How can I incorporate categorical independent variables into regression analyses using SUDAAN?

Authored by

stats writer

July 1, 2024

The process of incorporating categorical independent variables into regression analyses using SUDAAN involves first identifying the categorical variables in the dataset. These variables should then be coded into numerical values in order to be used in the regression analysis. Next, the SUDAAN software should be utilized to perform a weighted analysis, taking into account the complex survey design and potential sampling errors. The categorical variables can then be included in the regression model as dummy variables to examine their effects on the dependent variable. This method ensures accurate and reliable results by properly accounting for the survey design and potential biases.

How can I use categorical independent variables in regression analyses in SUDAAN? | SUDAAN FAQ

Using the class statement

**The class statement is available in SUDAAN 9. You use it in the
same way that you use the class statement works in SAS: you list
categorical variables on this statement so that those variables are not treated
as continuous variables by the program. Dummy variables (0/1 variables) do
not need to be listed on the class statement. If you include
srsex on the class statement, the results will exactly match those obtained
using the subgroup and levels statements. In this example, srsex is coded 1 = male and 2 = female, and**

racehpra is coded 1 = Latino, 2 = Pacific Islander, 2 = AIAN, 4 = Asian,
5 = African American, 6 = White and 7 = Other.

NOTE: The class statement in SUDAAN uses dummy coding (0 1, or
what SAS calls reference coding). The class statement in SAS uses
effect coding (-1 1). Assuming that you had a two-level categorical
variable, in SUDAAN the reference category is coded 0 0, while in SAS it is
coded -1 -1. This means that if you run an analysis in both SAS and SUDAAN
using the class statement, the coefficients of the dummies of the categorical
variable will not match. To get the results to match, use the param =
ref option on the class statement in SAS. You can tell what
kind of coding is being used by looking at the top of the output (in either
SUDAAN or SAS).

proc regress data=chis filetype=sas design = jackknife;
weight rakedw0;
jackwgts rakedw1--rakedw80 / adjjack=1;
model ab1 = srsex racehpra;
class racehpra;
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 8
Weighted mean response is 2.502603

Multiple R-Square for the dependent variable AB1: 0.043591

Frequencies and Values for CLASS Variables
by: SRSEX.

SRSEX Frequency Value

Ordered

Position:

1 23002 1

Ordered

Position:

2 32426 2

Frequencies and Values for CLASS Variables

by: RACEHPRA.

RACEHPRA Frequency Value

Ordered

Position:

1 9458 1

Ordered

Position:

2 219 2

Ordered

Position:

3 781 3

Ordered

Position:

4 3956 4

Ordered

Position:

5 2764 5

Ordered

Position:

6 36729 6

Ordered

Position:

7 1521 7

Variance Estimation Method: Replicate Weight Jackknife

Working Correlations: Independent

Link Function: Identity

Response variable AB1: AB1

by: Independent Variables and Effects.

Independent

Variables and Beta Lower 95% Upper 95%
Effects Coeff. SE Beta Limit Beta Limit Beta T-Test B=0

Intercept 2.398056 0.047290 2.303947 2.492165 50.710087
SRSEX 0.078808 0.011561 0.055801 0.101815 6.816660
RACEHPRA
LATINO 0.347818 0.041945 0.264346 0.431291 8.292332
PACIFIC ISLANDER 0.000548 0.115201 -0.228710 0.229806 0.004759
AIAN 0.221383 0.071965 0.078169 0.364598 3.076281
ASIAN -0.005809 0.043640 -0.092656 0.081038 -0.133110
AFRICAN AMERICAN 0.103354 0.044022 0.015748 0.190960 2.347804
WHITE -0.184855 0.041252 -0.266949 -0.102761 -4.481135
OTH SINGL/MULTI
RACE 0.000000 0.000000 0.000000 0.000000 .

Independent P-value

Variables and T-Test

Effects B=0

Intercept 0.000000

SRSEX 0.000000

RACEHPRA

LATINO 0.000000

PACIFIC ISLANDER 0.996215

AIAN 0.002869

ASIAN 0.894440

AFRICAN AMERICAN 0.021355

WHITE 0.000024

OTH SINGL/MULTI

RACE .

--

Contrast Degrees

of P-value

Freedom Wald F Wald F

--

OVERALL MODEL 8.000000 ******** 0.000000**

MODEL MINUS

INTERCEPT 7.000000 178.070752 0.000000

INTERCEPT . . .

SRSEX 1.000000 46.466848 0.000000

RACEHPRA 6.000000 183.206577 0.000000

Using the subgroup and levels statements

If you are using an earlier version of SUDAAN or if you want more control over the handling of your categorical variables, you can use the subgroup and levels statements. For each variable listed on the subgroup statement, you need to list the number of levels of categories that each variable has.

