Overview

A "balanced panel" consists of repeated observations on a set of measures for a sample of individuals or families. Only individuals or families who are in all periods or years are members of a balanced panel. Adding to these the individuals or families who are in the sample in less than all years creates an 'unbalanced panel'. Here we show how to create and explore a balanced panel of individuals using recently added features of the PSID Data Center.

Public access files from the PSID archive were first delivered 'in bulk' over the Internet in 1995. Since then there has been a growth of functionality in data delivery via the PSID website. Correspondingly, there has been an emerging use of the PSID data sources for instructional as well as research purposes. A key element in delivery functionality was added in the spring of 1996, when the PSID began providing the capacity for users to download not just the full 'in bulk' archive, but to use our on-line system to create data subsets, customized to the researcher's domain of interest as defined by year, level (family, individual), variables and value ranges of variables. More recently we have added subsetting functionality to the Data Center.

In addition, important generated variables, such as wealth, primary income components, and work hours, which have been created as a companion to Public Release files, are now part of the files accessible via the Data Center. These generated variables typically appear toward the end of the list for each year's family file. Until recently these were only stored as supplemental files by variable, under Supplemental Data Files. Now these variables have been added to the regular subsetting functionality within the Data Center by year. (They continue to also reside as supplemental files by variable - a rare case of file redundancy in the PSID.)

Improved Data Center functionality has lead to broader use by the research community resulting in a much increased number and range of publications. It also has attracted graduate student 'classroom' users, beyond the traditional use of PSID resources for Ph.D. dissertation research. Recently there has been a trend toward some use of PSID files in courses by undergraduates. To facilitate that use, we have written 'User Guide Tutorials'. Because of the complexity of the PSID archive we have decided to use a building blocks approach, starting first with a tutorial on creating cross-sectional files (Tutorial 1). Tutorial 1 provides an early section on different tools for navigating the Data Center archive to locate variables of interest for a research project. It also covers a number of other basics.

The current tutorial creates a balanced panel of individual adults drawing information from both the family and individual files. The ideas set out in the tutorial can also serve as a guide for users wishing to create a balanced panel of families (or heads of families). Our Data Center defines the same family as one headed by the same individual in a selected panel of years. Alternative 'same family' definitions (of a more extensive or restricted nature) can be considered by merging files created from the Data Center for different years. For a balanced panel of individuals, 'sameness' in this sense is not a concern, but there are issues of the individual's changing role as a head or wife across the selected years - as will be discussed below.

Our tutorial assumes that the user selects dBase or Excel as the output file format option, although in practice several options are available including SAS, STATA, and SPSS. The user reads the dBase file into Excel or works directly from an Excel file and performs transformations and computes basic descriptive statistics.

I. Using the PSID Data Center

A. The Nature of the Exercise
The exercise involves subsetting the data to a sample of women, some of whom are female heads and some of whom are spouses in families with a male head. For more on the spousal aspect of the PSID, refer back to Tutorial 1. The subset will include variables on labor market income in the 1995 and 1999 survey years which will be used to analyze transitions across different relative locations in the labor income distribution. The years of measurement (1994 and 1998) differ from the survey years (1995 and 1999) because respondents are asked about income in the prior calendar year. Labor market income includes earnings from jobs, overtime work, or professional practice. Labor income from work in unincorporated small businesses or running a farm owned by the family is also measured in the PSID, but for our tutorial we are looking at something which can be thought of as wage and salary labor market income.

The sample is restricted to those women who are under age 58 and over age 23 as of 1995. As part of the analysis the sample will be subdivided into two groups: women with at least a college education and women with less than a college education. We want to apply individual weights and get simple weighted averages. As we will see there are some complications in using weighted data on relative position in the labor income distribution.

There have been some minor changes in the data collection instruments over the years, but the consistency of the series is very high, 1995-1999. The specific questions from Section G (Income) in 1999 can also be viewed.

The conversion of the numerous labor and small business income responses into 'labor income' is an involved process and the basic features of the data processing and editing can be viewed through our website. The codebook for the resulting labor income variables also can be viewed, or one can click on the documentation icon for each variable (separately) when one's Data Cart is displayed (as shown with screenshot 4 below).

In the special labor income codebook mentioned above, scroll down to the labor income of the head and the wife, which are variable numbers 21 for the head of the family (head to be defined below) and 40 for the wife (defined below). The variable names are 'ER16463' and 'ER16465'. Various labor income components are available. The variable labels are 'Labor Income of the Head 1999' and 'Labor Income of the Wife 1999', respectively. A description of the codes for each variable is given below, along with a brief summary of the question text (from Section G in the CATI) on which the data are based. For additional discussion of the income variables one can refer to notes and the readme at https://simba.isr.umich.edu/Zips/ZipSupp.aspx#income94- (which is the URL where the labor income codebook is located). The specific wording of the questions for the underlying components is also available in Section G in the CATI. Since income is such an important element in the PSID, one can also review a recent assessment of income quality.

The labor income transition table we will be generating is based on quintile transitions from 1994 to 1998 and will be in the form of Table 1 below.


