# Monday 7th Febraury 2005

**Analysis of Pigs data**

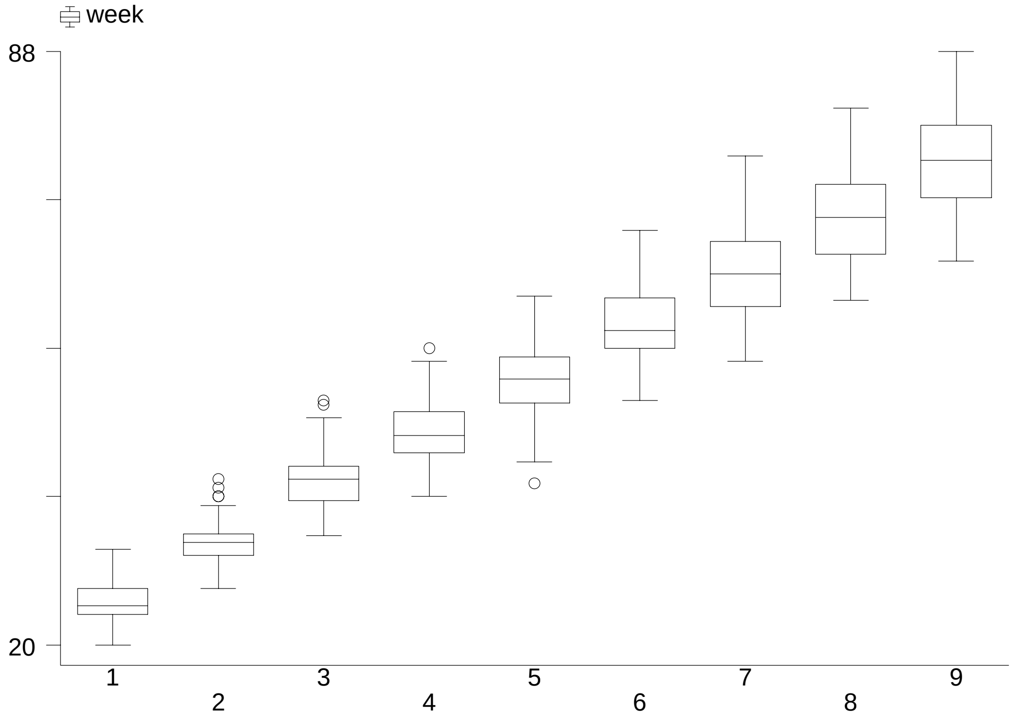
Data: Body weights of 48 pigs at 9 successive follow-up visits.

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.