By default, the last category (i.e., the highest numbered category) is used as the reference category when you have categorical predictors in a regression model.

proc regress data=chis filetype=sas design = jackknife;

weight rakedw0;
jackwgts rakedw1--rakedw80 / adjjack=1;
model ab1 = srsex racehpra;
subgroup srsex racehpra;
levels 2 7;
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 8
Weighted mean response is 2.502603

Multiple R-Square for the dependent variable AB1: 0.043591

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

Response variable AB1: AB1

by: Independent Variables and Effects.

Independent

Variables and Beta Lower 95% Upper 95%

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

Intercept 2.555672 0.040901 2.474277 2.637067 62.484879

SRSEX

MALE -0.078808 0.011561 -0.101815 -0.055801 -6.816660

FEMALE 0.000000 0.000000 0.000000 0.000000 .

RACEHPRA

LATINO 0.347818 0.041945 0.264346 0.431291 8.292332

PACIFIC ISLANDER 0.000548 0.115201 -0.228710 0.229806 0.004759

AIAN 0.221383 0.071965 0.078169 0.364598 3.076281

ASIAN -0.005809 0.043640 -0.092656 0.081038 -0.133110

AFRICAN AMERICAN 0.103354 0.044022 0.015748 0.190960 2.347804

WHITE -0.184855 0.041252 -0.266949 -0.102761

-4.481135

OTH SINGL/MULTI

RACE 0.000000 0.000000 0.000000 0.000000 .

Independent P-value

Variables and T-Test

Effects B=0

Intercept 0.000000

SRSEX

MALE 0.000000

FEMALE .

RACEHPRA

LATINO 0.000000

PACIFIC ISLANDER 0.996215

AIAN 0.002869

ASIAN 0.894440

AFRICAN AMERICAN 0.021355

WHITE 0.000024

OTH SINGL/MULTI

RACE .

Contrast Degrees
of P-value
Freedom Wald F Wald F

OVERALL MODEL 8.000000 ******** 0.000000**
MODEL MINUS
INTERCEPT 7.000000 178.070752 0.000000
INTERCEPT . . .
SRSEX 1.000000 46.466848 0.000000
RACEHPRA 6.000000 183.206577 0.000000

Changing the default reference category

To change the reference category, you can use the reflevel statement.
In this example, we have changed the reference category for both variables.
Also, we have used only the first four categories of the race variable,
racehpra. If you want to use only some of the

categories, you can
recode the variable such that the categories that you want to use are the first
ones (e.g., coded 1, 2, 3, etc.) and then just give the number of desired
categories on the levels statement.

```
proc regress data=chis filetype=sas design = jackknife;
weight rakedw0;
jackwgts rakedw1--rakedw80 / adjjack=1;
reflevel racehpra = 2 srsex = 1 ;
model ab1 = srsex racehpra;
subgroup srsex racehpra;
levels 2 4;
run;
```

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

Intercept 2.46 0.11 22.88 0.0000
SRSEX
MALE 0.00 0.00 . .
FEMALE 0.11 0.03 4.51 0.0000

RACEHPRA

LATINO 0.35 0.11 3.27 0.0016

PACIFIC ISLANDER 0.00 0.00 . .

AIAN 0.22 0.12 1.91 0.0598

ASIAN -0.01 0.11 -0.06 0.9562

Contrast Degrees

of P-value

Freedom Wald F Wald F

OVERALL MODEL 5 12285.21 0.0000

MODEL MINUS

INTERCEPT 4 56.79 0.0000

INTERCEPT . . .

SRSEX 1 20.36 0.0000

RACEHPRA 3 71.68 0.0000

Using only some of the categories in a categorical variable

You can specify just some of the levels of a categorical variable by

listing only the desired levels on the catlevel statement.

```
proc descript data=chis filetype=sas design = jackknife;
weight rakedw0;
jackwgts rakedw1--rakedw80 / adjjack=1;
var srsex racehpra racehpra racehpra;
catlevel 1 1 3 5;
setenv colwidth=12;
print nsum wsum total setotal;
run;
```

Variable		One
		1
SRSEX: MALE	Sample Size	55428
	Weighted Size	23847415.32
	Total	11631728.37
	SE Total	515.26
RACEHPRA:	Sample Size	55428

LATINO	Weighted Size	23847415.32
	Total	5643945.79
	SE Total	28469.00

RACEHPRA: AIAN	Sample Size	55428
	Weighted Size	23847415.32
	Total	85146.30
	SE Total	4008.83

RACEHPRA:	Sample Size	55428
AFRICAN	Weighted Size	23847415.32
AMERICAN	Total	1387993.65
	SE Total	11564.25

The coding of categorical variables

The numerical values used for the codes of a categorical variable are very important. Values of variables listed on the subgroup statement must be positive or else they are considered as

**missing. This means that
you cannot list a 0/1 dummy variable on the subgroup statement. In
most regression analyses, this is not a problem; you just include the variable
in the model and not on the subgroup statement (just as you would not
include a dummy variable on a class statement in SAS). However, in
other procedures, such as proc descript, you may want to include a dummy
variable on the subgroup statement. In this case, you would want to
recode the variable either using the recode statement or in a SAS data
step.**