Sample Code to Analyze CHIS Data

Intro

This page provides sample code for appropriately analyzing California Health Interview Survey (CHIS) data in several statistical software packages using the replicate weights method or the Taylor series linearization method. For a summary of how weighting and variance estimation work in analyzing CHIS Data, refer to **Weighting and Variance Estimation**.

In order to accurately estimate variance in analyses of CHIS data, a complex sample, either replicate weights or the Taylor series linearization method may be used.

Below are sample codes

- Replicate Weighting examples are available for:
  - Stata, R, SAS, SUDAAN.
- Taylor Series examples are available for:
  - Stata, R, SAS, SUDAAN, SPSS.
- Pooling examples are available for:
  - Stata, SAS, R.

Replicate Weights

Examples here illustrate how CHIS PUFs can be analyzed with replicate weights to produce valid variance estimates using statistical software packages.

For continuous variables, sample calculations of means and linear regression analysis are presented; for categorical variables, sample calculations of frequencies and logistic regression analysis are presented. Estimates and standard errors are identical across the software packages examined, but confidence intervals may differ because of different default methods of computation (for SAS, the default is Wald confidence intervals; for SUDAAN and Stata, it is logit confidence intervals).

Stata

Mean Calculation

In the following sample code, the distribution of BMI (bmi) is examined by race (racehpr2) and by race and sex (racehpr2*srsex).

```
*Sample design specification step*
use "DATASET LOCATION”
svyset [pw=rakedw0], jkrw(rakedw1-rakedw80, multiplier(1)) vce(jack) m
se

*Analysis*
svy: mean bmi, over(racehpr2)
svy: mean bmi, over(srsex racehpr2)
```

In Stata, the sample design specification step should be included before conducting any analysis.

Frequency Calculation

In the following sample code, the percentage of people who currently have asthma (astcur) is examined by race (racehpr2) and by race and sex (racehpr2*srsex). Weighted counts are also given.

```
*Sample design specification step*
use "DATASET LOCATION”
svyset [pw=rakedw0], jkrw(rakedw1-rakedw80, multiplier(1)) vce(jack) m
se

*Analysis*
svy: tabulate astcur racehpr2, col se ci
svy, subpop (if srsex==1): tab astcur racehpr2, col se ci
svy, subpop (if srsex==2): tab astcur racehpr2, col se ci
svy: tabulate astcur racehpr2, count format(%11.0fc)
```

In Stata, the sample design specification step should be included before conducting any analysis.

Stata V.10 and higher cannot accommodate 3 or more variables in the tab command.

Linear Regression

In the following sample code, Body Mass Index (bmi) is examined in relation to race (racehpr2), sex (srsex), and age (srage) while controlling for each other. Note that racehpr2 and srsex are categorical variables; White (racehpr2=6) and Male (srsex=1) are used as their reference categories.

```
*Sample design specification step*
use "DATASET LOCATION”
svyset [pw=rakedw0], jkrw(rakedw1-rakedw80, multiplier(1)) vce(jack) m
se

*Analysis*
recode racehpr2 (6=1) (1=2) (2=3) (3=4) (4=5) (5=6) (7=7), gen(race)
```

```
xi: svy: regress bmi i.srsex i.race srage
```

In Stata, the sample design specification step should be included before conducting any analysis.

Recoding is done in order to choose “White” (racehpr2=6) as the reference group.

Logistic Regression

In the following sample code, current asthma status (astcur) is examined, controlling for race (racehpr2), sex (srsex), and age (srage). SUDAAN and Stata require the dependent variables to be coded as 0 and 1 for logistic regression, so a new dependent variable ast is created and assigned 1 where astcur=1 (“Current asthma”) and 0 where astcur=2 (“No current asthma”). The category “No current asthma” is used as the reference in the analysis.

```
*Sample design specification step*
use "DATASET LOCATION”
svyset [pw=rakedw0], jkrw(rakedw1-rakedw80, multiplier(1)) vce(jack) m
se

*Analysis*
recode astcur (2=0) (1=1) (-9=.) (-1=.), gen (ast)

xi: svy: logit ast srage i.race i.srsex

xi: svy: logistic ast srage i.race i.srsex
```

In Stata, the sample design specification step should be included before conducting any analysis.

logit produces parameter estimates.

logistic produces odds ratios. Stata automatically chooses the lowest value of the categorical variable as the reference group for the independent and dependent variables.

R

Mean Calculation

In the following sample code, the distribution of BMI (bmi) is examined by race (racehpr2) and by race and sex (racehpr2*srsex).

```
# Sample design specification step

# Developed with 'R version 4.2.3' and 'survey version 4.1-1'
library(survey)

chis_design <- svrepdesign(data = data , weights = ~ rakedw0 ,
```

```
                            repweights = "rakedw[1-9]" ,
                            type = "other" , scale = 1 ,
                            rscales = 1 , mse = TRUE)
# Analysis

chis_design |>
  svyby( formula = ~ bmi , by = ~ racehpr2,
         FUN=svymean, vartype=c("se","ci"), deff = TRUE)

chis_design |>
  svyby( formula = ~ bmi , by = ~ racehpr2+srsex,
         FUN=svymean, vartype=c("se","ci"), deff = TRUE)
```

Frequency Calculation

In the following sample code, the percentage of people who currently have asthma (astcur) is examined by race (racehpr2) and by race and sex (racehpr2*srsex).

```
# Sample design specification step

# Developed with 'R version 4.2.3' and 'survey version 4.1-1'
library(survey)

chis_design <- svrepdesign(data = data , weights = ~ rakedw0 ,
                            repweights = "rakedw[1-9]" ,
                            type = "other" , scale = 1 ,
                            rscales = 1 , mse = TRUE)
# Analysis

# sub pop 1
chis_design |> subset(srsex=="1") |>
  svytable(formula=~astcur+racehpr2) |>
  prop.table()  # proportions

chis_design |> subset(srsex=="1") |>
  svytable(formula=~astcur+ racehpr2) |>
  # prop.table() |>
  summary()  # totals with chi-square stats

# sub pop 2
chis_design |> subset(srsex=="2") |>
  svytotal(x=~interaction(astcur, racehpr2))  # total as interaction

chis_design |> subset(srsex=="2") |>
  svymean(x=~interaction(astcur, racehpr2)) |>
  confint()  # confidence intervals
```

