

The SAS ROBREG9 Macro

Ellen Hertzmark and Donna Spiegelman

April 14, 2010

Abstract

The %ROBREG9 macro is a SAS version 9 macro that runs robust linear regression models showing both the model-based (assuming normality) and empirical standard errors, for situations where it is reasonable to use PROC REG (i.e. no repeated measures, continuous dependent variable). This macro can also calculate point and interval estimates of effect on the (unitless) percent change scale, which is often more widely interpretable.

Keywords: SAS, macro, PROC REG, empirical variance, robust variance

Contents

1	Description	2
2	Invocation	2
3	Examples	3
4	Details	14
5	References	15
6	Credits	15

1 Description

%ROBREG9 is a SAS version 9 macro that gives the empirical standard errors and p -values, equivalent to PROC MIXED empirical with TYPE=SIMPLE, when there are no repeated measures. Using this macro instead of PROC MIXED empirical with TYPE=SIMPLE will often result in a substantial reduction of CPU time.

2 a

nd DetailsInvocation

3 Invocation and Details

To call %ROBREG9, your program must know where to look for it. The most efficient way is to include the following line (or its equivalent) at the top of your program.

```
options mautosource sasautos='/usr/local/channing/sasautos';
```

After creating an analysis file, you call %ROBREG9 as follows:

```
%robreg9(  
  data=      name of data set on which the regression is to be run  
             REQUIRED  
  
  depend=   name of the dependent variable  
             REQUIRED  
  
  independ= list of the model variables  
             REQUIRED
```

byvar= "BY" variables, if any.
OPTIONAL

where= a subsetting statement
OPTIONAL

exp= whether you want to do the analysis on the log scale to
compute percent difference in the dependent variable.
default=F

estdat= the name of a data set containing "observations"
at which to compute predicted values.
Each observation in the data set must have a value
for every variable in the model.
OPTIONAL

test1= contrast that can be done.
to make sure that SAS understands what you want,
it is probably safest to put the test in %quote().
if we want to test whether a 1 gram decrease in fat
intake is equivalent to a 2 gram increase in
alcohol intake,
we write

$$\text{test1}=\%quote(2*\text{alco86n} = \text{tfat86n}),$$
or
$$\text{test1}=\%quote(2*\text{alco86n} - \text{tfat86n} = 0),$$
or just
$$\text{test1}=\%quote(2*\text{alco86n} - \text{tfat86n}),$$
(the '=0' is assumed)
The tests are then shown with the labels test1, test2, etc.
See Example 3 below.
OPTIONAL

...

test5= contrast that can be done

inc1= increment for a continuous variable so that the coefficient
relates to an 'interesting' difference in the covariate.
The form is

$$\text{inc1} = \langle \text{variable name} \rangle \langle \text{increment} \rangle.$$
inc1=age86 5,

means that the increment for age86 is 5 years.
See example 3 below.
The order of these parameters is not important
(i.e. they do not have to be in the same order
as the variables are listed in the model).
OPTIONAL

...
inc20= increment for a continuous variable...

4 Examples

Using a data set from HPFS, we examine the relationship between BMI and a number of possible correlates, cross-sectionally in 1986.

BMI86 is the individual's BMI in 1986
age86 is the individual's age (in years) in 1986
tfat86n is the individual's daily intake of total fat
in grams per day in 1986
alco86n is the individual's daily intake of alcohol
in grams per day in 1986
smk86 is the individual's smoking status in 1986
(0=non-smoker, 1=smoker)

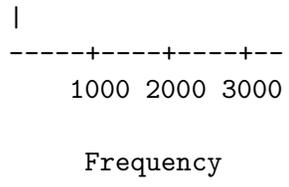
The basic data set is called ALL1X.

The trimmed data set ALL1 is a data set made from ALL1X by deleting observations with alcohol intake over 45 or fat intake over 125 or BMI outside the range of 18-45 or caloric intake outside the range of 1000-3200 .

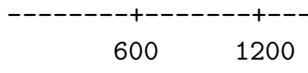
```
data all1; set all1x;
where alco le 45 and fat le 125 and 18 le bmi86 le 45 and 1000 le calor le 3200;
run;
```

Alcohol intake is highly skewed, and fat intake is also skewed, as shown by the stem-and-leaf plots below. Although highly skewed independent variables can lead to the presence of one or more underlying influential points, it should be noted that regression models never require normality assumptions on the *independent* variables.

Alcohol gm Midpoint		Freq	Cum. Freq	Percent	Cum. Percent
0	*****	3371	3371	30.33	30.33
4	*****	1957	5328	17.61	47.94
8	*****	1324	6652	11.91	59.85
12	*****	1236	7888	11.12	70.97
16	*****	984	8872	8.85	79.83
20	**	499	9371	4.49	84.32
24	*	243	9614	2.19	86.50
28	*	196	9810	1.76	88.27
32	*	218	10028	1.96	90.23
36	**	326	10354	2.93	93.16
40	*	201	10555	1.81	94.97
44	*	121	10676	1.09	96.06
48	*	104	10780	0.94	96.99
52		40	10820	0.36	97.35
56		49	10869	0.44	97.80
60		37	10906	0.33	98.13
64		46	10952	0.41	98.54
68		52	11004	0.47	99.01
72		23	11027	0.21	99.22
76		27	11054	0.24	99.46
80		14	11068	0.13	99.59
84		17	11085	0.15	99.74
88		8	11093	0.07	99.81
92		3	11096	0.03	99.84
96		4	11100	0.04	99.87
100		8	11108	0.07	99.95
104		1	11109	0.01	99.96
108		1	11110	0.01	99.96
112		0	11110	0.00	99.96
116		2	11112	0.02	99.98
120		0	11112	0.00	99.98
124		0	11112	0.00	99.98
128		0	11112	0.00	99.98
132		1	11113	0.01	99.99
136		0	11113	0.00	99.99
140		1	11114	0.01	100.00



Total Fat gm Midpoint		Freq	Cum. Freq	Percent	Cum. Percent
16		14	14	0.13	0.13
24	**	129	143	1.16	1.29
32	*****	416	559	3.74	5.03
40	*****	837	1396	7.53	12.56
48	*****	1218	2614	10.96	23.52
56	*****	1354	3968	12.18	35.70
64	*****	1413	5381	12.71	48.42
72	*****	1338	6719	12.04	60.46
80	*****	1152	7871	10.37	70.82
88	*****	872	8743	7.85	78.67
96	*****	661	9404	5.95	84.61
104	*****	536	9940	4.82	89.44
112	*****	384	10324	3.46	92.89
120	****	265	10589	2.38	95.28
128	**	175	10764	1.57	96.85
136	**	119	10883	1.07	97.92
144	*	78	10961	0.70	98.62
152	*	51	11012	0.46	99.08
160	*	42	11054	0.38	99.46
168		27	11081	0.24	99.70
176		10	11091	0.09	99.79
184		10	11101	0.09	99.88
192		4	11105	0.04	99.92
200		2	11107	0.02	99.94
208		3	11110	0.03	99.96
216		2	11112	0.02	99.98
224		0	11112	0.00	99.98
232		1	11113	0.01	99.99
240		0	11113	0.00	99.99
248		0	11113	0.00	99.99
256		0	11113	0.00	99.99
264		1	11114	0.01	100.00



White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrics* 1980; 48:817-838.

6 Credits

Written by Ellen Hertzmark and Donna Spiegelman for the Channing Laboratory. Questions can be directed to Ellen Hertzmark, stleh@channing.harvard.edu, (617) 432-4597.

7 See Also