1. Objective

- Introduction to SAS PROC MIXED
- Analyzing protein milk data using STATA
- Refit protein milk data using PROC MIXED

2. Introduction to SAS PROC MIXED

The MIXED procedure provides you with flexibility of modeling not only the means of yours data (as in the standard linear model) but also their variances and covariance as well (the mixed linear model).

- **THE MIXED LINEAR MODEL**
  - **The standardized linear model**
    - $\mathbf{Y=X\beta+\varepsilon}$
    - β is an unknown vector of fixed-effects parameters with known design matrix X.
    - ε is assumed to be independent and identically distributed Gaussian random variables with mean 0 and variance σ^2.
  - **The mixed linear model**
    - a generalized version of the standardized linear model as follows:
      $\mathbf{Y=X\beta+Z\gamma+\varepsilon}$
    - γ is an unknown vector of random-effects parameters with known design matrix Z
    - ε is the residuals vector whose elements are no longer required to be independent and homogeneous, and its variance is **R**.
    - The variance of **Y** is **V = ZGZ' + R**
    - For G and R, you must select some covariance structure.

- **SYNTAX: (details refer to SAS help)**

PROC MIXED < options > **;**
BY variables **;**
CLASS variables **;**
ID variables **;**
MODEL dependent = < fixed-effects > < / options > **;**
RANDOM random-effects < / options > **;**
REPEATED < repeated-effect > < / options > **;**
PARMS (value-list) ... < / options > **;**
PRIOR < distribution > < / options > **;**
CONTRAST 'label' < fixed-effect values ... >
< | random-effect values ... > , ... < / options > **;**
ESTIMATE 'label' < fixed-effect values ... >
< | random-effect values ... >< / options > **;**
LSMEANS fixed-effects < / options > **;**
MAKE 'table' **OUT=**SAS-data-set **;**
WEIGHT variable **;**
RUN;

Let's look at **method=options for PROC**, **CLASS**, **MODEL, RANDOM**, and **REPEATED**.

- **Method=option**

  The METHOD= option specifies the estimation method for the covariance parameters. The REML specification performs restricted maximum likelihood, and it is the default method. The ML specification performs maximum likelihood.

- **CLASS**
  - The CLASS statement names the classification variables to be used in the analysis.
  - If the CLASS statement is used, it must appear before the MODEL statement.
  - Classification variables can be either character or numeric.
- **MODEL**
  - MODEL dependent = < fixed-effects >< / options >;
  - The MODEL statement names a single dependent variable and the fixed effects, which determine the **X** matrix of the mixed model.
- **RANDOM**
  - RANDOM random-effects < / options >;
  - Define **Z**
  - Define $\boldsymbol{\gamma}$
  - Define **G**
- **REPEATED**
  - REPEATED < repeated-effect > < / options >;
  - Specify the R matrix in the mixed model.

3. **Dataset:** Protein milk data set (in the class website)

   Data description: Percentage protein content of milk samples at weekly intervals from each of 25 cows on barley diet, 27 cows on mixed diet and 27 cows on lupins diet.

4. **STATA output of the analysis**

- **Read data into STATA**

```
.log using c:\data\lab7sup,replace
. set mem 50m
(51200k)
. set matsize 800
. *read the data into STATA from three text files
. infile y1 y2 y3 y4 y5 y6 y7 y8 y9 y10 y11 y12 y13 y14 y15  y16 y17 y18 y19 using
c:\data\cows.barley.data, clear
(25 observations read)
. gen id=_n
. gen grp=1
. count
   25
. save c:\data\milk1, replace
file c:\data\milk1.dta saved
. infile y1 y2 y3 y4 y5 y6 y7 y8 y9 y10 y11 y12 y13 y14 y15  y16 y17 y18 y19 using
c:\data\cows.mixed.data, clear
```