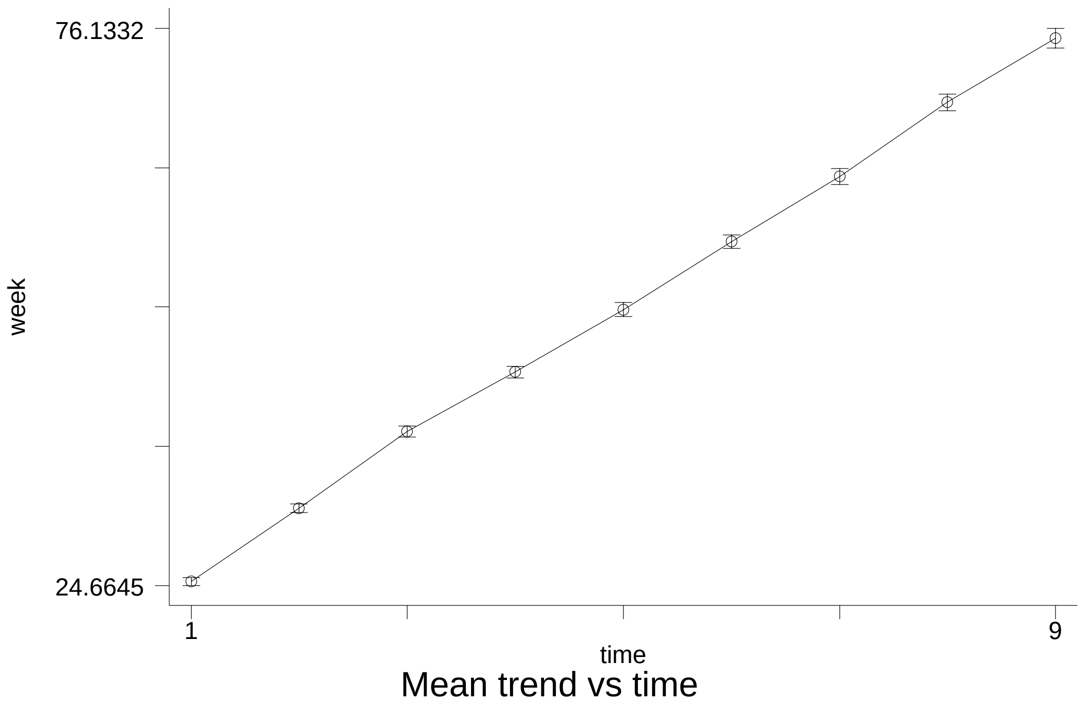
```
. set memory 40m
(40960k)
. set matsize 100

. reshape long week, i(Id) j(time)
(note: j = 1 2 3 4 5 6 7 8 9)

Data                               wide   ->   long
-----------------------------------------------------------------
Number of obs.                       48   ->    432
Number of variables                  10   ->      3
j variable (9 values)                     ->   time
xij variables:
               week1 week2 ... week9   ->   week
-----------------------------------------------------------------

. tsset Id time
      panel variable:  Id, 1 to 48
       time variable:  time, 1 to 9

. xtdes
     Id:  1, 2, ..., 48                              n =         48
   time:  1, 2, ..., 9                               T =          9
          Delta(time) = 1; (9-1)+1 = 9
          (Id*time uniquely identifies each observation)

Distribution of T_i: min    5%    25%    50%      75%     95%     max
                      9      9     9        9        9       9       9
    Freq.  Percent    Cum. |  Pattern
 ---------------------------+-----------
      48    100.00  100.00 |  111111111
 ---------------------------+-----------
      48    100.00         |  XXXXXXXXX
```

**Exploratory Data Analysis**

```
. sort time
. graph week, by(time) box
. ** STATA 8 command : graph box week, by(time) **
```




```
. ** mean trend plot **
. xtgraph week, ti("Mean trend vs time") bar(se)
```