<table>
<thead>
<tr>
<th>1994 Quintile</th>
<th>1998 Quintile</th>
</tr>
</thead>
<tbody>
<tr>
<td>(S0) 1 (xxxx.1)</td>
<td>yyyy.1</td>
</tr>
<tr>
<td>(S0) 2 (xxxx.2)</td>
<td>yyyy.2</td>
</tr>
<tr>
<td>(S0) 3 (xxxx.3)</td>
<td>yyyy.3</td>
</tr>
<tr>
<td>(S0) 4 (xxxx.4)</td>
<td>yyyy.4</td>
</tr>
<tr>
<td>(S0) 5 (xxxx.5)</td>
<td>yyyy.5</td>
</tr>
</tbody>
</table>

The labels on the left side, '1994 Quintile,' are for the 6 divisions (5 quintiles and a separate row for those with $0 income) of labor income for women age 24-57 as of 1995 (who by 1999 will be mostly age 28-61, depending on the exact interview dates in the respective years). We use an upper age of 61 as of 1999 since individuals begin leaving the labor force more or less permanently in the early 60's. The lower age of 24 as of 1995 is to allow for the fact that many complete their college education after age 21. The labels across the top for the columns, '1998 Quintile,' are the $0 and 5 Quintile divisions four years later. Quintiles indicate relative position, and we also want to keep track of the actual ($2000) income values to which they correspond. These values are denoted by the xxxx.1,..., for 1994 and by yyyy.1,..., for 1998.
We will be looking at all adult women in our selected age ranges, and some do not work in the labor market in a given year (1994 or 1998). Their labor income will be zero. Suppose 15 percent do not earn labor income in a given year (1994). Then 3/4 of the bottom quintile (xxxx.1.) would have a value of '$0'. This would distort our analysis and therefore we will do a separate analysis for these women. Beyond that, the increasing quintile values will represent the cross-sectional distribution in this particular sample of women who were successfully interviewed in both 1995 and 1999 to create this 'balanced panel'.

Now we have to get to work.

B. Getting Started

Labor market income of the head or wife is described in the codebook referenced above, but a few remarks are in order.

Before starting, a few words about variable code values. What is done if a respondent gives a report of 'I don't know?', or the respondent 'refuses' various elements in the Section G income sequences? These answers initially get codes of '999998' ('don't know', 999998 for a 6-digit variable like labor income) and '999999' (refused or otherwise not ascertained = N.A.). For the income variables we are using here, such missing data values have been assigned or imputed for those observations. So in reviewing code for these variables one should not find such values. A more subtle point is that for labor income of the wife there are three reasons for a labor income value of 0. One is that there is a wife in the family, but she happened to earn no labor income in the prior calendar year. Another reason for a value of 0 is that the family is headed by a male who has no wife present and a third reason is that the family is headed by a female and her labor income is therefore reported as the labor income of the head. In these three situations (wife present but labor income of zero, no wife present, or female head) the code values for labor income of the wife will be '0'. The circumstance of 'no wife' or 'female head' should be better interpreted as 'inapplicable' rather than zero labor income. We will consider more closely in the Section C below the importance of this distinction in analyzing the labor income of adult women, some of who are family heads and some of whom are wives/wives'.

One final point about PSID data codes (really!). Labor income and the other income variables are referred to as numeric variables. This is distinct from a categorical variable. For example, if male (a category) is coded as '1' and female is coded as '2' or if 'yes' is coded as 1 and 'no' is coded as '5' (a PSID tradition of mysterious origin) one would not want to say that 'no' has a five times greater value than yes or that women are (necessarily) twice men. So here the '1' the '2' and the '5' are simply index categories.

C. Using the Data Center

To obtain data we need to go to the PSID Data Center web site and its Set Theoretic Subsetting System. Our first goal will be to create a data cart that contains all of the variables that we need to do our analysis. As you can see in screenshot 1, there are several ways to put variables in your data cart. It is possible to select variables by file-for example, by going to the family file for a given year, such as 1999. It is also possible to select variables by using the PSID's cross-year index (the by index option). This tutorial will showcase these two methods of selecting variables. A third and fourth option exist: One can conduct a search by listing a keyword, such as "age," and instructing the Data Center to list all variables in the PSID related to that concept so that you can choose from the list; or one can select variables" by cart," which allows a user to retrieve a data cart that was created previously so that data can be added to it.

If you are a new PSID user, please note that the PSID requires users to register prior to downloading data in order to impart the Conditions of Use. Through your registration, it also becomes possible to convey the size of the user community to sponsors, allowing the PSID to continue to collect these important data. Registration is very easy, and it is required only once. If you do not already have a PSID login and password, you will want to register at the Data Center before you begin the next few sections of the tutorial. To do such, click on the word "login" that appears under today's date in the far right upper corner of the "Welcome to the Data Center" screen (screenshot 1). Once you have registered, you can return to the Data Center to get going on the tutorial.

Screenshot #1 Main Data Center screen
Before we begin to select the variables however, it is useful to think about what the PSID means by "family files" or "family-level" data. In particular, the term 'Family' may benefit from some discussion. Fundamentally, the PSID interviewer asks the respondent questions that apply to the whole family of one or more persons, including children because a family (married couple or single adult with children) shares a residence and economic resources, by definition. So the respondent (i.e., head) answers family-level questions such as the following about the family's housing situation: Does the family rent or own? If it owns does the family have a mortgage on that home? Such data are considered family-level data.

Because the study goes way back (designed in 1966-1967), it was common then for such household or family economic surveys to ask financial and income questions about the household head or 'breadwinner' who, back then, was commonly assumed to be the male partner in a married or cohabiting couple. Of course if no husband (or significant male other) was around, the household head or breadwinner would be the economically active adult female. This seemingly restrictive structure was carried forward even though the economic role of women has broadened and become far more influential, especially since about 1975, not long after the study began! And through time far more information about the wife or 'wife' in a cohabiting couple was asked. By now the same information is obtained for both adults in a married couple family. But the basic head-wife structure was necessarily maintained in terms of the data consistency. If not, one could not have the consistency over time to complete Table 1 and it turns out this is a key aspect of the relational data structure of the PSID.
Consider a world with only two simple cases for the PSID to handle. Married couples and single adults 'heading' a household - which could be just that single person (or possibly with dependents, such as young children). The precise definition of 'heading' a household is known to the PSID staff but cannot be normally explained to others in a finite lifetime. So, let's just pretend that someone in Ann Arbor has done this for you. If you are really compulsive you can immediately turn to a discussion at the Frequently Asked Questions Web page.

Given what we just said, how do you think one would get a good national sample of adult males as of 1999? Surprisingly simple. Just limit the subset to the families with a male head. Some will have a wife, but those who don't will be the single male family heads. No problem in concept, given the traditional male-head-by-default assumption. But, given what we just said, how do you think one would get a good national sample of adult females, married or otherwise, as of 1999? Still pretty simple in concept, but relational data structures are needed, which means ID links. That is, adult women are either single female heads or the wives in the married/cohabiting couples.

