Appendix: Data Extract, NLSY79

The Other Side of the Mountain:
Women's Employment and Earnings over the Family Cycle

Claudia Goldin, Sari Pekkala Kerr, and Claudia Olivetti

The following describes how we compiled the extract of the NLSY79 that we used in our paper, "The Other Side of the Mountain." We are grateful to Jennifer Walsh and Lucy Cheskin for their extraordinary work in putting together this complicated extract.

Work History Information from Week-by-Week Arrays

The NLSY79 provides a rich and informative work history record for each respondent through constructed week-by-week arrays. These arrays cover the entire duration of a respondent's participation in the survey, including years in which they were not interviewed, through a backfilling procedure undertaken by the NLSY. The two week-by-week arrays central to our analysis are the status array and the hours array. The status array reports the job number of the primary job worked each week, or another labor force status if applicable. These other statuses include "no information reported to account for week," "not working (unemployment vs. out of the labor force cannot be determined)", "associated with an employer but the periods not working for the employer are missing," "unemployed," "out of the labor force," and "active military service."[1] In addition to the status array, the hours array provides the actual hours worked each week across all jobs and facilitates our calculation of the total hours the respondent worked each year. As described below, this measure also assists with the interpolation of self-reported income. Furthermore, the hours array allows us to impose work history requirements on our sample based on actual hours and weeks worked. Such work history requirements are further discussed in the following section.

Interpolation of Self-Reported Income

Of particular interest to this analysis are the annual incomes reported by respondents. We consider separately the income from wages/salary alone and the total income including both wage/salary and any own business/farm income.[2] Notably, these reported values are backward looking, each referring to income earned in the calendar year prior to the interview. The fact that these measures are self-reported as an annual sum lends this measure a sense of completeness for the years in which the income is in fact reported. No further calculation is required to annualize income or sum income across jobs. However, given that this measure is only collected when the respondent is interviewed (unlike variables from the week-by-week arrays that are backfilled for any missed interview years), with any skipped interviews we lack a complete income history using the raw variables alone. This challenge is amplified after 1993 when the survey switches from an annual to a biennial administration. In order to recover income in non-survey years and years in which a respondent is not interviewed but works positive hours according to the week-

[1] See NLSY79 Appendix 18: Work History Data for an accounting of labor forces statuses and job numbering.
[2] When a respondent has both wage/salary and business/farm income, their total income is the sum of the two. If a respondent has only wage/salary income or only business/farm income, their total income is equivalent to their wage/salary or farm/business income respectively.

by-week array, we turn to a simple interpolation of missing values. This process unfolds over multiple steps:

1. Income is "unlagged" such that the reported income now refers to the current year, rather than the previous calendar year.
2. Income is set to zero if no hours are worked that year (according to the week-by-week array)
3. Income is set to missing if it is reported as a zero, but more than 300 hours are worked that year
4. Since respondents report the prior year's income, all respondents will have missing incomes in 2018, as their last reported income in 2018 pertains to 2017. We then assign respondents their 2017 income (reported in 2018) if their 2018 hours worked are within 30% of their 2017 hours in either direction.
5. Finally, we fill in the missing values through a simple linear/straight line interpolation, so long as the non-missing values that bookend the single missing income or series of missing incomes are positive and non-zero. Specifically, if the lead or lag value that would be used for interpolation is zero, the intervening missing incomes remain missing.

Although this method leaves a certain number of missing incomes that cannot be interpolated in this simple manner, it does allow us to recover a significant share of the missing incomes, particularly in the non-interview years after 1993. While the above steps describe the process for interpolating total income, the process for interpolating salary/wage income alone is nearly identical. The only difference in the processes appears in step 3. For salary/wage income, incomes reported as zero are set to missing if more than 300 hours are worked that year *and* the total income is also zero. This additional requirement is imposed as the positive hours worked could be associated with business/farm income, thus making it possible for a respondent to truly earn zero salary while working positive weeks (earning positive total income).

Sample Inclusion

Once we have interpolated the income variables via the aforementioned process, we then impose certain work history and income restrictions on the analysis sample. First, we only consider the respondents as working if they earn at least half of the equivalent that a full-time, full-year worker would make at the federal minimum wage applicable to that year. We then begin to follow a respondent's work history when they have worked positive hours for at least 26 weeks per year and 20 hours or more on average across the weeks with positive hours for two consecutive years. We continue to follow such respondents if they meet the following requirements for at least 20% of the time between their first eligible year (as defined above) and 2018:

1. At least 26 weeks worked with positive hours per year.
2. At least 20 hours worked per week on average, conditional on working positive hours.
3. Total income greater than what an individual would earn working for half of the contemporaneous federal minimum wage for 1,400 hours in the year in question.