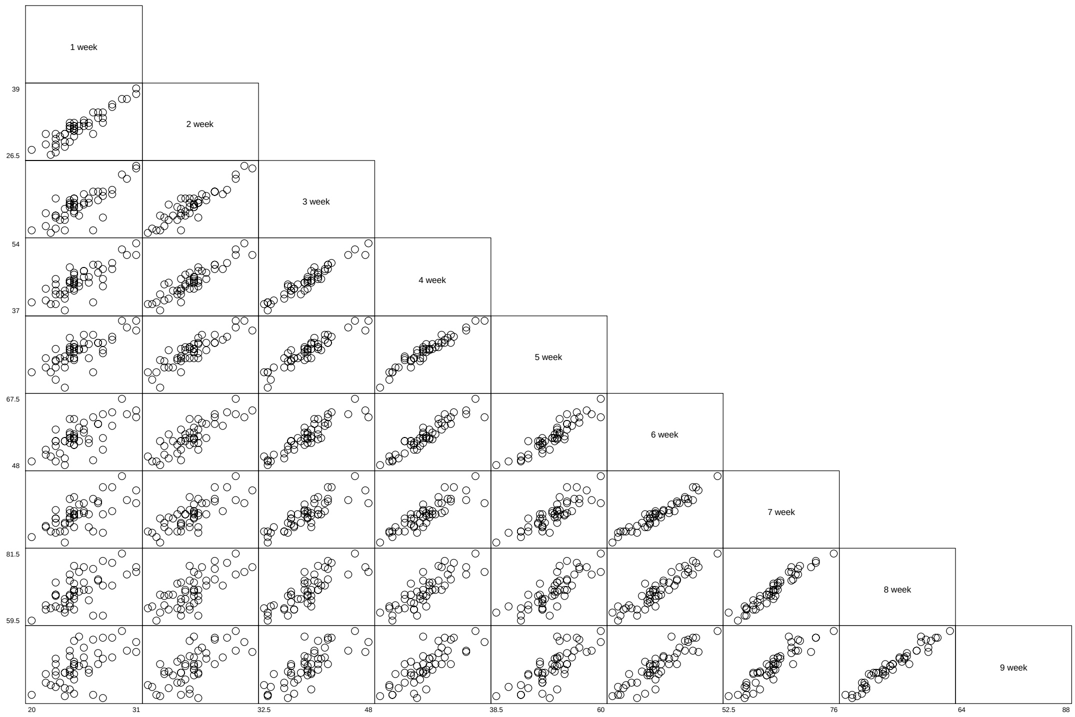

We can see the mean trend is quite linear across time.

Now we make scatter plots of the data, pair wise scatter plots for data at each two time points. For this we requite the data in wide format.

```
. reshape wide week, i(Id) j(time)
```

```
(note: j = 1 2 3 4 5 6 7 8 9)

Data                               long   ->   wide
-----------------------------------------------------------------
Number of obs.                       432   ->     48
Number of variables                    3   ->     10
j variable (9 values)               time   ->   (dropped)
xij variables:
                                   week   ->   week1 week2 ... week9
-----------------------------------------------------------------
```

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

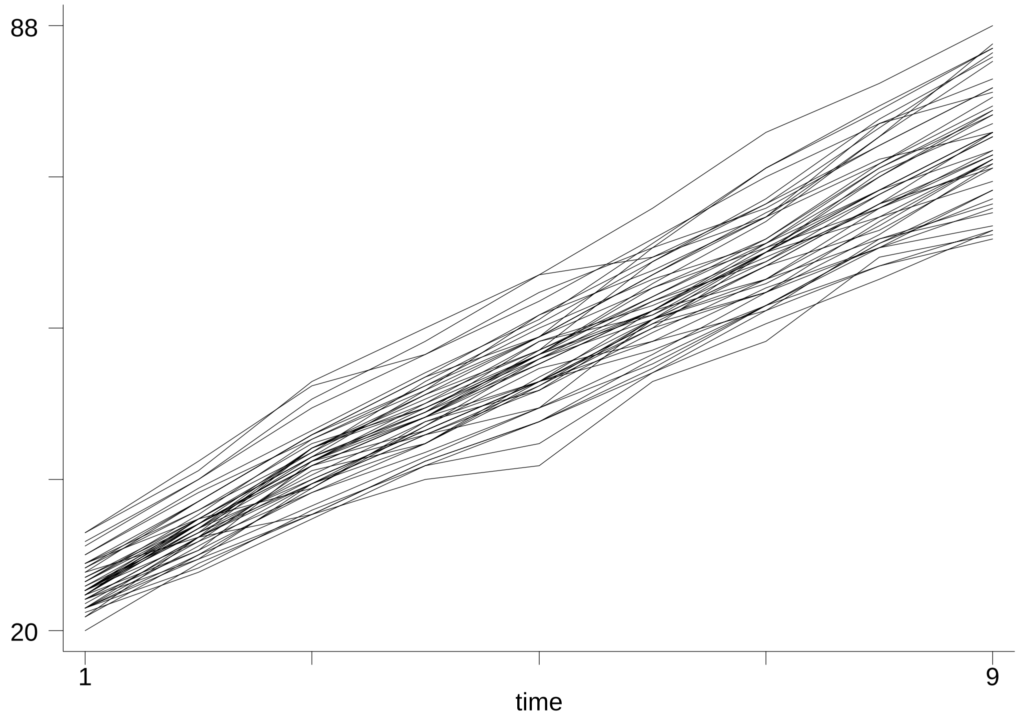



What do you conclude from the above scatter plot matrix?

A relatively constant positive linear trend for all pair wise scatter plots indicates that heavier pigs tend to remain heavier across all time points.

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.

```
. sort Id time
. graph week time, c(L) s(i)
. ** STATA 8 command : twoway line week time, c(L) s(i) **
```




What do you conclude from the graph?

A linear relationship between outcome and time is showed. Also, not much cross-over of these lines indicates that the relative order of pigs, ordered by their weights, remain unchanged over time, which confirms the conclusion we drew from the scatter matrix plot.