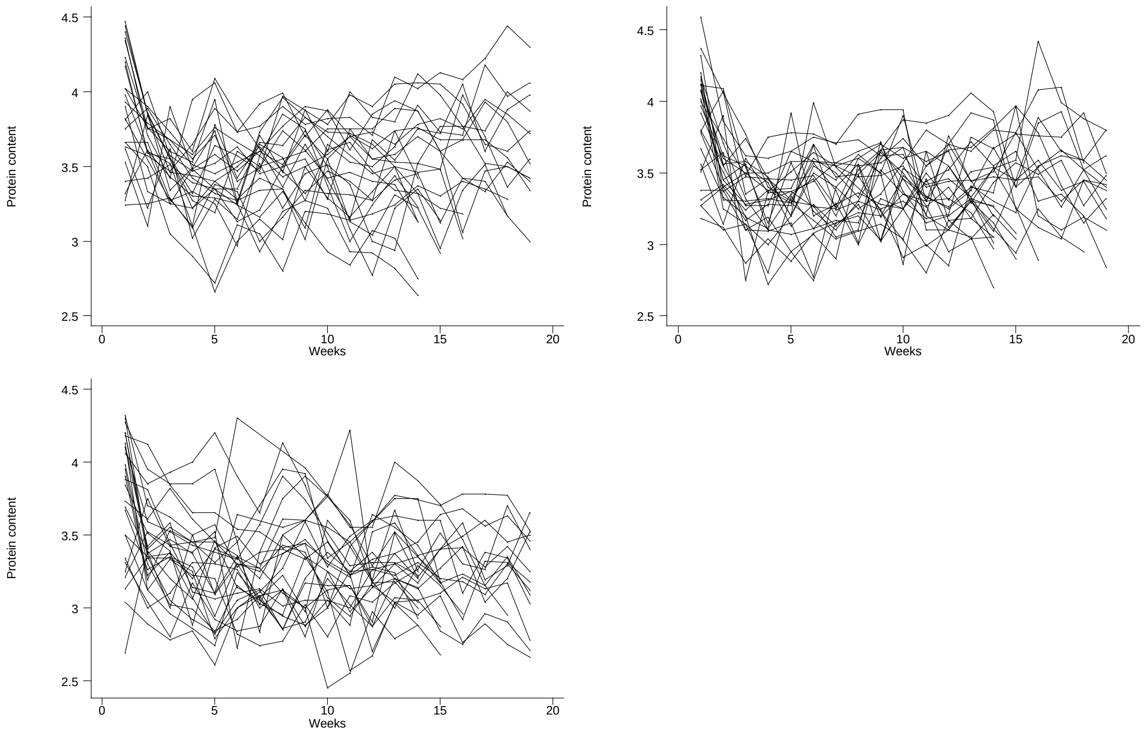
```
(27 observations read)
. gen id=_n+25
. gen grp=2
. count
   27
. save c:\data\milk2, replace
file c:\data\milk2.dta saved
. infile y1 y2 y3 y4 y5 y6 y7 y8 y9 y10 y11 y12 y13 y14 y15  y16 y17 y18 y19 using
c:\data\cows.lupins.data, clear
(27 observations read)
. gen id=_n+52
. gen grp=3
. append using "c:\data\milk1.dta" , nolabel
. append using "c:\data\milk2.dta", nolabel
. sort id
. save c:\data\milk, replace
file c:\data\milk.dta saved
```

