1 Robust estimation of standard error

Data set: http://biosun01.biostat.jhsph.edu/~fdominic/teaching/LDA/sitka.data

Objective: We will estimate the robust estimation of the standard error using matrix calculation in STATA and reproduce the result in LDA book P72-74.

STATA Output:

```
. set matsize 800
. set memory 2000 m
(2000k)
. use c:/data/sitka,clear

*only for 1988 data
. keep if days<300
(632 observations deleted)

*fit the saturate model
. anova logsize days  chamber

                           Number of obs =     395     R-squared     =  0.3926
                           Root MSE      = .627703     Adj R-squared =  0.3816

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  98.5497457     7  14.0785351      35.73     0.0000
                         |
                    days |  93.3623074     4  23.3405769      59.24     0.0000
                 chamber |  5.18743824     3  1.72914608       4.39     0.0047
                         |
                Residual |  152.482109   387  .394010617
              -----------+----------------------------------------------------
                   Total |  251.031854   394  .637136687

. predict yhat
(option xb assumed; fitted values)

*calculate the residul
. gen res=logsize-yhat

. keep tree res  days
. reshape group days 152 174 201 227 258
. reshape var res
. reshape cons tree
. reshape wide

.*Calculate the REML estimate of covariance
. matrix accum Cov=res152 res174 res201 res227 res258 , deviations noconstant
(obs=79)
. scalar adj=1/(_result(1)-4)
. matrix Cov=adj*Cov

. *REML estimate of covariance for 1988, P72
. matrix list Cov
symmetric Cov[5,5]
           res152     res174     res201     res227     res258
res152  .44776565
res174  .40565527  .39764053
res201  .37381363  .37315419   .3709305
res227  .36839556  .37366602  .37197783  .40196995
res258  .36718667  .37388806  .37091338  .40248523  .41478816
```

```
.*Calculate the OLS estimate of bete
. use c:/data/sitka,clear
. keep if days<300
(632 observations deleted)
. tab days, gen(t)

       days |      Freq.     Percent        Cum.
------------+-----------------------------------
        152 |         79       20.00       20.00
        174 |         79       20.00       40.00
        201 |         79       20.00       60.00
        227 |         79       20.00       80.00
        258 |         79       20.00      100.00
------------+-----------------------------------
      Total |        395      100.00

.*indicator for ozone
. gen tao=1-ozone

.*covariate of time
. gen x=days/100*tao

.*We fit GEE marginal model with robust estimate of standard error
. xtgee logsize t1 t2 t3 t4 t5 tao x, noconst i(tree) corr(exc) robust
Iteration 1: tolerance = 6.450e-13
GEE population-averaged model                   Number of obs      =       395
Group variable:                         tree    Number of groups   =        79
Link:                                identity   Obs per group: min =         5
Family:                              Gaussian                  avg =       5.0
Correlation:                    exchangeable                   max =         5
                                                Wald chi2(6)       =   6651.46
Scale parameter:                     .3881248   Prob > chi2        =    0.0000

                            (standard errors adjusted for clustering on tree)
------------------------------------------------------------------------------
             |             Semi-robust
     logsize |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          t1 |   4.060577   .0790991    51.34   0.000     3.905546    4.215609
          t2 |   4.470879   .0776508    57.58   0.000     4.318686    4.623071
          t3 |   4.842733   .0773668    62.59   0.000     4.691097    4.994369
          t4 |   5.178935   .0820593    63.11   0.000     5.018102    5.339769
