LONGITUDINAL DATA ANALYSIS
Homework I, 2005
SOLUTION

1. **Suppose A and B are both 2×2 matrices with**

$$\mathbf{A} = \begin{pmatrix} 6 & 3 \\ -2 & 5 \end{pmatrix}, \mathbf{B} = \begin{pmatrix} -4 & 10 \\ 7 & 6 \end{pmatrix}$$

(a) Verify that $|\mathbf{A}||\mathbf{B}| = |\mathbf{AB}|$.

$|\mathrm{A}| = 6 \times 5 - 3 \times (-2) = 36$; $|\mathrm{B}| = (-4) \times 6 - 10 \times 7 = -94$.
Therefore $|\mathrm{A}||\mathrm{B}| = 36 \times (-94) = -3384$.

$$\mathrm{AB} = \begin{pmatrix} 6 & 3 \\ -2 & 5 \end{pmatrix} \begin{pmatrix} -4 & 10 \\ 7 & 6 \end{pmatrix} = \begin{pmatrix} -3 & 78 \\ 43 & 10 \end{pmatrix}.$$

So, $|\mathrm{AB}| = (-3) \times 10 - 78 \times 43 = -3384 = |\mathrm{A}||\mathrm{B}|$

(b) Verify that $|\mathbf{A}| = 1/|\mathbf{A}^{-1}|$.

$$\mathrm{A}^{-1} = \frac{1}{36} \begin{pmatrix} 5 & -3 \\ 2 & 6 \end{pmatrix}.$$

Therefore, $|\mathrm{A}^{-1}| = \frac{5}{36} \cdot \frac{6}{36} - \left(-\frac{3}{36}\right) \frac{2}{36} = 1/36 = |\mathrm{A}|^{-1}$

(c) Verify that $\mathbf{tr}(\mathbf{AB}) = \mathbf{tr}(\mathbf{BA})$.

$$\mathrm{AB} = \begin{pmatrix} -3 & 78 \\ 43 & 10 \end{pmatrix}, \mathrm{tr}(\mathrm{AB}) = -3 + 10 = 7$$

$$\mathrm{BA} = \begin{pmatrix} -44 & 38 \\ 30 & 51 \end{pmatrix}, \mathrm{tr}(\mathrm{BA}) = -44 + 51 = 7$$

So, $\mathrm{tr}(\mathrm{AB}) = \mathrm{tr}(\mathrm{BA})$.

2. **Suppose that a_1 and a_2 are constants, and y_1 and y_2 are (possibly correlated) random variables with means μ_1 and μ_2 respectively. Show that $cov(a_1y_1, a_2y_2) = a_1a_2cov(y_1, y_2)$ by using the definition of covariance.**

Proof By the definition of covariance:

$$\begin{aligned} cov(a_1y_1, a_2y_2) &= E((a_1y_1 - a_1\mu_1)(a_2y_2 - a_2\mu_2)) \\ &= a_1a_1E((y_1 - \mu_1)(y_2 - \mu_2)) \\ &= a_1a_2cov(y_1, y_2) \end{aligned}$$