- **Reshape to long format**

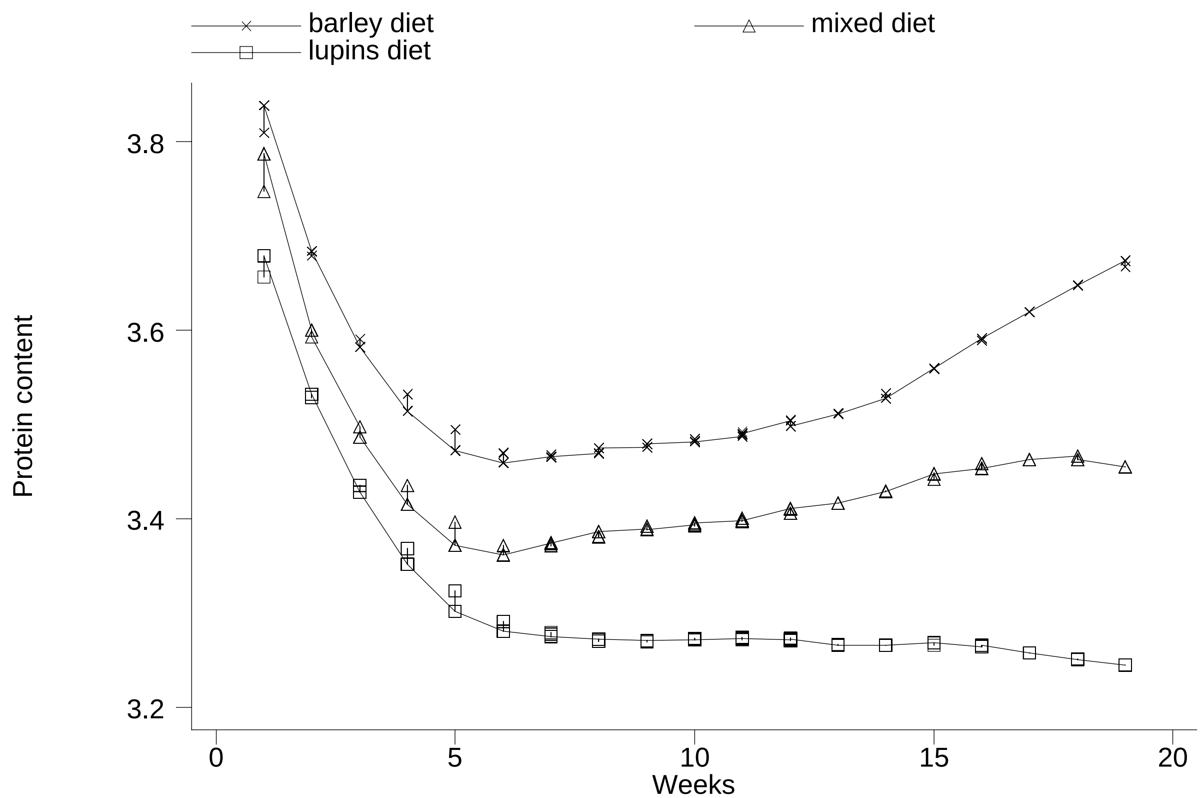
```
. reshape long y, i(id) j(t)
(note: j = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                       79   ->    1501
Number of variables                  21   ->       4
j variable (19 values)                    ->   t
xij variables:
                          y1 y2 ... y19   ->   y
-----------------------------------------------------------------------------

. *recode missing value
. replace y=. if y==0
(164 real changes made, 164 to missing)

. label var y "Protein content"
. label var t "Weeks"
. label define group 1 "barley diet" 2 "mixed diet" 3 "lupins diet"
. label value grp group
. save c:\data\milk, replace
file c:\data\milk.dta saved
```

- **Some plots**

```
. *spaghetti plots
. sort id t grp
. graph y t if grp==1, c(L) s(.) xlab ylab saving(g1,replace)
. graph y t if grp==2, c(L) s(.) xlab ylab saving(g2,replace)
. graph y t if grp==3, c(L) s(.) xlab ylab saving(g3,replace)
. graph using g1 g2 g3
```




```
.

. *smoothing plots
. gen y1 = y if grp == 1
(1076 missing values generated)
. gen y2 = y if grp == 2
(1042 missing values generated)
. gen y3 = y if grp == 3
(1048 missing values generated)
. ksm y1 t, gen(sm1) lowess bw(0.6) nograph
. ksm y2 t, gen(sm2) lowess bw(0.6) nograph
. ksm y3 t, gen(sm3) lowess bw(0.6) nograph
. label var sm1 "barley diet"
. label var sm2 "mixed diet"
. label var sm3 "lupins diet"

. sort t
. graph sm1 sm2 sm3 t, c(lll)s(x) xlab ylab l1("Protein content")
```

