**CODEBOOK FOR NATIONAL LONGITUDINAL SURVEY OF YOUTH TRANSLATED DATA**

**Overview**

This document describes how we've translated data from the National Longitudinal Survey of Youth 1979 and 1997 cohorts into a common key, as well as how we've imputed missing values.

The "key" listed under each question describes the values that we've chosen to use in our final data. (In other words, it may not exactly match the keys used in either the original 1979 or 1997 cohort data.)

The "exception handling" sections under each question describe how we've handled problems or exception codes specific to that variable. Some general exception codes also exist in the source data:

-1 REFUSAL

-2 DON'T KNOW

-3 INVALID SKIP

-4 VALID SKIP

-5 NON-INTERVIEW

**Race**

**Explanation:** This represents the respondent's race, as reported in their initial interview.  
 **Questions:** In the 1979 cohort, SAMPLE\_RACE. In the 1997 cohort, KEY!RACE\_ETHNICITY.

**Key:**

1 Black

2 Hispanic

3 Non-Black / Non-Hispanic

**Exception handling:** In the 1997 cohort, a separate category existed for "Mixed Race (Non-Hispanic)". This included only 83 respondents and is translated in our data into the "Non-Black / Non-Hispanic" category.

**Imputation of missing values:** Not applicable. This data exists for every member of each cohort.

**Gender**

**Explanation:** This represents the respondent's gender, as reported in their initial interview.

**Questions:** In the 1997 cohort, SAMPLE\_SEX. In the 1997 cohort, KEY!SEX.

**Key:**

1 Male

2 Female

**Imputation of missing values:** Not applicable. This data exists for every member of each cohort.

**Age**

**Explanation:** This is a discrete variable representing the respondent's age as of the interview date.

**Questions:** In the 1979 cohort, AGEATINT. In the 1997 cohort, CV\_AGE\_INT\_DATE.

**Imputation of missing values:** Not applicable. This was a constructed variable in the original data set and so exists for every recorded instance in each cohort.

**Highest Grade Completed**

**Explanation:** This is a discrete variable representing years of schooling, ranging from 0 (no education) to 20.

**Questions:** In the 1979 cohort, HGCREV79 to HGCREV16. In the 1997 cohort, YSCH-3112.

**Key:** In our SQL database, we've used the following codes:

0 NONE

1 1ST GRADE

2 2ND GRADE

3 3RD GRADE

4 4TH GRADE

5 5TH GRADE

6 6TH GRADE

7 7TH GRADE

8 8TH GRADE

9 9TH GRADE

10 10TH GRADE

11 11TH GRADE

12 12TH GRADE

13 1ST YEAR COLLEGE

14 2ND YEAR COLLEGE

15 3RD YEAR COLLEGE

16 4TH YEAR COLLEGE

17 5TH YEAR COLLEGE

18 6TH YEAR COLLEGE

19 7TH YEAR COLLEGE

20 8TH YEAR COLLEGE OR MORE

In our final data set used for machine learning, we've binned the values as follows:

0 LESS THAN ELEMENTARY SCHOOL (0-4)

5 ELEMENTARY SCHOOL (5-7)

8 MIDDLE SCHOOL (8-11)

12 HIGH SCHOOL (12)

13 SOME COLLEGE (13-15)

16 FOUR-YEAR COLLEGE DEGREE (16)

17 GRADUATE SCHOOL (17-20)

**Exception handling:** Across the 1979 and 1997 cohorts, several exception codes were utilized:

93 PRE-KINDERGARTEN

94 KINDERGARTEN

95 UNGRADED

Our translator converts these to 0.

**Imputation of missing values:** In instances where the highest grade level is not reported, we've carried forward the last previously reported value by the same respondent.

**Hours Worked in Last Calendar Year**

**Explanation:** This is a discrete variable representing the total number of hours the respondent worked in the past calendar year.

**Variables:** In the 1997 cohort, HRSWK-PCY. In the 1997 cohort, CVC\_HOURS\_WK\_YR\_ALL.97 to CVC\_HOURS\_WK\_YR\_ALL.16.

**Exception handling:** To deal with extreme outliers, we've capped the number of hours worked at the mean plus three standard deviations of all reported instances in the data set.

**Imputation of missing values:** In instances where hours worked in the last calendar year are not reported, we've imputed the median value of hours worked for all reported instances.

**Weeks Worked in Last Calendar Year**

**Explanation:** This is a discrete variable representing the total number of weeks in which the respondent worked at any job in the past calendar year.

**Variables:** In the 1979 cohort, WKSWK-PCY. In the 1997 cohort, CVC\_WKSWK\_YR\_ALL.97 to CVC\_WKSWK\_YR\_ALL.16.

**Imputation of missing values:** In instances where weeks worked in the last calendar year are not reported, we've imputed the median value of weeks worked for all reported instances.