Age of Youngest Child

We rely primarily on the dates of birth of all biological children of the respondent to measure the age of the youngest child in each year. Given our ability to follow families over multiple decades in the NLSY79, it is important to note that the youngest child in a given year is not necessarily the same child as the

"youngest" child in the previous or following year. Unlike in cross-sectional data, this variable does not measure the age of the last child ever born to the respondent by 2018, but rather the age of the biological child who is the youngest alive in the given year. As such, the "age of youngest child" does not necessarily increase linearly across the survey years. For example, when the first child is born, the age of the youngest child will be 0, 1, etc. However, when the second child is born, the age of the youngest child returns to 0, and so on. We also account for the rare instances in which a child passes away, particularly if that child was formerly the youngest of the respondent's children. The age of the youngest child appears as a categorical interaction with gender in the main estimation, where those men and women with no biological children become the comparison group.

College Graduation and Advanced Degrees

The NLSY79 provides multiple measures of educational attainment. The first is the highest grade completed. This measure is collected beginning in 1979 and we rely specifically on the revised version constructed by the NLSY79 which resolves instances where the respondent's highest grade completed appears to regress in the self-reported version of the variable.[3] The second measure is the highest degree completed since the date of the last interview. This question, however, does not appear on the NLSY79 questionnaire until 1988. While the highest degree completed is our preferred measure of educational attainment, we use the highest grade completed to complete the educational record in instances where a degree appears to be "skipped" and to fill in degrees earned before 1988, as described below.

Furthermore, since the NLSY79-provided highest degree completed variable only reports a value in interview years (and only if a new degree has been earned since the last interview), this variable alone does not provide a running account of educational attainment in each year between 1979 and 2018. Instead, we construct a running highest degree completed variable that captures the highest degree earned by the respondent in each year, leveraging the information contained in both the highest degree and the highest grade completed variables. Specifically, we fill forward and backward the highest degree completed so that the running variable reflects the appropriate degree from the year of completion of that degree until the following degree is earned (if any following degree exists). To ensure that college graduation is properly identified, to address certain irregularities in the degree reporting, and to fill in the record prior to 1988, we make the following adjustments:

- If the respondent reports a degree higher than a BA/BS in 1988 (i.e. the first year when they were asked about their degree status), we assign a BA to the first year when the highest grade completed is equal to at least four years of college, if possible.
- If the respondent reports a High School/Associate's degree and later reports a Master's Degree or higher, but not any college degrees in between, we assume that they have completed an undergraduate degree. If possible, we assign a BA to the first year when the highest grade completed is equal to four years or college or greater.
- In a few instances, the respondent may report that their highest degree completed as of 1988 is HS/AA with a graduation year some time prior to 1988, while their fourth year of college, according to the highest grade completed variable, falls sometime between the year of HS graduation and 1988. In such cases where the highest grade completed information is also relatively complete, increases relatively linearly, and a professional degree is earned sometime

[3] See the NLSY79 Education Topical Guide for more information.

after 1988, we assign a college degree to the first year in which they completed four years of college.

- We also made some other rare adjustments to clean out inconsistent degree sequences (i.e. if the highest degree appeared to be downgrading over time) and to deal with respondents that had long interview gaps but later returned with an advanced degree. Stata programs detailing these minor changes are available from the authors upon request.

Upon constructing this complete year-by-year accounting of the highest degree completed, we then determined the year in which each respondent graduated from college with an undergraduate degree (or in some cases a higher degree if the timing of the BA could not be assigned or the BA was earned before the respondent's entry into the analysis sample).

The advanced degree dummy variable was constructed using the running highest degree variable. It turns into value one the first year when the respondent was observed with a master's degree (or higher), and remains zero otherwise.

Current Marital Status

Current marital status is determined according to the start and end dates provided for the respondent's first four marriages, if applicable. The respondent is considered "currently married" from the year the marriage begins until the end of that marriage. They are no longer considered married in the year the marriage ends, unless a new marriage begins in the same year (or if the marriage began and ended in the same year). Given that this measure defines the respondent's marital status through start and end dates alone, it does not account for cases where the respondent has separated from their spouse if they have not yet ended their marriage.