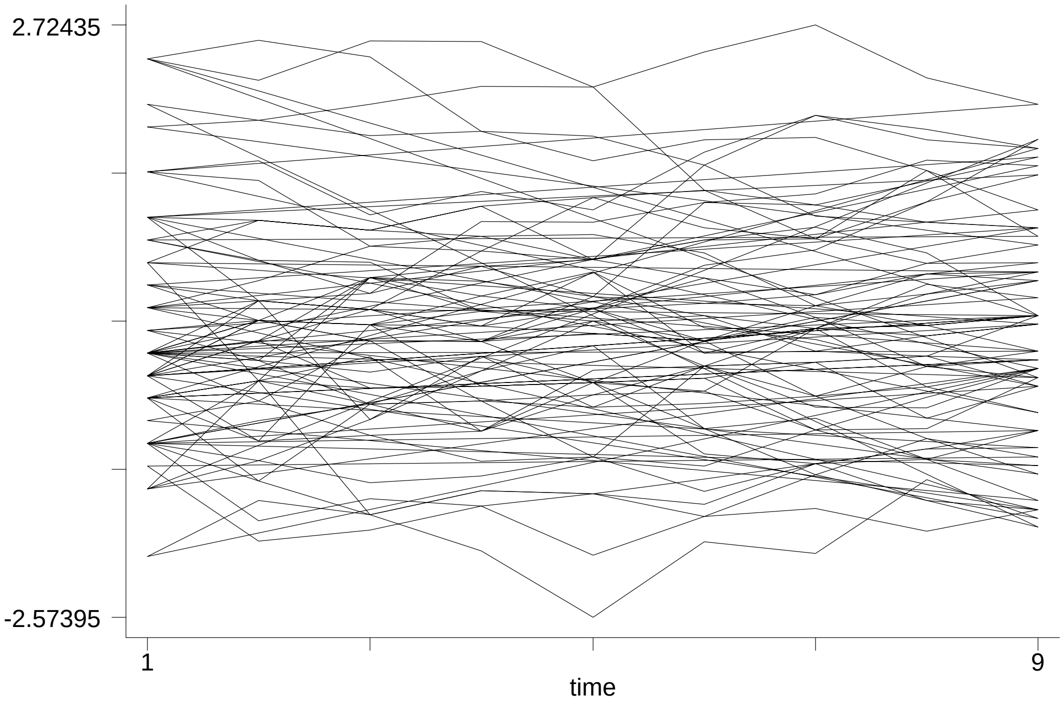
The above figure is enough to explore the growth data. It is hard to pick out individual response profiles. We can add a second display, obtained form first standardizing each observation. This is achieved by, subtracting the mean, and dividing by the standard deviation of the 48 observations at each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

```
. sort time
. by time: sum week
```

---

```
-> time = 1
   Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       week |      48    25.02083   2.468866        20        31
--more—
. reshape wide week, i(Id) j(time)

. gen sweek1 = (week1 - 25.02)/2.47
. gen sweek2 = (week2 - 31.78)/2.79
. gen sweek3 = (week3 - 38.86)/3.54
. gen sweek4 = (week4 - 44.39)/3.73
. gen sweek5 = (week5 - 50.16)/4.53
```

```
. gen sweek6 = (week6 - 56.45)/4.45
. gen sweek7 = (week7 - 62.46)/4.97
. gen sweek8 = (week8 - 69.30)/5.42
. gen sweek9 = (week9 - 75.22)/6.34

. reshape long week sweek, i(Id) j(time)

. sort Id time
. graph sweek time, c(L) s(i)
. ** STATA 8 command : twoway line sweek time, c(L) s(i) **
```