Linear Regression

In the following sample code, Body Mass Index (bmi) is examined in relation to race (racehpr2), sex (srsex), and age (srage) while controlling for each other. Note that racehpr2 and srsex are categorical variables; White (racehpr2=6) and Male (srsex=1) are used as their reference categories.

```
# Sample design specification step

# Developed with 'R version 4.2.3' and 'survey version 4.1-1'
library(survey)

chis_design <- svrepdesign(data = data , weights = ~ rakedw0 ,
                           repweights = "rakedw[1-9]" ,
                           type = "other" , scale = 1 ,
                           rscales = 1 , mse = TRUE)
# Analysis

# use relevel(ref=,...) inside the formula to pick reference category
# can alternatively pre process data before the formula call

chis_design |>
  svyglm(formula = bmi ~ relevel(as.factor(srsex),ref="1") +
           relevel(as.factor(racehpr2),ref="6") +
           srage,
         family=stats::gaussian(),
         rescale=TRUE) |>
         summary()
```

Logistic Regression

In the following sample code, current asthma status (astcur) is examined, controlling for race (racehpr2), sex (srsex), and age (srage). SUDAAN and Stata require the dependent variables to be coded as 0 and 1 for logistic regression, so a new dependent variable ast is created and assigned 1 where astcur=1 (“Current asthma”) and 0 where astcur=2 (“No current asthma”). The category “No current asthma” is used as the reference in the analysis.

```
# Sample design specification step

# Developed with 'R version 4.2.3' and 'survey version 4.1-1'
library(survey)

chis_design <- svrepdesign(data = data , weights = ~ rakedw0 ,
                           repweights = "rakedw[1-9]" ,
                           type = "other" , scale = 1 ,
                           rscales = 1 , mse = TRUE)
```

```
# Analysis

# use stats::update() to transform ASTCUR after the survey design spec
# can alternatively pre process data before the survey design spec

# use relevel(ref=,...) inside the formula to pick reference category
# can alternatively pre process data before the formula call

chis_design |>
  stats::update(ast = ifelse(astcur==2,0,1)) |>
  svyglm(formula = ast ~ srage +
           relevel(as.factor(racehpr2),ref="6") +
           relevel(as.factor(srsex),ref="1"),
         family=quasibinomial(),
         rescale=TRUE) |>
         summary()
```

SAS

Mean Calculation

In the following sample code, the distribution of BMI (bmi) is examined by race (racehpr2) and by race and sex (racehpr2*srsex).

```
PROC SORT DATA = data;
BY racehpr2;
RUN;

PROC SURVEYMEANS DATA = data VARMETHOD=JACKKNIFE;
WEIGHT rakedw0;
REPWEIGHT rakedw1-rakedw80/JKCOEFS=1;
VAR bmi;
BY racehpr2; RUN;

PROC SORT DATA = data;
BY racehpr2 srsex;
RUN;

PROC SURVEYMEANS DATA = data VARMETHOD=JACKKNIFE;
WEIGHT rakedw0;
REPWEIGHT rakedw1-rakedw80/JKCOEFS=1;
VAR bmi;
BY racehpr2 srsex;
RUN;
```

Jackknife coefficients are necessary for accurate variance calculations, and jackknife coefficients of 1 in SAS will produce equal variance calculations as those produced in SUDAAN. However, for SAS V.9.2(TS1M0) and earlier, a value of 1 will not be accepted; as a substitute, 0.9999 can be entered. Without this specification, the default value of the jackknife coefficients will be [(# replicate weights - 1)/# replicate weights]; for CHIS, this would be [(80 - 1)/80] = 0.9875.

Frequency Calculation

In the following sample code, the percentage of people who currently have asthma (astcur) is examined by race (racehpr2) and by race and sex (racehpr2*srsex).

```
PROC SURVEYFREQ DATA = data VARMETHOD=JACKKNIFE;
WEIGHT rakedw0;
REPWEIGHT rakedw1-rakedw80/JKCOEFS=1;
TABLES racehpr2*astcur/row;
RUN;

PROC SURVEYFREQ DATA = data VARMETHOD=JACKKNIFE;
WEIGHT rakedw0;
REPWEIGHT rakedw1-rakedw80/JKCOEFS=1;
TABLES srsex*racehpr2*astcur/row;
RUN;
```

One caveat in creating multiple tables in one PROC SURVEYFREQ procedure is that the procedure takes the smallest applicable sample sizes among all variables. Therefore, creating one table per one PROC SURVEYFREQ procedure is recommended:

Jackknife coefficients are necessary for accurate variance calculations, and jackknife coefficients of 1 in SAS will produce equal variance calculations as those produced in SUDAAN. However, for SAS V.9.2(TS1M0) and earlier, a value of 1 will not be accepted; as a substitute, 0.9999 can be entered. Without this specification, the default value of the jackknife coefficients will be [(# replicate weights - 1)/# replicate weights]; for CHIS, this would be [(80 - 1)/80] = 0.9875.

Linear Regression

In the following sample code, Body Mass Index (bmi) is examined in relation to race (racehpr2), sex (srsex), and age (srage) while controlling for each other. Note that racehpr2 and srsex are categorical variables; White (racehpr2=6) and Male (srsex=1) are used as their reference categories.

```
PROC SURVEYREG DATA = data VARMETHOD=JACKKNIFE;
WEIGHT rakedw0;
REPWEIGHT rakedw1-rakedw80/JKCOEFS=1;
FORMAT racehpr2 racehprf. srsex srsex.;
CLASS racehpr2 srsex;
MODEL bmi = srsex racehpr2 srage/SOLUTION;
RUN;
```

Jackknife coefficients are necessary for accurate variance calculations, and jackknife coefficients of 1 in SAS will produce equal variance calculations as those produced in SUDAAN. However, for SAS V.9.2(TS1M0) and earlier, a value of 1 will not be accepted; as a substitute, 0.9999 can be entered. Without this specification, the default value of the jackknife coefficients will be [(# replicate weights - 1)/# replicate weights]; for CHIS, this would be [(80 - 1)/80] = 0.9875

When the values are formatted either in the data step or in the procedure, SAS automatically picks the category of the categorical variables whose label is alphabetically last as a reference group.

