How do I utilize categorical variables in SUDAAN?

Authored by

stats writer

July 1, 2024

Categorical variables in SUDAAN refer to variables that have discrete values or categories. These variables can be utilized in SUDAAN by specifying them as the independent or dependent variable in statistical analyses. SUDAAN allows for the inclusion of categorical variables in various procedures, such as regression, ANOVA, and chi-square tests. Additionally, SUDAAN provides options for recoding and weighting categorical variables to accurately represent the population being studied. By utilizing categorical variables in SUDAAN, researchers can obtain accurate and precise estimates for complex survey data.

How do I use categorical variables in SUDAAN? | SUDAAN FAQ

Dummy variables

**Suppose that you are running a regression or a logistic regression, and
instead of getting the results that you expect, you find that SUDAAN considers a large portion of
your data to be missing. You use proc print or something to see
your data set, and it looks fine to you. What has happened? You may have
included in your model and on the subgroup statement dummy variables that
are coded 0/1. While this is the standard
way of coding dummy variables, SUDAAN considers the cases coded
as 0 to be missing for variables that are listed on the subgroup**

**statement. In other words, to SUDAAN, non-positive values in variables
that are used as categorical independent variables are considered to be missing. Hence, when SUDAAN does a listwise deletion of missing
data, a large portion of your cases may be deleted, possibly to the point of
making the model unestimatible. (Please see pages 165-166 of the SUDAAN
manual for a complete description regarding the use of the subgroup
statement, including valid values for subgroups, and below for the example using
the subgroup statement.) Consider the example below in which srsex is coded 1/2 and
newvar1 is coded 0/1. As you can see, an
error is printed in the log and the number of cases used in the analysis is
about 4050 fewer than there should be (the 4050 cases that are coded 0 in the
data step). You have several ways of dealing with this problem.
Perhaps the easiest is to not list the 0/1 variable on the subgroup**

**statement. In many ways the subgroup statement in SUDAAN is like
the class statement in SAS. In the same way that you would not list
a 0/1 variable on the class statement in SAS, you do not list a 0/1
variable on the subgroup statement in SUDAAN. Another solution is
the recode the 0/1 variable to be a 1/2 variable. If you have a variable
that is 0/1/2, then you need to recode it. You can do this in a data step
before running the procedure.**

```
data temp01;
set temp1;
newvar1 = 0;
if _n_ ge 4050 then newvar1 = 1;
run;

proc regress data=temp01 filetype=sas design = jackknife;
weight rakedw0;
jackwgts rakedw1--rakedw80 / adjjack=1;
```

model ab1 = srsex newvar1 ;
subgroup srsex newvar1;
levels 2 2;
run;

The following message is displayed in the log.

Opened SAS data file TEMP01 for reading.

DATA WARNING:
The matrix for estimable parameters is singular.
The model may be overspecified. You should reduce the number of
variables on the right-hand side and refit the model before attempting
to draw any conclusions.

DATA WARNING
:
Degrees of freedom for OVERALL contrast are less than maximum number of estimable parameters
You may wish to rerun this job with a tolerance (TOL) of 1.000000e-007 and 1.000000e-005

The erroneous output is shown below.

Number of observations read : 55428 Weighted count: 23847415
Observations used in the analysis : 51339 Weighted count: 22067131
Denominator degrees of freedom : 80

Maximum number of estimable parameters for the model is 3
Weighted mean response is 2.504337

Multiple R-Square for the dependent variable AB1: 0.001315

Variance Estimation Method: Replicate Weight Jackknife
Working Correlations: Independent
Link Function: Identity
Response variable AB1: AB1

Independent P-value
Variables and Beta T-Test
Effects Coeff. SE Beta T-Test B=0 B=0

Intercept 2.54 0.01 323.02 0.0000

SRSEX

MALE -0.08 0.01 -6.56 0.0000

FEMALE 0.00 0.00 . .

NEWVAR1

1 0.00 0.00 . .

2 0.00 0.00 . .

Contrast Degrees

of P-value

Freedom Wald F Wald F

OVERALL MODEL 2 75483.96 0.0000

MODEL MINUS

INTERCEPT 1 43.01 0.0000

INTERCEPT . . .

SRSEX 1 43.01 0.0000

NEWVAR1 . . .