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
```

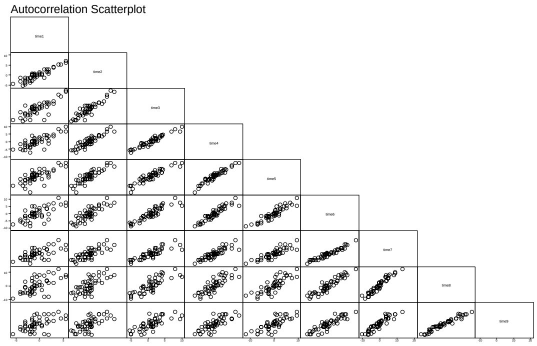
**Table 1**

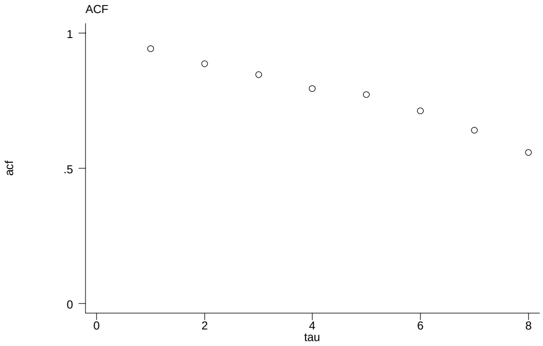
| | time1 | time2 | time3 | time4 | time5 | time6 | time7 | time8 | time9 |
|---|---|---|---|---|---|---|---|---|---|
| time1 | 1.0000 | | | | | | | | |
| time2 | 0.9156 | 1.0000 | | | | | | | |
| time3 | 0.8015 | 0.9118 | 1.0000 | | | | | | |
| time4 | 0.7958 | 0.9084 | 0.9582 | 1.0000 | | | | | |
| 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

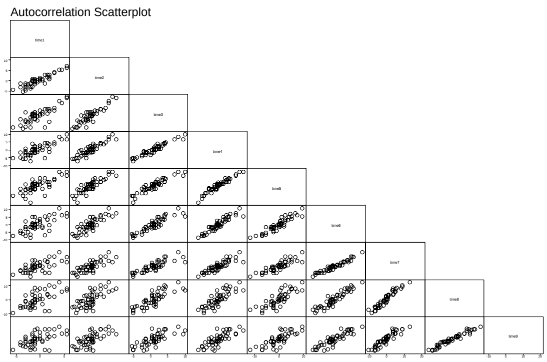
1. .9425781
2. .8870165
3. .8462396
4. .7962576
5. .7724156
6. .7121489
7. .6407955
8. .5581002

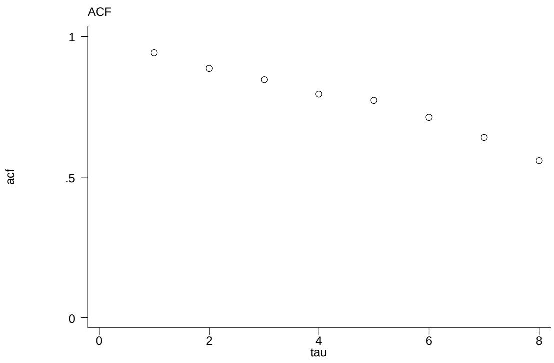



What is the correlation structure taking away the covariate’s effect?

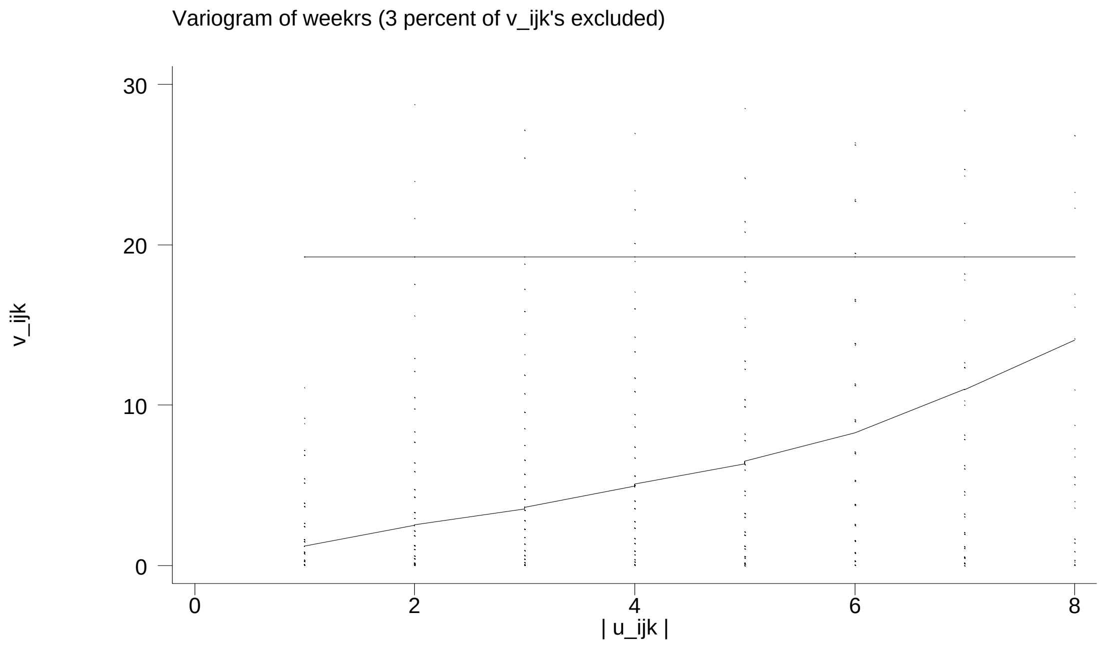
```
. regress week time
. predict weekrs, resid
. autocor weekrs time Id
```




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.

