ROBUST ESTIMATOR FOR LINEAR REGRESSION

Ashok Mehta
Syracuse University, Syracuse, NY

Consider the following linear regression model
\[ y = x\beta + u \] (1)
where \( y \) is an \((n \times 1)\) vector of dependent variables, \( x \) is an \((n \times p)\) matrix of independent variables, \( \beta \) is a \((p \times 1)\) vector of parameters and \( u \) is an \((n \times 1)\) vector of error terms. It is assumed that error terms have a symmetric distribution.

Equation (1) is generally estimated by ordinary least squares (OLS) by minimizing the function \( L = \sum (y_i - x_i\beta)^2 \). The OLS estimates of the \( \beta \) vector are very sensitive to outliers, i.e., parameter estimates change drastically if an extreme value is dropped from the data set. Let \( \hat{\beta}_{i0} \) be the OLS estimate of \( \beta \) when the \( i \) th observation \((y_i, x_i)\) is dropped from the data set, and \( \hat{\beta} \) be the usual OLS estimate for the entire data set. Li (1983) has shown that the change in the parameter estimate caused by the omission of one observation is given by the statistic
\[ \hat{\beta} - \hat{\beta}_{i0} = \left( (x^Tx)^{-1}x_i / (1 - h_i) \right) u_i, \] (2)
where \( u_{i0} \) is the \( i \)th residual of the OLS regression calculated from the entire data set, and \( h_i = x_i(x^Tx)^{-1}x_i^T \) (i.e., the diagonal element of the matrix \( x(x^Tx)^{-1}x^T \)).

From equation (2) it is evident that the sensitivity of the OLS estimate is a direct function of (a) residual for the \( i \)th observation, and (b) extreme values of \( x_i \). Therefore, in order to reduce the instability of the OLS estimate one has to control these two factors. This can be achieved by introducing the following two functions:

(i) \( P(u) \), which increases gradually in comparison with the squares of residuals in \( L \).
(ii) \( \gamma(x) \), provides a lower weightage to the outlier values in the data set.

The resulting modified objective function is
\[ \Sigma \gamma^2(x)P(u) \] (3)
where \( P(.) \) is a convex function and \( \gamma(.) \) is jointly determined during the estimation process.

It is basically a different set of choices of \( P(.) \) and \( \gamma(.) \) that results into various robust estimators. Krasker & Welsch (1982) used \( P(u) = u^4 \) and \( \gamma(x) = w(u, d(x)) \) in equation (3). Their objective function is \( \Sigma (w(u, d(x)))^2 \). [For complete discussion refer to Krasker & Welsch (1982) and Furno & Baum (1989)].

The weights \( w(u, d(x)) \) are defined as:
\[ w((u, d(x))) = \min \{ 1, a/|u| \} \] (5)
\[ t = a_0/d(x) \] (6)
where \( a \) is the bound for constraining extreme values in both residuals and independent variables. Krasker & Welsch has suggested that a reasonable value for \( a \) is \( 1.5s_{1/2} \). The weights take value one for \( t < a \), and for the larger values of \( t \) the weights are inversely related to standardized residuals in absolute value, and to \( d(x) \).

The term \( d(x) \) is referred as robust distance in the literature and is defined as
\[ d(x) = (x^TA^{-1}x)^{1/2} \]

The \((p \times p)\) matrix \( A \) is the robust second moment matrix of the independent variables and is given by:
\[ \Sigma w^2(d(x))x_i^Tx_i = A \] (7)

PROGRAMMING IN SAS® FOR THE COMPUTATION

The estimation is a four stage process.

