Wednesday 13 February 2002

Analysis of pigs data

Data: Body weights of 48 pigs at 9 successive follow-up visits.

This is an equally spaced data. It is always a good habit to reshape the data, so we can easily switch form wide to long or long to wide depending on the required analysis. The data is in the wide format; let's reshape it into long format.

```
. reshape groups time 1-9 ## Since we have observations at 9 time points

. reshape vars week ## declare the variables

. reshape cons Id ## declare constants

. reshape long (wide)
```

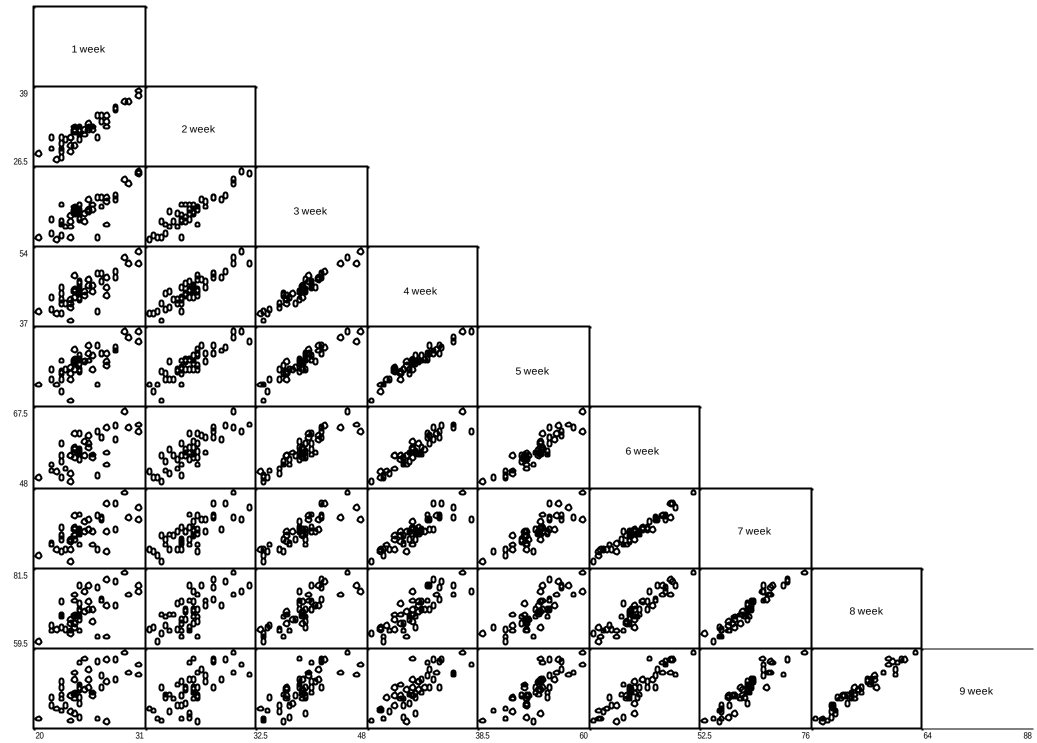
While reshaping the data from any format, the first three steps are the same. The lat command, `reshape long` or `reshape wide` depends on the structure of the original data.

Now we have the data in long format, to go into the wide form, just type `reshape wide.` You don't have to repeat the first three commands.

Exploratory analysis

Lets just make some scatter plots of the data. First we plot the scatter plot matrix. For this we require the data in wide format.

```
. reshape wide

. graph week1 week2 week3 week4 week5 week6 week7 week8 week9, matrix half
```