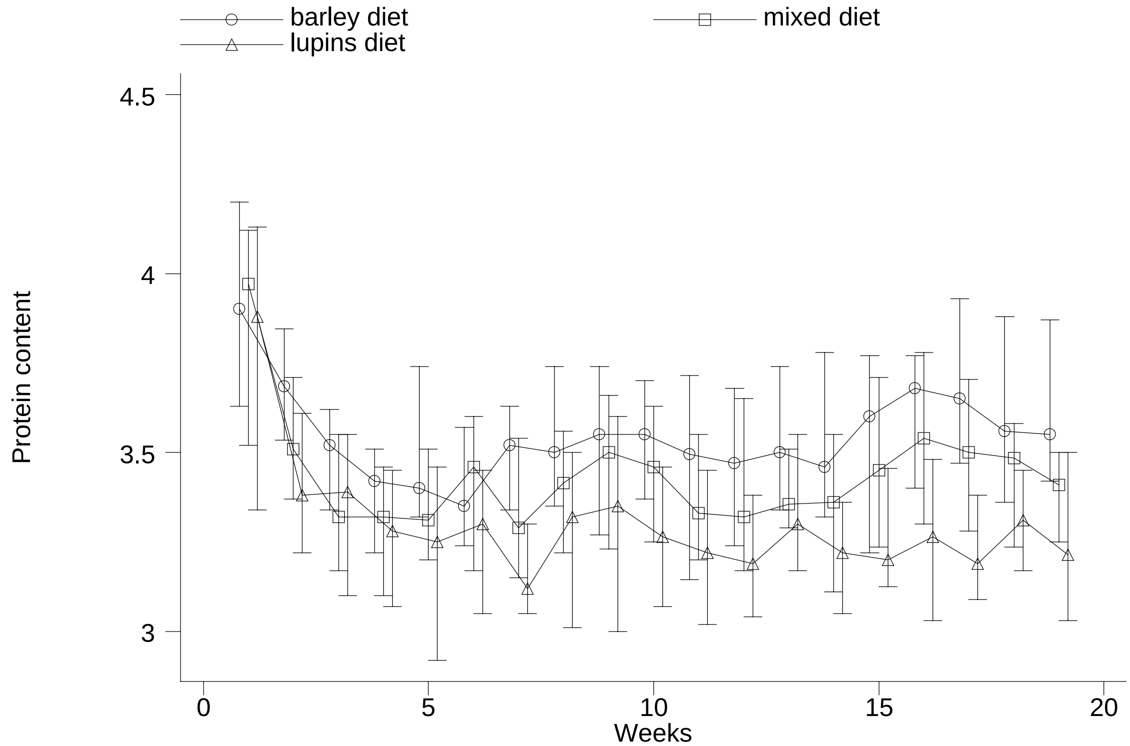



```
. drop sm1 sm2 sm3 y1 y2 y3

. save c:\data\milk1, replace
file c:\data\milk1.dta saved
```

- **Set to longitudinal data**

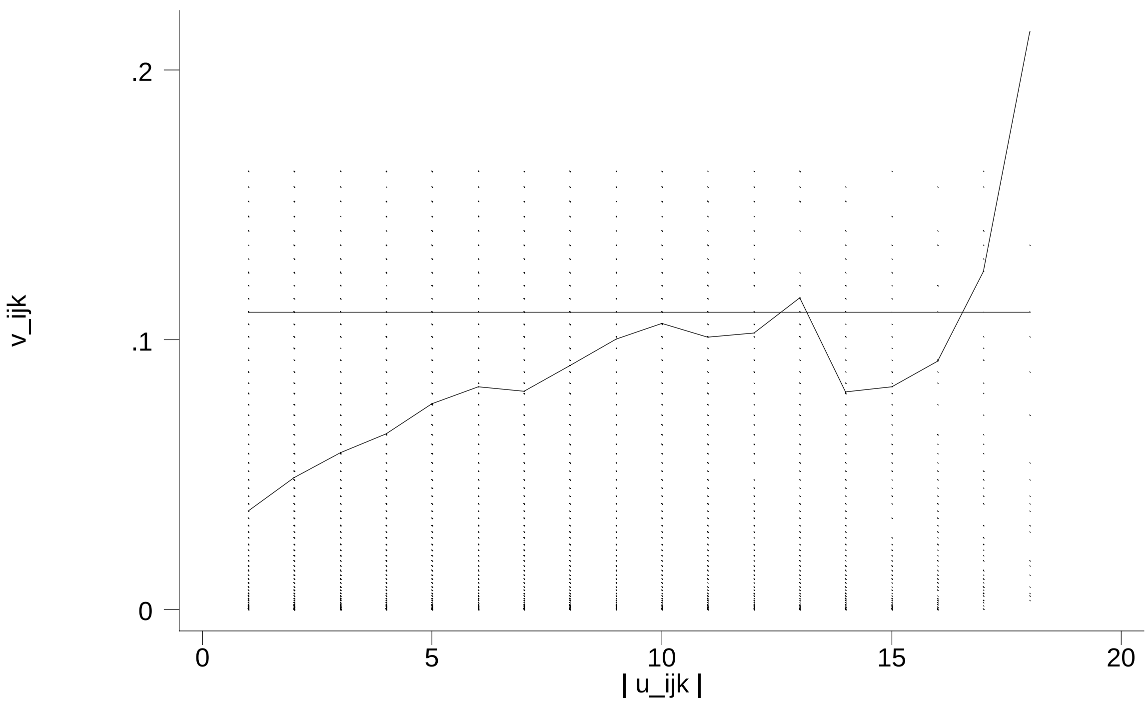
```
. tsset id t
       panel variable:  id, 1 to 79
        time variable:  t, 1 to 19

. *description of the longitudinal data
. xtdes

      id:  1, 2, ..., 79                                    n =         79
       t:  1, 2, ..., 19                                    T =         19
           Delta(t) = 1; (19-1)+1 = 19
           (id*t uniquely identifies each observation)
Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                        19      19      19        19        19      19      19

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------------------
       79    100.00  100.00 |  1111111111111111111
 ---------------------------+---------------------
       79    100.00         |  XXXXXXXXXXXXXXXXXXX
```

```
. xtgraph y, group(grp) av(median) bar(iqr) xlab ylab offset(.2)
```