SOLUTION option provides the parameter estimates when using a CLASS statement.

Logistic Regression

In the following sample code, current asthma status (astcur) is examined, controlling for race (racehpr2), sex (srsex), and age (srage). SUDAAN and Stata require the dependent variables to be coded as 0 and 1 for logistic regression, so a new dependent variable ast is created and assigned 1 where astcur=1 ("Current asthma") and 0 where astcur=2 ("No current asthma"). The category "No current asthma" is used as the reference in the analysis.

```
PROC SURVEYLOGISTIC DATA = data VARMETHOD=JACKKNIFE;
FORMAT astcur astcurf. racehpr2 racehprf. srsex srsex.;
WEIGHT rakedw0;
REPWEIGHT rakedw1-rakedw80/JKCOEFS=1;
CLASS astcur (REF="NO CURRENT ASTHMA") racehpr2 (REF="WHITE") srsex (R
EF="MALE")/PARAM=REF;
MODEL astcur = racehpr2 srsex srage;
RUN;
```

In PROC SURVEYLOGISTIC, the reference category of the independent and dependent variables may be specified in a CLASS statement. PARAM=REF is specified to ensure dummy coding of the categorical independent variables.

Jackknife coefficients are necessary for accurate variance calculations, and jackknife coefficients of 1 in SAS will produce equal variance calculations as those produced in SUDAAN. However, for SAS V.9.2 (TS1M0) and earlier, a value of 1 will not be accepted; as a substitute, 0.9999 can be entered. Without this specification, the default value of the jackknife coefficients will be [(# replicate weights - 1)/# replicate weights]; for CHIS, this would be [(80 - 1)/80] = 0.9875.

SUDAAN

Mean Calculation

In the following sample code, the distribution of BMI (bmi) is examined by race (racehpr2) and by race and sex (racehpr2*srsex).

```
PROC DESCRIPT DATA = data FILETYPE=SAS DESIGN=JACKKNIFE;
```

```
WEIGHT rakedw0;
JACKWGTS rakedw1-rakedw80/ADJJACK=1;
VAR bmi;
TABLES racehpr2 racehpr2*srsex;
SUBGROUP racehpr2 srsex;
LEVELS 7 2;
RUN;
```

Frequency Calculation

In the following sample code, the percentage of people who currently have asthma (astcur) is examined by race (racehpr2) and by race and sex (racehpr2*srsex).

```
PROC CROSSTAB DATA = data FILETYPE=SAS DESIGN=JACKKNIFE;
WEIGHT rakedw0;
JACKWGTS rakedw1-rakedw80/ADJJACK=1;
TABLES racehpr2*astcur srsex*racehpr2*astcur;
SUBGROUP astcur racehpr2 srsex;
LEVELS 2 7 2;
RUN;
```

Linear Regression

In the following sample code, Body Mass Index (bmi) is examined in relation to race (racehpr2), sex (srsex), and age (srage) while controlling for each other. Note that racehpr2 and srsex are categorical variables; White (racehpr2=6) and Male (srsex=1) are used as their reference categories.

```
PROC REGRESS DATA = data FILETYPE=SAS DESIGN=JACKKNIFE;
WEIGHT rakedw0;
JACKWGTS rakedw1-rakedw80/ADJJACK=1;
SUBGROUP racehpr2 srsex;
LEVELS 7 2;
REFLEVEL racehpr2=6 srsex=1;
MODEL bmi = racehpr2 srsex srage;
RUN;
```

Logistic Regression

In the following sample code, current asthma status (astcur) is examined, controlling for race (racehpr2), sex (srsex), and age (srage). SUDAAN and Stata require the dependent variables to be coded as 0 and 1 for logistic regression, so a new dependent variable ast is created and assigned 1 where astcur=1 (“Current asthma”) and 0 where astcur=2 (“No current asthma”). The category “No current asthma” is used as the reference in the analysis.

```
DATA newdata;
SET data;
IF astcur=1 THEN ast=1;
```

```
ELSE IF astcur=2 THEN ast=0;
RUN;

PROC RLOGIST data = newdata FILETYPE=SAS DESIGN=JACKKNIFE;
WEIGHT rakedw0;
JACKWGTS rakedw1-rakedw80/ADJJACK=1;
SUBGROUP racehpr2 srsex;
LEVELS 7 2;
REFLEVEL racehpr2 = 6 srsex = 1;
MODEL ast = racehpr2 srsex srage; RUN;
```

Taylor Series Linearization Method

Examples here illustrate how CHIS data can be analyzed with Taylor series linearization to produce valid variance estimates using statistical software packages. The required variables for the Taylor series linearization method (tsvarstr and tsvrunit) have not been included in the CHIS Public Use Files.

For continuous variables, sample calculations of means and linear regression analysis are presented; for categorical variables, sample calculations of frequencies and logistic regression analysis are presented. Estimates and standard errors are identical across the software packages examined, but confidence intervals may differ because of different default methods of computation (for SAS, the default is Wald confidence intervals; for SUDAAN and Stata, it is logit confidence intervals).

Stata

Mean Calculation

In the following sample code, the distribution of BMI (bmi) is examined by race (racehpr2) and by the interaction between race and sex (racehpr2*srsex).

```
*Sample design specification step*
use "DATASET LOCATION”
svyset tsvrunit [pw=rakedw0], strata (tsvarstr)

*Analysis*
svy: mean bmi, over(racehpr2)
svy: mean bmi, over(srsex racehpr2)
```

In Stata, the sample design specification step should be included before conducting any analysis.

