Conducting Trend Analyses of YRBS Data

September 2024

Where can I get more information? Visit www.cdc.gov/yrbss or call 800–CDC–INFO (800–232–4636).

Introduction

Purpose

The purpose of this document is to describe a method for conducting trend analyses of Youth Risk Behavior Survey (YRBS) data to identify and describe changes in the prevalence of risk behaviors over time.

Contents

Background

National, state, territorial, tribal, and local school district YRBSs are periodically conducted using similar sampling methods, survey administration procedures, and questionnaires to produce a series of independently selected cross-sectional data sets which can be compared across years. Trend analyses using YRBS data allow analysts to identify changes in the prevalence of risk behaviors over time, and to describe that change. This document assumes the reader has a working knowledge of YRBS data as well as a thorough knowledge of complex survey data and analysis, statistical software programming, regression analysis, and trend analysis using Joinpoint software from the National Cancer Institute (NCI). This document is not intended to teach statistical methodology.

Readers unfamiliar with YRBS data should review:

- the latest national data documentation,
- Methodology of the Youth Risk Behavior Surveillance System—2013,
- Software for Analysis of YRBS Data,
- Combining YRBS Data Across Years and Sites,
- Interpretation of YRBS Trend Data, and
- the Frequently Asked Questions page on the YRBS website.

Data Requirements

The first step in conducting this type of trend analysis is to assemble a trend data set by combining multiple YRBS data sets that contain data on the variables you wish to analyze. In general, a trend data set should be created by combining (concatenating) data sets from surveys for the time period of interest. For example, a trend analysis to determine progress in achieving Healthy People objectives for years 2017 through 2021 might include national YRBS data from the years 2017 through 2021. In addition, the data set must include a variable (e.g., SURVYEAR) that identifies the survey year for each observation. Finally, the dependent variables of interest (i.e., the risk behaviors to be tested for trend) must be based on questions and response options consistently worded across all the survey years included in the trend analysis. Important guidance in combining YRBS data sets is available in the document "How to Combine YRBS Data Across Years and Sites."

Software Requirements

The Division of Adolescent and School Health (DASH) conducts all complex survey analyses using SAS-callable SUDAAN (Research Triangle Institute, Research Triangle Park, NC). While other statistical software packages are designed to analyze complex survey data , this document describes trend analysis in the context of SAS-callable SUDAAN. In addition, Joinpoint software, a free trend analysis program offered through NCI, is used when there are significant non-linear (quadratic, cubic, etc.) trends to identify the year(s) where the trend changes. More information about Joinpoint, as well as the software download information, can be found on the NCI website.

3-Step Analysis

Step 1: Conduct Regression Analysis

Trend analyses for dichotomous risk behaviors, such as current smoking, are conducted using logistic regression. Similar analyses of continuous risk behaviors are conducted using linear regression. The logistic regression model used by DASH regresses the risk behavior (dependent variable) on continuous linear and non-linear time variables. The dependent variable in public use data sets will need to be recoded because SUDAAN will expect two levels, and they must be coded as 0 or 1. When recoding variables, keep the missing values coded as missing (.) so they can be appropriately excluded. In addition, models typically control for sex, race/ethnicity (four levels), and grade in school (grade = 5 set to missing [.]) using categorical variables. If any control variable is missing for a record, that record will be excluded from the SUDAAN procedure. Testing for linear and non-linear trends is accomplished using the following method:

1) Test for linear trends using a model that contains only a linear time variable (plus variables controlling for sex, race/ethnicity, and grade).
2) Test for quadratic trends by re-running the model with both linear and quadratic time variables.
3) Test for cubic trends by re-running the model a third time and include linear, quadratic, and cubic time variables.

In each case, only the highest-order time variable in the model is valid and can be accurately interpreted. All time variables (linear, quadratic, cubic, etc.) are treated as continuous and are created by coding each year with orthogonal coefficients calculated using PROC IML in SAS (see example SAS code for linear and quadratic time variables). Note that a year variable is specified in the NEST statement of SUDAAN procedures when conducting trend analyses.

If the p-value for the linear time variable is less than the *a priori* significance level (DASH typically uses $\alpha=0.05$), then there is evidence of a linear change. If the associated beta for the significant linear time variable is negative (i.e., less than 0), there is evidence of a linear decrease. Similarly, if the associated beta is positive (i.e., greater than 0), there is evidence of a linear increase.

If the p-value for the quadratic time variable is less than the *a priori* significance level (DASH typically uses $\alpha=0.05$), then there is evidence of a quadratic change. When quadratic changes are detected, the next step is to calculate the adjusted (e.g., for sex, race/ethnicity, and grade) prevalence and standard error by year, and then export these values into Joinpoint software to determine the critical year(s) or "joinpoints" where the non-linear trends change.

Step 2: Use Joinpoint software to determine critical values (for significant quadratic or cubic time components)

When a significant quadratic change has been detected, the next step is to calculate the adjusted prevalence (predicted marginal) and associated standard error (SE) for each year in the analysis. In SUDAAN, this is easily done by adding the "PREDMARG" statement to the RLOGIST procedure code. Sample SUDAAN code for obtaining adjusted prevalence can be found below and in Bieler, et al.,[1] a paper on estimating model-adjusted risks and risk ratios in complex survey data. The year, adjusted prevalence estimate, and SE can be copied to a plain text tab-delimited file and then imported into Joinpoint software.