1. Using the "by file" option to add earnings data to your data cart

Let's begin selecting data by focusing on the labor income data that you want to record in the transitions matrix. We are studying the 1995-1999 balanced panel, so you want earnings information from 1995 and 1999. Generally, information about labor market activities is collected for the head of the family and, if present, the wife of the head of the family. There are sections with extensive detail for the head and the wife on these topics. Traditionally, and today, this has been questionnaire Section B (employment of heads who are active in the labor market - primarily working now or actively looking for work - as of the interview date) and Section C (employment of heads who are not currently working in the labor market). Sections D and E are parallel to B and C, but for the wife/wife'. The information obtained as responses to these questions is considered family-level data. We want information about earnings of heads and earnings of wives, since--as previously noted--adult women may be heads or wives of households. The ER16463 AND ER16465 variables, and similar ones for the year 1995, can be found toward the end of the variable list for the family files of the relevant year. (Note that because the PSID is a survey that focuses on income, there actually are a number of income-related variables in addition to ER6980, ER6984, ER16463 and ER16465. For example, in many years such as 1968 through 1993 and again beginning in 1997, there is also a variable called head's "labor income." Through 2001, this measure incorporates information about income received from participating in farm and business activities, in addition to information like wage and salary earnings. This broader measure of labor income would be appropriate if one wanted data for longer periods of time than the head labor income variable is available.) To get your earnings data: Click "by file" and a list of the different PSID files from which you can select data will appear (as shown in screenshot 2). Next, click on the "+" sign next to the phrase "PSID-Family level." This action should cause the list of family-level files to expand. Click on the "+" sign next to "PSID main family data" and a list of all the years in which the PSID collected data should appear. From this list, we will want to go to the year 1999 and then to the year 1995, so we can locate the head and wife earnings variables for our two years of interest. Clicking on the "+" next to 1999 opens up a box that lists several variables, as shown below in screenshot 3. This box lists all the variables available for a given family in the year. If you scroll down toward the end you should see the variables "ER16463" and "ER16465". Highlighting these two variables will instruct the Data Center to select them for you. (FYI: To select more than one variable you need to hold down on the control key as you click the cursor to highlight the variables after your first choice. When selecting non-adjacent variables in the variable list box, you can skip those variables you do not want to select by scrolling to the desired variable and highlighting--all while holding the control key down.)

Screenshot 2 : Selecting data using the "BY FILE" option--different file options available
Screenshot 3: Choosing from the yearly family files
Using the drill-down interface, scroll through the variables contained in various data files. Double click on the variable to view its codebook. Highlight variables you want to include in your extract, holding the ctrl key to choose multiple variables. Then press Add To Cart to add variables to your data cart.
You can select the variables for head and wife earnings in 1995 by repeating some of the steps you took above. Click on the "+" sign next to the year 1995 and a box listing the data available in the 1995 family file will appear. Scroll down towards the end and highlight ER6980 and ER6984.

**Now the most important step of all:** Click the "add to cart" box that appears just above the start of the list of all the years for which PSID data are available. Once you do that, the Data Center officially adds the 4 earnings variables that you have selected to a shopping cart (or "data cart" in PSID parlance) for you. A screen should appear that looks much like screenshot 2 with the exception that the a box containing the words "variables added to your cart" appears as well. Screenshot 4 (below) provides an illustration.

Screenshot 4: Determining what variables have been added to your cart

Next we want to obtain other variables that will help us in our study. For example, we will want a variable identifying the sex of the individuals in our sample so that we
can isolate the women. And, as noted earlier, we will need to know whether a given women was head or wife in the family so that we use the appropriate earnings variable for her labor income. We also will want information about the women's ages, since our transitions table is only going to include people under age 62. And, we'll want to know the woman's employment status (is she working outside the home, or a stay-at-home mom or not employed for some other reason). Then, for background, we will pull information about individuals' education. Finally, we will select two variables that can be used to weight the data to make sure the results that we obtain are nationally representative. Because we will need data for two different calendar years for each variable of interest, the exercise provides an opportunity to showcase the data selection "by index" feature of the PSID Data Center.

To add additional variables to your data cart, move your cursor up to the words "Data Center" in the upper left portion of the screen so that the words "variable selection" materialize. Then move your cursor down to this phrase ("variable selection") and click on the "by index" option. This should take you to a new screen that has a series of boxes with calendar years in the right portion of the screen. (See screenshot #5.)

Screenshot 5: Selecting data by using the cross-year index

We will start by gathering individual-level data from the index (i.e., from the "Individual Data Index" option). These are variables that pertain to an individual in the PSID without regard to the individual's relationship to the head of the household. A characteristic like an individual's sex, for example, is recorded in the PSID's individual-level data files. The age of an individual is a similar type of variable. As noted earlier, most of the data in the PSID is collected at the level of the family. However, there are occasions in which data are recorded at the level of the individual, meaning that info about a specific attribute is kept for each individual in the PSID-not just heads and wives--so the information cannot be reasonably stored in a family-file using the head/wife format that was described earlier. To illustrate this principle more clearly, consider a variable like "birth year." The PSID actually contains the birth year of every individual who appears in the survey. This means it has the
birth years of the head of a household in any given year, the wife, and any children. Because an individual's birth year does not change over time, and because birth year information is collected for so many individuals in the PSID, it makes little sense to store this information in the family files. (That would mean reproducing it for each individual for every wave of the PSID.)