```
. variogram weekrs
```



We can conclude 4 things from this graph:

1. The variogram starts at 0, that means, there is no measurement error.
2. It doesn't reach the total at the end, that means, we need to model a random intercept for the correlation.
3. We observe an ascending line, that means, there is series correlation in the data, the correlation between measurements from the same pig gets smaller when the time interval becomes larger.
4. No need to model random slope, since we don't observe multiple lines.

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

$$Y_{ij} = \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + \varepsilon_{ij},$$

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 $\mathbf{X}$ be the matrix of explanatory variables.

$$Y \sim MVN(X\beta, \sigma^2 V)$$

**The Uniform Correlation Model**
In this model we assume that there is positive correlation between any two measurements.

**The Exponential Correlation Model**
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    75.88   0.000     6.049038    6.370754
       _cons |   19.35561   .4605447    42.03   0.000     18.45041    20.26081
------------------------------------------------------------------------------
```

2. Independent correlation model

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

Iteration 1: tolerance = 2.217e-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)       =   5784.19
Scale parameter:                  19.20076      Prob > chi2        =    0.0000
Pearson chi2(432):                 8294.73      Deviance           =   8294.73
Dispersion (Pearson):             19.20076      Dispersion         =  19.20076
------------------------------------------------------------------------------
        week |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   6.209896   .0816513    76.05   0.000     6.049862    6.369929
       _cons |   19.35561   .4594773    42.13   0.000     18.45505    20.25617
------------------------------------------------------------------------------

. xtcorr
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
```

**. xtgls week time, i(Id) corr(ind)**

```
Cross-sectional time-series FGLS regression
Coefficients:  generalized least squares
Panels:        homoscedastic
Correlation:   no autocorrelation

Estimated covariances      =         1          Number of obs      =       432
Estimated autocorrelations =         0          Number of groups   =        48
Estimated coefficients     =         2          No. of time periods=         9
                                                Wald chi2(1)       =   5784.19
Log likelihood             = -1251.251          Prob > chi2        =    0.0000
------------------------------------------------------------------------------
        week |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   6.209896   .0816513    76.05   0.000     6.049862    6.369929
       _cons |   19.35561   .4594773    42.13   0.000     18.45505    20.25617
------------------------------------------------------------------------------
```

## 3. Uniform correlation model

**. xtgee week time, i(Id) corr(exc)**

```
Iteration 1: tolerance = 5.934e-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:                exchangeable                    max =         9
                                             Wald chi2(1)       =  25337.48
Scale parameter:                19.20076     Prob > chi2        =    0.0000
------------------------------------------------------------------------------
        week |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   6.209896   .0390124   159.18   0.000     6.133433    6.286359
       _cons |   19.35561   .5974055    32.40   0.000     18.18472    20.52651
------------------------------------------------------------------------------
```