STAGE I

Obtain the OLS estimates of the parameters using PROC REG and output RESIDUALS and DFFITS in a data set. DFFITS is the robust distance mentioned earlier in our discussion.
STAGE II

Define a new data set which incorporates the data set outputted in Stage I. Furthermore, define new variables: \( k, \) \( kwl, \) and \( w \) in the data step where

\[
k = k \cdot \left( \frac{\text{number of variables}}{\text{number of observations}} \right)^{1/2}
\]

\[
kwl = \frac{k}{\text{abs (DFFITS)}}
\]

if \( kwl > 1 \) then \( w = 1, \) else \( w = kwl. \)

Again run PROC REG with weight as \( w \) and output RESIDUALS and DFFITS in a new data set. Remember to name these statistics differently from Stage I names as DFFITS1 and RES1.

STAGE III

Define a new data set incorporating the data set outputted in Stage II. Keep only the variables used in the model, RES, RES1 AND DFFITS1 in the new data set. Define DFFITS again as DFFITS = DFFITS1 and then modify it as DFFITS = DFFITS*(RES/RES1). Also define \( k, \) \( kw \), and \( w \) as in Stage II. Run PROC REG with \( w \) as weight and output DFFITS and RESIDUAL as DFFITS3 and RES3 respectively.

STAGE IV

Finally, create a macro in which Stage III output is written as input data set for Stage II. This macro is executed in a loop until one of the following conditions is met:

(i) Predefined number of iterations.
(ii) Maximum absolute value of difference between two successive set of parameter estimates is smaller than a predefined number.

One can get an estimated covariance matrix for the Krasker-Welsch parameter estimates by using COVB option in the model statement of the PROC REG in the last stage.

[See Appendix for the SAS code]

APPLICATION OF KRASKER-WELSCH ESTIMATOR

Krasker-Welsch estimation can be used in regression analysis wherever there is a need to examine the possibility of errors in the data. It is also useful when caution is needed in specifying linear models.

As an illustration we estimated the Krasker-Welsch estimates for the following linear model:

\[
\text{Capital Flight} = f(\text{Domestic Inflation Rate, Domestic Interest rate, Adjusted Foreign Interest Rate})
\]

where adjusted foreign interest rate = (foreign currency + expected depreciation of domestic currency).

We used quarterly data for Mexico from 1976Q2 to 1987Q1 for the estimation purposes. We constrained the number of iterations by ending the loop in Stage IV by imposing upper limit on iteration as 50 or the maximum absolute difference between two successive set of estimates as less than/equal to 0.05. The results are reported in Table 1.

Note that the program produces the OLS estimates, Krasker-Welsch estimates and covariance matrix for the final Krasker-Welsch estimates. It also gives weights used in the last iteration, which allow us to identify the outliers. Weights less than 1 suggesting outlier and need for investigation.

In our example two observations have values less than 1. On examining the data it was found that these two observations were corresponding to the maximum and minimum capital flight values.

The program was executed on an IBM 3090 mainframe computer. It is a time consuming program and also requires enormous amount of disk space to store the saslog and output files. In order to run the program on micro-computers it is advisable to restrict the number of iterations to 10 or 15.

CONCLUSION

In this paper we developed a SAS program to estimate Krasker-Welsch robust estimator for a linear multiple regression model. The program also identifies outliers and also provides estimates of the covariance matrix of
the KIasker-Welsch estimates. This program can be further modified to estimate a linear simultaneous equation model.

REFERENCES


The author would like to thank Dr. Ron Kalinoski and Nelson Pardee for their help and support.

SAS and SAS/STAT are registered trademarks of SAS Institute, Inc., Cary, NC.

Contact:

Dr. Ashok Mehta
Research Consulting Services
250 Machinery Hall
Syracuse University
Syracuse, NY 13244

Tel: (315) 443-2196
TABLE 1

OLS REGRESSION COEFFICIENT ESTIMATES

<table>
<thead>
<tr>
<th>OBS VAR</th>
<th>LABEL _ COLREG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FIN ~</td>
</tr>
<tr>
<td>2</td>
<td>INTERCEPT INTERCEPT</td>
</tr>
<tr>
<td>3</td>
<td>PERCPI</td>
</tr>
<tr>
<td>4</td>
<td>TBILLINT</td>
</tr>
</tbody>
</table>