3 Exploratory Data Analysis

(a) read data from back.raw

```
. clear
. set memory 40m
(40960k)
. set matsize 800
. *log using c:\data\midterm.log, replace
.  infile  id  group  pnvrs1  pnvas1  anvas1  alvas1  time1   pnvrs2  pnvas2  anvas2
alvas2 time2  pnvrs3 pnvas3 anvas3 alvas3 time3  pnvrs4 pnvas4 anvas4 alvas4
time4 using c:\data\back.raw
(27 observations read)

. *reshape to the long format
. reshape long pnvrs pnvas anvas alvas time, i(id) j(set)
(note: j = 1 2 3 4)

Data                               wide   ->   long
---------------------------------------------------------------------------
Number of obs.                       27   ->     108
Number of variables                  22   ->       8
j variable (4 values)                     ->   set
xij variables:
               pnvrs1 pnvrs2 ... pnvrs4   ->   pnvrs
               pnvas1 pnvas2 ... pnvas4   ->   pnvas
               anvas1 anvas2 ... anvas4   ->   anvas
               alvas1 alvas2 ... alvas4   ->   alvas
                  time1 time2 ... time4   ->   time
---------------------------------------------------------------------------


. *make sure it is long format
. list id pnvrs pnvas anvas alvas time in 1/10

            id      pnvrs      pnvas      anvas      alvas       time
  1.         1          2         29          9         27         65
  2.         1          2         22         12         48        298
  3.         1          2         33          6          9        545
  4.         1          2         11         32          8        785
  5.         2          2         31         13         91         90
  6.         2          1          0         14         12        342
  7.         2          1          1          6         29        575
  8.         2          1          6          3         81        855
  9.         3          2         10         15         53        270
 10.         3          2         20         35         46        374

. *recode missing values
. for var pnvrs pnvas anvas alvas time: replace X = . if X == -9
->  replace pnvrs = . if pnvrs == -9
(1 real change made, 1 to missing)
->  replace pnvas = . if pnvas == -9
(2 real changes made, 2 to missing)
->  replace anvas = . if anvas == -9
(4 real changes made, 4 to missing)
->  replace alvas = . if alvas == -9
(3 real changes made, 3 to missing)
->  replace time = . if time == -9
(1 real change made, 1 to missing)

. *convert to cross-sectional time-series data
```

```
. tsset id time
       panel variable:  id, 1 to 27
        time variable:  time, 25 to 890, but with gaps

. iis id
. tis time
```

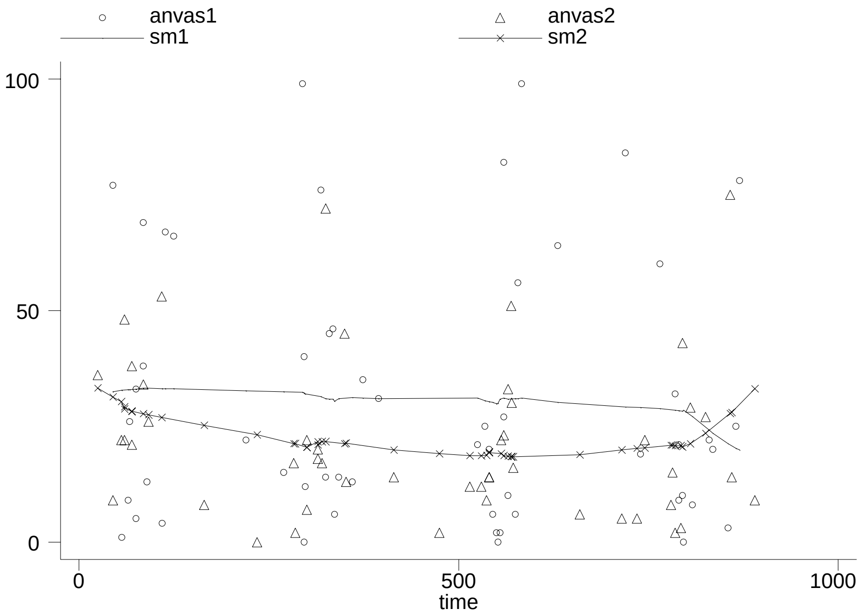
(b) describe the data

```
. xtdes, patterns(0)

      id:  1, 2, ..., 27                                     n =         27
    time:  25, 42, ..., 890                                  T =         90
           Delta(time) = 1; (890-25)+1 = 866
           (id*time uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         4       4       4         4         4       4       4
```

There are 27 subjects in this dataset. The measurements are obtained at 90 different times since the treatment. The panel variable and the time variable can uniquely identify each observation. Each subject is tested 4 times. In this data, the time varying variables are time since treatment (in minutes) and the scores of the four test, pain VRS, pain VAS, anxiety VAS, and alertness VAS. The baseline variable is the treatment group, either intercostals/epidural analgesic (Group 1), or morphine infusion analgesic (Group 2). The data is balanced but not equally spaced.