```
. xtcorr
Estimated within-Id correlation matrix R:
       c1      c2      c3      c4      c5      c6      c7      c8      c9
r1  1.0000
r2  0.7717  1.0000
r3  0.7717  0.7717  1.0000
r4  0.7717  0.7717  0.7717  1.0000
r5  0.7717  0.7717  0.7717  0.7717  1.0000
r6  0.7717  0.7717  0.7717  0.7717  0.7717  1.0000
r7  0.7717  0.7717  0.7717  0.7717  0.7717  0.7717  1.0000
r8  0.7717  0.7717  0.7717  0.7717  0.7717  0.7717  0.7717  1.0000
r9  0.7717  0.7717  0.7717  0.7717  0.7717  0.7717  0.7717  0.7717  1.0000
```

**. xtreg week time, i(Id) pa** (equivalent to the previous one)

```
Iteration 1: tolerance = 6.283e-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:                 exchangeable                    max =        9
                                              Wald chi2(1)       =  25337.48
Scale parameter:                 19.20076     Prob > chi2        =    0.0000
------------------------------------------------------------------------------
       week |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       time |   6.209896   .0390124   159.18   0.000     6.133433    6.286359
      _cons |   19.35561   .5974055    32.40   0.000     18.18472    20.52651
------------------------------------------------------------------------------
```

## 5. Random effect model with random intercept

**. 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   158.97   0.000     6.133333    6.286458
      _cons |   19.35561    .603139    32.09   0.000     18.17348    20.53774
```

```
-------------+----------------------------------------------------------------
     sigma_u |  3.8912528
     sigma_e |  2.0963561
         rho |  .77505203   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

**. xtreg week time, i(Id) mle**

```
Fitting constant-only model:
Iteration 0:   log likelihood = -13399.842
Iteration 1:   log likelihood = -8045.1554
Iteration 2:   log likelihood = -5080.5329
Iteration 3:   log likelihood = -3459.5929
Iteration 4:   log likelihood =  -2593.189
Iteration 5:   log likelihood = -2148.9623
Iteration 6:   log likelihood = -1938.3661
Iteration 7:   log likelihood = -1853.2441
Iteration 8:   log likelihood = -1830.3843
Iteration 9:   log likelihood = -1827.3012
Iteration 10:  log likelihood =  -1827.212
Iteration 11:  log likelihood = -1827.2118

Fitting full model:
Iteration 0:   log likelihood = -1014.9757
Iteration 1:   log likelihood = -1014.9268
Iteration 2:   log likelihood = -1014.9268

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   159.18   0.000     6.133433    6.286359
       _cons |   19.35561   .5974055    32.40   0.000     18.18472    20.52651
-------------+----------------------------------------------------------------
    /sigma_u |    3.84935   .4058114     9.49   0.000     3.053974    4.644725
    /sigma_e |   2.093625   .0755471    27.71   0.000     1.945555    2.241694
-------------+----------------------------------------------------------------
         rho |    .771714   .0393959                      .6876303    .8413114
------------------------------------------------------------------------------
```

```
Likelihood ratio test of sigma_u=0: chibar2(01)=  472.65 Prob>=chibar2 = 0.000
```

## 6. Exponential correlation model

**. xtgls week time, igls corr(ar1) i(Id) force**

```
Iteration 1: tolerance = 0

Cross-sectional time-series FGLS regression

Coefficients:  generalized least squares
Panels:        homoscedastic
Correlation:   common AR(1) coefficient for all panels  (0.9161)

Estimated covariances      =         1          Number of obs      =       432
Estimated autocorrelations =         1          Number of groups   =        48
Estimated coefficients     =         2          No. of time periods=         9
                                                Wald chi2(1)       =   7867.89
Log likelihood             = -806.5921          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        week |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   6.272057     .07071    88.70   0.000     6.133467    6.410646
       _cons |   18.84279   .6001148    31.40   0.000     17.66658    20.01899
------------------------------------------------------------------------------
```

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

```
Iteration 1: tolerance = .02513276
Iteration 2: tolerance = .00009237
Iteration 3: tolerance = 4.366e-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)       =   6254.91
Scale parameter:                    19.26754    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        week |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   6.272089   .0793052    79.09   0.000     6.116654    6.427524
       _cons |   18.84218   .6745715    27.93   0.000     17.52004    20.16431
------------------------------------------------------------------------------
```