When using concatenated data across adults, adolescents, and/or children, use tsvrunit; when using separate data files, delete the commands associated with tsvrunit.

Frequency Calculation

In the following sample code, the percentage of people who currently have asthma (astcur) is examined by race (racehpr2) and by race and sex (racehpr2*srsex). Weighted counts are also given.

```
*Sample design specification step*
use "DATASET LOCATION"
svyset tsvrunit [pw=rakedw0], strata (tsvarstr)

*Analysis*
svy: tabulate astcur racehpr2, col se ci
svy, subpop (if srsex==1): tab astcur racehpr2, col se ci
svy, subpop (if srsex==2): tab astcur racehpr2, col se ci
svy: tabulate astcur racehpr2, count format(%11.0fc)
```

In Stata, the sample design specification step should be included before conducting any analysis.

When using concatenated data across adults, adolescents, and/or children, use tsvrunit; when using separate data files, delete the commands associated with tsvrunit.

Stata V.10 and higher cannot accommodate 3 or more variables in the tab command.

Linear Regression

In the following sample code, Body Mass Index (bmi) is examined in relation to race (racehpr2), sex (srsex), and age (srage) while controlling for each other. Note that racehpr2 and srsex are categorical variables, and White (racehpr2=6) and Male (srsex=1) are used as their reference categories.

```
*Sample design specification step*
use "DATASET LOCATION"
svyset tsvrunit [pw=rakedw0], strata (tsvarstr)

*Analysis*
recode racehpr2 (6=1) (1=2) (2=3) (3=4) (4=5) (5=6) (7=7), gen(race)

xi: svy: regress bmi i.srsex i.race srage
```

In Stata, the sample design specification step should be included before conducting any analysis.

When using concatenated data across adults, adolescents, and/or children, use tsvrunit; when using separate data files, delete the commands associated with tsvrunit.

Reordering is done in order to choose “White” (racehpr2=6) as the reference group.

Logistic Regression

In the following sample code, current asthma status (astcur) is examined, controlling for race (racehpr2), sex (srsex), and age (srage). As SUDAAN and Stata require the dependent variables coded as 0 and 1 for logistic regression, a new dependent variable ast is created and assigned 1 where astcur=1 ("Current asthma") and 0 where astcur=2 ("No current asthma"). The category "No current asthma" is used as the reference in the analysis.

```
*Sample design specification step*
use "DATASET LOCATION"
svyset tsvrunit [pw=rakedw0], strata (tsvarstr)

*Analysis*
recode astcur (2=0) (1=1) (-9=.) (-1=.), gen (ast)

xi: svy: logit ast srage i.race i.srsex

xi: svy: logistic ast srage i.race i.srsex
```

In Stata, the sample design specification step should be included before conducting any analysis.

When using concatenated data across adults, adolescents, and/or children, use tsvrunit; when using separate data files, delete the commands associated with tsvrunit.

logit produces parameter estimates.

logistic produces odds ratios. Stata automatically chooses the lowest value of the categorical variable as the reference group for the independent and dependent variables.

R

The Taylor Series estimation examples in *R* using the *survey* package are exactly the same as the Replicate Weighting examples. However, the single important difference is telling *R* to change the survey design setup at the very beginning as we demonstrate below.

```
library(survey)

# instead of ?svrepdesign() for replicate weights
# chis_design <- svrepdesign(data=your_chis_data, ... )

# use ?svydesign() for taylor series
chis_design <- svydesign(id=~tsvrunit,
                         strata=~tsvarstr,
                         weights=~rakedw0,
                         data=your_chis_data,
                         nest=TRUE)
```

```
# Now, all downstream estimation functions are the same
svymean(design=chis_design, … )
```

Below are the Taylor series examples, correctly using *svydesign()* instead of *svrepdesign()*

Mean Calculation

In the following sample code, the distribution of BMI (bmi) is examined by race (racehpr2) and by the interaction between race and sex (racehpr2*srsex).

```
# Sample design specification step

# Developed with 'R version 4.2.3' and 'survey version 4.1-1'
library(survey)

chis_design <- svydesign(id=~tsvrunit,
                         strata=~tsvarstr,
                         weights=~rakedw0,
                         data=your_chis_data,
                         nest=TRUE)
# Analysis

chis_design |>
  svyby( formula = ~ bmi , by = ~ racehpr2,
         FUN=svymean, vartype=c("se","ci"), deff = TRUE)

chis_design |>
  svyby( formula = ~ bmi , by = ~ racehpr2+srsex,
         FUN=svymean, vartype=c("se","ci"), deff = TRUE)
```

Frequency Calculation

In the following sample code, the percentage of people who currently have asthma (astcur) is examined by race (racehpr2) and by race and sex (racehpr2*srsex).

```
# Sample design specification step

# Developed with 'R version 4.2.3' and 'survey version 4.1-1'
library(survey)

chis_design <- svydesign(id=~tsvrunit,
                         strata=~tsvarstr,
                         weights=~rakedw0,
                         data=your_chis_data,
                         nest=TRUE)
# Analysis
```

```
# sub pop 1
chis_design |> subset(srsex=="1") |>
  svytable(formula=~astcur+racehpr2) |>
  prop.table()  # proportions

chis_design |> subset(srsex=="1") |>
  svytable(formula=~astcur+ racehpr2) |>
  # prop.table() |>
  summary()  # totals with chi-square stats

# sub pop 2
chis_design |> subset(srsex=="2") |>
  svytotal(x=~interaction(astcur, racehpr2))  # total as interaction

chis_design |> subset(srsex=="2") |>
  svymean(x=~interaction(astcur, racehpr2)) |>
  confint()  # confidence intervals
```

