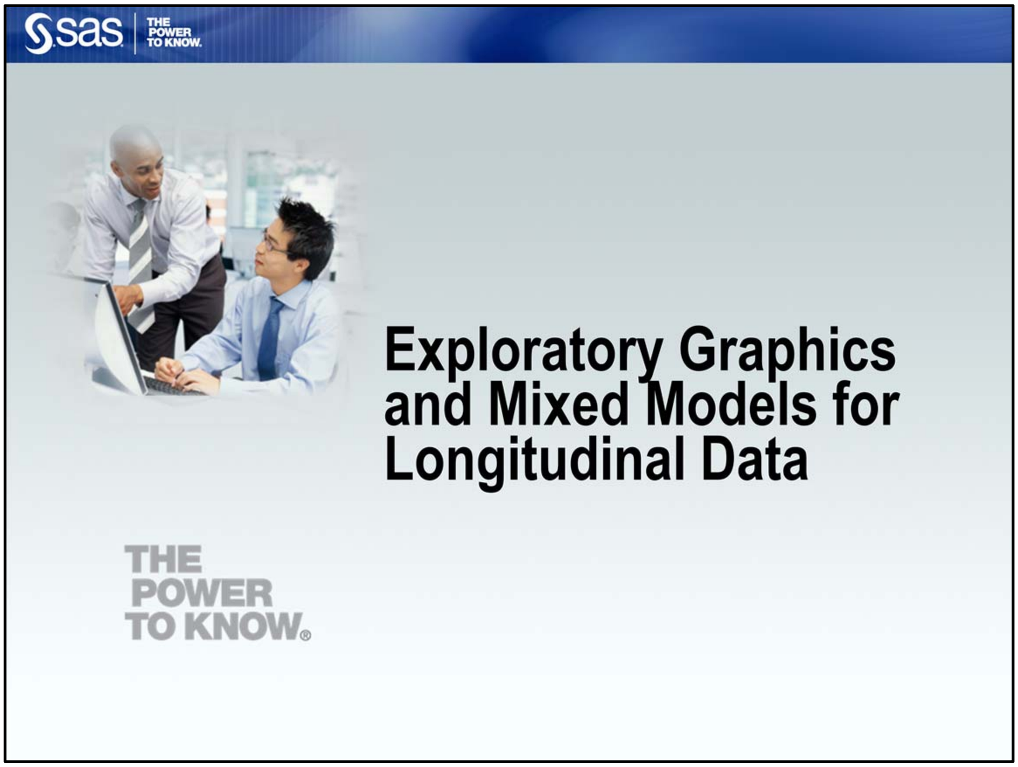
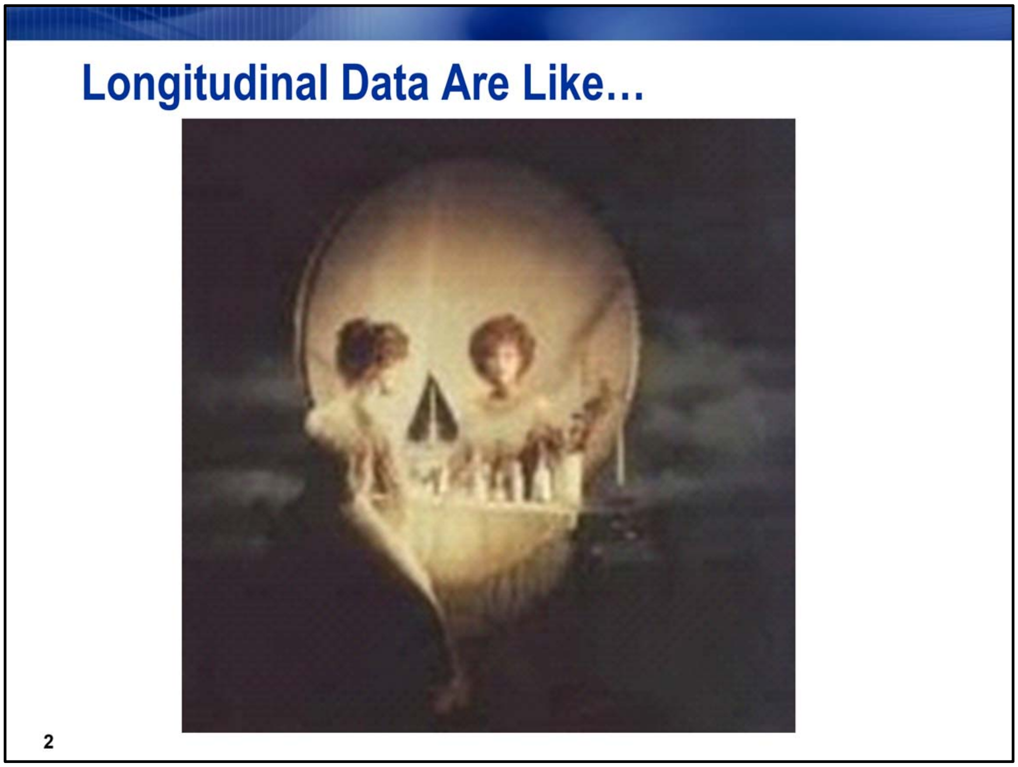
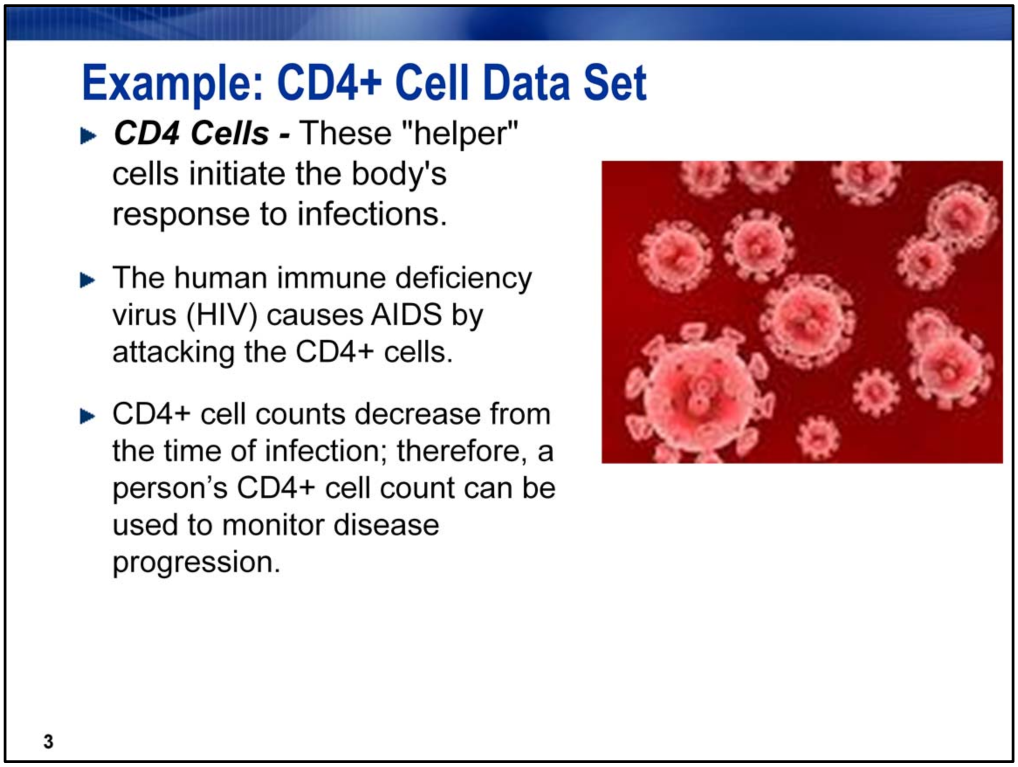