(a) sex

Click on the "+" sign next to the phrase "Individual Data Index" so that the list of available individual data expands. The expansion reveals a number of types of data. Let's start by scrolling down to the S's so that we can select the variable that reveals the sex of the individuals in our dataset. You can click on the box with the downward arrow signs to the far right of the word "sex" to see the documentation for the variable in order to confirm that it indicates whether an individual is male or female. To select the variable, use your cursor to check the next box to the right on this same line. That is a box that corresponds to the year 1968. For most variables in the Individual Data Index one can select from a range of years (if one wants the info for 1995 and for 1999 for example). Sex is an exception because it presumably does not change over time.

(b) age, sequence number, relationship to head, education and statistical weights

Next scroll back toward the top of the page to the word "age." We want our individual's ages for 1995 and 1999, so click on the boxes for 1995 and 1999 in the age row. Note how, in addition to listing the years at the top of the index, this page also allows you to see the year corresponding to a given box by positioning the cursor on the bottom right portion of the box. This second feature comes in handy if one is dealing with variables further down in the alphabet (where it is not always possible to see the labels row).

Next scroll down to the R's and stop when you reach the row containing the variable labeled "relationship to head." We need to select this variable for 1995 and 1999, so click on the relevant boxes in that row. Do the same for "Sequence number." These two variables will be used to determine which women in our sample are current wives and which are current heads (so that we know which variable to use to get their earnings information).

Now go up to the E's so that you can select two variables that measure the individual's education level in 1995. Since we want to see if the transitions differ by level of education, you will need a variable on education which is gathered from two separate sequences as part of an education update in 1995. Education is not asked each year, since, for adults, it does not change that rapidly or commonly. In 1985 and 1995 there were educational inventories for each family member. For people age 5-49, the respondent is asked if they currently attend school. From this skip pattern, if they currently (1995) attend school, the grade/year of school is asked. The resulting variable is ER33222 M6 HIGHEST GRADE OR YEAR IN NOW 95 [01 = first grade, 02 = second grade, ..., 12 = 12th grade, ..., 14 = two years college or Associate's Degree, ..., 15 = three years of college or more, no degree, 16 = graduated, Bachelor's Degree, 17 = at least one year of postgraduate or more, 93 = ungraded, 94 = preschool, 95 = kindergarten]. For those who are not currently in school, the highest grade or year completed is asked. The resulting variable is ER33227 M10 HIGHEST GRADE OR YEAR COMPLETED 95 [01 = first grade, 02 = second grade, ..., 12 = 12th grade, ..., 14 = two years college or Associate's Degree, ..., 15 = three years of college or more, no degree, 16 = graduated, Bachelor's Degree, 17 = at least one year of postgraduate or more, 93 = ungraded, 94 = preschool, 95 = kindergarten]. Both education variables also may take on the values "98" and "99" These reflect instances in which the respondent answered "don't know" or where the education level was not ascertained (for example, if the respondent refused to answer). They are not actual years of schooling. To construct a years of education variable for heads and wives one can sum the individual level variables, ER33222 and ER33227. This is because the skip pattern routes them to M6 or M10 and they are given a value of '0' or Inap. for the route not followed. To select your education variables: If you click on the "+" sign next to the word education, you see that there are several options. We want two variables that are part of an individual's education history, so click on the "+" sign next to the word "history." Then, note that if you go down to the phrase "grade highest" you can click on a "+" sign to reveal two variables. The first is for those who were in school during the survey year, the second was for those who weren't. Notice that these education variables present another instance in which there is only one box that you can select from (one calendar year).

Finally, and perhaps most important, you will the PSID's statistical weights in order to do your analysis. The PSID sample is representative of the U.S., but because of the long evolution, weights are needed to allow for the fact that certain groups of individuals are over or underrepresented relative to the U.S. population. We will use the 1995 weights which are ER33275 CORE IND WEIGHT 95 in the list of 1995 Individual Data, and we also need ER33546 INDIVIDUAL WEIGHT 99 from the 1999
list in order to limit our sample. To get the weights: The relevant statistical weight variables can be found in the W section of the Individual Data Index: Go to the word "weight" and click on the "+" sign to the left of this word. After doing such, look at the first row. It contains the weights from the years 1997 onward, so you can click on the box for 1999 in this row. To get the 1995 weight you need to go down to the phrase "individual core" and you will see that there are 2 rows of information for this phrase. The first row presents boxes spanning the years 1968 to 1992 from which you can check. To get the weight for 1995 however, one has to go to the second row however. Why are there two rows? This is a case in which the variable may have changed slightly over time. If there is a slight change in a variable from one survey year to the next, this is indicated by providing separate rows of information for the variable, with the years for which the question asked was exactly the same all included in one row, and any years in which there was some slight change recorded in another row.

(c) variables from the cross-year index's "Family Data Index"

While you are probably anxious to get to some analysis, there are a few more variables that you need to select using the cross-year index. Note that if you scroll down to the end of the list of variables in the "Individual Data Index," you see the words "Family Data Index." We need to select some variables from this portion of the cross-year index too, so click on the "+" sign next to the phrase. You will get an expanded list of variables if you do this. Go down to the section for "marital status" and click on the "+" sign next to that phrase so you can see the full list of choices under this option. We want information about the head's marital status, so we need to click on the "+" next to head. Then you see the option of selecting a variable that gives a head's "present marital status," which would be his or her marital status in any given year of the survey. We want marital status as "reported by the respondent." Choose the years 1995 and 1999 here. (FYI: The marital status variables, ER13021 and ER5013 are coded as follows: 1 = married; 2 = never married; 3 = widowed; 4 = divorced; 5 = separated. So they provide an example of the categorical variable data-type discussed in section B)