INITIAL K-W REGRESSION COEFFICIENT ESTIMATES

<table>
<thead>
<tr>
<th>OBS VAR</th>
<th>LABEL _ COLWREG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FIN ~</td>
</tr>
<tr>
<td>2</td>
<td>INTERCEPT INTERCEPT</td>
</tr>
<tr>
<td>3</td>
<td>PERCPI</td>
</tr>
<tr>
<td>4</td>
<td>TBILLINT</td>
</tr>
</tbody>
</table>

MAX ABS. DIFF. BETWEEN OLS AND INT. KW PARAMS = 209.5672

ITERATION 1: K-W COEFFICIENT ESTIMATES

<table>
<thead>
<tr>
<th>OBS VAR</th>
<th>LABEL _ COLWREG1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FIN ~</td>
</tr>
<tr>
<td>2</td>
<td>INTERCEPT INTERCEPT</td>
</tr>
<tr>
<td>3</td>
<td>PERCPI</td>
</tr>
<tr>
<td>4</td>
<td>TBILLINT</td>
</tr>
</tbody>
</table>

MAX ABS. DIFF. BETWEEN INT. KW AND ITER 1 PARAM = 281.3999

COVARIANCE OF ESTIMATES

<table>
<thead>
<tr>
<th>COVB</th>
<th>INTERCEPT</th>
<th>PERCPI</th>
<th>FIN ~</th>
<th>TBILLINT</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEPT</td>
<td>81297.86</td>
<td>-292847</td>
<td>-841.584</td>
<td>-255.897</td>
</tr>
<tr>
<td>PERCPI</td>
<td>-292847</td>
<td>16558208</td>
<td>-14527.6</td>
<td>-34751.1</td>
</tr>
<tr>
<td>FIN ~</td>
<td>-841.584</td>
<td>-14527.6</td>
<td>90.6385</td>
<td>3.119059</td>
</tr>
<tr>
<td>TBILLINT</td>
<td>-255.897</td>
<td>-34751.1</td>
<td>3.119059</td>
<td>102.3037</td>
</tr>
</tbody>
</table>

ITERATION 2: K-W COEFFICIENT ESTIMATES

<table>
<thead>
<tr>
<th>OBS VAR</th>
<th>LABEL _ COLWREG1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FIN ~</td>
</tr>
<tr>
<td>2</td>
<td>INTERCEPT INTERCEPT</td>
</tr>
<tr>
<td>3</td>
<td>PERCPI</td>
</tr>
<tr>
<td>4</td>
<td>TBILLINT</td>
</tr>
</tbody>
</table>

MAX ABS. DIFF. BETWEEN ITER 2 AND PREV. ITER. PARAMS =47.38925

COVARIANCE OF ESTIMATES

<table>
<thead>
<tr>
<th>COVB</th>
<th>INTERCEPT</th>
<th>PERCPI</th>
<th>FIN ~</th>
<th>TBILLINT</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEPT</td>
<td>91088.9</td>
<td>-312862</td>
<td>-934.552</td>
<td>-324.605</td>
</tr>
<tr>
<td>PERCPI</td>
<td>-312862</td>
<td>20334194</td>
<td>-16631.7</td>
<td>-38088.2</td>
</tr>
<tr>
<td>FIN ~</td>
<td>-934.552</td>
<td>-16631.7</td>
<td>102.0479</td>
<td>3.978559</td>
</tr>
<tr>
<td>TBILLINT</td>
<td>-324.605</td>
<td>-38088.2</td>
<td>3.978559</td>
<td>113.4162</td>
</tr>
</tbody>
</table>

1403
**TABLE 1 (cont.)**

ITERATION 3 TO 49 have been deleted by the author.