**Limited in Kind of Work Because of Health**

**Explanation:** This is a binary variable representing whether the respondent was limited in the kind of work they could perform in the past calendar year due to health conditions.

**Variables:** In the 1979 cohort, Q11-4. In the 1997 cohort, YHEA-1005.

**Key:**

1 YES

0 NO

**Exceptions and imputation of missing values:** This variable included a considerable number of missing values, as for the 1997 cohort this question was asked only from 2007 on.

Because the vast majority (~95%) of respondents did not report a health limitation, we typically imputed 0 for missing values. However, in instances where a respondent's last available response for this question was 1, we've imputed 1 for subsequent missing values.

**Limited in Amount of Work Because of Health**

**Explanation:** This is a binary variable representing whether the respondent was limited in the amount of work they could perform in the past calendar year due to health conditions.

**Variables:** In the 1979 cohort, Q11-5. In the 1997 cohort, YHEA-1006.

**Key:**

1 YES

0 NO

**Exceptions and imputation of missing values:** This variable included a considerable number of missing values, as for the 1997 cohort this question was asked only from 2007 on.

Because the vast majority (~95%) of respondents did not report a health limitation, we typically imputed 0 for missing values. However, in instances where a respondent's last available response for this question was 1, we've imputed 1 for subsequent missing values.

**Household Size**

**Explanation:** This is a discrete variable representing the respondent's household size.

**Variables:** In the 1979 cohort, FAMSIZE. In the 1997 cohort, CV\_HH\_SIZE.

**Data limitations:** This represents a subtly different question in each cohort. In 1979, it represents family size; in 1997, it represents household size.

**Imputation of missing values:** In instances where household size is not reported, we've carried forward the last previously reported value by the same respondent.

**Region**

**Explanation:** This is a categorical variable representing the region of the respondent's household.

**Variables:** In the 1979 cohort, REGION. In the 1997 cohort, CV\_CENSUS\_REGION.

**Key:**

1 NORTHEAST

2 NORTH CENTRAL

3 SOUTH

4 WEST

**Imputation of missing values:** Not applicable. Region was reported for every instance for which we could determine whether an income shock had occurred.

**Urban-Rural**

**Explanation:** This is a binary variable representing whether the respondent lives in an urban or rural area.

**Variables:** In the 1979 cohort, URBAN-RURAL. In the 1997 cohort, CV\_URBAN-RURAL.

**Key:**

0 Rural

1 Urban

**Exception handling:** In the 1997 data, a "2" value is included for "UNKNOWN." Our translator converts those responses to the broader "-2" exception code (i.e., the standard "DON'T KNOW" exception).

**Imputation of missing values:** In instances where urban or rural status is not reported, we've carried forward the last previously reported value by the same respondent.

**Marital Status**

**Explanation:** This is a categorical variable representing the respondent's marital status as of the interview date.

**Variables:** In the 1979 cohort, MARSTAT-KEY. In the 1997 cohort, CV\_MARSTAT\_COLLAPSED.

**Key:**

0 NEVER MARRIED

1 MARRIED

2 SEPARATED

3 DIVORCED

4 WIDOWED

**Issues:** In the 1979 cohort, the code "6" indicated widowed. That has been translated to "4" in our data to match the 1997 key.

In the 1979 cohort, a "5" code indicated "REMARRIED", but the 1997 data has no comparable code. In our data, these responses have been translated to the broader "1" key for "MARRIED".

**Exception handling:** In instances where marital status is not reported, we've carried forward the last previously reported value by the same respondent. In instances where no prior information on a respondent's marital status exists, we've imputed that they were unmarried.

**Current Pregnancy**

**Explanation:** This is a binary variable indicating whether the respondent is currently pregnant.

**Variables:** In the 1979 cohort, FFER-13 through 1990 and Q11-5C afterward. In the 1997 cohort, YSAQ-315 through 2001 and YSAQ2-315 afterward.

**Key:**

1 Yes

0 No

**Issues:** This data exists only for female respondents.

**Imputation of missing values:** In instances where pregnancy status is not reported, we've imputed what is by far the most common value in our data set (i.e., 0, or not pregnant). We've imputed the same value for male respondents.

**Number of Children Ever Born**

**Explanation:** This is a discrete variable indicating the number of biological children ever born to the respondent.

**Variables:** In the 1979 cohort, NUMKID80 through NUMKID16. In the 1997 cohort, YSAQ-324 through 2001 and YSAQ2-324 afterward.

**Issues:** There's a subtle difference between the 1979 question, which asks about number of children ever born, and the 1997 question, which asks specifically about live births.