Example

```
PROC RLOGIST DATA=WORK.VARSET DESIGN=WR FILETYPE=SAS;
              NEST SURVYEAR STRATUM PSU/PSULEV=3 MISSUNIT;
              WEIGHT WEIGHT;
              CLASS SEX RACE GRADE YEAR;
              MODEL QL54 = SEX RACE GRADE YEAR;
PREDMARG YEAR;
              PRINT/BETAFMT=F8.5 SEBETAFMT=F8.5 P_BETAFMT=F8.5;
              RUN;
```

Once values have been imported into Joinpoint, use Joinpoint to determine the location of 1 joinpoint for a significant quadratic or 2 joinpoints for a significant cubic trend. The last analytical step is to return to SUDAAN to test each line segment (i.e., before and after each joinpoint) for linearity.

Step 3: Test segments

After determining the joinpoint(s) for significant non-linear trends (e.g., quadratic, cubic, etc.) the resulting line segments are tested for linearity in SUDAAN. To test both sides of a significant quadratic trend with a single joinpoint, two data sets are created – the first includes all years from the first available year up to and including the joinpoint year; the second includes all years beginning with the joinpoint year up to the last available year. For example, if a trend analysis covers the past decade (2013 to 2023) and the Joinpoint is determined to be 2015, the first data set would contain 2013, 2015, and 2017; the second data set would contain 2019, 2021, 2023, and 2023. Orthogonal coefficients for each time segment (2013-2017 and 2019-2023) are calculated using PROC IML in SAS.

With each data set, run a logistic regression analysis containing a linear, but not a higher order (e.g., quadratic or cubic) time component. If the linear component is significant at the *a priori* level, this indicates either a significant decrease or a significant increase for that time period. If the associated beta for the significant linear time component is negative (i.e., less than 0), there is evidence of a linear decrease. Similarly, if the associated beta is positive (i.e., greater than 0), there is evidence of a linear increase.

Example

Testing line segments for linearity to the left and right of a joinpoint (2017) in the event of a significant quadratic trend (QN22: Ever physically dating violence)

Year	2013	2015	2017	2019	2021	2023
Prevalence	10.3%	9.6%	8.0%	8.2%	8.5%	10.4%

```
/*Produce orthogonal coefficients for linear trend testing to the left and right of the joinpoint year 2015*/

PROC IML; X={2013 2015 2017}; XP=ORPOL(X,1); PRINT XP; RUN; QUIT;
PROC IML; X={2017 2019 2021 2023}; XP=ORPOL(X,1); PRINT XP; RUN; QUIT;

DATA VARSET;   SET DATA1.TRENDWK;  /* Testing for linearity to the left of the joinpoint. */
      IF (YEAR=2013 OR YEAR=2015 OR YEAR=2017);

        IF YEAR=2013 THEN T3L_L=-0.707107;
ELSE IF YEAR=2015 THEN T3L_L=0;
ELSE IF YEAR=2017 THEN T3L_L=0.707107;

proc sort data = work.varset;
      by survyear stratum psu;

PROC RLOGIST DATA=WORK.VARSET DESIGN=WR FILETYPE=SAS;
      NEST SURVYEAR STRATUM PSU/PSULEV=3 MISSUNIT;
      WEIGHT WEIGHT;
      CLASS SEX RACE GRADE;
      MODEL QL22 = SEX RACE GRADE T3L_L;
      PRINT/BETAFMT=F8.5 SEBETAFMT=F8.5 P_BETAFMT=F8.5;
      RUN;
```

In this example, T3L_L had a p-value 0.00015 and beta= -0.2121. Therefore, there was a "significant linear decrease in the prevalence of students experiencing physical dating violence during 2013-2017."

```
DATA VARSET;   SET DATA1.TRENDWK;  /* Testing for linearity to the right of the joinpoint. */
      IF (YEAR=2017 OR YEAR=2019 OR YEAR=2021 OR YEAR=2023);

        IF YEAR=2017 THEN T4L_R=-0.670820;
ELSE IF YEAR=2019 THEN T4L_R=-0.223607;
ELSE IF YEAR=2021 THEN T4L_R=0.223607;
ELSE IF YEAR=2023 THEN T4L_R=0.670820;

proc sort data = work.varset;
      by survyear stratum psu;

PROC RLOGIST DATA=WORK.VARSET DESIGN=WR FILETYPE=SAS;
      NEST SURVYEAR STRATUM PSU/PSULEV=3 MISSUNIT;
      WEIGHT WEIGHT;
      CLASS SEX RACE GRADE;
      MODEL QL22 = SEX RACE GRADE T4L_R;
      PRINT/BETAFMT=F8.5 SEBETAFMT=F8.5 P_BETAFMT=F8.5;
      RUN;
```

In this example, T4L_R had a p-value=0.00001 and beta= 0.2455. Therefore, there was a "significant linear increase" in the prevalence of students experiencing physical dating violence during 2017-2023."

In summary, there was a significant linear decrease in the prevalence of experiencing physical dating violence during 2013-2017, followed by a significant linear increase during 2017-2023.

References

1) Bieler GS, Brown GG, Williams RL, Brogan DJ. Estimating model-adjusted risks, risk differences, and risk ratios from complex survey data. *American Journal of Epidemiology* 2010;171(5):618-23.