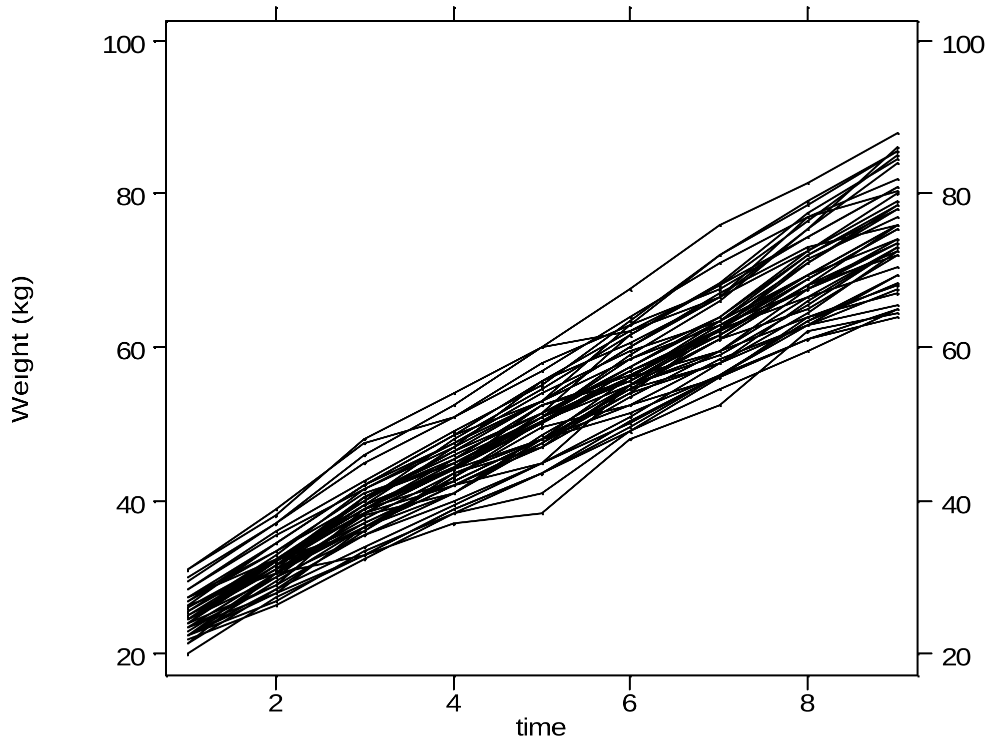


What do you conclude from the above scatter plot matrix?

To see if the pigs gained weight over time lets plot the line (spaghetti) plot. For this we need the data in the long form.

```
. reshape long

. sort Id time
```

```
. graph week time, c(L) s(i) xlab(2 4 6 8) ylab rlab
```



What do you conclude from the graph?

The above figure is enough to explore the growth data. It is hard to pick out individual response profiles. We can add a second display, obtained form first standardizing each observation. This is achieved by, subtracting the mean, and dividing by the standard deviation of the 48 observations at each time (week). For this we would need the data in wide format.