Exploratory Graphics and Mixed Models for Longitudinal Data



Longitudinal Data Are Like…

2

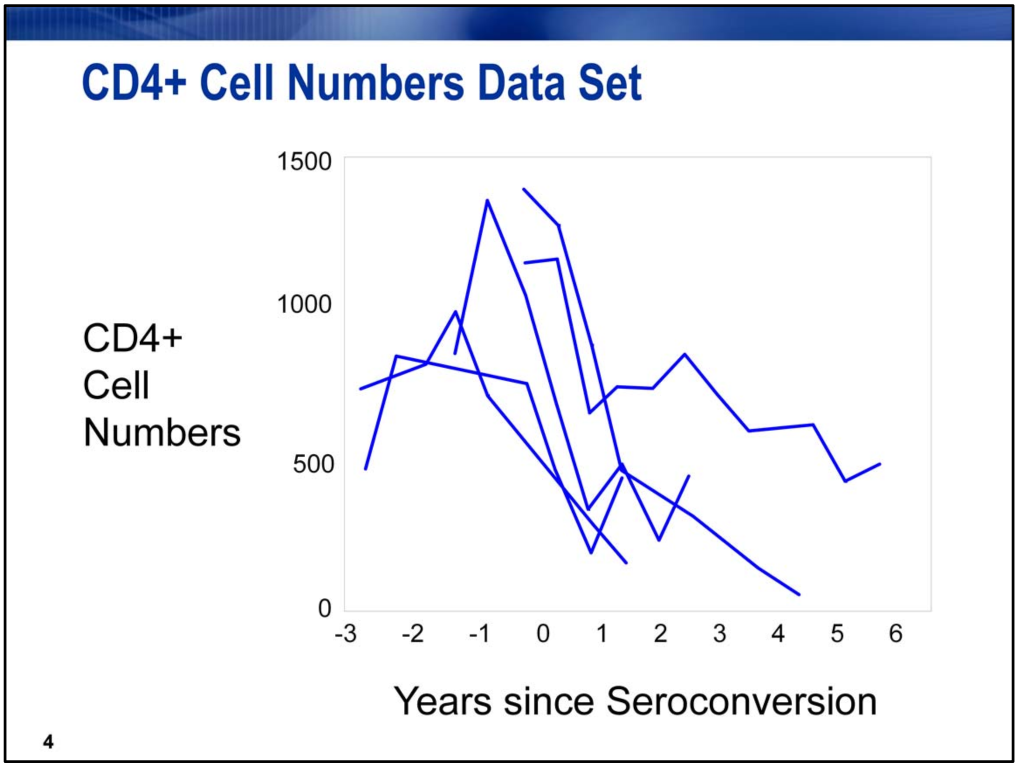


Example: CD4+ Cell Data Set

- ***CD4 Cells -*** These "helper" cells initiate the body's response to infections.
- The human immune deficiency virus (HIV) causes AIDS by attacking the CD4+ cells.
- CD4+ cell counts decrease from the time of infection; therefore, a person's CD4+ cell count can be used to monitor disease progression.

CD4 cells are a type of white blood cell, that play an important role in the immune system, particularly in the adaptive immune system. These cells cannot kill infected host cells or pathogens. Rather, they help other immune cells—they activate and direct other immune cells

