Noname manuscript No.
(will be inserted by the editor)

NlsyLinks

An R package for research with the NLSY (National Longitudinal Survey of Youth)

William H Beasley ·

Received: date / Accepted: date

Abstract The text of your abstract. 150 – 250 words.

Keywords key · dictionary · word ·

1 Introduction

1.1 Abstract

1.2 Introduction

- NLSY Structure
- Benefits of Accounting for Kinships
  - BG
  - D'Onofrio-type research
- Terminology

1.2.1 Structural and Topical information

The NlsyLinks package offers two types of datasets: topical and structural. *Topical datasets* contain predictor and outcome variables typically used to test a focused hypotheses. For instance, the NLSY79 Gen2 variables Rqqq.qq and Rqqq.qq are critical when studying the relationship between conduct disorder and menarche (*e.g.*, Rodgers et al, 2015), but are not relevant to many hypotheses outside these fields.

William H Beasley
Howard Live Oak, LLC
E-mail: wibeasley@hotmail.com

In contrast, variables in *structural datasets* are not typically directly stated in the hypotheses, yet are essential to many NLSY-related investigations including:

- familial relationships (*e.g.*, Subjects 301 and 302 are half-brothers; Subjects 301 and 403 are first-cousins),
- subject characteristics (*e.g.*, Subject 301 is a Native American female; Subject 607 died from heart disease in 2005; Subject 802 is part of the military over-sample), and
- subject-survey characteristics (*e.g.*, Subject 301 was 15 years old for the 1981 Survey; Subject 301 did not respond to the 1990 survey; Subject 702 completed the NLSY-C survey in 1996 and the NLSY-YA survey in 1998).

The NlsyLinks includes small topical datasets which allows the vignettes and examples to be reproducible and more realistic. The structural datasets are intended to be the authoritative representations, and are the product of two NIH grants (for a complete history of the familial relationships, see Rodgers et al., 2016).

1.2.2 Terminology

The package pertains to multiple generations of the 'Nlsy79' and multiple generations of the 'Nlsy97'. Because the NlsyLinks package structures information within and between generations of the NLSY simultaneously, it requires slightly unconventional NLSY terminology to reduce ambiguity.

The 'Nlsy79 sample' refers to both the original 12,686 subjects interviewed in 1979, and their 11,500+ children (termed 'Nlsy79 Gen1' and 'Nlsy79 Gen2', respectively). Data for the 'Nlsy79 Gen1' comes from the original NLSY79 study, while data for the 'Nlsy Gen2' comes from both the NLSY-C study and the NLSY-YA study ('C' stands for children, and 'YA' stands for young adult). More specifically, the Gen2 subjects are the biological offspring of the Gen1 mothers; they initially completed the NLSY-C survey until roughly age 14, and then completed the NLSY-YA survey (the oldest 'Young Adult' respondent was QQ in the 2016 survey). Although the NLSY does not interview 'Nlsy79 Gen0' (the parents of Gen1) or 'Nlsy79 Gen3' (the children of Gen2), it does contain direct and indirect information about them.

The terminology for the 'Nlsy97' sample is similar yet simpler than the 'Nlsy79', because the explicit respondents come from a single generation (*i.e.*, the 'Nlsy97 Gen1'). A few variables reflect Gen0 and Gen2. In contrast to the Nlsy79, the Nlsy97 contains more information about the housemates.

They both are nationally-representative samples.

Common NLSY79 terminology refers second generation subjects as 'children' when they are younger than age 15 (NLSYC), and as 'young adults' when they are 15 and older (NLSY79-YA); though they are the same respondents, different funding mechanisms and different survey items necessitate the distinction. This cohort is sometimes abbreviated as 'NLSY79-C', 'NLSY79C',

'NLSY-C' or 'NLSYC'. This packages uses 'Gen2' to refer to subjects of this generation, regardless of their age at the time of the survey.

Ambiguous Categories: In some cases, the kinship of relatives can be safely narrowed to two categories, but not one. The most common scenario involves the *ambiguous siblings* (among Nlsy Gen2 siblings); due to the sampling design, we know that the two participants share the same mother, so they are either full siblings (R=.50) or half siblings (R=.25). For the qqq% that we could not satisfactorily classify, we categorize them as "ambiguous siblings" and assign them R=.375. Another common scenario involves the *ambiguous twins* (among Nlsy97 and both generations of Nlsy79). For qqq% of the twins we comfortably classify their relationship as either MZ or DZ; the remaining qqq% percent are classified as "ambiguous twins" and assigned R=0.75. Empirically, this approach has been successful in the previous 20 years of our previous research, and is discussed further in (Rodgers, qqq).

1.3 Retrieving Data with the NLS Investigator

– This will use much of the existing NLS Investigator vignette.

When a researcher pursues a new idea, we suggest to start by exploring what the NLSY can offer by poking around the (a) vast online documentation and (b) NLS Investigator. The documentation online (www.qqq), and has general information (*e.g.*, how to connect the nationally representative sample was collected), topical information (*e.g.*, what medical and health information has been collected across survey waves and subject ages), and descriptive summaries (*e.g.*, attrition over time for different race and ethnic groups). This material has helpful suggestions which variables are available and appropriate.

With these hints, it's time to identify and download the specific variables from the NLS Investigator. The NLS Investigator is described briefly here (see the NLS Investigator vignette for more detailed instruction). Researchers new to the NLSY should expect at least a dozen round trips as they iteratively improve and complete their set of variables. First, select the 'Study', such as 'NLSY79 Child & Young Adult' (which corresponds to 'Nlsy79 Gen2' in our terminology. Second, select your desired variables, out of the tens of thousands available ones.

1.4 ACE DF Analysis of One Generation

– This very basic analysis that should provide an initial feel for the inputs, mechanics, and goals of the analysis.

1.5 ACE SEM of Two Generations

– This is a more moderately difficult analysis with a more common estimation mechanism.

– Benefits of cross-generational analysis

1.6 More Advanced ACE Analyses

– When the analysis grows beyond a single outcome at one time point, researchers are better off using the modeling software itself (*e.g.*, OpenMx, lavaan, Mplus) than then wrappers provided by `NlsyLinks`, or any other package.

1.7 Data Manipulation and Non-Biometric Analyses

– Even if the investigation doesn't involve family structure, `NlsyLinks` functions and dataset can make the research can be more efficient.

1.8 SAS Analogues

– The `NlsyLinks` datasets can be used in any statistical package, and we demonstrate that here with SAS.
– Downloadable from qqq.-qqqq

1.9 Additional Resources

Your text comes here. Separate text sections with [1].

References

1. R. Mislevy, in *Educational Assessment*, ed. by R.L. Brennan (American Council on Education and Praeger Publishers, 2006), chap. 8