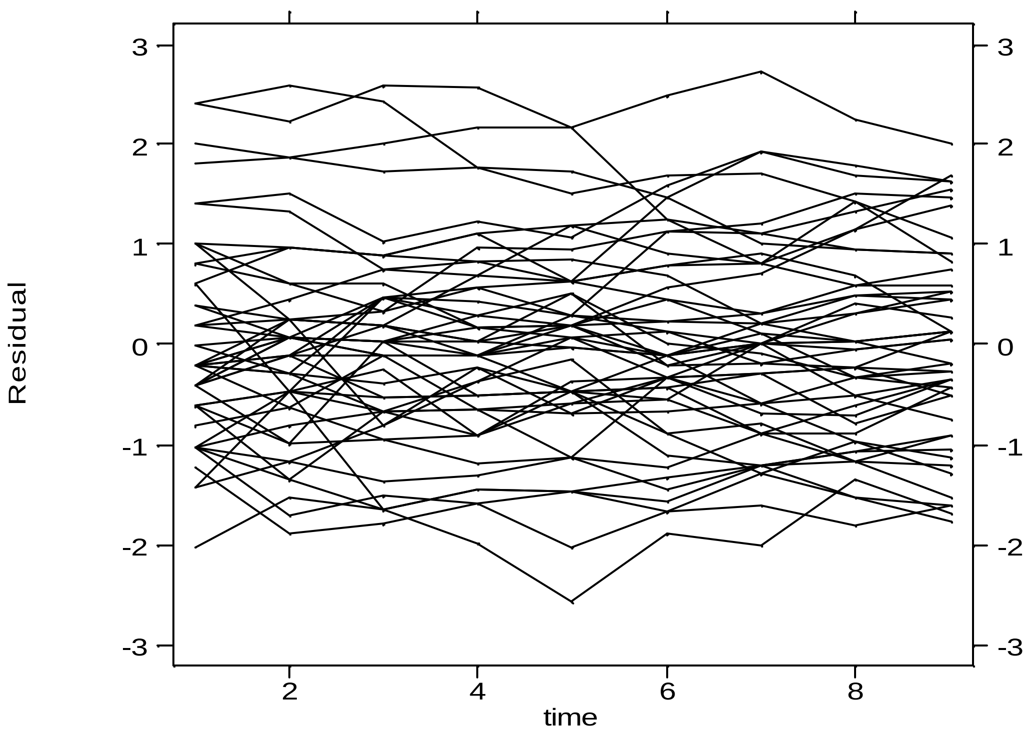
```
. reshape wide
```

Now do the following for each of week1, week2, …, week9

```
. sum week1

Variable |     Obs        Mean   Std. Dev.       Min        Max

---------+-----------------------------------------------------

   week1 |      48    25.02083   2.468866         20         31
. gen Sweek1 = (week1 – 25.02)/2.47
```

After you do this, we will have 9 new variables. To make the plot, again reshape long.

```
. sort Id time

. graph  Sweek time, c(L) s(.) xlab(2 4 6 8) ylab(-3 -2 -1 0 1 2 3)
rlab(-3 -2  -1 0 1 2 3)
```



The plot is able to highlight the degree of *tracking*, animals tend to maintain their relative size over time.

Exploring the correlation structure

Auto-correlation function. For this we require the data in long format.

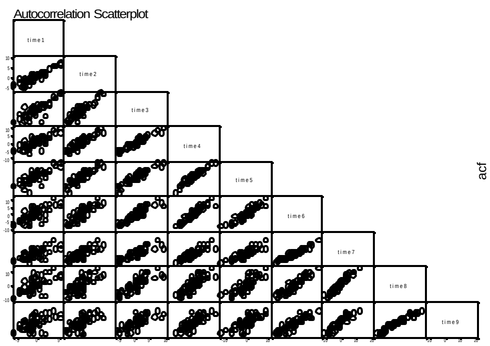
```
. autocor week time Id
```

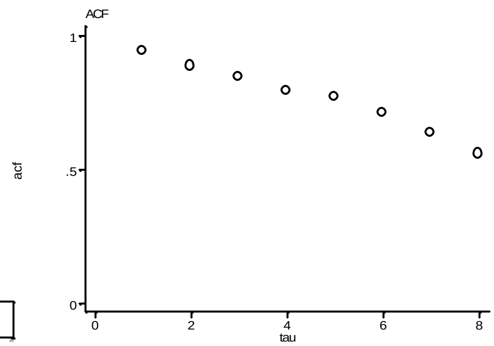
Table 1

	time1	time2	time3	time4	time5	time6	time7	time8	time9
time1	1.0000								
time2	0.9156	1.0000							
time3	0.8015	0.9118	1.0000						
time4	0.7958	0.9084	0.9582	1.0000					
time5	0.7494	0.8809	0.9280	0.9621	1.0000				
time6	0.7051	0.8353	0.9058	0.9327	0.9219	1.0000			
time7	0.6551	0.7759	0.8435	0.8681	0.8546	0.9633	1.0000		
time8	0.6255	0.7133	0.8167	0.8293	0.8104	0.9280	0.9586	1.0000	
time9	0.5581	0.6638	0.7689	0.7856	0.7856	0.8893	0.9170	0.9695	1.0000

Table 2

	acf
1.	.9425781
2.	.8870165
3.	.8462396
4.	.7962576
5.	.7724156
6.	.7121489
7.	.6407955
8.	.5581002





Notice from the main diagonal of the scatter plot matrix there is positive correlation between repeated observations on the same animal that are 1 week apart. The degree of correlation decreases as the observations are moved farther from the diagonals. Also the correlation is reasonably consistent along the diagonal in the matrix. This indicates that the correlation depends on the time between observations than their absolute times. The estimated correlation matrix for this data is given in table 1. The correlations show some tendency of decrease with increasing time lag. Assuming stationarity, a single correlation estimate can be obtained for each distinct value of the time separation or lag, $|t_{ij} - t_{ik}|$. This corresponds to pulling observation pairs along the diagonals of the scatter plot matrix. The autocorrelation function takes the value as in table 2.