          t5 |    5.31669   .0838617    63.40   0.000     5.152325    5.481056
         tao |  -.2216775   .2429765    -0.91   0.362    -.6979027    .2545478
           x |    .213851   .0789394     2.71   0.007     .0591327    .3685694
------------------------------------------------------------------------------

.*Incorrect standard error using OLS
. reg logsize t1 t2 t3 t4 t5 tao x, noconst
  Source |       SS       df       MS                  Number of obs =     395
---------+------------------------------               F(  7,   388) = 3381.85
   Model |  9353.83562     7  1336.26223               Prob > F      =  0.0000
Residual |  153.309291   388  .395127038               R-squared     =  0.9839
---------+------------------------------               Adj R-squared =  0.9836
   Total |  9507.14491   395  24.0687213               Root MSE      =  .62859

------------------------------------------------------------------------------
 logsize |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      t1 |   4.060577     .07937     51.160   0.000       3.904528    4.216626
      t2 |   4.470879   .0756955     59.064   0.000       4.322054    4.619703
      t3 |   4.842733   .0739281     65.506   0.000       4.697383    4.988083
```

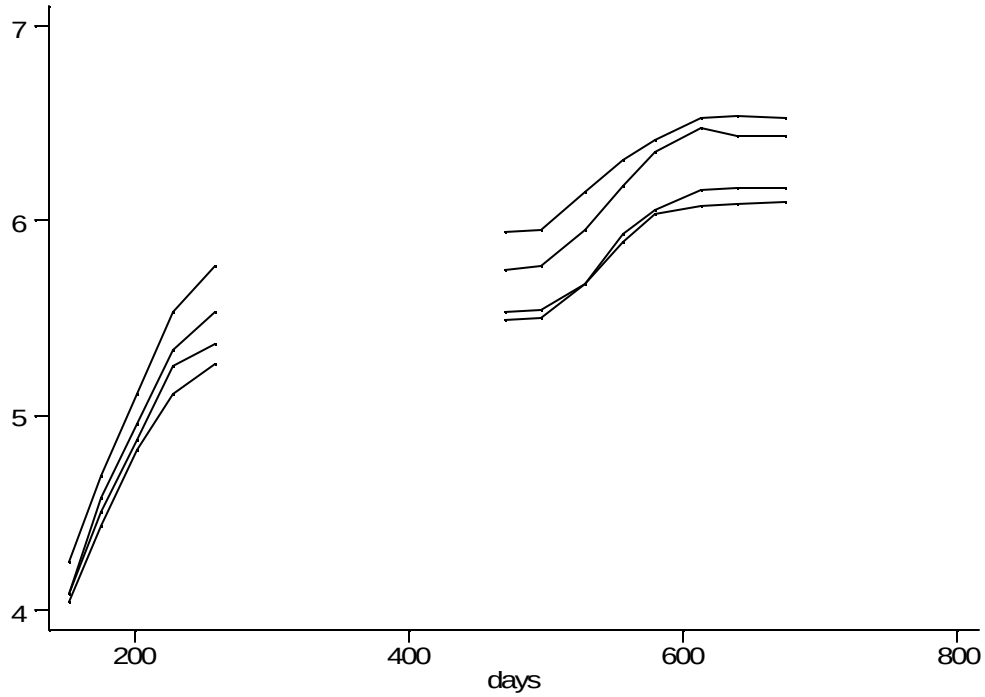
```
      t4 |   5.178935    .075257     68.817   0.000       5.030973    5.326898
      t5 |    5.31669   .0805032     66.043   0.000       5.158413    5.474967
     tao |  -.2216775   .3729256     -0.594   0.553      -.9548853    .5115304
       x |    .213851   .1811625      1.180   0.239      -.1423321    .5700341
------------------------------------------------------------------------------

.*Now create design matrix of X for 1988 data
. mkmat t1 t2 t3 t4 t5 tao x, matrix(X)

.*Now create response matrix of Y for 1988 data
. mkmat logsize, matrix(Y)

.*OLS estimate of bete
. matrix bete=syminv(X'*X)*X'*Y
. matrix list bete

bete[7,1]
        logsize
 t1   4.0605772
 t2   4.4708787
 t3   4.8427332
 t4   5.1789353
 t5   5.3166904
tao  -.22167746
  x   .21385103

.*Calculate the robust estimate of standard error of bete
. matrix diag=I(79)
. matrix V=diag#Cov
. matrix RW=syminv(X'*X)*X'*V*X*syminv(X'*X)

.*Robust estimate of covariance matrix for the coefficients
. matrix list RW
symmetric RW[7,7]
             t1          t2          t3          t4          t5          tao
 t1   .00811725
 t2   .00750314   .00735327
 t3   .00700059    .0069839   .00694552
 t4    .0068362   .00693316   .00694892   .00736432
 t5   .00670666   .00686775   .00692369   .00741347   .00767666
tao  -.00950986  -.00854075  -.00735139  -.00620608  -.00484052   .05120942
  x   .00116445   .00069542   .00011979  -.00043452  -.00109543  -.01391991

              x
  x     .006737

.* Calculate the robust standard error
. matrix var=vecdiag(RW)
. matrix var=var'
. svmat var, name(varbete)
. keep varbete1
. drop in 8/ 395
(388 observations deleted)
. gen se=sqrt(varbete1)
. svmat bete, name(bete)
. mkmat bete1 se, matrix(Coef)
. matrix Coef=Coef'
. matrix list Coef
Coef[2,7]
               r1          r2          r3          r4          r5          r6
bete1   4.0605774   4.4708786   4.8427334   5.1789355   5.3166904  -.22167745
   se    .0900958   .08575123   .08333977    .0858156   .08761657   .22629499
```

```
              r7
bete1   .21385103
   se   .08207925

. matrix colnames Coef= bete1 bete2 bete3 bete4 bete5 tao gamma

. *Table4.3 for 1988 P75
. matrix list Coef

Coef[2,7]
            bete1       bete2       bete3       bete4       bete5         tao
bete1   4.0605774   4.4708786   4.8427334   5.1789355   5.3166904  -.22167745
   se    .0900958   .08575123   .08333977    .0858156   .08761657   .22629499

            gamma
bete1   .21385103
   se   .08207925

.