**Imputation of missing values:** In instances where number of children is not reported, we've carried forward the last previously reported value by the same respondent.

**Total Income from Wages and Salaries**

**Explanation:** This is a discrete variable representing the respondent's total wage and salary income from the prior year.

**Variables:** In the 1979 cohort, Q13-5 through 1981, then Q13-5\_TRUNC\_REVISED through 2000, then Q13-5\_TRUNC afterward. In the 1997 cohort, YINC-1700.

**Issues:** Very high salary values are topcoded in the NLSY data to the mean of the incomes reported by the top 2% of respondents.

To enable meaningful comparisons of dollar values over time, our translator converts all figures to 2018 dollars.

**Industry:**

**Explanation:** This is a categorical variable representing the respondent's industry, using 1990 Census codes.

**Key:** Values are translated according to the crosswalk in the industry\_crosswalk.csv file.

**Variables:** In the 1979 cohort, INDALL-EMP.01. In the 1997 cohort, YEMP\_INDCODE-2002.01.

**Issues:** The sheer number of industry codes included in our data set is a challenge for machine learning. To deal with this, in our final data set we've "binned" industry data into the same bins utilized in the 1990 Census data, as outlined in industry\_crosswalk.csv (e.g., all codes from 10-39 are translated to 10, representing " AGRICULTURE, FORESTRY, AND FISHERIES").

**Occupation:**

**Explanation:** This is a categorical variable representing the respondent's occupation, using 2010 Census codes.

**Key:** Values are translated according to the crosswalk in the ocupation\_crosswalk.csv file.

**Variables:** In the 1979 cohort, OCCALL-EMP.01. In the 1997 cohort, YEMP\_OCCCODE-2002.01.

**Issues:** The sheer number of occupation codes included in our data set is a challenge for machine learning. To deal with this, in our final data set we've "binned" occupation data into the same bins utilized in the 2010 Census data, as outlined in occupation\_crosswalk.csv (e.g., all codes from 10-499 are translated to 10, representing " MANAGEMENT, BUSINESS, SCIENCE, AND ARTS").

**Inflation:**

**Explanation:** This is a discrete variable, rounded to the nearest percentage point, representing inflation in each year.

**Issues:** Because inflation measurements typically are not available until after a year is completed, we've used the *prior* year's inflation information in our predictions of an individual's income shock risk.

**GDP Growth:**

**Explanation:** This is a discrete variable, rounded to the nearest percentage point, representing U.S. growth in gross domestic product in each year.

**Issues:** Because GDP growth measurements typically are not available until after a year is completed, we've used the *prior* year's GDP information in our predictions of an individual's income shock risk.

**Unemployment Rate:**

**Explanation:** This is a discrete variable, rounded to the nearest percentage point, representing U.S. unemployment rate in each year.

**Issues:** Because unemployment rate measurements typically are not available until after a year is completed, we've used the *prior* year's unemployment rate information in our predictions of an individual's income shock risk.

**Regional Unemployment Rate:**

**Explanation:** This is a discrete variable, rounded to the nearest percentage point, representing the unemployment rate in the respondent's reported census region in each year.

**Issues:** Because unemployment rate measurements typically are not available until after a year is completed, we've used the *prior* year's unemployment rate information in our predictions of an individual's income shock risk.

**Imputation of missing values:** We were not able to identify regional unemployment data in the early years included in our data set. In those years, we've set each region's unemployment rate equal to the national unemployment rate in that year.

**Prior Income:**

**Explanation:** This is a constructed variable representing the respondent's inflation-adjusted income two years *prior* to the current year.

**Imputation of missing values:** In instances where prior income is not available, we've imputed that the respondent's prior income was equal to their current income.

**Income Change:**

**Explanation:** This is a constructed variable representing the respondent's percentage increase or decrease in inflation-adjusted income relative to two years previously.

**Issues:** Respondents who reported $0 in income in the prior year had an undefined/infinite percentage increase in income. Others who had very low income in prior years had an extraordinarily high percentage increase (e.g., a respondent with $10 in income in the prior year and $80,000 in income in the current year would have an 800,000% increase).

To deal with these exceptional cases, we've capped income change at 500%, equal to the 99.5th percentile of all values in our data set.

**Imputation of missing values:** In instances where prior income is not available, we've imputed that the income change is 0%, equal to the median income change in our data set.

**Shock:**

**Explanation:** This is a binary constructed variable indicating whether the respondent has suffered an income shock (that is, a 20% of greater reduction in inflation-adjusted income, relative to two years previously).

**Key:**

1 Yes

0 No

-1 Unknown

**Imputation of missing values:** Not relevant. Because this was our primary outcome variable, we've dropped all instances where the income shock outcome could not be determined (e.g., instances where the respondent's income was unknown for either the current or prior year).