Now we just need information about the employment status of the head and wife for each year of interest (1995 and 1999). Scroll back up in the "Family Data Index" until you see the word "employment." Clicking on the "+" sign next to that word expands the range of options. We want information about the current employment status (not information about previous employers). To see which variable will meet your needs you can position the cursor over the phrase "employment status" anywhere that it appears, and an extended explanation of the variable will appear. This helps us determine that it's the third option on "employment status" (the one that does not include the word "with") that we are interested in. Click on the "+" sign next to this option and you'll see that you can select information about current employment status for both heads and wives. For heads in 1995 click on the box in the "head" row. For years beginning in 1997 we see that the PSID begins to ask about employment status more than one time. Accordingly, for 1999 you need to click the box in the "head 1st mention" row. To get wives' employment information, look at the options under the word "wife" (in this same employment status section). For 1995 you can check the box in the row that just says "wife," but for 1999 you have to select from the row with the wife's "1st mention." Why do you need the employment status variables? These data will tell us whether the person reports primarily 'keeping house' as of 1999. We want to be looking at labor income transitions of adult women only for those who have a labor force connection or interest. Women who are keeping house and not working may be just discouraged workers, so excluding such cases from our labor income mobility tutorial is not so obvious if one reason for being persistently with no labor income is that the job market is seen as unattractive. For our analysis, we will exclude those women who are 'keeping house;' therefore, we need to extract this indicator of labor market involvement as part of our data subset. (The user may want to explore on his or her own to see if the results look very different when such cases are included.) Since some of the women will be wives and some of the women will be family heads, we need to get an indicator of housekeeping as a primary activity for both wives and (female) heads. When we get to the point where you view your datacart you will see that these are the variables ER13205, B1 EMPLOYMENT STATUS(1ST MENTION) - HD and ER13717, D1 EMPLOYMENT STATUS(1ST MENTION) - WF, and ER5067 B1 EMPLOYMENT STATUS - HD and ER5561 D1A EMPLOYMENT STATUS-WF. If you look at the [online CATI documentation](https://www.bls.gov/cps/onlinecati.html) for these variables, you can see that the codes include '1' for 'WORKING NOW', '6' for 'KEEPING HOUSE', and other categories such as '4' for 'RETIRED' and '7' for 'STUDENT'.

**HUGELY important step:** Now we are done identifying the variables we would like to obtain from the cross-year index. You have checked a lot of boxes by now. To ask the Data Center to put all of the checked variables in your data cart, you have to scroll up to the top of the page to the box containing the word "add to cart." (If you don't do that, all this work will have been for nothing, or practice!)

### 3. Checking your data cart

At this point it makes sense to check the list of variables contained in your data cart to make sure that you really have selected all the variables that you wanted. You should now be facing a screen that has a box with the phrase "variables added to the cart" as an option for you to click on. Click it to see what your data cart contains. You can then select "expand the list" to call up the full list of variables. The chart below shows the full list of variables that you should have (although it does not organize them by year as the Data Center screen will).
Table 2. List of variables needed for the tutorial exercise

<table>
<thead>
<tr>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>ER16463</td>
</tr>
<tr>
<td>ER16465</td>
</tr>
<tr>
<td>ER6980</td>
</tr>
<tr>
<td>ER6984</td>
</tr>
<tr>
<td>ER32000</td>
</tr>
<tr>
<td>ER33502</td>
</tr>
<tr>
<td>ER33503</td>
</tr>
<tr>
<td>ER33504</td>
</tr>
<tr>
<td>ER33202</td>
</tr>
<tr>
<td>ER33203</td>
</tr>
<tr>
<td>ER33222</td>
</tr>
<tr>
<td>ER33227</td>
</tr>
<tr>
<td>ER33275</td>
</tr>
<tr>
<td>ER33546</td>
</tr>
<tr>
<td>ER13021</td>
</tr>
<tr>
<td>ER13205</td>
</tr>
<tr>
<td>ER13717</td>
</tr>
<tr>
<td>ER5013</td>
</tr>
<tr>
<td>ER5067</td>
</tr>
<tr>
<td>ER5561</td>
</tr>
</tbody>
</table>

If your list looks complete you are ready to hit "check out." This will take you to a screen that allows you to select output options, which govern how your data is sent to you (for example, whether you want an Excel file or ASCII data with SAS statements, et cetera). NOTE that before you can see your choices for output options, a screen appears asking you to login. If you have used the Data Center before you should already have a login name (basically your e-mail address) and a password. If this is the case but you don't remember your password, you can instruct the Data Center to re-send it to you via e-mail. After you login, you should see a screen like screenshot 6 (below).

Screenshot 6: Output options
4. Choosing your output options
At this point you want to tell the Data Center whether you'd like a codebook, and what format to send it in. For the purposes of this tutorial example, for data output options you want excel (or the dbase option). Next note that the output options screen allows you the option of entering subsetting criteria. This is a way to restrict your dataset to specific cases, for example if you are only interested in people in a certain age range, or if you are only interested in people of a certain sex. In this box you want to enter the following command(s):

\((\text{ER32000}=2) \text{ and } (\text{ER33502}<21) \text{ and } (\text{ER33504}<62) \text{ and } (\text{ER33504}>27) \text{ and } (\text{ER5067 ne 6}) \text{ and } (\text{ER5561 ne 6}) \text{ and } (\text{ER33203 in (10,20,22)}) \text{ and } (\text{ER33503 in (10,20,22)}) \text{ and } (\text{ER33275}>0) \text{ and } (\text{ER33546}>0) \text{ and } (\text{ER33222}<93) \text{ and } (\text{ER33227}<93)\)

Why do you do this? Having some key info about how some of your variables are coded will help you understand why. First, the SEX OF INDIVIDUAL or ER32000 variable is coded as 1 = male, 2 = female. We want a sample of females, so we can restrict ER32000 to the value of 2. The RELATION TO HEAD VARIABLES (ER33503 AND ER33203) are coded as follows: code value 10 = head, code value 20 = wife, and code value 22 = 'wife' for cohabiting couples. There are other code values for other family members (dependent children in the household, for example). Since we want individuals who are either heads or wives, we need to tell the Data Center to restrict our output to cases taking on the codes relevant to those states only. (FYI: the 22 captures the special case of a partner if the head is living permanently with someone but they are not legally married.) The additional restriction on the ER33502 ensures that only female cases where the individual resides with the family are included in your dataset (ER33502<21). The format on these subsetting statements is important and there is on-line help if needed. A typo in this box will return the message 'Internal Server Error' when you go to create your analysis file.