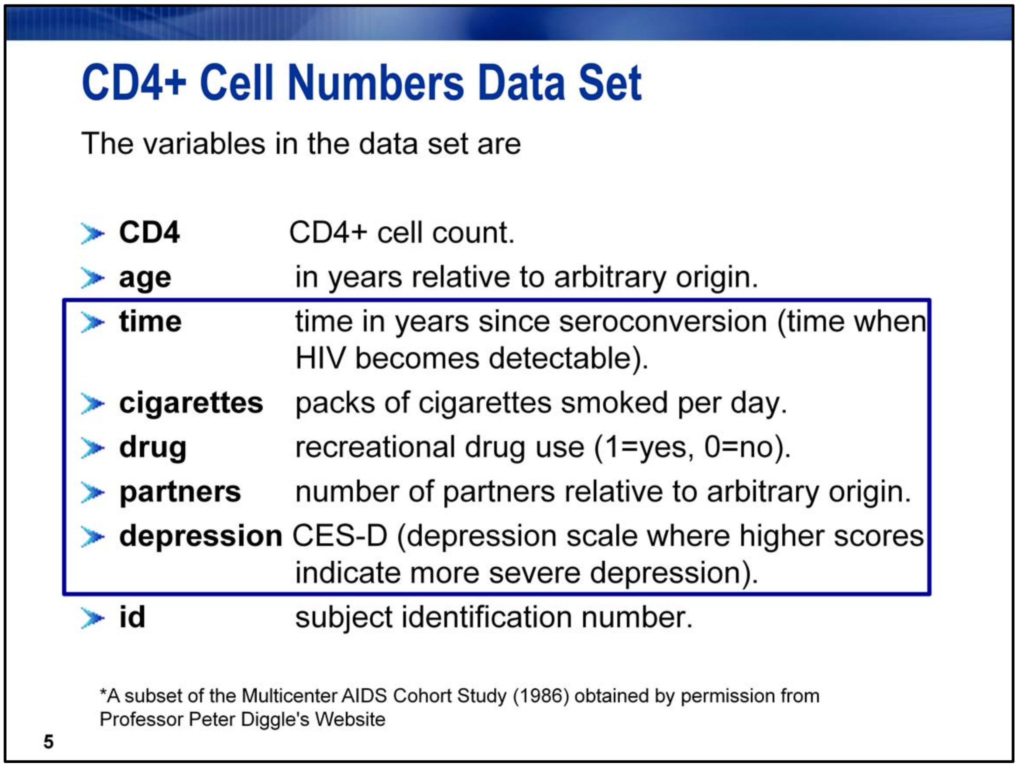
HIV usually progresses to AIDS, defined as possessing a CD4+ lymphocyte count under 200 cells/µl or HIV infection plus co-infection with an AIDS defining opportunistic infection.



CD4+ Cell Numbers Data Set

*This data set is a subset of the Multicenter AIDS Cohort Study (1986) obtained by permission from Professor Peter Diggle's Website.

The progression of CD4 cells over time will be used to study the progression of the HIV virus in infected individuals. Shown are the profile plots for 5 patients; note the downward trend in CD4 count over time. We'll see how to obtain this plot for all patients.



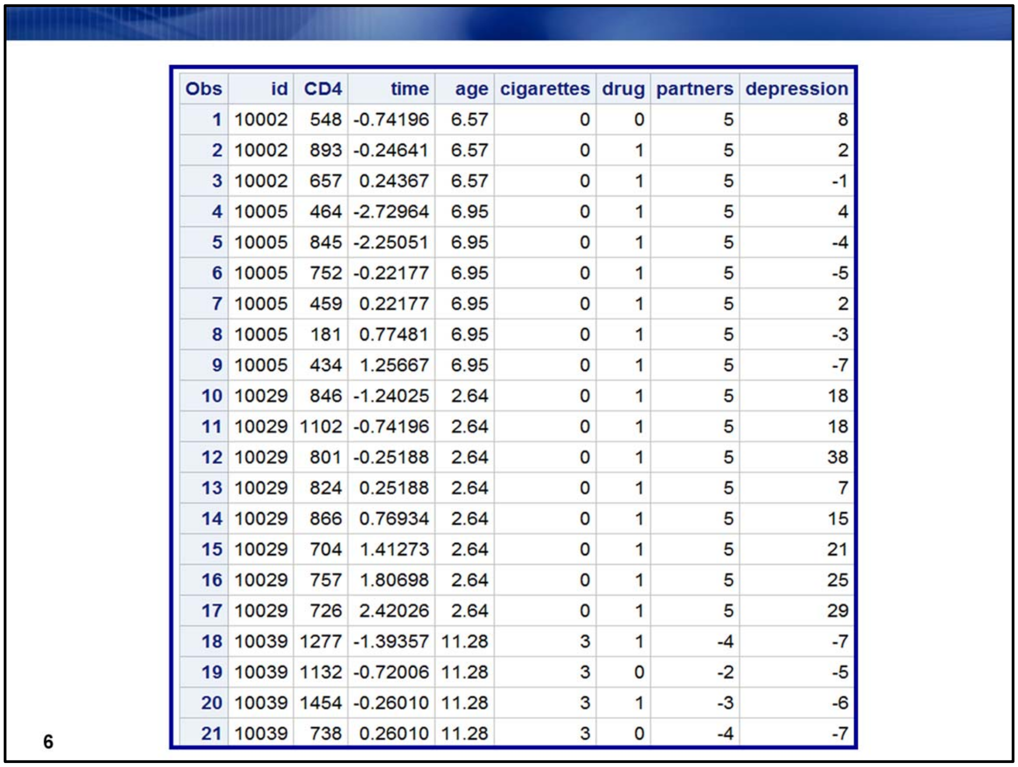
CD4+ Cell Numbers Data Set

The variables in the data set are

- **CD4** CD4+ cell count.
- **age** in years relative to arbitrary origin.
- **time** time in years since seroconversion (time when HIV becomes detectable).
- **cigarettes** packs of cigarettes smoked per day.
- **drug** recreational drug use (1=yes, 0=no).
- **partners** number of partners relative to arbitrary origin.
- **depression** CES-D (depression scale where higher scores indicate more severe depression).
- **id** subject identification number.

*A subset of the Multicenter AIDS Cohort Study (1986) obtained by permission from Professor Peter Diggle's Website

Point out the time-dependent variables – the box.