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

Variogram of y (13 percent of v_ijk's excluded)




```
. * variogram indicates exponential correlation
```

- **Specify the mean model**

```
*create design matrix for mean model
. tab grp, gen(bg)

        grp |      Freq.     Percent        Cum.
------------+-----------------------------------
barley diet |        475       31.65       31.65
 mixed diet |        513       34.18       65.82
lupins diet |        513       34.18      100.00
------------+-----------------------------------
      Total |       1501      100.00

. gen b1=t

. replace b1=3 if t>3
(1264 real changes made)
. gen b2= (t-3)*(t>3)
. gen b3= (t-3)*(t-3)*(t>3)
. sort id t

. prais y bg1 bg2 bg3 b1 b2 b3, noconst

Number of gaps in sample:  88   (gap count includes panel changes)
(note: computations for rho restarted at each gap)

Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.6050
Iteration 2:  rho = 0.6096
Iteration 3:  rho = 0.6097
Iteration 4:  rho = 0.6097
Iteration 5:  rho = 0.6097

Prais-Winsten AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =    1337
-------------+------------------------------           F(  6,  1331) = 8614.52
       Model |   2930.0779     6  488.346317           Prob > F      =  0.0000
    Residual |  75.4527555  1331  .056688772           R-squared     =  0.9749
-------------+------------------------------           Adj R-squared =  0.9748
       Total |  3005.53066  1337  2.24796609           Root MSE      =  .23809

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bg1 |   4.140753   .0519486    79.71   0.000     4.038843    4.242664
         bg2 |   4.036527   .0515201    78.35   0.000     3.935457    4.137596
         bg3 |   3.930836   .0515166    76.30   0.000     3.829774    4.031899
          b1 |   -.221672   .0186047   -11.91   0.000    -.2581697   -.1851742
          b2 |   .0030986    .009073     0.34   0.733    -.0147003    .0208975
          b3 |   -.000164   .0005728    -0.29   0.775    -.0012876    .0009596
-------------+----------------------------------------------------------------
         rho |   .6096952
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    0.700758
Durbin-Watson statistic (transformed) 2.058911

. xtgee y bg1 bg2 bg3 b1 b2 b3, noconst i(id) t(t) corr(ar1)

note:  observations not equally spaced
       modal spacing is delta t = 1
       8 groups omitted from estimation

Iteration 1: tolerance = .00520863
```

```
Iteration 2: tolerance = .00020207
Iteration 3: tolerance = 8.292e-06
Iteration 4: tolerance = 3.451e-07

GEE population-averaged model                   Number of obs      =      1211
Group and time vars:                    id t    Number of groups   =        71
Link:                               identity    Obs per group: min =        14
Family:                             Gaussian                   avg =      17.1
Correlation:                           AR(1)                   max =        19
                                                Wald chi2(5)       =  31728.78
Scale parameter:                    .0897981    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bg1 |   4.132398   .0550036    75.13   0.000     4.024593    4.240203
         bg2 |   4.051468   .0541372    74.84   0.000     3.945361    4.157575
         bg3 |   3.912359   .0543971    71.92   0.000     3.805743    4.018975
          b1 |  -.2222159   .0197475   -11.25   0.000    -.2609202   -.1835115
          b2 |  -.0030645   .0095473    -0.32   0.748    -.0217768    .0156478
          b3 |   .0002933   .0006033     0.49   0.627    -.0008892    .0014757
------------------------------------------------------------------------------

. *create a text file for SAS
. outfile  id t y grp bg1 bg2 bg3 b1 b2 b3 using c:\data\milk.txt,nolabel replace
. log close
```

