];end;
if AGE[YEAR-1967]-1=45 then do; LABOR45=LABOR_INC[YEAR-1967];end;
if AGE[YEAR-1967]-1=46 then do; LABOR46=LABOR_INC[YEAR-1967];end;
if AGE[YEAR-1967]-1=47 then do; LABOR47=LABOR_INC[YEAR-1967];end;
if AGE[YEAR-1967]-1=48 then do; LABOR48=LABOR_INC[YEAR-1967];end;
if AGE[YEAR-1967]-1=49 then do; LABOR49=LABOR_INC[YEAR-1967];end;
if AGE[YEAR-1967]-1=50 then do; LABOR50=LABOR_INC[YEAR-1967];end;

```

```
end;
```

```

SUM_LABOR40_50=sum(of LABOR40-LABOR50);
N_LABOR40_50=N(of LABOR40-LABOR50);
AVR_LABOR40_50=SUM_LABOR40_50/N_LABOR40_50;

```

```
run;
```

```

/*CREATE A SONS' DATASET*/
data DATA_SON(keep=ID68_SON PN_SON AGE2000_SON indwgt LABY_2001_SON
AVR_LABOR40_50_SON);
set SWEEP;
/*NEED FOR ALL EXAMPLES*/

```

```
ID68_SON=ID68;
```

```

PN_SON=pn;
AGE2000_SON=AGE2001-1;

/*EXAMPLES 1 AND 2: FIXED YEAR, 2001 for example*/
LABY_2001_SON=LABY2001;

/*EXAMPLE 3: AVERAGE when person was 40 to 50*/
AVR_LABOR40_50_SON=AVR_LABOR40_50;

run;

/*NOW LET'S CREATE A DATASET FOR DADS*/
data DATA_DAD(keep= DAD_ID LABY_1973_DAD   AVR_LABOR40_50_DAD);
set SWEEP;
/*WE WILL NEED FOR ALL EXAMPLES*/
DAD_ID=1000*id68+pn;

/*EXAMPLE 1 : FIXED YEAR, 1973 for example*/
LABY_1973_DAD=LABY1973;

/*EXAMPLES 2 and 3: AVERAGE when person was 40 to 50*/
AVR_LABOR40_50_DAD=AVR_LABOR40_50;

run;
#####

/*MERGE DATA_SON, FIMS AND DATA_DAD DATASETS*/
proc sort data=FIMS;by ID68_SON PN_SON; run;
proc sort data=DATA_SON;by ID68_SON PN_SON; run;

data SON_FIMS;
merge FIMS(in=b) DATA_SON(in=a);
by ID68_SON PN_SON;
if a=1 and b=1;
run;

proc sort data=SON_FIMS;by DAD_ID; run;
proc sort data=DATA_DAD;by DAD_ID; run;

data SON_FIMS_DAD;
merge SON_FIMS(in=a) DATA_DAD(in=b);
by DAD_ID;
if a=1 and b=1;
run;

#####
/*NOW WE ARE REDY TO DO ANALYSIS*/

/*EXAMPLE 1*/

data SON_FIMS_DAD_1;
set SON_FIMS_DAD;
if 36<=AGE2000_SON<=54;
if LABY_2001_SON >0;
if LABY_1973_DAD>0;

```

```

LOG_LABY_2001_SON = LOG(LABY_2001_SON);
LOG_LABY_1973_DAD = LOG(LABY_1973_DAD);

run;

proc reg data=SON_FIMS_DAD_1;
Title "EXAMPLE 1";
model LOG_LABY_2001_SON=LOG_LABY_1973_DAD;
weight indwgt;
run;
*****;

/*EXAMPLE 2*/

data SON_FIMS_DAD_2;
set SON_FIMS_DAD;
if 36<=AGE2000_SON<=54;
if LABY_2001_SON >0;
if AVR_LABOR40_50_DAD>0;
LOG_LABY_2001_SON = LOG(LABY_2001_SON);
LOG_AVR_LABOR40_50_DAD = LOG(AVR_LABOR40_50_DAD);

run;

proc reg data=SON_FIMS_DAD_2;
Title "EXAMPLE 2";
model LOG_LABY_2001_SON =LOG_AVR_LABOR40_50_DAD;
weight indwgt;
run;

*****;

*****;

/*EXAMPLE 3*/

data SON_FIMS_DAD_3;
set SON_FIMS_DAD;

if AVR_LABOR40_50_SON>0;
if AVR_LABOR40_50_DAD>0;

LOG_AVR_LABOR40_50_SON = LOG(AVR_LABOR40_50_SON);
LOG_AVR_LABOR40_50_DAD = LOG(AVR_LABOR40_50_DAD);

run;

proc reg data=SON_FIMS_DAD_3;
Title "EXAMPLE 3";
model LOG_AVR_LABOR40_50_SON =LOG_AVR_LABOR40_50_DAD;
run;

```