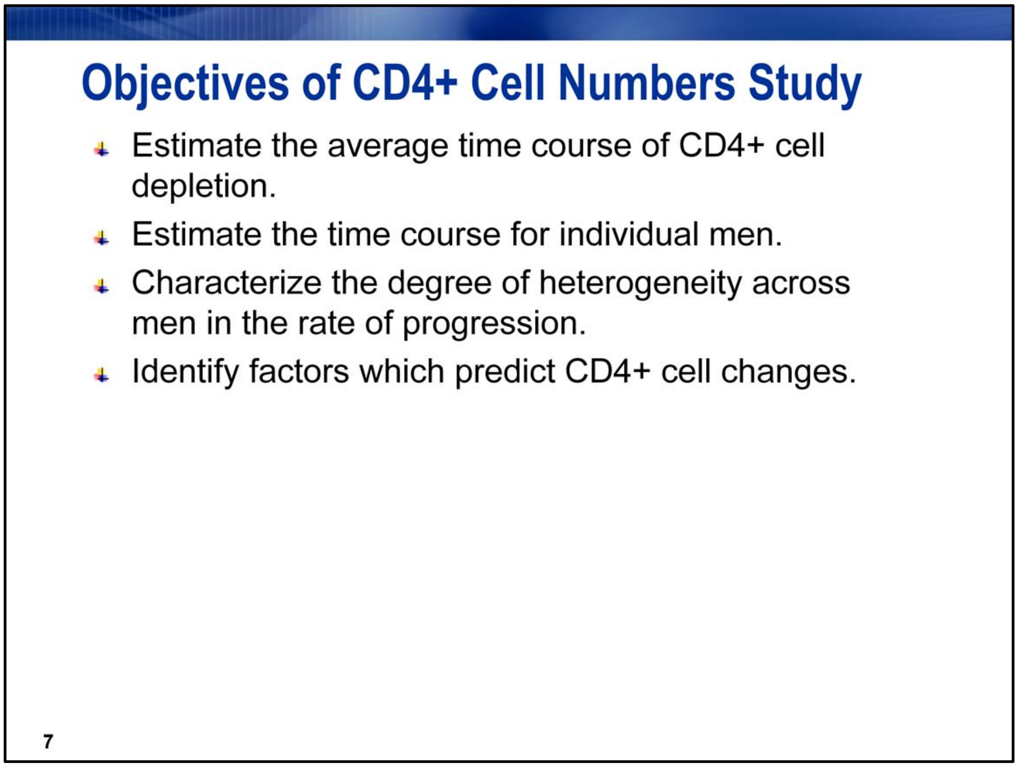
Obs	id	CD4	time	age	cigarettes	drug	partners	depression
1	10002	548	-0.74196	6.57	0	0	5	8
2	10002	893	-0.24641	6.57	0	1	5	2
3	10002	657	0.24367	6.57	0	1	5	-1
4	10005	464	-2.72964	6.95	0	1	5	4
5	10005	845	-2.25051	6.95	0	1	5	-4
6	10005	752	-0.22177	6.95	0	1	5	-5
7	10005	459	0.22177	6.95	0	1	5	2
8	10005	181	0.77481	6.95	0	1	5	-3
9	10005	434	1.25667	6.95	0	1	5	-7
10	10029	846	-1.24025	2.64	0	1	5	18
11	10029	1102	-0.74196	2.64	0	1	5	18
12	10029	801	-0.25188	2.64	0	1	5	38
13	10029	824	0.25188	2.64	0	1	5	7
14	10029	866	0.76934	2.64	0	1	5	15
15	10029	704	1.41273	2.64	0	1	5	21
16	10029	757	1.80698	2.64	0	1	5	25
17	10029	726	2.42026	2.64	0	1	5	29
18	10039	1277	-1.39357	11.28	3	1	-4	-7
19	10039	1132	-0.72006	11.28	3	0	-2	-5
20	10039	1454	-0.26010	11.28	3	1	-3	-6
21	10039	738	0.26010	11.28	3	0	-4	-7

6

Note that TIME has been centered—this was done based on time of seroconversion. At the first instance that the antibody marker is present in the blood, seroconversion is deemed to have occurred between that observation and the previous.

AGE was gathered at time of entry and is time independent; it has been centered relative to an arbitrary origin.

CIGARETTES and DRUG are 0/1 indicators, and PARTNERS has been centered relative to an arbitrary origin. These are time-dependent variables.



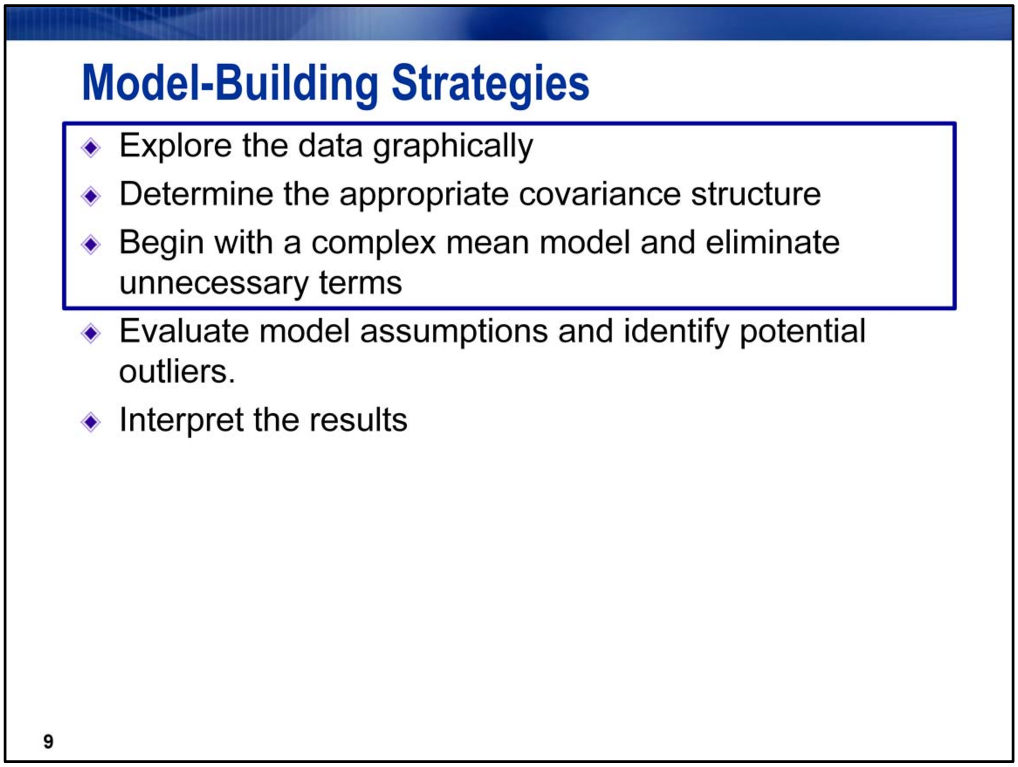
Objectives of CD4+ Cell Numbers Study

- Estimate the average time course of CD4+ cell depletion.
- Estimate the time course for individual men.
- Characterize the degree of heterogeneity across men in the rate of progression.
- Identify factors which predict CD4+ cell changes.