5. Refit the model using SAS

```
LIBNAME lab 'c:\data';

DATA milk;
INFILE 'c:\data\milk.txt';
INPUT id t y grp bg1 bg2 bg3 b1 b2 b3;
RUN;

*ML ESTIMATE WITH AR1 CORRELATION STRUCTURE*;
PROC MIXED DATA=milk METHOD = ML;
CLASS id;
MODEL y = bg1 bg2 bg3 b1 b2 b3/ NOINT SOLUTION CL;
REPEATED /SUBJECT=ID TYPE = AR(1);
RUN;
```

```
                Covariance Parameter Estimates

                    Cov Parm     Subject    Estimate

                    AR(1)        id           0.6415
                    Residual                 0.09503

                          Fit Statistics

                 -2 Log Likelihood             -16.0
                 AIC (smaller is better)        -0.0
                 AICC (smaller is better)        0.1
                 BIC (smaller is better)        18.9
```

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
bg1	4.1414	0.05317	76	77.89	<.0001	0.05	4.0355	4.2473
bg2	4.0406	0.05259	76	76.83	<.0001	0.05	3.9358	4.1453
bg3	3.9287	0.05261	76	74.68	<.0001	0.05	3.8240	4.0335
b1	-0.2216	0.01853	1255	-11.96	<.0001	0.05	-0.2579	-0.1852
b2	0.001592	0.009366	1255	0.17	0.8651	0.05	-0.01678	0.01997
b3	-0.00010	0.000591	1255	-0.17	0.8647	0.05	-0.00126	0.001059

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
bg1	1	76	6066.31	<.0001
bg2	1	76	5903.19	<.0001
bg3	1	76	5577.56	<.0001
b1	1	1255	143.02	<.0001
b2	1	1255	0.03	0.8651
b3	1	1255	0.03	0.8647

```
*REML ESTIMATE WITH AR1 CORRELATION STRUCTURE*;
PROC MIXED DATA=milk METHOD = REML;
CLASS id;
MODEL y = bg1 bg2 bg3 b1 b2 b3/ NOINT SOLUTION CL;
REPEATED /SUBJECT=ID TYPE = AR(1);
RUN;
```

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
AR(1)	id	0.6465
Residual		0.09645

Fit Statistics

-2 Res Log Likelihood	28.9
AIC (smaller is better)	32.9
AICC (smaller is better)	32.9
BIC (smaller is better)	37.6

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
bg1	4.1413	0.05353	76	77.36	<.0001	0.05	4.0347	4.2479
bg2	4.0405	0.05294	76	76.33	<.0001	0.05	3.9351	4.1460
bg3	3.9287	0.05295	76	74.19	<.0001	0.05	3.8233	4.0342
b1	-0.2215	0.01858	1255	-11.92	<.0001	0.05	-0.2579	-0.1850
b2	0.001579	0.009438	1255	0.17	0.8672	0.05	-0.01694	0.02010
b3	-0.00011	0.000595	1255	-0.18	0.8579	0.05	-0.00127	0.001061

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
bg1	1	76	5984.63	<.0001
bg2	1	76	5825.74	<.0001
bg3	1	76	5504.42	<.0001
b1	1	1255	142.13	<.0001
b2	1	1255	0.03	0.8672
b3	1	1255	0.03	0.8579

```
*ML ESTIMATE WITH EXPONENTIAL CORRELATION STRUCTURE*;
PROC MIXED DATA=milk METHOD = ML;
CLASS id;
MODEL y = bg1 bg2 bg3 b1 b2 b3/ NOINT SOLUTION CL;
REPEATED /SUBJECT=ID TYPE = SP(POW) (t);
RUN;
```

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
SP(POW)	id	0.6415
Residual		0.09503

Fit Statistics

-2 Log Likelihood	-16.0
AIC (smaller is better)	-0.0
AICC (smaller is better)	0.1
BIC (smaller is better)	18.9

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
bg1	4.1414	0.05317	76	77.89	<.0001	0.05	4.0355	4.2473
bg2	4.0406	0.05259	76	76.83	<.0001	0.05	3.9358	4.1453
bg3	3.9287	0.05261	76	74.68	<.0001	0.05	3.8240	4.0335
b1	-0.2216	0.01853	1255	-11.96	<.0001	0.05	-0.2579	-0.1852
b2	0.001592	0.009366	1255	0.17	0.8651	0.05	-0.01678	0.01997
b3	-0.00010	0.000591	1255	-0.17	0.8647	0.05	-0.00126	0.001059

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
bg1	1	76	6066.31	<.0001
bg2	1	76	5903.19	<.0001
bg3	1	76	5577.56	<.0001
b1	1	1255	143.02	<.0001
b2	1	1255	0.03	0.8651
b3	1	1255	0.03	0.8647