Linear Regression

In the following sample code, Body Mass Index (bmi) is examined in relation to race (racehpr2), sex (srsex), and age (srage) while controlling for each other. Note that racehpr2 and srsex are categorical variables, and White (racehpr2=6) and Male (srsex=1) are used as their reference categories.

```
# Sample design specification step

# Developed with 'R version 4.2.3' and 'survey version 4.1-1'
library(survey)

chis_design <- svydesign(id=~tsvrunit,
                         strata=~tsvarstr,
                         weights=~rakedw0,
                         data=your_chis_data,
                         nest=TRUE)
# Analysis

# use relevel(ref=,...) inside the formula to pick reference category
# can alternatively pre-process data before the formula call

chis_design |>
  svyglm(formula = bmi ~ relevel(as.factor(srsex),ref="1") +
           relevel(as.factor(racehpr2),ref="6") +
           srage,
         family=stats::gaussian(),
         rescale=TRUE) |>
```

```
        summary()
```

Logistic Regression

In the following sample code, current asthma status (astcur) is examined, controlling for race (racehpr2), sex (srsex), and age (srage). As SUDAAN and Stata require the dependent variables coded as 0 and 1 for logistic regression, a new dependent variable ast is created and assigned 1 where astcur=1 (“Current asthma”) and 0 where astcur=2 (“No current asthma”). The category “No current asthma” is used as the reference in the analysis.

```
# Sample design specification step

# Developed with 'R version 4.2.3' and 'survey version 4.1-1'
library(survey)

chis_design <- svydesign(id=~tsvrunit,
                         strata=~tsvarstr,
                         weights=~rakedw0,
                         data=your_chis_data,
                         nest=TRUE)

# Analysis

# use stats::update() to transform ASTCUR after the survey design spec
# can alternatively pre-process data before the survey design spec

# use relevel(ref=,...) inside the formula to pick reference category
# can alternatively pre-process data before the formula call

chis_design |>
  stats::update(ast = ifelse(astcur==2,0,1)) |>
  svyglm(formula = ast ~ srage +
           relevel(as.factor(racehpr2),ref="6") +
           relevel(as.factor(srsex),ref="1"),
         family=quasibinomial(),
         rescale=TRUE) |>
         summary()
```

SAS

Mean Calculation

In the following sample code, the distribution of BMI (bmi) is examined by race (racehpr2) and by the interaction between race and sex (racehpr2*srsex).

```
PROC SURVEYMEANS DATA = data mean stderr NOMCAR VARMETHOD=TAYLOR;
STRATA tsvarstr; CLUSTER tsvrunit;
WEIGHT rakedw0;
VAR bmi;
DOMAIN racehpr2 racehpr2*srsex;
RUN;
```

If conducting a domain analysis, the DOMAIN statement is necessary for accurate variance estimation. Using BY or WHERE statements will not produce valid variance estimates for the subpopulation/domain. In SAS, the NOMCAR option presents the assumption that missing values are not completely at random. This, along with the DOMAIN statement, is the appropriate approach for domain analyses, which uses the entire sample for variance estimation.

When using concatenated data across adults, adolescents, and/or children, use tsvrunit; when using separate data files, delete the commands associated with tsvrunit.

Frequency Calculation

In the following sample code, the percentage of people who currently have asthma (astcur) is examined by race (racehpr2) and by race and sex (racehpr2*srsex).

```
PROC SURVEYFREQ DATA = data NOMCAR VARMETHOD=TAYLOR;
STRATA tsvarstr; CLUSTER tsvrunit;
WEIGHT rakedw0;
TABLE racehpr2*asctur/row;
RUN;

PROC SURVEYFREQ DATA = data NOMCAR VARMETHOD=TAYLOR;
STRATA tsvarstr; CLUSTER tsvrunit;
WEIGHT rakedw0;
TABLE srsex*racehpr2*astcur/row;
RUN;
```

One caveat in creating multiple tables in one PROC SURVEYFREQ procedure is that the procedure takes the smallest applicable sample sizes among all variables. Therefore, creating one table per one PROC SURVEYFREQ procedure is recommended.

If conducting a domain analysis, the DOMAIN statement is necessary for accurate variance estimation. Using BY or WHERE statements will not produce valid variance estimates for the subpopulation/domain. The NOMCAR option presents the assumption that missing values are not completely at random. This, along with the DOMAIN statement, is the appropriate approach for domain analyses, which uses the entire sample for variance estimation.

When using concatenated data across adults, adolescents, and/or children, use tsvrunit; when using separate data files, delete the commands associated with tsvrunit.

Linear Regression

In the following sample code, Body Mass Index (bmi) is examined in relation to race (racehpr2), sex (srsex), and age (srage) while controlling for each other. Note that racehpr2 and srsex are categorical variables, and White (racehpr2=6) and Male (srsex=1) are used as their reference categories.

```
PROC SURVEYREG DATA = data NOMCAR VARMETHOD=TAYLOR;
STRATA tsvarstr; CLUSTER tsvrunit;
WEIGHT rakedw0;
FORMAT racehpr2 racehprf. srsex srsex.;
CLASS racehpr2 srsex;
MODEL bmi = srsex racehpr2 srage/SOLUTION;
RUN;
```

If conducting a domain analysis, the DOMAIN statement is necessary for accurate variance estimation. Using BY or WHERE statements will not produce valid variance estimates for the subpopulation/domain. The NOMCAR option presents the assumption that missing values are not completely at random. This, along with the DOMAIN statement, is the appropriate approach for domain analyses, which uses the entire sample for variance estimation.

When using concatenated data across adults, adolescents, and/or children, use tsvrunit; when using separate data files, delete the commands associated with tsvrunit.

SOLUTION option provides the parameter estimates when using a CLASS statement.