Before we proceed with the analysis, lets look at some theory.
y_{ij} $j = 1, 2, \ldots, n$ be the sequence of observed measurements on the *ith* of the m subjects and t_j, $j = 1, 2, \ldots, n$ be the corresponding times at which the measurements are taken on each unit. Associated with each y_{ij} are the values, x_{ijk}, $k = 1, 2, \ldots, p$ of p explanatory variables. We assume that y_{ij} are realizations of random variables Y_{ij} which follow the regression model

$$Y_{ij} = b_1 x_{ij1} + \cdots + b_p x_{ijp} + e_{ij},$$

In the classical linear model we assume the errors to be mutually independent normal random variables. In our context, the longitudinal structure of the data means that we expect the errors to be correlated within subjects.
Let $\boldsymbol{y_i} = (y_{i1}, y_{i2}, \ldots, y_{in})$ be the observed sequence of measurements on the *ith* subject and $\boldsymbol{y} = (\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y_m})$ be the complete set of $N = nm$ observations. Let **X** be the matrix of explanatory variables.

$$Y \sim MVN(Xb, s^2 V)$$

<u>The Uniform Correlation Model</u>

In this model we assume that there is positive correlation between any two measurements.

<u>The Exponential Correlation Model</u>

The correlation between a pair of measurements on the same unit decays towards zero as the time separation between measurements increases.
The exponential correlation model is sometimes called the *first order autoregressive model*.

For the data on pigs we fit a couple of models. We require the data in long format

1. Ordinary least squares ignoring correlation

```
.regress week time

Source |       SS       df       MS                  Number of obs =     432
---------+------------------------------              F(  1,   430) = 5757.41
   Model |  111060.882     1  111060.882              Prob > F      =  0.0000
Residual |  8294.72677   430  19.2900622              R-squared     =  0.9305
---------+------------------------------              Adj R-squared =  0.9303
   Total |  119355.609   431  276.927167              Root MSE      =   4.392

------------------------------------------------------------------------------
    week |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    time |   6.209896   .0818409     75.878   0.000       6.049038    6.370754
   _cons |   19.35561   .4605447     42.028   0.000       18.45041    20.26081
------------------------------------------------------------------------------
```

2. Independent correlation model

```
. xtgee week time, i(Id) corr(ind)

Iteration 1: tolerance = 1.848e-15

GEE population-averaged model                   Number of obs      =       432
Group variable:                         Id      Number of groups   =        48
Link:                             identity      Obs per group: min =         9
Family:                           Gaussian                     avg =       9.0
Correlation:                   independent                     max =         9
                                                Wald chi2(1)       =   5757.41
Scale parameter:                  19.29006      Prob > chi2        =    0.0000

Pearson chi2(430):                 8294.73      Deviance           =   8294.73
Dispersion (Pearson):             19.29006      Dispersion         =  19.29006

------------------------------------------------------------------------------
    week |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    time |   6.209896   .0818409     75.878   0.000       6.049491    6.370301
   _cons |   19.35561   .4605447     42.028   0.000       18.45296    20.25826
------------------------------------------------------------------------------

xtcorr   ## estimates the correlation matrix

Estimated within-Id correlation matrix R:

        c1      c2      c3      c4      c5      c6      c7      c8      c9
r1  1.0000
r2  0.0000  1.0000
r3  0.0000  0.0000  1.0000
r4  0.0000  0.0000  0.0000  1.0000
r5  0.0000  0.0000  0.0000  0.0000  1.0000
r6  0.0000  0.0000  0.0000  0.0000  0.0000  1.0000
r7  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  1.0000
r8  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  1.0000
r9  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  1.0000
```

