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

%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
                  test1=%quote(2*alco86n = tfat86n),
          or      test1=%quote(2*alco86n - tfat86n = 0),
          or just test1=%quote(2*alco86n - 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
              inc1 = <variable name> <increment>.
         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                                  Cum.               Cum.
Midpoint                          Freq      Freq   Percent   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
```

```
|
-----+----+----+--
    1000 2000 3000

     Frequency
```

```
Total Fat gm                            Cum.              Cum.
Midpoint                         Freq   Freq  Percent  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
          |
          --------+-------+---
                600     1200
```

```
                Frequency
```

[[

NOTE also that we include the predictors as linear continuous variables. Unless linearity of the association is carefully investigated and verified, linear continuous variables should not be entered in models. We do this here only to illustrate.

Example 1. Basic macro call – untrimmed data

The basic macro call (using only the three required parameters) is

```
title2 '1986--untrimmed data';
%robreg9(data=all1x, depend=bmi86, independ=age86 tfat86n alco86n smk86);
```

The results are

[[

```
/udd/stleh/helpme/pkb/robrbase.sas         14:16 Wednesday, April 14, 2010  57
1986--untrimmed data


Data set is all1x   Dependent variable is bmi86

  # obs=8465 , R-squared=0.0093

                                                        emp lower emp upper
                    Model-    Model- Empirical Empirical  95% conf  95% conf
varname   Estimate based SE based P      SE         P       bound     bound

INTERCEPT  23.3589  0.19601  0.0000  0.20143    0.0000    22.9641  23.7537
AGE86       0.0169  0.00341  0.0000  0.00354    0.0000     0.0099   0.0238
TFAT86N     0.0080  0.00110  0.0000  0.00117    0.0000     0.0057   0.0103
ALCO86N     0.0037  0.00203  0.0682  0.00207    0.0737    -0.0004   0.0077
SMK86      -0.1361  0.11731  0.2462  0.12320    0.2694    -0.3775   0.1054
```

[[

The macro tells you the number of observations and the value of R-squared. Then it gives the point estimates of the coefficients and both the model-based and empirical standard errors and p-values.

Example 2. Untrimmed data with WHERE and BYVAR parameters

This is the same example, but restricting to men under 65 years old stratified by smoking status.

The macro call is

```
%robreg9(data=all1x, depend=bmi86, independ=age86 tfat86n alco86n ,
byvar=smk86, where=age86 lt 65);
```

The results are

```
[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[

/udd/stleh/helpme/pkb/robrbase.sas         14:16 Wednesday, April 14, 2010  58
1986--untrimmed data with WHERE parameter and BY variable


Data set is all1x   Dependent variable is bmi86
where  age86 lt 65

smk86=.   # obs=91 , R-squared=0.0692

                                                              emp lower emp upper
                    Model-   Model- Empirical Empirical  95% conf  95% conf
varname   Estimate based SE based P      SE         P       bound     bound

INTERCEPT  28.8442  2.03200  0.0000  1.72191    0.0000    25.4693   32.2192
AGE86      -0.0472  0.03930  0.2326  0.03364    0.1602    -0.1132    0.0187
TFAT86N    -0.0211  0.00962  0.0306  0.00790    0.0075    -0.0366   -0.0056
ALCO86N     0.0100  0.01521  0.5114  0.01578    0.5253    -0.0209    0.0410


smk86=0   # obs=7153 , R-squared=0.0136
```

```
                                                         emp lower emp upper
                    Model-   Model- Empirical Empirical  95% conf  95% conf
varname   Estimate based SE based P     SE        P        bound     bound

INTERCEPT  22.7953  0.24110  0.000   0.23907    0.0000    22.3267  23.2639
AGE86       0.0268  0.00451  0.000   0.00448    0.0000     0.0180   0.0356
TFAT86N     0.0092  0.00119  0.000   0.00126    0.0000     0.0068   0.0117
ALCO86N     0.0040  0.00229  0.082   0.00226    0.0779    -0.0004   0.0084


smk86=1   # obs=563 , R-squared=0.0005

                                                         emp lower emp upper
                    Model-   Model- Empirical Empirical  95% conf  95% conf
varname   Estimate based SE based P     SE        P        bound     bound

INTERCEPT  24.8982  0.91156  0.0000  1.28098    0.0000    22.3874  27.4089
AGE86      -0.0050  0.01673  0.7673  0.02421    0.8379    -0.0524   0.0425
TFAT86N     0.0016  0.00436  0.7097  0.00468    0.7283    -0.0075   0.0108
ALCO86N    -0.0014  0.00610  0.8170  0.00643    0.8262    -0.0140   0.0112

[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[
```

NOTE that the macro has told you that the analysis data set was restricted using a WHERE parameter.

NOTE that there is a group of men for whom SMK86 is unknown. Since we are probably not interested in results in this small group, we could use the WHERE parameter to exclude them. In that case, the macro call would have

```
        where = age86 lt 65 and smk86 ne .
```

Example 3. Trimmed data with increments and estimating points (ESTDAT) and a test

The data set ESTDAT was made using the following code.

