Abstract

A SAS program, Robust, has been developed to compute parameter estimates from an Interactively Reweight Least Square procedure. This program uses a WLS regression from the PROC REG procedure, the case weight is calculated by Huber's function (C=1.345) or Tukey's biweight function (C=4.6845). The macro procedure is shown to be very flexible and allows user input data set, run bivariate or multiple robust regression model. In addition, for comparison, the OLS estimates, $R^2$, adjusted $R^2$, and total sample size are also listed in the output.

Introduction

The Robust regression statistics procedure is designed to work well, even if the underlying assumptions are somewhat in error. A classic OLS is the unbiased estimator, given the data is normal distributed. A couple of important cases of Robust estimates, P. Huber's M-estimates and John Tukey's biweight estimator, play a major role in the research. The Interactively Reweight Least Square approach (proposed by Guoying Li, in Hoaglin, Mosteller, and Tukey, 1985) started with OLS method and integrated convergence of Huber estimation and Tukey's biweight estimation. The Huber iteration provides robust starting values for the biweight.

The IRLS involves the following steps:

1. Obtain starting values. Use OLS to compute residuals and a MAD-based scale estimate.
   
   where MAD = Median[|residual-median (residual)|].
   
   we use:
   
   scale estimate is $s = \frac{\text{MAD}}{0.6745}$
   
   PROC univariate;
   
   VAR $y$ - media;
   
   Output out = m
   
   Median = MAD;
   
   To GED value MAD

2. Huber estimation:

   Use Huber function with $C = 1.345$.

   Define case weights ($w_i$):
   
   $w_i = 1$ if $|e_i|/s_i \leq 1.345$
   
   $w_i = \frac{1.345/|e_i|/s_i}{1.345}$ if $|e_i|/s_i > 1.345$

Run WLS regression as shown in the following:

```
PROC Reg outest = new;
model $y = x$;
weight weight;
output out = $y$
predicted = yhat residual = $y - y$
```

iteration continues until the maximum change in weights value from one iteration to the next one is less than .05.

3. Tukey estimation:

   Calculate case weight by using the biweight function with $c = 4.685$.

   
   $w_i = \left\{\begin{array}{ll}
   1 - \frac{(e_i/|s_i|)^2}{4.685^2} & \text{if} |e_i|/s_i \leq 4.685 \\
   0 & \text{if} |e_i|/s_i > 4.685
   \end{array}\right.$

Run WLS regression model as shown above. Calculate the maximum change in weights from one iteration to the next, if the value is less than 0.01, iteration stops.

The outputs (Appendix 2.) provides the final set of IRLS estimates.

Macro Robust Description and Use

Attached is a listing of the SAS Macro Robust program to compute the OLS estimate and IRLS (Robust) estimates.

The program requires the user to input the following:

1) dataset name
2) dependent variable
3) sets of independent variable.

Example

An example is from regression with graphics by Lawrence C. Hamilton for Hydrocarbon pollution potential mortality rate on 60 cities of the Metropolitan area and results from robust regression of mortality on $\log_e$ (Hydrocarbon Pollution).

The results are shown in Appendix 2.
Discussion

Using robust estimation gives better statistics properties than OLS when your underlying data assumption is not normally distributed. As described by Lawrence C. Hamilton (1992), OLS is simpler and preferable if OLS and IRLS produce the same results; if we get larger discrepancy between the two methods, we should lean toward the robust results.