Two types of models can be fit to this data set in PROC MIXED: models with a REPEATED statement and those with a RANDOM statement.

Only the RANDOM coefficient models can give the second and third bullets (time course for individual men and characterize the degree of heterogeneity across men in the rate of progression). Both models can estimate the average time course of CD4 cell depletion and identify factors that predict CD4 changes.

How do we tackle this?

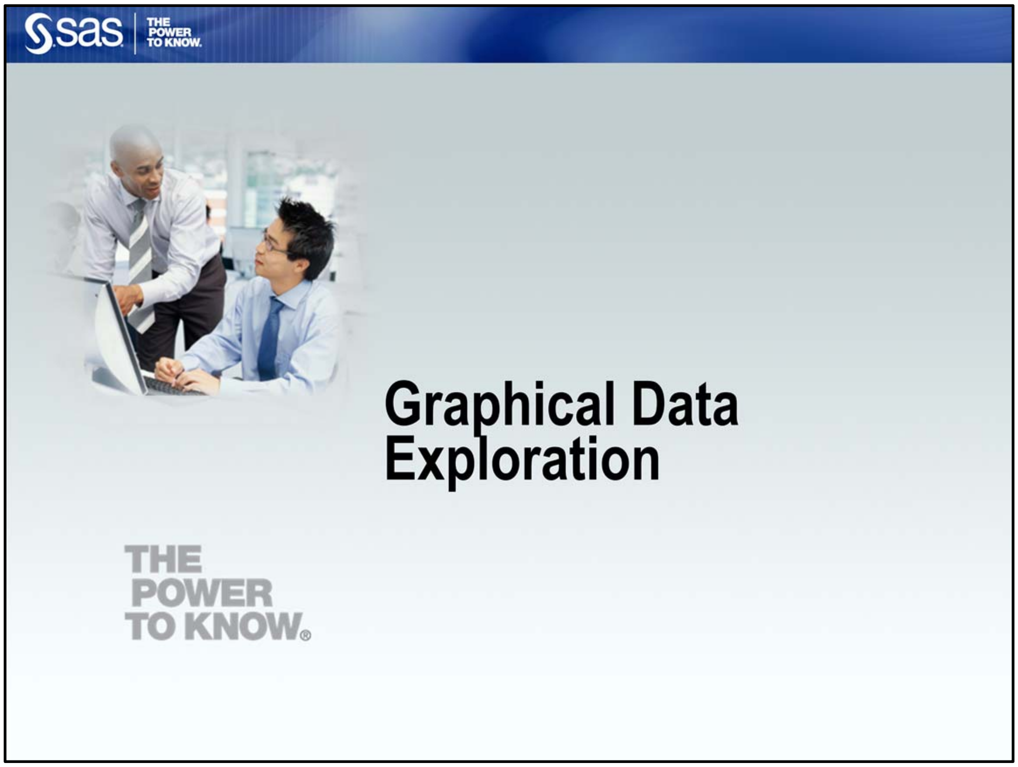


Model-Building Strategies

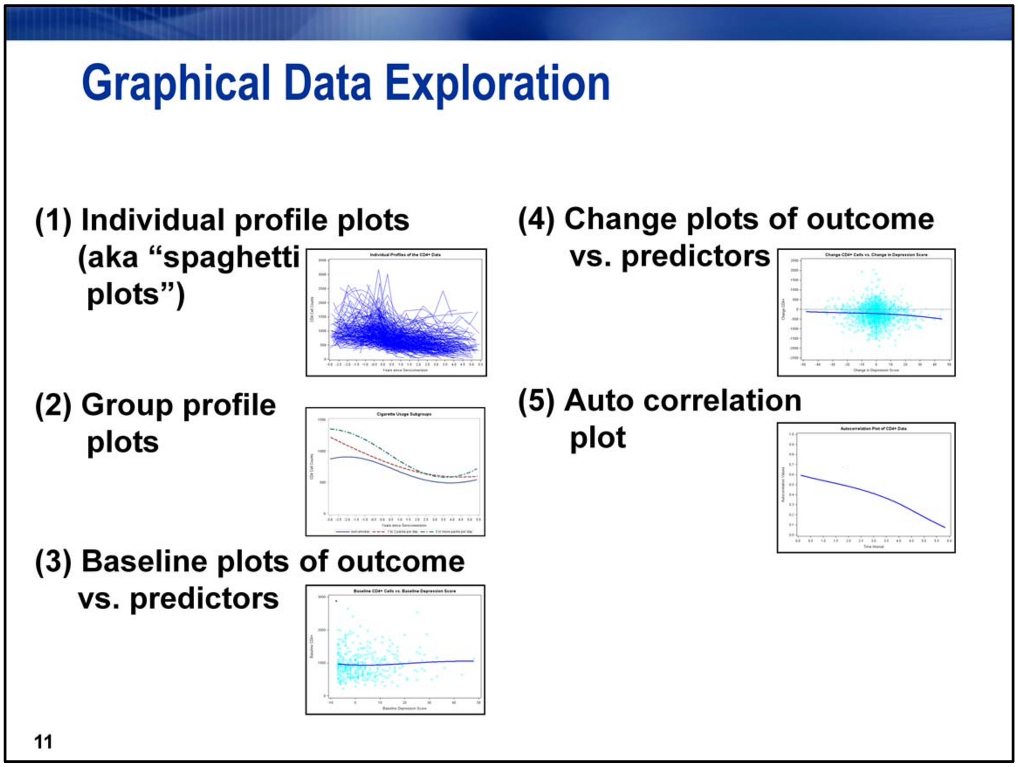
- Explore the data graphically
- Determine the appropriate covariance structure
- Begin with a complex mean model and eliminate unnecessary terms
- Evaluate model assumptions and identify potential outliers.
- Interpret the results

9

Our discussion will cover only the first three bullets



Graphical Data Exploration



Graphical Data Exploration

(1) Individual profile plots (aka “spaghetti plots”)