ITERATION 50: K-W COEFFICIENT ESTIMATES

<table>
<thead>
<tr>
<th>OBS</th>
<th>VAR</th>
<th>LABEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FIN'T</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>INTERCEPT</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>PRCPI</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>TBILLINT</td>
<td></td>
</tr>
</tbody>
</table>

MAX ABS. DIFF. BETWEEN ITER 50 AND PREV. ITER.Parms = 101.2348

COVARIANCE OF ESTIMATES

<table>
<thead>
<tr>
<th>COVB</th>
<th>INTERCEPT</th>
<th>PRCPI</th>
<th>FIN'T</th>
<th>TBILLINT</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEPT</td>
<td>86194.57</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRCPI</td>
<td>-302880</td>
<td>19450450</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIN'T</td>
<td>-888.209</td>
<td>-155581.1</td>
<td>96.37419</td>
<td></td>
</tr>
<tr>
<td>TBILLINT</td>
<td>-290.13</td>
<td>-36426.9</td>
<td>3.551349</td>
<td>107.8702</td>
</tr>
</tbody>
</table>

WEIGHTS ASSOCIATED WITH THE FINAL ITERATION

<table>
<thead>
<tr>
<th>OBS</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00000</td>
</tr>
<tr>
<td>2</td>
<td>1.00000</td>
</tr>
<tr>
<td>3</td>
<td>1.00000</td>
</tr>
<tr>
<td>4</td>
<td>1.00000</td>
</tr>
<tr>
<td>5</td>
<td>1.00000</td>
</tr>
<tr>
<td>6</td>
<td>1.00000</td>
</tr>
<tr>
<td>7</td>
<td>1.00000</td>
</tr>
<tr>
<td>8</td>
<td>1.00000</td>
</tr>
<tr>
<td>9</td>
<td>1.00000</td>
</tr>
<tr>
<td>10</td>
<td>1.00000</td>
</tr>
<tr>
<td>11</td>
<td>1.00000</td>
</tr>
<tr>
<td>12</td>
<td>1.00000</td>
</tr>
<tr>
<td>13</td>
<td>1.00000</td>
</tr>
<tr>
<td>14</td>
<td>1.00000</td>
</tr>
<tr>
<td>15</td>
<td>1.00000</td>
</tr>
<tr>
<td>16</td>
<td>1.00000</td>
</tr>
<tr>
<td>17</td>
<td>1.00000</td>
</tr>
<tr>
<td>18</td>
<td>1.00000</td>
</tr>
<tr>
<td>19</td>
<td>1.00000</td>
</tr>
<tr>
<td>20</td>
<td>1.00000</td>
</tr>
<tr>
<td>21</td>
<td>1.00000</td>
</tr>
<tr>
<td>22</td>
<td>1.00000</td>
</tr>
<tr>
<td>23</td>
<td>1.00000</td>
</tr>
<tr>
<td>24</td>
<td>1.00000</td>
</tr>
<tr>
<td>25</td>
<td>0.80383</td>
</tr>
<tr>
<td>26</td>
<td>1.00000</td>
</tr>
<tr>
<td>27</td>
<td>1.00000</td>
</tr>
<tr>
<td>28</td>
<td>0.81995</td>
</tr>
<tr>
<td>29</td>
<td>1.00000</td>
</tr>
<tr>
<td>30</td>
<td>1.00000</td>
</tr>
<tr>
<td>31</td>
<td>1.00000</td>
</tr>
<tr>
<td>32</td>
<td>1.00000</td>
</tr>
<tr>
<td>33</td>
<td>1.00000</td>
</tr>
<tr>
<td>34</td>
<td>1.00000</td>
</tr>
<tr>
<td>35</td>
<td>1.00000</td>
</tr>
<tr>
<td>36</td>
<td>1.00000</td>
</tr>
<tr>
<td>37</td>
<td>1.00000</td>
</tr>
<tr>
<td>38</td>
<td>1.00000</td>
</tr>
<tr>
<td>39</td>
<td>1.00000</td>
</tr>
<tr>
<td>40</td>
<td>1.00000</td>
</tr>
<tr>
<td>41</td>
<td>1.00000</td>
</tr>
<tr>
<td>42</td>
<td>1.00000</td>
</tr>
<tr>
<td>43</td>
<td>1.00000</td>
</tr>
<tr>
<td>44</td>
<td>1.00000</td>
</tr>
</tbody>
</table>
APPENDIX