```
/* data set of points at which want to estimate bmi */
```

```
data estdat;
age86=60;  tfat86n=70;  alco86n=5;  smk86=0;  output;
age86=60;  tfat86n=50;  alco86n=5;  smk86=0;  output;
age86=60;  tfat86n=70;  alco86n=0;  smk86=0;  output;
age86=65;  tfat86n=60;  alco86n=0;  smk86=0;  output;
age86=65;  tfat86n=60;  alco86n=0;  smk86=1;  output;
run;
```

ESTDAT could also have been made by reading a file.

The macro call is

```
%robreg9(data=all1, depend=bmi86, independ=age86 tfat86n alco86n smk86,
inc1=age86 5, inc2=tfat86n 5, inc3=alco86n 10, estdat=estdat,
test1=%quote(tfat86n=2*alco86n) );
```

The increments correspond to 'interesting' changes in the values of the variables, such as 5 years of age, 5 grams of fat, 10 grams of alcohol (1 drink).

In addition, we are interested in testing whether the effects of alcohol and fat are inversely proportional to their caloric contributions, so we do a test. Since fat is twice as energy-dense as alcohol, we multiply the coefficient of alcohol by 2 to test whether a 2 gram increase in alcohol is the same as a 1 gram increase in fat. Note that we used `%quote` on the test condition, because it contains an `=`. We could also have used `%str`. The results are

```
[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[
/udd/stleh/helpme/pkb/robrbase.sas        14:16 Wednesday, April 14, 2010  59
1986--trimmed data, with increments and estimating points
testing whether fat effect is twice as large as alcohol effect

Data set is all1   Dependent variable is bmi86

  # obs=7775 , R-squared=0.0075

                                                        emp lower emp upper
                    Model-   Model- Empirical Empirical  95% conf  95% conf
varname   Estimate based SE based P     SE        P       bound     bound
```

INTERCEPT	23.3339	0.20660	0.0000	0.20401	0.0000	22.9340	23.7337
AGE86	0.0903	0.01751	0.0000	0.01758	0.0000	0.0558	0.1247
TFAT86N	0.0397	0.00675	0.0000	0.00664	0.0000	0.0267	0.0527
ALCO86N	-0.0058	0.02890	0.8414	0.02924	0.8433	-0.0631	0.0515
SMK86	-0.0662	0.12538	0.5974	0.12879	0.6071	-0.3187	0.1862

[[

/udd/stleh/helpme/pkb/robrbase.sas 14:16 Wednesday, April 14, 2010 60
1986--trimmed data, with increments and estimating points
testing whether fat effect is twice as large as alcohol effect

Data set is all1 Dependent variable is bmi86

estimates at specific data values

age86	Total Fat gm	Alcohol gm	smk86	Predicted Value of bmi86	Lower Bound of 95% C.I. for Mean
60	70	5	0	24.9699	24.8766
60	50	5	0	24.8111	24.7069
60	70	0	0	24.9728	24.8672
65	60	0	0	24.9837	24.8526
65	60	0	1	24.9174	24.6471

Upper Bound of 95% C.I. for Mean	Lower Bound of 95% C.I.(Individual Pred)	Upper Bound of 95% C.I.(Individual Pred)
25.0632	19.6858	30.2540
24.9153	19.5268	30.0954
25.0784	19.6885	30.2571
25.1147	19.6988	30.2685
25.1878	19.6273	30.2076

```
[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[

/udd/stleh/helpme/pkb/robrbase.sas        14:16 Wednesday, April 14, 2010  61
1986--trimmed data, with increments and estimating points
testing whether fat effect is twice as large as alcohol effect

Data set is all1   Dependent variable is bmi86


results of tests

                                       p for        p for
                                      ols std     empirical
Obs    Test          testing            err        std err

 1     test1    tfat86n=2*alco86n    0.3804      0.3855

[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[
```

NOTE: Since the p-value for the test is not significant, we say that there is no evidence that alcohol and fat affect BMI through any mechanism other than their energy content.

Example 4. Trimmed data with a contrast and exponentiated coefficients

Sometimes the linear model for the conditional mean as a function of the model covariates fits better on the log scale (multiplicative model). Here our dependent variable is lbmi86=log(bmi86). Again using the trimmed data set ALL1, we demonstrate other features of ROBREG9.

Our model is now

```
log(bmi)=intercept + b1*age86 + b2*tfat86n + b3*alco86n + smk86
```

Because the model predicts the dependent variable on the log scale, but we are really interested in the original scale, we use

```
exp=T
```

to give the percent difference in BMI for each covariate. The increment parameters can be used here to get percent differences for 'interesting' changes in the continuous covariates.

The macro call is

```
%robreg9(data=all1, depend=lbmi86, exp=T, independ=age86 tfat86n alco86n smk86,
inc1=age86n 5, inc2=tfat86n 5, inc3=alco86n 10);
```

The results are

```
[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[

/udd/stleh/helpme/pkb/robrbase.sas         14:41 Wednesday, April 14, 2010  63
1986-trimmed data
outcome is log(bmi), so we use EXP=T
using test1 parameter
Data set is all1   Dependent variable is lbmi86

exponentiated

  # obs=7775 , R-squared=0.0077

               Percent      Model-     Empirical    Lower 95%    Upper 95%
varname      difference    based P         P        CL % diff    CL % diff

INTERCEPT       2226.2      0.0000      0.0000       2189.8       2263.2
AGE86              0.4      0.0000      0.0000          0.2          0.5
TFAT86N            0.2      0.0000      0.0000          0.1          0.2
ALCO86N            0.0      0.9719      0.9722         -0.2          0.2
SMK86             -0.3      0.5092      0.5286         -1.3          0.7
[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[
```

5 References

Huber, P.J.: The behavior of maximum likelihood estimates under nonstandard conditions. Proc Fifth Berkeley Symposium Math. Statist. Prob., 1967; 1:221-233.

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