(c) explore the anxiety `VAS with respect to time and to treatment group.`

```
. sum pnvas anvas alvas
    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------
       pnvas |     106     29.0283    20.8379          0         86
       anvas |     104    26.46154    23.9534          0         99
       alvas |     105    54.86667   27.23405          2         97

. gen anvas1 = anvas if group == 1
(53 missing values generated)
. gen anvas2 = anvas if group == 2
(59 missing values generated)
. ksm anvas1 time, gen(sm1) lowess bw(0.8) nograph
. ksm anvas2 time, gen(sm2) lowess bw(0.8) nograph
. sort time
. graph anvas1 anvas2 sm1 sm2 time, c(..ll)s(oT.x) xlab ylab saving(c,replace)
```



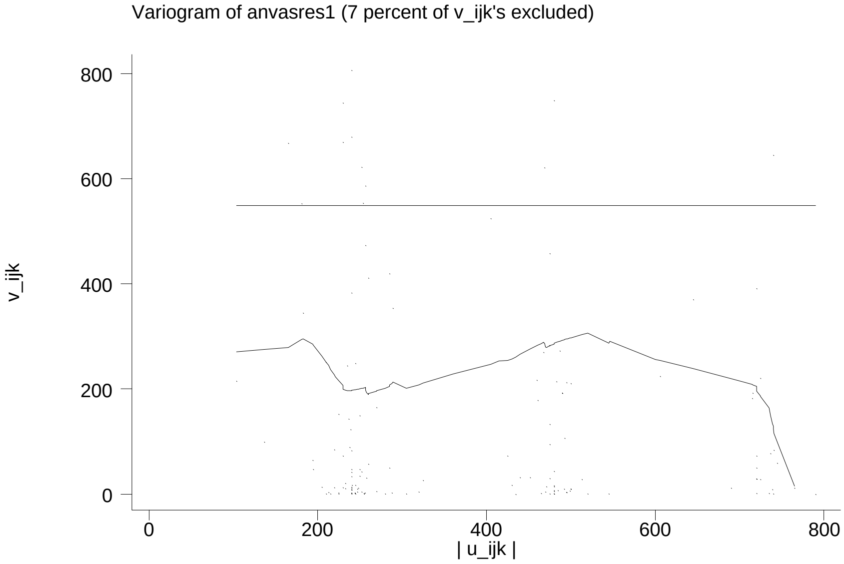
The figure above shows the score of anxiety VAS in the two treatment groups across the time since treatment (in minutes). The open squares represent the scores of anxiety VAS in Group 1 intercostals/epidural analgesic), while the open triangles show the scores of anxiety VAS in Group 2 (morphine infusion analgesic). There are two smoothed lines showing the marginal tread to the scores of anxiety VAS across time in the figure. The top line represents the marginal trend for Group 1 while the bottom line for Group 2. In the beginning, both groups have similar scores of anxiety VAS. After that, it seems that the scores of anxiety VAS in Group 1are relatively stable over time. On the other hand, the scores of anxiety VAS in Group 2 decrease over time.

(d) explore the correlation structure.

```
. xi: reg anvas i.group*time
i.group           _Igroup_1-2         (naturally coded; _Igroup_1 omitted)
i.group*time      _IgroXtime_#        (coded as above)

      Source |       SS       df       MS              Number of obs =     104
-------------+------------------------------           F(  3,   100) =    1.50
       Model |  2538.57675     3  846.192249           Prob > F      =  0.2202
    Residual |  56559.2694   100  565.592694           R-squared     =  0.0430
-------------+------------------------------           Adj R-squared =  0.0142
       Total |  59097.8462   103  573.765497           Root MSE      =  23.782

------------------------------------------------------------------------------
       anvas |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Igroup_2 |  -7.997515   9.193631    -0.87   0.386    -26.23742    10.24239
        time |  -.0068926    .012107    -0.57   0.570    -.0309126    .0171274
_IgroXtime_2 |  -.0020711   .0174927    -0.12   0.906    -.0367761    .0326338
       _cons |   33.79133   6.347495     5.32   0.000     21.19808    46.38458
------------------------------------------------------------------------------

. predict anvasres1, resid
(4 missing values generated)

. xtsumcorr anvasres1

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
anvasr~1 overall |  3.64e-08   23.43329  -32.39845   69.22707 |     N =     104
         between |             19.62853  -30.60852   56.40872 |     n =      27
         within  |             13.02375  -34.83516   32.30041 | T-bar = 3.85185
   corr. between |             17.95069                       |
   corr. within  |             15.06292                       |
```

```
         rho     |              .5868  (betw. fract. of total) |
```

```
. variogram anvasres1, bw(0.8)
Computing smooth lowess model for v in ulag
```



The figure above shows the variogram after removing the time and treatment group effects. It is clear that the variogram does not monotonically increase over time lag. We can notice that the time lag is highly clustered. Therefore we can group time using hour as the unit for the time since treatment instead of using minute.