OPTIONS NODATE LS = 72 NOCENTER LS "" DQUOTE;
/
* STAGE I: GETTING OLS ESTIMATES *
* *******************************************************/
PROC REG DATA = ASHOK.DATA OUTEST = EST1;
MODEL CFLOW = PERCPI FINT TBILLINT;
OUTPUT OUT = DATA.STEP1 DFFITS = DFFITS RESIDUAL = RES;
TITLE "OLS REGRESSION ESTIMATES";
PROC TRANSPOSE DATA = EST1 (KEEP = INTERCEP PERCPI FINT TBILLINT)
OUT = A2 (RENAME = ( _NAME_ = VAR COL1 = COLREG));
PROC SORT;
PROC PRINT;
TITLE "OLS REGRESSION COEFFICIENT ESTIMATES";
/******* STAGE II: INITIAL K-W ESTIMATES ***********/
DATA B:
SET DATA.STEP1;
K1 4; /*NUMBER OF VARIABLES*/
K2 = K1-1;
K3 = K1-1;
K4 = 1.5*(4)**0.5;
/*KRASKEI-WEISCH CONSTANT = 1.5*P**0.5 WHERE P = NUMBER OF PARAMETERS TO BE ESTIMATED*/
X4 = K4*(K1/K2)**0.5;
DFFITS = DFFITS1;
DFFITS = DFFITS* (RES1/RES);
WEL = K4/ABS(DFFITS);
IF WEL > 1 THEN W = 1; ELSE W = WEL;
PROC REG OUTEST = EST2 NOPRINT;
MODEL CFLOW = PERCPI FINT TBILLINT;
WEIGHT W;
OUTPUT OUT = DATA.STEP2 DFFITS = DFFITS2 RESIDUAL = RRS2;
PROC TRANSPOSE DATA = EST2 (KEEP = INTERCEP PERCPI FINT TBILLINT)
OUT = C2 (RENAME = ( _NAME_ = VAR COL1 = COLWRKG));
PROC SORT DATA = C2; BY VAR;
PROC PRINT;
TITLE "INITIAL K-W REGRESSION COEFFICIENT ESTIMATES";
DATA STEP3:
MERGE B2 C2; BY VAR;
ADIFF = ABS(COLWRKG - COLWREG);
DATA FINALB (KEEP = ADIFF);
MERGE B2 C2; BY VAR;
ADIFF = ABS(COLWRKG - COLWREG);
PROC TRANSPOSE DATA = FINALB OUT = FINALB1;
DATA FINALB2 (KEEP = KW);
SET FINALB2;
KW = MAX(COL1, COL2, COL3, COL4);
FILE PRINT;
PUT "MAX ABS. DIFF. BETWEEN OLS AND INT. PARMS = KW;";
/* STAGE III: ITERATION 1 FOR K-W ESTIMATES */
* *****************************************************/
DATA C:
SET DATA.STEP2 (KEEP = DFFITS1 CFLOW TBILLINT PRECPI RRS1 RRS2);
K1 = 4; /*NUMBER OF VARIABLES*/
K3 = K1-1;
K2 = 44; /*NUMBER OF OBSERVATIONS*/
K4 = 1.5*(4)**0.5;
/*KRASKEI-WEISCH CONSTANT = 1.5*P**0.5 WHERE P = NUMBER OF PARAMETERS TO BE ESTIMATED*/
K4 = K4*(K1/K2)**0.5;
DFFITS = DFFITS1;
DFFITS = DFFITS* (RRS1/RRS);
WEL = K4/ABS(DFFITS);
IF WEL > 1 THEN W = 1; ELSE W = WEL;
PROC REG OUTEST = EST3 NOPRINT;
MODEL CFLOW = PERCPI FINT TBILLINT;
WEIGHT W;
OUTPUT OUT = DATA.STEP3 DFFITS = DFFITS2 RESIDUAL = RRS2;
PROC TRANSPOSE DATA = EST3 (KEEP = INTERCEP PERCPI FINT TBILLINT)
OUT = C2 (RENAME = ( _NAME_ = VAR COL1 = COLWRKG));
PROC SORT DATA = C2; BY VAR;
PROC PRINT;
TITLE "ITERATION 1: K-W COEFFICIENT ESTIMATES";
DATA STEP3:
MERGE B2 C2; BY VAR;
ADIFF = ABS(COLWRKG - COLWREG);
DATA FINALC (KEEP = ADIFF);
MERGE B2 C2; BY VAR;
ADIFF = ABS(COLWRKG - COLWREG);
PROC TRANSPOSE DATA = FINALC OUT = FINALC1;
DATA FINALC2 (KEEP = KW);
SET FINALC1;
KW = MAX(COL1, COL2, COL3, COL4);
FILE PRINT;
PUT "MAX ABS. DIFF. BETWEEN INT. KW AND ITER 1 PARAM = KW;";
/* STAGE IV: SETTING UP MACRO TO ITERATE */
* STAGE II AND STAGE III **************/
%MACRO WELSCH;
%LET Y = 1;
%LET COUNT = 1;
%DO WHILE (&Y = 1);
%LET COUNT = %EVAL (&COUNT +1);
%LET Y = 1;
DATA C:
SET DATA.STEP3 (KEEP = CFLOW PENCPI FINT PRECPI RRS1 RRS2 DFFITS);
K1 = 4; /*NUMBER OF VARIABLES*/
K3 = K1-1;
K2 = 44; /*NUMBER OF OBSERVATIONS*/
K4 = 1.5*(4)**0.5;
*/
WELSCH CONSTANT = 1.5*P**0.5 WHERE P = NUMBER OF
PARAMETERS TO BE ESTIMATED