Next, recalling that Table 1 is for people age 61 and under as of 1999 you will need the variable ER 33504, AGE OF INDIVIDUAL 1999 to be less than 62, hence the command: "(ER33504<62)" in the subsetting line. Additionally, since we want women age 24 or older as of 1995, we ultimately are imposing a requirement that the 1999 age be greater than 27 (ER33504>27).

Now note that the analysis group needed for Table 1 is cases where the same adult woman was present in the study in both 1995 and 1999 and was in the labor market and was not primarily a housekeeper as of 1995. To make sure the subsample is just for those with some reasonable labor force interest as of 1995, so you want to add (ER5067 ne 6) and (ER5561 ne 6). (Although this selection may be better handled as part of analysis after data exploration, we make the selection at this point to simplify the tutorial. If we wanted, we could also limit the subsample to those with labor force interest as of 1999. However, as we discussed above, some of those who began keeping house between 1995 and 1999 are in fact discouraged workers and we may want to include them in our analysis.) One final note on this subject. We have included all individuals who say they are working even if (for whatever reason) they say they are also retired or unemployed or a student. The user may want to also exclude these cases from their analysis but we will not. To follow the tutorial, it is assumed that you have only made the original market interest restrictions for 1995 and none for 1999.

Finally, our goal is to characterize transitions of a national sample of women, however the data files include some observations which are not a part of the national sample. These cases have an individual weight of '0'. To exclude them we need to make the selections (ER33275>0) and(ER33546>0). In the meanwhile, to calculate years of education, we need to exclude the missing vales,'93(Ungraded)','98(Don't know)', and '99(Not ascertained/Refused)', with respect to education ((ER33222<93) or (ER33275>0)). This data could simplify the tutorial. If we wanted, we could also limit the subsample to those with labor force interest as of 1999. However, as we discussed above, some of those who began keeping house between 1995 and 1999 are in fact discouraged workers and we may want to include them in our analysis.) One final note on this subject. We have included all individuals who say they are working even if (for whatever reason) they say they are also retired or unemployed or a student. The user may want to also exclude these cases from their analysis but we will not. To follow the tutorial, it is assumed that you have only made the original market interest restrictions for 1995 and none for 1999.

Next, recalling that Table 1 is for people age 61 and under as of 1999 you will need the variable ER 33504, AGE OF INDIVIDUAL 1999 to be less than 62, hence the command: "(ER33504<62)" in the subsetting line. Additionally, since we want women age 24 or older as of 1995, we ultimately are imposing a requirement that the 1999 age be greater than 27 (ER33504>27).

Now you are ready to move on to the next box--the one titled "subsetting of individuals?" Here you should select "all individuals." The final steps here are to tell the Data Center whether it is ok to compress the files (meaning you will have to unzip them using a program like Winzip when you receive them). Additionally, you want to tell the Data Center to send the data to you via e-mail (meaning you will receive an e-mail with a link to a site where you can download your dataset and the codebook). Finally, you may want to name your data cart. This comes in handy if you decide to do another study with the PSID data later. When you have several data carts it's nice to be able to distinguish between them. (Maybe you want to enter "tutorial3" in the cart name line?) Should you click on the box making the files publicly available? That depends. If you want other people to be able to retrieve your data cart, or if you want to be able to get to it without having to enter your password, you should check this box.

Now hit "submit" and the Data Center will begin to create your customized dataset for you. A "creating your data" statement should appear below the submit box. Once the datacenter is done you will receive an e-mail providing a link to your dataset. If you selected the dbase option and you are a PC user you should right click (other users may be required to use alternates to right click) on the blue text of dBase Data File (DBF) and select open; the file will be transferred to you. You should do the same for the Variable Labels. Then you need to save the dBase Data File as an Excel file. Excel will work with a dBase file directly, but doesn't save you work. To save your work, save the file as an Excel Workbook rather than a DBF file. Instructions for using Excel come next.

II.
Using Excel (2000) on your Output Subset*
You should have an Excel file with the variable names arrayed across the top row and the variable values running from row 2 to row 1634 (there should be 1,633 observations in this subset from the PSID Data Center. Steps to take.

1. In order to make it easier to identify the variables represented in each column, you may want to change the variable names that are listed in your dataset. In the table depicted in step 3 (below) you can see the labels for each variable along with the formal or official variable names. You may wish to use these labels to rename your variables (or you can simply come up with intuitive names on your own). For more instructions for how to change variable names, and how to import labels for your variables, go to the tutorial section of the PSID web site.

2. Our dataset contains 1633 observations, and we will have occasion to scroll through it. A convenient Excel option that makes it efficient to scroll through the data is "freeze panes." To use this option highlight the far left non-data 'row button' at row 2. If you then click on the 'Window' menu and then 'freeze panes', then the top two rows of labels will remain visible as you scroll down through the data rather than disappearing from sight.

3. Looking at your dataset, you should find 26 variables in the following order, with the variable names listed in the first row of your spreadsheet:

<table>
<thead>
<tr>
<th>Column</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>ER30001</td>
<td>ER30002</td>
<td>ER32000</td>
<td>ER33201</td>
<td>ER33202</td>
<td>ER33203</td>
<td>ER33222</td>
<td>ER33227</td>
</tr>
<tr>
<td>Label</td>
<td>1968 PSID ID number</td>
<td>PSID person number</td>
<td>Sex of Individual</td>
<td>1995 Interview Number</td>
<td>Sequence Number 95</td>
<td>Relation to Head 95</td>
<td>What grade 95</td>
<td>Highest Grade/Year completed 95</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>ER33275</td>
<td>ER5001</td>
<td>ER5013</td>
<td>ER5067</td>
<td>ER5561</td>
<td>ER6980</td>
<td>ER6984</td>
<td>ER33051</td>
<td>ER33052</td>
</tr>
<tr>
<td>Weight 95</td>
<td>Release Number 95</td>
<td>Head Marital Status 95</td>
<td>Employment Status of the Head 95</td>
<td>Employment Status of the Wife 95</td>
<td>Labor Income of the Head in 95</td>
<td>Labor Income of the Wife 95</td>
<td>1999 Interview Number</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R</th>
<th>S</th>
<th>T</th>
<th>U</th>
<th>V</th>
<th>W</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>ER33503</td>
<td>ER33504</td>
<td>ER33546</td>
<td>ER13001</td>
<td>ER13021</td>
<td>ER13205</td>
<td>ER13717</td>
<td>ER16463</td>
<td>ER16465</td>
</tr>
</tbody>
</table>