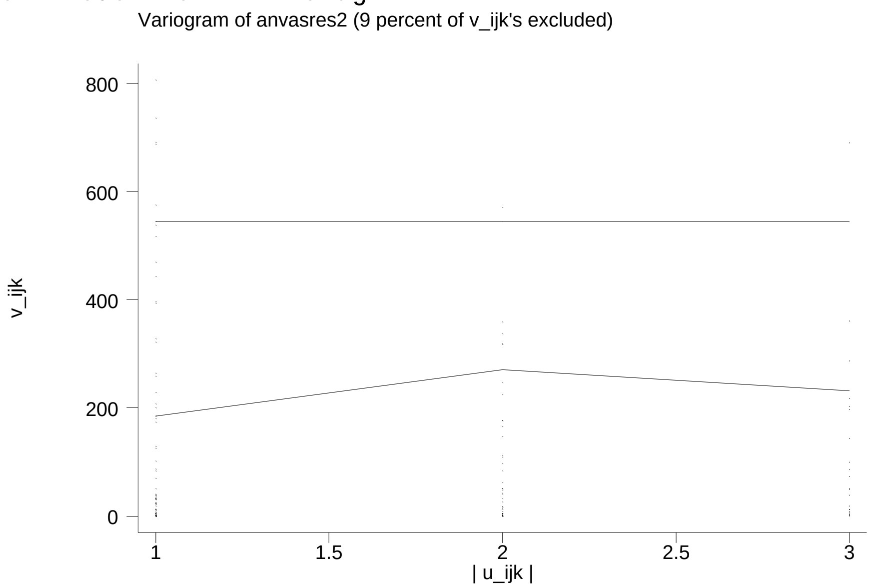
```
. gen time2hrs = round(time/60,1)
(1 missing value generated)
. tab  time2hrs set
```

time2hrs	set 1	2	3	4	Total
0	1	0	0	0	1
1	17	0	0	0	17
2	6	0	0	0	6
3	1	0	0	0	1
4	1	1	0	0	2
5	1	15	0	0	16
6	0	8	0	0	8
7	0	2	0	0	2
8	0	0	1	0	1
9	0	0	19	0	19
10	0	0	5	0	5
11	0	0	2	0	2
12	0	0	0	5	5
13	0	0	0	13	13
14	0	0	0	7	7
15	0	0	0	2	2
Total	27	26	27	27	107

From the table above, the time since treatment is clustered at 1, 5, 9 and 13 hours. Also it is highly related to the variable “set” (the correlation between time2hrs and set is 0.9830 which can be shown using “corr time2hrs set” in STATA). Thus we can use set as the time variable in variogram.

```
. tis set
. sort group set
. by group set: egen anvasmn = mean(anvas)
```

```
. gen anvasres2 = anvas - anvasmn
(4 missing values generated)

.  xtsumcorr anvasres2

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
anvasr~2 overall | -9.17e-08   23.33507  -33.15385         69 |     N =     104
         between |             19.62466  -30.46703   56.53297 |     n =      27
         within  |             12.85709  -34.06868   32.78846 | T-bar = 3.85185
   corr. between |             17.98343                       |
   corr. within  |             14.87017                       |
         rho     |             .5939  (betw. fract. of total) |

.  variogram anvasres2, discrete
Computing ANOVA model for v in ulag
```



The overall variance is 23.335^2 = 544.22. From the variogram, the between-subject "trait" variance is less than 200. The variogram is almost flat, suggesting that after removing the time and treatment effect, uniform correlation might be reasonable.

4 Confirmatory Data Analysis

(a)

The scientific purpose of the study is to compare two treatments (either intercostals/epidural analgesic, group 1, or morphine infusion analgesic, group 2) with respect to alertness. Since rating scores of alertness were obtained at different time since treatment, a model of alertness with treatment group, time, and their interaction may capture the scientific goal. The model is

$$E\,(\,alvas_{ij}\,) = \beta_0 + \beta_1\, group_i + \beta_2\ time_{ij} + \beta_3\, group_i * \ time_{ij}$$

where alvas is the rating score of alertness, group is treatment group indicator, time is the test time since treatment (in minutes), and βs are coefficients.