3. Exponential correlation model

```
. xtgee week time, i(Id) corr(ar1) t(time)

Iteration 1: tolerance = .02516015
Iteration 2: tolerance = .00009265
Iteration 3: tolerance = 4.393e-07

GEE population-averaged model                   Number of obs      =       432
Group and time vars:                Id time     Number of groups   =        48
Link:                              identity     Obs per group: min =         9
Family:                            Gaussian                    avg =       9.0
Correlation:                          AR(1)                    max =         9
                                                Wald chi2(1)       =   6255.06
Scale parameter:                   19.35735     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
    week |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    time |   6.272119   .0793047     79.089   0.000       6.116685    6.427553
   _cons |   18.84162   .6759819     27.873   0.000       17.51672    20.16652
------------------------------------------------------------------------------

. xtcorr

Estimated within-Id correlation matrix R:

        c1      c2      c3      c4      c5      c6      c7      c8      c9
r1  1.0000
r2  0.9172  1.0000
r3  0.8413  0.9172  1.0000
r4  0.7716  0.8413  0.9172  1.0000
r5  0.7077  0.7716  0.8413  0.9172  1.0000
r6  0.6491  0.7077  0.7716  0.8413  0.9172  1.0000
r7  0.5954  0.6491  0.7077  0.7716  0.8413  0.9172  1.0000
r8  0.5461  0.5954  0.6491  0.7077  0.7716  0.8413  0.9172  1.0000
r9  0.5008  0.5461  0.5954  0.6491  0.7077  0.7716  0.8413  0.9172  1.0000
```

4. Between effects WLS (usually done in case of unbalanced data instead of OLS)

```
. xtreg week time, be wls i(Id)
```

5. Random-effects model

```
. xtreg week time, re i(Id)

Random-effects GLS regression                   Number of obs      =       432
Group variable (i) : Id                         Number of groups   =        48

R-sq:  within  = 0.9851                         Obs per group: min =         9
       between = 0.0000                                        avg =       9.0
       overall = 0.9305                                        max =         9

Random effects u_i ~ Gaussian                   Wald chi2(1)       =  25271.50
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
    week |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    time |   6.209896   .0390633    158.970   0.000       6.133333    6.286458
   _cons |   19.35561    .603139     32.091   0.000       18.17348    20.53774
---------+--------------------------------------------------------------------
 sigma_u |  3.8912529
 sigma_e |  2.0963559
     rho |  .77505208   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

6. xtreg, mle

```
. xtreg week time, i(id) mle

Random-effects ML regression                    Number of obs      =       432
Group variable (i) : Id                         Number of groups   =        48

Random effects u_i ~ Gaussian                   Obs per group: min =         9
                                                               avg =       9.0
                                                               max =         9

                                                LR chi2(1)         =   1624.57
Log likelihood  = -1014.9268                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
    week |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    time |   6.209896   .0390124    159.178   0.000       6.133433    6.286359
   _cons |   19.35561   .5974055     32.399   0.000       18.18472    20.52651
---------+--------------------------------------------------------------------
/sigma_u |    3.84935   .4058114      9.486   0.000       3.053974    4.644725
/sigma_e |   2.093625   .0755471     27.713   0.000       1.945555    2.241694
---------+--------------------------------------------------------------------
     rho |    .771714   .0393959                          .6876303    .8413114
------------------------------------------------------------------------------
Likelihood ratio test of sigma_u=0:  chi2(1) =   472.65   Prob > chi2 = 0.0000
```

Note that, `xtreg, re` and `xtreg, mle` give the same estimates.

The following table might be useful for future reference. To learn more explore the `xt` set of commands in STATA

If I want to do	STATA command
OLS	`regress Y X`
Between-effects model (WLS)	`xtreg Y X, be i(id) wls` if data is unbalanced specify wls instead of OLS, default is OLS
Random-effects model (mixed)	`xtreg Y X, re i(id)`
Fixed effect model (within)	`xtreg Y X, fe i(id)`
MLE random effect model	`xtreg Y X, i(id) mle`

Population average model	`xtreg Y X, pa i(id)`
	This command is equivalent to
	`xtgee Y X, f(gaussian) link(id) corr(exc)`

Generalized least squares	`xtgls Y X, i(id) corr(ar1 or ind)` See STATA help before you use this command

Population averaged panel data using GEE	`xtgee Y X, f(family) l(link) corr(correlation) i(id)`
	for example:
	`xtgee Y X1 X2, f(gauss) l(id) corr(exc) i(id)`
	would estimate population-averaged linear regression with equal correlation comparable to random-effects regression fro `xtreg`