4. As we said above, in order to do analysis by level of education, we need to create a combined variable by summing ER33222 and ER33227. To do this, right-click on the gray column 'I' box and select 'Insert'. This should create a new blank column 'I' to the right of the column containing ER33227, label this column 'Years of Education'. In cell I2, enter the formula [=SUM(G2:H2)] to sum the two education variables. Then highlight the cells downward to I1634. This can be done quickly by placing the cursor on the lower right hand corner of the cell and highlighting downward by pulling the cursor down and stopping at I1634. This should automatically 'fill down', calculating the cell-specific values of the years of education. If not, and the cells are only highlighted and not filled in, click on 'edit, 'fill' and 'down'.

5. The basis of our analysis is labor income of women so we also need a new variable bringing together the variables on female heads and wives/'wives'. To do this, in cell AB2, enter the command [=IF(F2=10,O2,P2)] and 'fill down' to AB1634 as before. This will return the cell-specific value of the head's labor income if the selected woman is head(F2) or the wife's labor income if the selected woman is wife. Label this column 'LABOR INCOME WOMEN 95'. To exploring the Quintile of Labor Income later, we first need to sort our sample in ascending order by the labor income variable we just created. To do this, go to the 'data' menu and select sort. Choose to sort by 'LABOR INCOME WOMEN 95'(column AB) and make sure ascending is selected. This will sort the sample starting with $0
going up through the respondent with the highest labor income.

6. Weights! PSID weights are aligned to national totals. Many statistical packages take such weights and normalize them (for you) to 1 in statistical work. Here we will do our own normalizing, so the sum of all the weights = 1. This will allow us to incorporate the weights as we classify the members of our sample into quintiles. But first we need to exclude those with labor income of $0 since they will be separate from the quintiles (we will reincorporate them once we begin the analysis). In cell AC2 enter the command [=IF(AB2=0,0,J2)], label 'Weight 95 - excluding 0' and 'fill down'. This will return a zero for those with no labor income and the 1995 individual weight for those with positive labor income. Now in cell AD2 enter the command [=AC2/(SUM($ACS2:$ACS1634))], label 'Renormalizing' and 'fill down'. Notice the $ signs, which are necessary to keep the formula in the denominator from changing when you fill down. This command divides the weight from AC2 by the sum of all the weights in column AC, which normalizes it. Now if you put the command [=AD2] in cell AE2 and the command [ = AD3+AE2] in cell AE3 and then fill down from AE3 to AE1634 you will have the cumulative weight counting up to 1.00 in cell AE1634. The number in column AE is the percentile (percentage with lower labor income) for each respondent. This allows us to create quintiles. In cell AF2 enter the command [=IF(AE2=0,0,IF(AE2<0.2,1,IF(AE2<0.4,2,IF(AE2<0.6,3,IF(AE2<0.8,4,5))))) and fill down. This returns a zero if there was $0 labor income, a one if the percentile is under .20, a two if it is between .2 and .4, . . . , a five if it is between .8 and 1.0. Label this column 'Quintile 95'. So now you have your 1994 quintiles.

7. Now in order to analyze transitions between relative locations in the distribution between 1994-1998 we need to do the same thing for the 1999 data. Because we will be sorting by a different set of responses, this must be done in a new sheet. So right-click on the highlighted tab at the bottom of the page (the one listing your worksheet's name) and select 'Move or Copy.' Make sure to check 'create a copy' and select ok. Now double-click on the new tab and rename it 'Quintiles 99' to distinguish it from the 1995 sheet which you should rename 'Quintiles 95'. Now make the following changes:

a) Enter the command in cellAB2=[IF(S2=10,Z2,AA2)], rename 'Labor Income of Women 99', 'fill down' and sort in ascending order by Labor Income of Women 99

b) Enter the command in cellAC2=[IF(AB2=0,0,J2)], rename 'Weight 99 - excluding 0' and 'fill down'

c) The command in cell AD2 should be [=AC2/(SUM($ACS2:$ACS1634))], the command [=AD2] in cell AE2 and the command [ = AD3+AE2] in cell AE3. AND the command in cellAF2shouldbe [=IF(AE2=0,0,IF(AE2<0.2,1,IF(AE2<0.4,2,IF(AE2<0.6,3,IF(AE2<0.8,4,5)))))(Those are not necessarily required.)

d) Rename column AF 'Quintile 99'

8. We want to compare the 1994 quintile for each respondent to their 1998 quintile. In order to do this, we need to find a way to link the 1995 data to the 1999 data. This is because once we sort by labor income, the respondents are no longer in the same order. The way we will link them is by creating two new sheets. In each one we will re-sort the data by '1995 Interview number' and then we can combine them since they will be in the same order. So insert a new sheet and rename it '95 re-sort'. In cell A1 of the new sheet, enter the command ['=Quintiles 95!'D1] and fill down to A1634 (If this formula returns '#REF!' it is because the label for your 95 quintiles sheet is different from ours so either change the formula or change the label so they match). You also need 'Years of Education' ['=Quintiles 95!'I1], 'WEIGHT 95' ['=Quintiles 95!'J1] (more on this later!), and 'Quintile 95' ['=Quintiles 95!'AF1]. Once you have filled down all of these columns, highlight cell A1 and sort in ascending by '1995 Interview number'. The interview #s in column A should go from 1-10437 and the rest of the data will correspondingly be resorted. Now insert a new column between 'Weight 95' and 'Quintile 95'. Label this column 'Adjusted Weight 95' and in the blank cell D2 enter the command [=C2/(SUM($CS2:$CS1634)) and fill down. This will normalize the weights (including the weights for those with $0 labor income who we excluded before). Insert a second blank sheet and label it '99 re-sort'. In this sheet you only need 3 columns: '1995 Interview #', 'Years of Education', and 'Quintile 99'. To get the values for these columns, just follow the same steps as for the 95 re-sort except in each formula change 'Quintiles 95' to 'Quintiles 99'. Once you have sorted by 1995 Interview #, the first two columns in each sheet should be exactly the same and only the quintiles will be different.