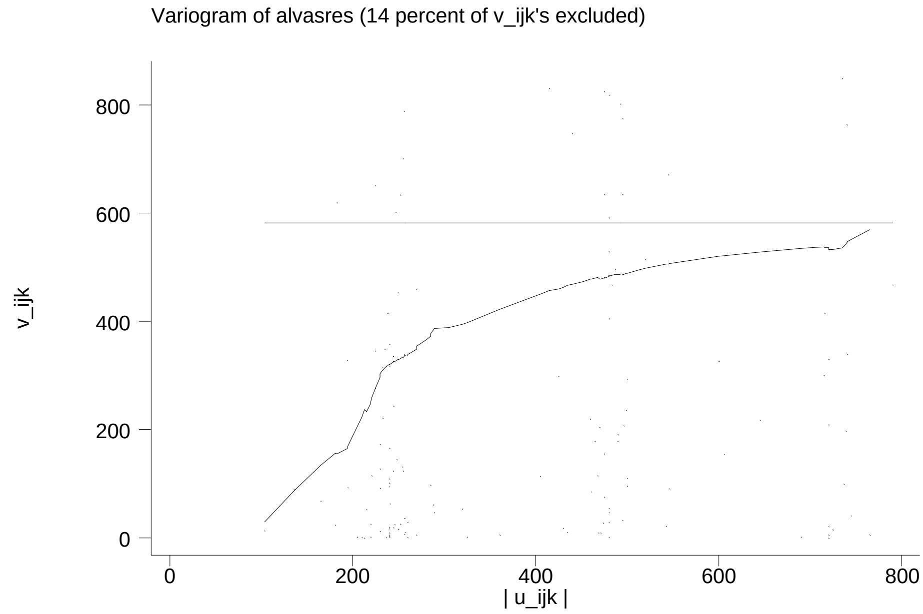
(b)

```
. xi:reg alvas i.group*time
i.group           _Igroup_1-2         (naturally coded; _Igroup_1 omitted)
i.group*time      _IgroXtime_#        (coded as above)

      Source |       SS       df       MS              Number of obs =     105
-------------+------------------------------           F(  3,   101) =    9.29
       Model |  16675.0656     3  5558.35521           Prob > F      =  0.0000
    Residual |  60461.0677   101  598.624433           R-squared     =  0.2162
-------------+------------------------------           Adj R-squared =  0.1929
       Total |  77136.1333   104   741.69359           Root MSE      =  24.467

------------------------------------------------------------------------------
       alvas |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Igroup_2 |   27.82285   9.399639     2.96   0.004       9.1765    46.46921
        time |  -.0174555   .0124556    -1.40   0.164     -.042164     .007253
_IgroXtime_2 |  -.0149187   .0179537    -0.83   0.408    -.0505341    .0206967
       _cons |   52.67963   6.530218     8.07   0.000     39.72544    65.63383
------------------------------------------------------------------------------
```

(c)

```
. predict alvasres, resid
(3 missing values generated)
. tsset id time
       panel variable:  id, 1 to 27
        time variable:  time, 25 to 890, but with gaps

. xtsumcorr alvasres

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
alvasres overall |  4.63e-08    24.11133  -42.21107    47.4845 |     N =     105
         between |              16.90955  -32.24049   33.04132 |     n =      27
         within  |              17.56376   -45.9137   41.28148 | T-bar = 3.88889
   corr. between |              13.04003                       |
   corr. within  |              20.28088                       |
         rho     |              .2925  (betw. fract. of total) |
```

```
. variogram alvasres, bw(0.8)
Computing smooth lowess model for v in ulag
```



The variogram of *alvas* (alertness) is shown above. The total variance is $24.111^2 = 581.34$. The between-subject “trait” variance and the measurement error variance are quite small.

(d)

The variogram of *alvas* shown in (c) suggests an exponential model for the correlation structure. The variogram increases monotonically with time lag, and saturates at a level very close to the total variance, a typical variogram pattern for an exponential correlation model.