(2) Group profile plots

(3) Baseline plots of outcome vs. predictors

(4) Change plots of outcome vs. predictors

(5) Auto correlation plot

11

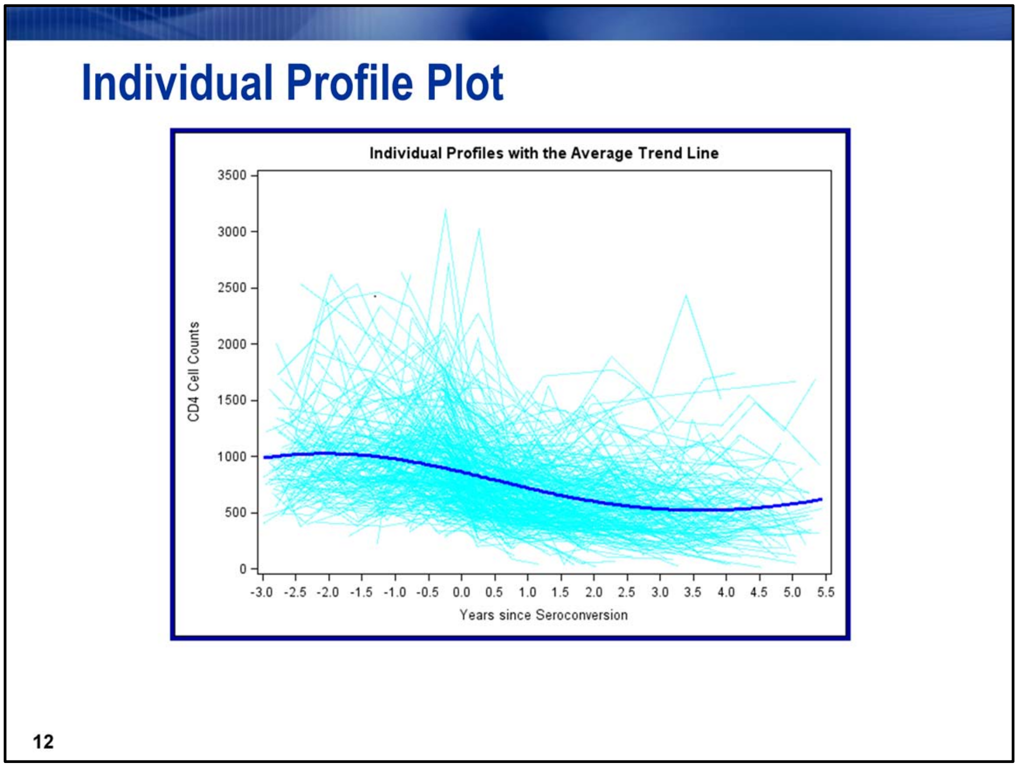
The individual profiles can have so many points as to provide no useful information. We will superimpose a trend line using a spline to discern any pattern in CD4 counts over time.

Group profile plots: are looking for differences in CD4 counts across the groups and differences in the changes over time for the groups.

Baseline plots: to see what is going on at the time the patients entered the study

Change plots: to examine the relationship between changes in values of the predictor variables to changes in the count of CD4 cells.

Autocorrelation plots**:** to gain insight into the correlations over time to inform the choice of correlation structure



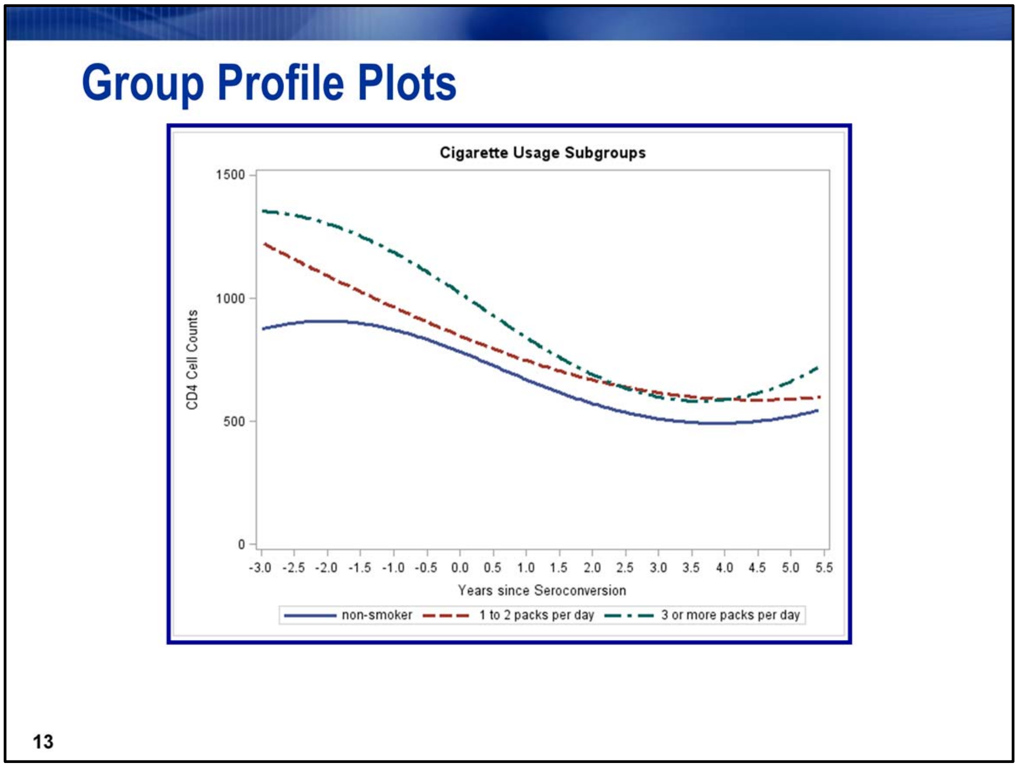
Individual Profile Plot

12

General coding idea: **Individual profile plots. General idea: PROC SGPLOT with SERIES and PBSPLINE statement;

RESULTS: Cubic relationship of depletion of CD4 cell counts over time.

```
proc sgplot data=sasuser.aids nocycleattrs noautolegend;
  series y=cd4 x=time / group=id transparency=0.5
      lineattrs=(color=cyan pattern=1);
  pbspline y=cd4 x=time / nomarkers smooth=50 nknots=5
      lineattrs=(color=blue pattern=1 thickness=3);
  xaxis values=(-3 to 5.5 by 0.5) label='Years since Seroconversion';
  yaxis values=(0 to 3500 by 500) label='CD4 Cell Counts';
  title 'Individual Profiles with the Average Trend Line';
run;
quit;
```