.
.
.*For 1989 data, we perform the similar analysis as above
. clear
. set matsize 800
. set memory 20000 m
(20000k)
. use c:/data/sitka,clear
. keep if days>300
(395 observations deleted)
. anova logsize days  chamber

                           Number of obs =     632     R-squared     =  0.1902
                           Root MSE      = .638854     Adj R-squared =  0.1772

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  59.5426223    10  5.95426223      14.59     0.0000
                         |
                    days |  41.5389651     7  5.93413788      14.54     0.0000
                 chamber |  18.0036572     3  6.00121906      14.70     0.0000
                         |
                Residual |  253.451873   621  .408135062
              -----------+----------------------------------------------------
                   Total |  312.994496   631  .496029312

. predict yhat
(option xb assumed; fitted values)
. gen res=logsize-yhat
. keep tree res  days
. reshape group days 469 496 528 556 579 613 639 674
. reshape var res
. reshape cons tree
. reshape wide
.
. matrix accum Cov=res469 res496 res528 res556 res579 res613 res639 res674
,deviations noconstant
(obs=79)
.scalar adj=1/(_result(1)-4)
. matrix Cov=adj*Cov

. *REML estimate of Covariance for 1989, P72
. matrix list Cov
```

```
symmetric Cov[8,8]
           res469     res496     res528     res556     res579     res613
res469  .45855646
res496  .45529381  .45222036
res528  .42771715  .42523222  .40954589
res556  .41708041  .41496013  .39721642  .39643825
res579  .43259198  .43021456   .4112053  .41024006  .43418411
res613  .42151602  .41934546  .40355341  .40226918  .42206962  .41665615
res639  .40744428  .40543899  .38891854   .3875753   .4047885  .40038313
res674   .4178666   .4157419    .400477  .39996871  .41808627  .41294601

           res639     res674
res639  .39402208
res674  .40283392  .41773502
. use c:/data/sitka,clear
. keep if days>300
(395 observations deleted)
. tab days, gen(t)

       days |      Freq.     Percent        Cum.
------------+-----------------------------------
        469 |         79       12.50       12.50
        496 |         79       12.50       25.00
        528 |         79       12.50       37.50
        556 |         79       12.50       50.00
        579 |         79       12.50       62.50
        613 |         79       12.50       75.00
        639 |         79       12.50       87.50
        674 |         79       12.50      100.00
------------+-----------------------------------
      Total |        632      100.00

. gen tao=1-ozone

. xtgee logsize t1 t2 t3 t4 t5 t6 t7 t8 tao, noconst i(tree) corr(exc) robust
Iteration 1: tolerance = 8.437e-13
GEE population-averaged model                   Number of obs      =       632
Group variable:                         tree    Number of groups   =        79
Link:                                identity   Obs per group: min =         8
Family:                              Gaussian                  avg =       8.0
Correlation:                     exchangeable                  max =         8
                                                Wald chi2(8)       =  11300.50
Scale parameter:                     .4023932   Prob > chi2        =    0.0000

                            (standard errors adjusted for clustering on tree)
------------------------------------------------------------------------------
             |             Semi-robust
     logsize |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          t1 |   5.504014   .0903405    60.93   0.000      5.32695    5.681078
          t2 |   5.516293   .0900977    61.23   0.000     5.339704    5.692881
          t3 |   5.679963   .0883702    64.27   0.000     5.506761    5.853166
          t4 |   5.901356   .0878268    67.19   0.000     5.729218    6.073493
          t5 |   6.040723   .0899908    67.13   0.000     5.864344    6.217102
          t6 |   6.127305   .0888877    68.93   0.000     5.953088    6.301522
          t7 |   6.128444   .0877665    69.83   0.000     5.956425    6.300464
          t8 |   6.130976   .0894018    68.58   0.000     5.955752      6.3062
         tao |   .3541158    .147512     2.40   0.016     .0649975     .643234
------------------------------------------------------------------------------

.*Incorrect standard error using OLS
. reg logsize t1 t2 t3 t4 t5 t6 t7 t8 tao, noconst
```

```
  Source |       SS       df       MS                 Number of obs =     632
---------+------------------------------              F(  9,   623) = 6189.73
   Model |  22740.1767     9   2526.6863              Prob > F      =  0.0000
Residual |  254.312517   623  .408206287              R-squared     =  0.9889
---------+------------------------------              Adj R-squared =  0.9888
   Total |  22994.4892   632  36.3836855              Root MSE      =  .63891

------------------------------------------------------------------------------
 logsize |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      t1 |   5.504014   .0739337     74.445   0.000       5.358825    5.649203
      t2 |   5.516293   .0739337     74.611   0.000       5.371103    5.661482
      t3 |   5.679963   .0739337     76.825   0.000       5.534774    5.825153
      t4 |   5.901356   .0739337     79.820   0.000       5.756166    6.046545
      t5 |   6.040723   .0739337     81.705   0.000       5.895533    6.185912
      t6 |   6.127305   .0739337     82.876   0.000       5.982116    6.272495
      t7 |   6.128444   .0739337     82.891   0.000       5.983255    6.273634
      t8 |   6.130976   .0739337     82.925   0.000       5.985787    6.276165
     tao |   .3541158   .0546439      6.480   0.000       .2468073    .4614243
------------------------------------------------------------------------------
.