9. Now we can combine this data and get the information to create our table. So create a new sheet and label it '95-99 combo'. In cell A1, enter the formula [=95 re-sort!'A1]. In cell B1 enter [=95-99 re-sort!'B1]. Then enter [=95 re-sort!'D1] in cell C1, enter [=95 re-sort!'E1] in cell D1, and enter [=99 re-sort!'C1] in cell E1. This will provide the labels for the columns and, once you fill down to row 1634, you will have the 1995 interview number, the education level, the adjusted weight, and both quintiles for each woman in our sample. (Note: Notice that the 1995 weights are being used as opposed to the 1999 weights. The reason for this is that our motivation is to follow the women in each category in 1995 and see where they end up in 1999 as opposed to seeing where the women in each category in 1999 came from.) Finally, before moving on, you may want to re-format the cells in column C. Column C contains the values for the adjusted weight variable. If
this column looks like it is full of zeroes, you will want to re-format the cells to instruct excel to display numbers revealing up to 2 or 3 decimal points in this column. If you do this, you will see that you actually have positive non-zero numbers in most of the cells. (To format a column select "format" then "cell" then choose "number" from the category list and enter your desired number in the box asking for the number of decimal places that you wish to be revealed.)

This next part is a little tedious but necessary for our analysis, so stay with us.

10. You need 36 new columns. They should be labeled: (0,0),(0,1),(0,2),(0,3),(0,4),(0,5), . . . . , (5,5). These columns represent the 36 cells in our table identified by their corresponding 1994 and 1998 Quintile. Now in cell F2 enter[=IF(AND($D2=0,$E2=0),$C2,""')] and fill down. This will return the weight for those women who had $0 labor income in both 1994 and 1998 and a blank for all others. Now copy this formula into the other 35 columns changing each one so it corresponds with the correct quintiles (For example, the formula in row 2 of the column labeled(3,5) should be[=IF(AND($D2=3,$E2=5),$C2,""')] and so on.

11. Now, in row 1636, sum each column. This is done in column F with the formula [=SUM(F2:F1634)]. (If all the sums appear to be zero, highlight these cells in row 1636 right-click and select 'Format Cells'. Under the 'number' tab highlight 'Percentage' and specify 2 decimal places. This should allow you to see positive, non-zero values.) Now you have the percentage of all the women in our sample who belong in each cell. The value in cell F1636 should be 3.95%. To confirm this, why not open our 'answer' spreadsheet (Note: Netscape users must right-click on this link and select 'save link as' in order to view the answer spreadsheet).

However, we want to see the transition from 1994 to 1998 so we want to see what percentage of each 1994 quintile ended up in each 1998 group along with where those who started with $0 labor income ended up. In other words, we want each row of our table to sum to 100%. So in cell F1638 enter the command [=F1636/SUM($F1636:$K1636)] which divides the percentage in F1636 by the sum of the percentages for the 6 groups that started with $0 labor income in 1994. Copy this formula in each column of row 1646 up until you reach column K. At this point you will want to change the formula slightly for the next group of six columns. (You want the denominator in your formula to sum over the next six columns that comprise the first quintile in 1995.) Continue this for each group of six columns up until cell AO1638, where the formula should be [=AO1636/SUM($AJ1636:$AO1636)].

Now you can use these percentages from row 1638 to complete the table and see the transitions across the labor income distribution.

12. We also want to do some analysis based on education level. To do this you need to copy the sheet labeled '95-99 combo' twice. One of these will be for the sample of women with less than a college education (<16 years), naming '95-99 < college' and the other is for those with at least a college education (>15 years), naming '95-99 > college'. First we will deal with less than college. In the first new sheet, add two blank columns to the right of 'Adjusted Weight '95'. Label the new column D 'Weight - Education<16' and the new column E 'Normalized Weight'. In D2, enter the formula [=IF(B2<16,C2,0)] and in E2 enter the formula [=D2/(SUM($D$2:$D$1634))] and fill down to row 1634. So now in column E you have a normalized weight for the women with less than 16 years of education and a zero for the rest. This column of weights can be used to find labor income transitions for this sub-sample. Go to cell H2 where the formula should be[=IF(AND($SF2=0,$G2=0),$SC2,"")]. We need to use the normalized weights so we want to change C2 to E2. We will do this for the cells from H2-AQ2 and then fill down. So when cell H2 is highlighted, press Ctrl+H. In the 'find what' box type in C2 and in the 'replace with' box type in E2. When you press the 'replace' button, the C2 in the formula will be replaced by E2 and cell I2 will be highlighted. If a different cell is highlighted, close the replace box, highlight cell I2, re-open the replace box by pressing Ctrl+H, and make sure 'By Rows' is selected. So now press 'Replace' for each cell until you get to AQ2. Now highlight cells H2-AQ2 and fill down to row 1634. The percentages in rows 1636 and 1638 will be recalculated based on these weights and can be used to complete a version of Table 1 for women with less than a college education.

Now do the exact same thing for the second new sheet except change the formula in cell D2 to (B2>15) instead of (B2<16) before you fill down and change the column label to 'Weight - Education>15'. Now you can complete Table 1 for those with at least a college education and do a comparison based on education.