This problem is caused by the dummy variable newvar1. If you

compare the number of cases used by SUDAAN for the analysis above, 51339, you will see that the 4050 cases coded as 0 in the data step above are missing. Although in the example below we have recoded the problem variable in a data step, you could also use the recode statement in SUDAAN to temporarily recode the variable. If you have many variables that need to be recoded, you may want to use an array in a data step. These options are perhaps most useful when you really want to have the dummy variable listed on the subgroup statement, such as when you are using proc crosstabs. As mentioned above, you could also list only the categorical variables coded with non-zero values on the subgroup statement.

```
data temp01a;
set temp01;
if newvar1 = 0 then newvar2 = 1;
if newvar1 = 1 then newvar2 = 2;
```

```
run;

proc regress data=temp01a filetype=sas design = jackknife;
weight rakedw0;
jackwgts rakedw1--rakedw80 / adjjack=1;
model ab1 = srsex newvar2 ;
subgroup srsex newvar2;
levels 2 2;
run;
```

Number of observations read : 55428 Weighted count: 23847415
Observations used in the analysis : 55383 Weighted count: 23829382
Denominator degrees of freedom : 80

Maximum number of estimable parameters for the model is 3
Weighted mean response is 2.502603

Multiple R-Square for the dependent variable AB1: 0.001221

Variance Estimation Method: Replicate Weight

Jackknife

Working Correlations: Independent

Link Function: Identity

Response variable AB1: AB1

Independent P-value

Variables and Beta T-Test

Effects Coeff. SE Beta T-Test B=0 B=0

Intercept 2.54 0.01 326.26 0.0000

SRSEX

MALE -0.07 0.01 -6.39 0.0000

FEMALE 0.00 0.00 . .

NEWVAR2

1 -0.02 0.02 -0.97 0.3343

2 0.00 0.00 . .

Contrast Degrees

of P-value

Freedom Wald F Wald F

OVERALL MODEL 3 60042.62 0.0000

MODEL MINUS

INTERCEPT 2 20.71 0.0000

INTERCEPT . . .

SRSEX 1 40.85 0.0000

NEWVAR2 1 0.94 0.3343

The subgroup statement

In this example we have a 0/1 variable (newvar1) and we are not
using it on the subgroup statement. If you want to have the table
broken out by the values of newvar1, then you need to recode it to be a
1/2 variable and include it on the subgroup statement and include the
number of levels on the levels statement.

```
proc descript data=temp01 filetype=sas design =
jackknife;
weight rakedw0;
jackwgts rakedw1--rakedw80 / adjjack=1;
```

var srsex racehpra newvar1;

subgroup srsex racehpra;

levels 2 2;

run;

Number of observations read : 55428 Weighted count : 23847415

Denominator degrees of freedom : 80

Variance Estimation Method: Replicate Weight Jackknife

by: Variable, Self-reported gender.

| | |

| Variable | | Self-reported gender

| | | Total | MALE | FEMALE |

| | | | | |

| Self-reported | Sample Size | 55428 | 23002 | 32426 |

| gender | Weighted Size | 23847415.32 | 11631728.37 | 12215686.95 |

	Total	36063102.27	11631728.37	24431373.90
	Mean	1.51	1.00	2.00
	SE Mean	0.00	0.00	0.00

Race - UCLA	Sample Size	9677	4084	5593
CHPR Definition	Weighted Size	5705917.88	2866894.01	2839023.87
	Total	5767889.98	2897175.85	2870714.13
	Mean	1.01	1.01	1.01
	SE Mean	0.00	0.00	0.00

NEWVAR1	Sample Size	55428	23002	32426
	Weighted Size	23847415.32	11631728.37	12215686.95
	Total	22084052.10	10772176.06	11311876.04
	Mean	0.93	0.93	0.93
	SE Mean	0.00	0.00	0.00

| | |

| Variable | | Race - UCLA CHPR Definition

| | | Total | LATINO | PACIFIC |

| | | | | ISLANDER |

| | | | | |

| Self-reported | Sample Size | 9677 | 9458 | 219 |

| gender | Weighted Size | 5705917.88 | 5643945.79 | 61972.10 |

| | Total | 8544941.75 | 8451279.40 | 93662.35 |

| | Mean | 1.50 | 1.50 | 1.51 |

| | SE Mean | 0.01 | 0.01 | 0.04 |

| | | | | |

| Race - UCLA | Sample Size | 9677 | 9458 | 219 |

| CHPR Definition | Weighted Size | 5705917.88 | 5643945.79 | 61972.10 |

| | Total | 5767889.98 | 5643945.79 | 123944.19 |

| | Mean | 1.01 | 1.00 | 2.00 |

| | SE Mean | 0.00 | 0.00 | 0.00 |

| | | | | |

| NEWVAR1 | Sample Size | 9677 | 9458 | 219 |

| | Weighted Size | 5705917.88 | 5643945.79 | 61972.10 |

| | Total | 5275702.50 | 5218781.91 | 56920.60 |

| | Mean | 0.92 | 0.92 | 0.92 |

| | SE Mean | 0.00 | 0.00 | 0.03 |