Logistic Regression

In the following sample code, current asthma status (astcur) is examined, controlling for race (racehpr2), sex (srsex), and age (srage). As SUDAAN and Stata require the dependent variables coded as 0 and 1 for logistic regression, a new dependent variable ast is created and assigned 1 where astcur=1 (“Current asthma”) and 0 where astcur=2 (“No current asthma”). The category “No current asthma” is used as the reference in the analysis.

```
PROC SURVEYLOGISTIC DATA = data NOMCAR VARMETHOD=TAYLOR;
FORMAT astcur astcurf. racehpr2 racehprf. srsex srsex.;
STRATA tsvarstr;
CLUSTER tsvrunit;
WEIGHT rakedw0;
CLASS astcur (REF=“NO CURRENT ASTHMA”) racehpr2 (REF=“WHITE”) srsex (R
EF=“MALE”)/PARAM=REF;
MODEL astcur = racehpr2 srsex srage;
RUN;
```

If conducting a domain analysis, the DOMAIN statement is necessary for accurate variance estimation. Using BY or WHERE statements will not produce valid variance estimates for the subpopulation/domain. The NOMCAR option presents the assumption that missing values are

not completely at random. This, along with the DOMAIN statement, is the appropriate approach for domain analyses, which uses the entire sample for variance estimation.

In PROC SURVEYLOGISTIC, the reference category of the independent and dependent variables may be specified in a CLASS statement. PARAM=REF is specified to ensure dummy coding of the categorical independent variables.

When using concatenated data across adults, adolescents, and/or children, use tsvrunit; when using separate data files, delete the commands associated with tsvrunit.

SUDAAN

Mean Calculation

In the following sample code, the distribution of BMI (bmi) is examined by race (racehpr2) and by the interaction between race and sex (racehpr2*srsex).

```
PROC SORT DATA = data;
BY tsvarstr tsvrunit;
RUN;

PROC DESCRIPTIVE DATA = data FILETYPE=SAS DESIGN=WR;
NEST tsvarstr tsvrunit;
WEIGHT rakedw0;
CLASS racehpr2 racehpr2*srsex;
VAR bmi;
TABLES racehpr2 racehpr2*srsex;
SUBGROUP racehpr2 srsex;
LEVELS 7 2;
RUN;
```

When using concatenated data across adults, adolescents, and/or children, use tsvrunit; when using separate data files, delete the commands associated with tsvrunit.

Frequency Calculation

In the following sample code, the percentage of people who currently have asthma (astcur) is examined by race (racehpr2) and by race and sex (racehpr2*srsex).

```
PROC SORT DATA = data;
BY tsvarstr tsvrunit;
RUN;

PROC CROSSTAB DATA = data FILETYPE=SAS DESIGN=WR;
NEST tsvarstr tsvrunit;
WEIGHT rakedw0;
TABLES racehpr2*astcur srsex*racehpr2*astcur;
SUBGROUP astcur racehpr2 srsex;
```

```
LEVELS 2 7 2;
RUN;
```

When using concatenated data across adults, adolescents, and/or children, use tsvrunit; when using separate data files, delete the commands associated with tsvrunit.

Linear Regression

In the following sample code, Body Mass Index (bmi) is examined in relation to race (racehpr2), sex (srsex), and age (srage) while controlling for each other. Note that racehpr2 and srsex are categorical variables, and White (racehpr2=6) and Male (srsex=1) are used as their reference categories.

```
PROC SORT DATA = data;
BY tsvarstr tsvrunit;
RUN;

PROC REGRESS DATA = data FILETYPE=SAS DESIGN=WR;
NEST tsvarstr tsvrunit;
WEIGHT rakedw0;
SUBGROUP racehpr2 srsex;
LEVELS 7 2;
REFLEVEL racehpr2=6 srsex=1;
MODEL bmi = racehpr2 srsex srage;
RUN;
```

When using concatenated data across adults, adolescents, and/or children, use tsvrunit; when using separate data files, delete the commands associated with tsvrunit.

Logistic Regression

In the following sample code, current asthma status (astcur) is examined, controlling for race (racehpr2), sex (srsex), and age (srage). As SUDAAN and Stata require the dependent variables coded as 0 and 1 for logistic regression, a new dependent variable ast is created and assigned 1 where astcur=1 (“Current asthma”) and 0 where astcur=2 (“No current asthma”). The category “No current asthma” is used as the reference in the analysis.

```
DATA newdata;
SET data;
IF astcur=1 THEN ast=1;
ELSE IF astcur=2 THEN ast=0;
RUN;

PROC RLOGIST data = newdata FILETYPE=SAS DESIGN=WR;
WEIGHT rakedw0;
NEST tsvarstr tsvrunit;
SUBGROUP racehpr2 srsex;
LEVELS 7 2;
```

```
REFLEVEL racehpr2 = 6 srsex = 1;
MODEL ast = racehpr2 srsex srage; RUN;
```

SPSS

Mean Calculation

In the following sample code, the distribution of BMI (bmi) is examined by race (racehpr2) and by the interaction between race and sex (racehpr2*srsex).

```
*Sample design specification step*
*   Analysis Preparation Wizard.

CSPLAN ANALYSIS
 /PLAN FILE='\\PATH FOR COMPLEX SURVEY PLAN FILE\FILENAME.csaplan'
 /PLANVARS ANALYSISWEIGHT=RAKEDW0
 /PRINT PLAN
 /DESIGN STRATA= TSVARSTR CLUSTER=TSVRUNIT
 /ESTIMATOR TYPE=WR.

*Analysis*
*   Complex Samples Descriptives.
CSDESCRIPTIVES
 /PLAN FILE = '\\PATH FOR COMPLEX SURVEY PLAN FILE\FILENAME.csaplan'
 /SUMMARY VARIABLES = bmi
 /SUBPOP TABLE = racehpr2 DISPLAY=LAYERED
 /MEAN
 /STATISTICS SE CV POPSIZE CIN (95)
 /MISSING SCOPE = ANALYSIS CLASSMISSING = EXCLUDE.
*   Complex Samples Descriptives.
CSDESCRIPTIVES
 /PLAN FILE = '\\PATH FOR COMPLEX SURVEY PLAN FILE\FILENAME.csaplan'
 /SUMMARY VARIABLES = bmi
 /SUBPOP TABLE = racehpr2 BY sex DISPLAY=LAYERED
 /MEAN
 /STATISTICS SE CV POPSIZE CIN (95)
 /MISSING SCOPE = ANALYSIS CLASSMISSING = EXCLUDE.
```

In SPSS, the sample design specification step should be included before conducting any analysis.