. mkmat t1 t2 t3 t4 t5 t6 t7 t8 tao, matrix(X)
. mkmat logsize, matrix(Y)
. matrix bete=syminv(X'*X)*X'*Y
. matrix list bete
bete[9,1]
       logsize
 t1   5.504014
 t2  5.5162925
 t3  5.6799634
 t4  5.9013557
 t5  6.0407229
 t6  6.1273052
 t7  6.1284444
 t8   6.130976
tao  .35411576

. matrix diag=I(79)
. matrix V=diag#Cov
. matrix RW=syminv(X'*X)*X'*V*X*syminv(X'*X)
. matrix list RW
symmetric RW[9,9]
             t1          t2          t3          t4          t5          t6
 t1   .00822969
 t2   .00818839   .00814949
 t3   .00783932   .00780787   .00760931
 t4   .00770468   .00767784   .00745324   .00744339
 t5   .00790103   .00787094   .00763031   .00761809   .00792118
 t6   .00776083   .00773335   .00753345    .0075172   .00776783   .00769931
 t7    .0075827   .00755732    .0073482    .0073312   .00754909   .00749332
 t8   .00771463   .00768774   .00749451   .00748808   .00771741   .00765235
tao  -.00766357  -.00766357  -.00766357  -.00766357  -.00766357  -.00766357

             t7          t8         tao
 t7    .0074128
 t8   .00752434   .00771297
tao  -.00766357  -.00766357   .02421689

. matrix var=vecdiag(RW)
. matrix var=var'
. svmat var, name(varbete)
. keep varbete1
. drop in 10/ 632
```

```
(623 observations deleted)
. gen se=sqrt(varbete1)
. svmat bete, name(bete)
. mkmat bete1 se, matrix(Coef)
. matrix Coef=Coef'
. matrix list Coef

Coef[2,9]
              r1         r2         r3         r4         r5         r6
bete1   5.504014  5.5162926  5.6799636  5.9013557  6.0407228   6.127305
   se  .09071766  .09027453  .08723134  .08627506  .08900102  .08774571

              r7         r8         r9
bete1  6.1284442  6.1309762  .35411575
   se  .08609764   .0878235  .15561777

. matrix colnames Coef= bete1 bete2 bete3 bete4 bete5 bete6 bete7 bete8 tao

. *Table4.3 for 1989 P75
. matrix list Coef

Coef[2,9]
           bete1      bete2      bete3      bete4      bete5      bete6
bete1   5.504014  5.5162926  5.6799636  5.9013557  6.0407228   6.127305
   se  .09071766  .09027453  .08723134  .08627506  .08900102  .08774571

           bete7      bete8        tao
bete1  6.1284442  6.1309762  .35411575
   se  .08609764   .0878235  .15561777


. *plots
. use c:/data/sitka,clear

. egen mean1 = mean(logsize) if chamber==1 & days<300, by(days)
(892 missing values generated)

. egen mean2 = mean(logsize) if chamber==2 & days<300, by(days)
(892 missing values generated)

. egen mean3 = mean(logsize) if chamber==3 & days<300, by(days)
(967 missing values generated)

. egen mean4 = mean(logsize) if chamber==4 & days<300, by(days)
(962 missing values generated)

. egen mean5 = mean(logsize) if chamber==1 & days>300, by(days)
(811 missing values generated)

. egen mean6 = mean(logsize) if chamber==2 & days>300, by(days)
(811 missing values generated)

. egen mean7 = mean(logsize) if chamber==3 & days>300, by(days)
(931 missing values generated)

. egen mean8 = mean(logsize) if chamber==4 & days>300, by(days)
(923 missing values generated)

. graph mean* days, c(llllllll) xlab ylab s(iiiiiiii)
```




```
. egen meanc1 = mean(logsize) if ozone==0 & days<300, by(days)
(902 missing values generated)
. egen meanc2 = mean(logsize) if ozone==0 & days>300, by(days)
(827 missing values generated)
. egen meant1 = mean(logsize) if ozone==1 & days<300, by(days)
(757 missing values generated)
. egen meant2 = mean(logsize) if ozone==1 & days>300, by(days)
(595 missing values generated)

. graph meanc1 meanc2 meant1 meant2 days, c(llll) xlab ylab s(iiii)
```