K4 = K4*(K1/K2)**0.5;
DIFFITS = DIFFITS2;
DIFFITS = DIFFITS*(RES2/RES1);
WEL = K4/ABS(DIFFITS);
IF WEL > 1 THEN W = 1; ELSE W = WEL;

PROC REG OUTEST = EST3;
MODEL CFLOW = PERCPI FINT TBILLINT /COVB
NORPRINT;
WEIGHT W;
OUTPUT OUT = DATA.STEP4 DIFFITS = DIFFITS3 RESIDUAL = RES3;
PROC TRANSPOSE DATA = EST3 (KEEP = INTERCEP PERCPI FINT TBILLINT)
OUT = C2 (RENAME = (NAME = VAR COL1 = COLWREG1));
PROC SORT DATA = C2;
BY VAR;
PROC PRINT;
TITLE "ITERATION &COUNT: K-M COEFFICIENT
ESTIMATES";
DATA FINALC (KEEP = ADIFF);
MERGE B2 C2; BY VAR;
ADIFF = ABS(COLWREG - COLWREG1);
PROC TRANSPOSE DATA = FINALC OUT = FINALC1;
DATA FINALC2 (KEEP = KW);
SET FINALC1;
KW = MAX(COL1, COL2, COL3, COL4);
IF KW LE 0.05 THEN CALL SYMPUT("Y", "0");
IF &COUNT = 50 THEN CALL SYMPUT
("Y", "0");
FILE PRINT;
PUT //"MAX ABS. DIFF. BETWEEN ITER &COUNT
AND PREV. ITER. PARM = "KW;",
STOP; RUN;
DATA DATA.STEP3;
SET DATA.STEP4
(KEEP = CFLOW PERCPI TBILLINT FINT RES3 RES2 DIFFITS3);
RENAME RES2 = RES1 RES3 = RES2 DIFFITS3 = DIFFITS2;
&END;
%MEND WELSCH;
RUN;
&WELSCH;
RUN;
PROC PRINT DATA = DATA.STEP4;
VAR W;
TITLE "WEIGHTS ASSOCIATED WITH THE FINAL
ITERATION";
RUN;