When using concatenated data across adults, adolescents, and/or children, use tsvrunit; when using separate data files, delete the commands associated with tsvrunit.

Frequency Calculation

In the following sample code, the percentage of people who currently have asthma (astcur) is examined by race (racehpr2) and by race and sex (racehpr2*srsex).

```
*Sample design specification step*
*   Analysis Preparation Wizard.

CSPLAN ANALYSIS
 /PLAN FILE='\\PATH FOR COMPLEX SURVEY PLAN FILE\FILENAME.csaplan'
 /PLANVARS ANALYSISWEIGHT=RAKEDW0
 /PRINT PLAN
 /DESIGN STRATA= TSVARSTR CLUSTER=TSVRUNIT
 /ESTIMATOR TYPE=WR.

*Analysis*
*   Complex Samples Crosstabs.
CSTABULATE
 /PLAN FILE = '\\PATH FOR COMPLEX SURVEY PLAN FILE\FILENAME.csaplan'
 /TABLES VARIABLES = astcur BY racehpr2
 /SUBPOP TABLE = srsex DISPLAY=LAYERED
 /CELLS POPSIZE COLPCT
 /STATISTICS SE CV CIN (95)
 /MISSING SCOPE = TABLE CLASSMISSING = EXCLUDE.
```

In SPSS, the sample design specification step should be included before conducting any analysis.

When using concatenated data across adults, adolescents, and/or children, use tsvrunit; when using separate data files, delete the commands associated with tsvrunit.

Linear Regression

In the following sample code, Body Mass Index (bmi) is examined in relation to race (racehpr2), sex (srsex), and age (srage) while controlling for each other. Note that racehpr2 and srsex are categorical variables, and White (racehpr2=6) and Male (srsex=1) are used as their reference categories.

```
*Sample design specification step*
* Analysis Preparation Wizard.

CSPLAN ANALYSIS
 /PLAN FILE='\\PATH FOR COMPLEX SURVEY PLAN FILE\FILENAME.csaplan'
 /PLANVARS ANALYSISWEIGHT=RAKEDW0
 /PRINT PLAN
 /DESIGN STRATA= TSVARSTR CLUSTER=TSVRUNIT
 /ESTIMATOR TYPE=WR.

*Analysis*
RECODE
srsex
 (1=2) (2=1) INTO newsex.
```

```
VARIABLE LABELS newsex ‘NEWSEX’.
EXECUTE.

RECODE
racehpr2
 (1=1) (2=2) (3=3) (4=4) (5=5) (6=7) (7=6) INTO newrace.

VARIABLE LABELS newrace ‘NEWRACEHPR2’.
EXECUTE.

* Complex Samples General Linear Model.
CSGLM bmi BY newsex newrace WITH srage
 /PLAN FILE = ‘\\PATH FOR COMPLEX SURVEY PLAN FILE\FILENAME.csaplan’
 /MODEL newsex newrace srage
 /INTERCEPT INCLUDE=YES SHOW=YES
 /STATISTICS PARAMETER SE CINTERVAL
 /PRINT SUMMARY VARIABLEINFO SAMPLEINFO
 /TEST TYPE=F PADJUST=LSD
 /MISSING CLASSMISSING=EXCLUDE
 /CRITERIA CILEVEL=95.
```

In SPSS, the sample design specification step should be included before conducting any analysis.

When using concatenated data across adults, adolescents, and/or children, use tsvrunit; when using separate data files, delete the commands associated with tsvrunit.

SPSS CSGLM automatically chooses the highest value of the categorical independent variables as the reference groups. Therefore, recoding categorical variables is necessary to select the desired reference categories if they are different than the categories with the highest values.

Logistic Regression

In the following sample code, current asthma status (astcur) is examined, controlling for race (racehpr2), sex (srsex), and age (srage). As SUDAAN and Stata require the dependent variables coded as 0 and 1 for logistic regression, a new dependent variable ast is created and assigned 1 where astcur=1 (“Current asthma”) and 0 where astcur=2 (“No current asthma”). The category “No current asthma” is used as the reference in the analysis.

```
*Sample design specification step*
* Analysis Preparation Wizard.

CSPLAN ANALYSIS
 /PLAN FILE=‘\\PATH FOR COMPLEX SURVEY PLAN FILE\FILENAME.csaplan’
 /PLANVARS ANALYSISWEIGHT=RAKEDW0
 /PRINT PLAN
 /DESIGN STRATA= TSVARSTR CLUSTER=TSVRUNIT
 /ESTIMATOR TYPE=WR.
```

```
*Analysis*

RECODE  srsex
 (1=2) (2=1) INTO newsex.
VARIABLE LABELS newsex 'NEWSEX'.
EXECUTE.

RECODE   racehpr2
 (1=1) (2=2) (3=3) (4=4) (5=5) (6=7) (7=6) INTO newrace.  VARIABLE LAB
ELS newrace 'NEWRACEHPR2'.  EXECUTE.

* Complex Samples Logistic Regression.

CSLOGISTIC astcur BY newsex newrace WITH srage
 /PLAN FILE = '\\PATH FOR COMPLEX SURVEY PLAN FILE\FILENAME.csaplan'
 /MODEL newsex newrace srage
 /INTERCEPT INCLUDE=YES SHOW=YES
 /STATISTICS PARAMETER EXP SE CINTERVAL
 /TEST TYPE=F PADJUST=LSD
 /ODDSRATIOS FACTOR=[newsex]
 /ODDSRATIOS FACTOR=[newrace]
 /ODDSRATIOS COVARIATE=[srage]
 /MISSING CLASSMISSING=EXCLUDE
 /CRITERIA MXITER=100 MXSTEP=5 PCONVERGE=[1e-006 RELATIVE]
  LCONVERGE=[0] CHKSEP=20 CILEVEL=95
/PRINT SUMMARY VARIABLEINFO SAMPLEINFO.
```

In SPSS, the sample design specification step should be included before conducting any analysis.

