Problem set 1: Empirical Methods for Applied Microeconomics

General instructions. Please work in a group no larger than 3. When you write up your results, please let me know who is in your group. (Only turn in 1 completed homework.). Present your answers in a concise way (typed is highly preferred). Please include relevant Stata output and well-commented do files and ado files for all the exercises (or equivalent in the package of your choice.) Please do NOT include lots of undigested log files.

Put the do files in an appendix and make clear reference to the regression output and/or figures.

Problem 1

Measurement error in the March CPS

Download the 2008 March CPS (available at the NBER's page on CPS supplement data (Go to the bottom of the page, March 2008, the box with "A" is the zipped data, that with "P" is the technical documentation for the data, and that with "T" is the data dictionary). The Stata dictionary and do files for reading in the data are at Programs. You can probably adapt these for your needs. There are also SAS and SPSS files.

We are going to investigate item non-response for wage and salary income earnings, Social Security income, and public assistance income in the CPS, using the insights from Bollinger and Hirsch's work. Create a data set with all persons 15 or older (the CPS only asks people 15 and older about receipt of income). We are going to explore the allocation/imputations for several variables for the person records.

(i) What share of those 15 and older have an allocated sex or age?

(ii) The wage and salary values like the other income values are collected by a question that says "Did you/name have any income from source "X" last year?" If the respondent says that person received income from that source, they are asked an amount. It is a bit more complicated for wage and salary income. So, there are 2 questions for the income sources, a yes/no question that is 1 if the person got that income, 2 if they did not, and 0 for out of universe individuals (children 0–14 not asked the question). We are going to look at allocations for 3 income sources for individuals during calendar year 2007: wage and salary, Social Security, and public assistance income.

What share of those 15 or older had the yes/no variables imputed for wage and salary income, for Social Security income, and for public assistance?

What share of those 15 or older had the amount imputed given that they reported that type of income (whether by true report or imputation)?

(iii) Now we are going to see whether this varies by whether the individual responded to the question or it was answered by a proxy. Each household has 1 respondent. You can identify this person by the their line number. The CPS has a hierarchical record structure, where there is a household line, then a family line, and then individual lines. The cpsmar08.do (or .sas) file appended the household and family information to the person records. So, the respondent is the person for whom the respondent number (h_respnm) is the same as their line number (a_lineno). I checked and this identifies the respondent for most records. Does the share of proxy reports vary by age and sex for the 15 and older population? Does it vary across states?

Does it vary in these ways for the yes/no questions for receipt of wages and salaries, Social Security, and public assistance?

What about for the levels of the income variables for those who had them?

(iv) What is the relationship between item non-response for demographic variables and yes/no income variables? What about for the level of income variables?

(v) Model item-nonresponse as a function of some demographics (e.g., race, Hispanic ethnicity, marital status, gender, education, age), month in sample, region, and whether the individual was the household respondent. First create a 0-1 variable for the imputation of the income variables (levels or yes/no). Then, run some linear regressions with robust standard errors. Do you find the same relationships for wage and salary income generally speaking as in the Bollinger and Hirsch articles? Is it different for Social Security income or public assistance income?

Now, use the March Supplement weight (marsupwt). Do the non-response adjusted weights matter for the relationships you've found? Do the weights vary with item non-response? (Does it matter what kind of weights you use with robust standard errors?)

Now, we will crudely account for the sample design by "clustering" the standard errors on state (gestfips). Does this make a difference? Give some reasons why you might think this wouldn't be sufficient to balance things.

(vi) Run at least one non-linear (e.g., logit or probit) specification and calculate the marginal effects of gender and whether the individual was a household respondent. Do the estimated marginal effects differ? Consider one interaction (say gender and respondent). Compare the marginal effect from the non-linear specification to that from the linear specification (be careful to account for the main terms). Does the functional form matter?

(vii) Compute inverse propensity score weighted estimate for differences across groups for the natural log of the income variables (as suggested by Bollinger and Hirsch (2006), in their table 2). (I will briefly explain in class what this is.) Compare them to the unadjusted estimates of differences across groups. Are your conclusions different for SS or public assistance than for wage and salary income?

(viii) I have put a comma separated file on the class website with administrative data on total spending on public assistance and the total caseload to compare to the analogous variables here. There are two caseload measures. One is recipients (people) and one is recipiency units (families getting benefits). Construct a measure of underreporting by comparing the state total spending or number of recipient units and recipients to the administrative totals. (Use the number of adults reporting PA income as a proxy for the number of recipiency units. Use the number of individuals in subfamilies with adults getting public assistance as a proxy for the number of recipients.) Does underreporting vary by state for receipt? For caseloads?

(ix) Suppose your boss asked what to do about the non-response. What would you do to assess and correct for non-response if you could collect auxiliary data (1/2 page maximum)?