Group Profile Plots

13

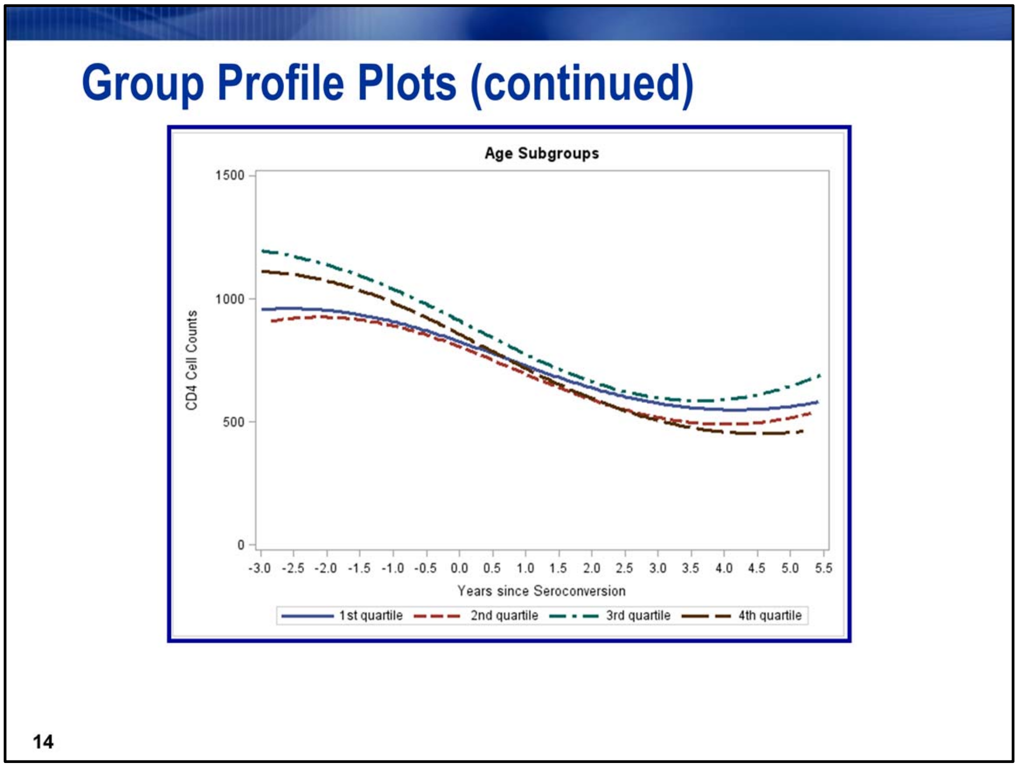
****General idea for coding**: create groups using programming statements; then create plot with PROC SGPLOT with a PBSPLINE statement with GROUP=option.

RESULTS: CD4 cell counts appear to differ by group; the rate of change may differ across groups indicating there may be a cigarette*time interaction

```
data sasuser.aids;
   set sasuser.aids;   ciggroup=1*(cigarettes=0)+2*(0<cigarettes<=2)+
                  3*(2<cigarettes<=4); run;

…proc format;

proc sgplot data=sasuser.aids;
   pbspline y=cd4 x=time / group=ciggroup nomarkers smooth=50
                  nknots=5 lineattrs=(thickness=3) name="cigarette";
   xaxis values=(-3 to 5.5 by 0.5) label='Years since
                                          Seroconversion';
   yaxis values=(0 to 1500 by 500) label='CD4 Cell Counts';
   keylegend "cigarette";    format ciggroup cgroup.;
   title 'Individual Profiles with Cigarette Usage Subgroups';
run;
```

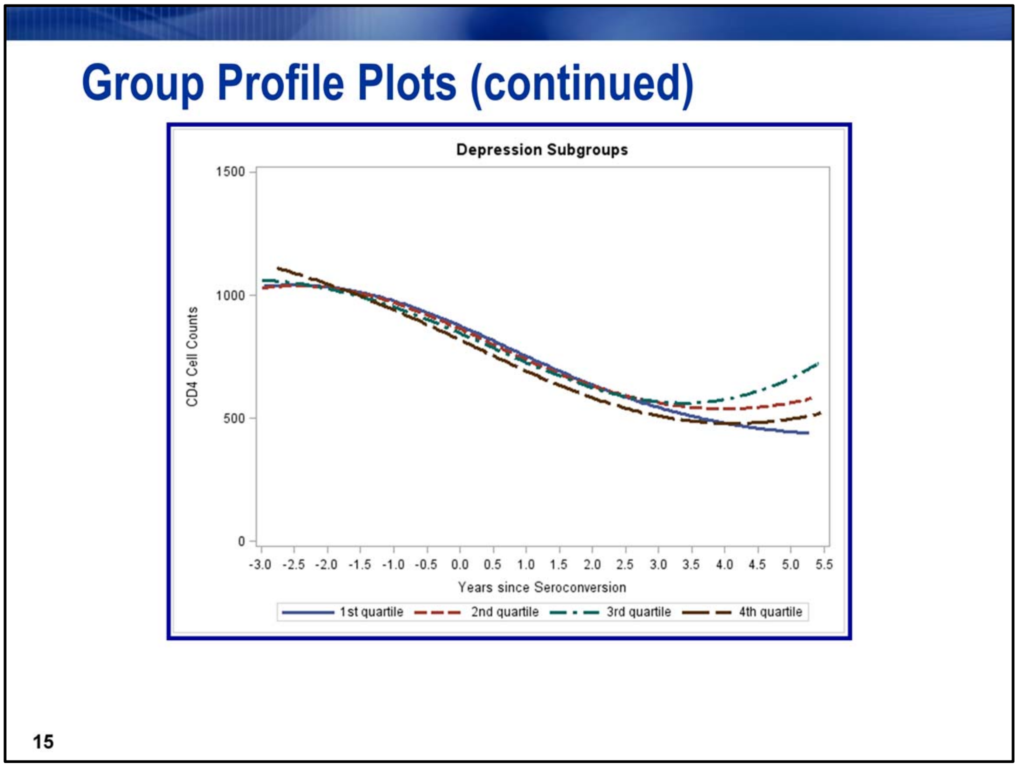


Group Profile Plots (continued)

14

****General idea for coding:** create quartiles for **age** using PROC RANK, then create the plot with PROC SGPLOT with the PBSPLINE statement with GROUP=AGEGROUP option.