References


The author may be contacted at:
University of Southern California School of Medicine
Department of Preventive Medicine
Institute for Health Promotion and Diseases Prevention Research
1000 S. Fremont Ave., Suite 641
Alhambra, CA 91803-1358
(818) 457-4064, (818) 457-4100/FAX

Appendix 1

*************************************************************************************
* ROBUST.SAS
* Programmer: Eric Wang
* Input: SAS data set, dependent var, independent var
* Output: OLS IRLS estimates
*************************************************************************************

options is=120 ps=60 nodate nonumber center MISSING='';
libname 'c:sugi93';

%macro media;
proc univariate noprint data=r ;
var r y ;
output out=s
median=mad ;
data m ;
if _N_ = 1 then set s ;
set r ;
dif=abs(r y -mad );
proc univariate noprint data=m ;
var dif;
output out=s1

median=mad ;
data m1 ;
if _N_ = 1 then set s1 ;
set r ;
%mend;

*************************************************************************************
%macro getmax;
data maximum(keep=max);
length diff maxdiff max 4 ;
set m1 end=end file ;
diff = abs(w-w0);retain maxdiff ;
if (maxdiff < dif ) then maxdiff = diff ;
if end_file then do; max=100*maxdiff;
output;
end;
* proc print;
data m1 ;
if _N_ = 1 then set maximum ;
set m1 ;
call symput('m',left(put(max,8.)));
%mend;

%macro huba;
proc reg data=m1(drop=max) outest=new;
model y =x ;
weight w ;
output out=r(drop=w)
predicted=yphat residual=r_y ;
%media
s0=mad/0.6745 ;
if ( abs(r_y )/s0 < = 1.345 ) then w =1 ;
if ( abs(r_y )/s0 > 1.345 ) then w =1.345/(abs(r_y )/s0 ) ;
drop r_y ;
%getmax
w0=w0;
* proc print;
*var w0 w max;
*************************************************************************************
%macro tukey;
proc reg data=m1(drop=max) outest=new;
model y =x ;
weight w ;
output out=r(drop=w)
predicted=yphat residual=r_y ;
%media

962
\[ s_0 = \text{mad} / 0.6745; \]
if \((\text{abs}(r_y) / s_0 \leq 4.685)\) then \(w = (1 - ((r_y / s_0) / 4.685)^2)^2; \)
if \((\text{abs}(r_y) / s_0 > 4.685)\) then \(w = 0; \)
drop \(r_y; \)

\[
\text{%getmax} \;
\]
\[
w_0 = w; \;
\]
*proc print; \;
*var w0 w max; \;
******************************************************************************;
%mend; \;

%macro ROBUS(data=,dev=,indev=); \;
****************************************************************************** OLS **********;
%macro ROBUS(data=,dev=,indev=); \;
proc reg data=&data outest=&innew covout; \;
model &dev= &indev; \;
output out=r predicted=yphat residual=r_y; \;
%mend; \;
%media \;
******************************************************************************;
w0=0; \;

\[
s_0 = \text{mad} / 0.6745; \)
if \((\text{abs}(r_y) / s_0 \leq 1345)\) then \(w = 1; \)
if \((\text{abs}(r_y) / s_0 > 1345)\) then \(w = 1345 / (\text{abs}(r_y) / s_0); \)
drop \(r_y; \)

%getmax \;
%huba; \;
%do %while(&m > 5); \;
%huba; \;
%end; \;
%tukey; \;
%do %while(&m > 1); \;
%tukey; \;
%end; \;
%mend; \;

**Appendix 2**

Hydrocarbon Pollution Potential, Mortality Rate, and Result from OLS and ROBUST regression of Mortality on Log(Hydrocarbon Pollution):

**OLS ESTIMATE**

| PARAMETER | ESTIMATE | STANDARD ERROR | PROB > |T| | PARAMETER | ESTIMATE | STANDARD ERROR | PROB > |T| |
|-----------|----------|----------------|--------|---| |-----------|----------|----------------|--------|---| | |-----|-----| |-----|-----| |
| INTERCEPT | 918.43   | 20.5           | 0.000**| 891.84 | 17.0 | 6000** |
| X         | 7.969    | 6.87           | 0.246  | 19.410 | 6.01 | .001*    |
| TOTAL N   | 60000    |                |        | 60000 |      |          |
| R         | 0.023    |                |        | 0.153 |      |          |
| ADJR      | 0.006    |                |        | 0.138 |      |          |

OLS and robust regressions of mortality rate on log hydrocarbon pollution.