When using concatenated data across adults, adolescents, and/or children, use tsvrunit; when using separate data files, delete the commands associated with tsvrunit.

SPSS CSLOGISTIC automatically chooses the highest value of the categorical variable as the reference group for the independent variables as well as the dependent variable. Therefore, recoding categorical variables is necessary to select the desired reference categories if they are different than the categories with highest values.

Sample Code to Pool Multiple Data Cycles

For background on pooling two or more data files, see the page on **Pooling CHIS Data**.

The following Stata, SAS, and R codes show how to combine CHIS 2017 and 2018 data files and to create weights accounting for the multiple files. In addition to the sample code, we also provide a SAS macro for the user interested in analyzing CHIS data in SAS. This SAS macro will do all the work of weight adjustments automatically, and also generate other necessary

information needed in later analysis. You can download this in the section CHIS **Pooling Macro** or refer to **video tutorial** on youtube.

R

```
library(openxlsx)
library(haven)
library(dplyr)
library(survey)
library(tidyverse)

# FOLLOWING CODE ONLY WORKS FOR SINGLE YEAR DATA SET FILES NOT TWO-YEA
RS

# Please download all SAS files you want to use and put them all into
one folder location:
folder_location <- "YOUR DIRECTORY/FOLDER LOCATION"

# Copy and paste below for as many years as you want to pool together
and change respective values:
adult17 <- read_sas(paste0(folder_location,"/adult_2017.sas7bdat")) %>
%
  rename_all(tolower) %>%
  mutate(year = 2017)

adult18 <- read_sas(paste0(folder_location,"/adult_2018.sas7bdat")) %>
%
  rename_all(tolower) %>%
  mutate(year = 2018)

# Put all imported data sets into this list

my_chis_list <- list(adult17, adult18)

# DO NOT CHANGE ANYTHING INSIDE THE FUNCTION UNLESS ABSOLUTELY NECESSA
RY

pooling <- function(chis_list) {

  for(i in 1:length(chis_list)) {

    chis_list[[i]][ , paste0("fnwgt", 0:80)] <- chis_list[[i]][ , past
e0("rakedw", 0:80)]

    chis_list[[i]] <-
      chis_list[[i]] %>%
      rename_at(vars(paste0("fnwgt", c(1:80))), ~ paste0("fnwgt", c(1:
80) + 80*(i-1)))
```

```
  }

  chis_list <- chis_list %>%
    map(. %>% mutate(across(everything(), .fns = as.character)))

  merged <-
    bind_rows(chis_list) %>%
    data.frame(., row.names = NULL)

  merged <-
    merged  %>%
    mutate_all(type.convert, as.is = TRUE)

  merged <-
    merged  %>%
    mutate(across(starts_with("fnwgt"), ~ ifelse(is.na(.), fnwgt0, .))
)

  merged <-
    merged %>%
    mutate(across(starts_with("fnwgt"), ~ ./length(chis_list)))

  merged

}

# Store pooled data, resulting data will either be in numeric or chara
cter format (no factors at all).

combined <- pooling(my_chis_list)

# Set up survey design for analysis

chis_design <- svrepdesign(data = combined,
                                  weights =  ~ fnwgt0,
                                  repweights = "fnwgt[1-9]",
                                  type = "other",
                                  scale = 1,
                                  rscales = 1,
                                  mse = TRUE)
```

Stata

```
log using "folder location\data_step.log", replace
```

```
***CHIS 2017 Adult data****
use "your folder location\CHIS 2017 data"

gen year=2017

gen fnwgt0=rakedw0/2

for new fnwgt1-fnwgt160: gen X=0

foreach i of numlist 1/80{
            local j=`i'-0
            replace fnwgt`i'=rakedw`j'/2

}

foreach i of numlist 81/160{
            replace fnwgt`i'=rakedw0/2

}

save adult17 , replace

***CHIS 2018 Adult data****
use "folder location\CHIS 2018 data"

gen year=2018

gen fnwgt0=rakedw0/2

for new fnwgt1-fnwgt160: gen X=0

foreach i of numlist 1/80{
            replace fnwgt`i'=rakedw0/2

}

foreach i of numlist 81/160{
            local j=`i'-80
            replace fnwgt`i'=rakedw`j'/2
}

/*this step concatenates the data files*/
append using adult17

save "folder location\combined.dta", replace

svyset [pw=fnwgt0], jkrw(fnwgt1-fnwgt160, multiplier(1)) vce(jack) mse
```

SAS / SUDAAN

```
data combined; /*this step concatenates the data files*/
    set libname.chis_2017 (in=in17) libname.chis_2018 (in=in18);

    if in17 then year=2017;
    else if in18 then year=2018;

    ***Create new weight variables;
    fnwgt0 = rakedw0/2;
    array a_origwgts[80] rakedw1-rakedw80;
    array a_newwgts[160] fnwgt1-fnwgt160;
    do i = 1 to 80;
        if year=2017 then do;
            a_newwgts[i] = a_origwgts[i]/2;
            a_newwgts[i+80] = rakedw0/2;
        end;
        else if year=2018 then do;
            a_newwgts[i]    = rakedw0/2;
            a_newwgts[i+80] = a_origwgts[i]/2;
        end;

    end;
run;

proc surveyfreq data = combined varmethod=jackknife;
    weight fnwgt0;
    repweight fnwgt1-fnwgt160/jkcoefs=1;
    table ins;
run;
```

Jackknife coefficients are necessary for accurate variance calculations, and jackknife coefficients of 1 in SAS will produce equal variance calculations as those produced in SUDAAN. However, for SAS V.9.2(TS1M0) and earlier, a value of 1 will not be accepted; as a substitute, 0.9999 can be entered. Without this specification, the default value of the jackknife coefficients will be [(# replicate weights - 1)/# replicate weights]; for CHIS, this would be [(80 - 1)/80] = 0.9875.