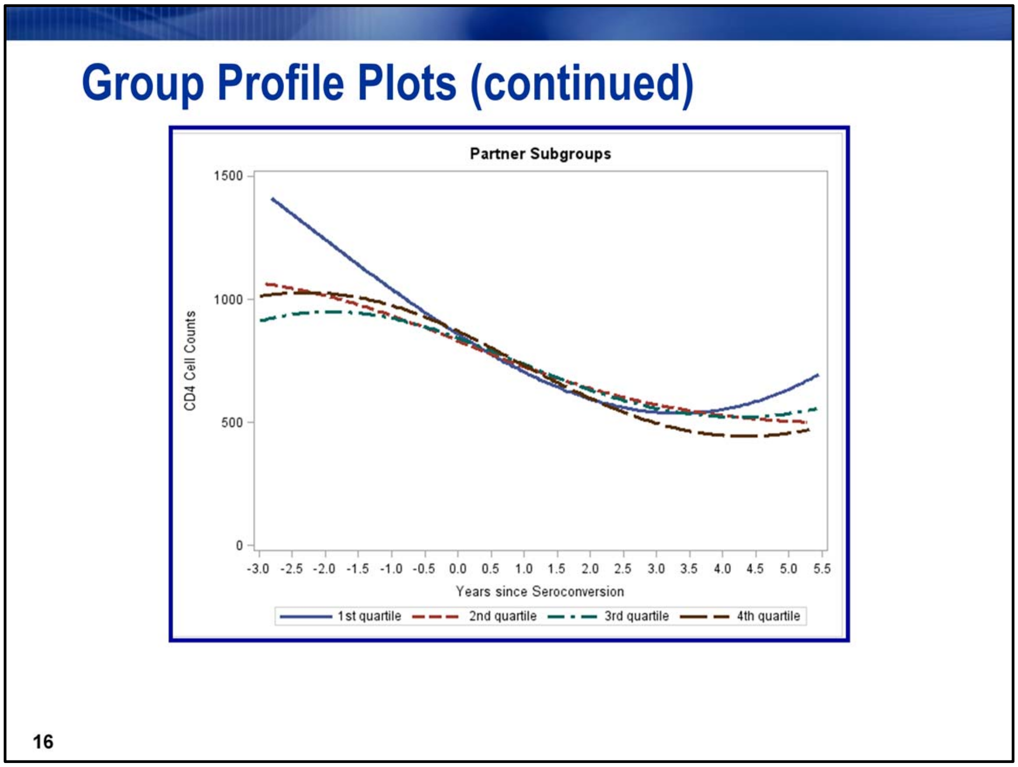
CD4 cell counts may differ by **age** and there may be a **time*age** interaction



Group Profile Plots (continued)

15

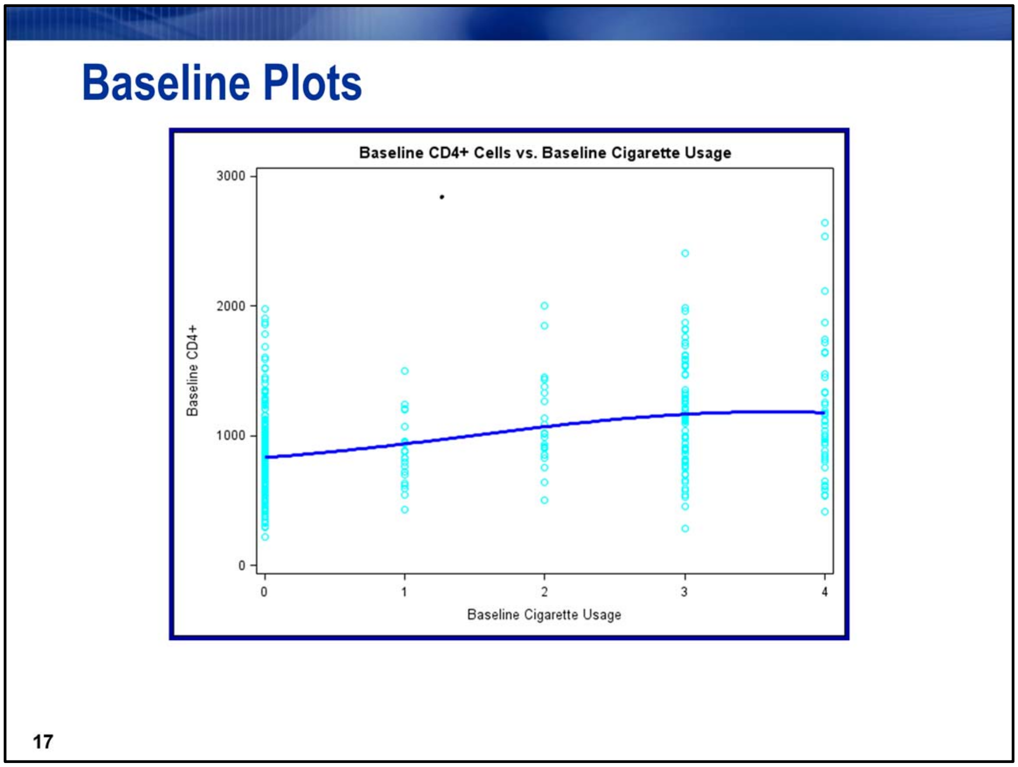
Results: CD4 counts for the groups do not appear to differ, except beyond year 3.5



Group Profile Plots (continued)

16

Results: The first quartile may appear to differ from the others with respect to CD4 cell counts, especially at the first time point. We'll see later that the uptick for the first time point comes from only 1 observation.



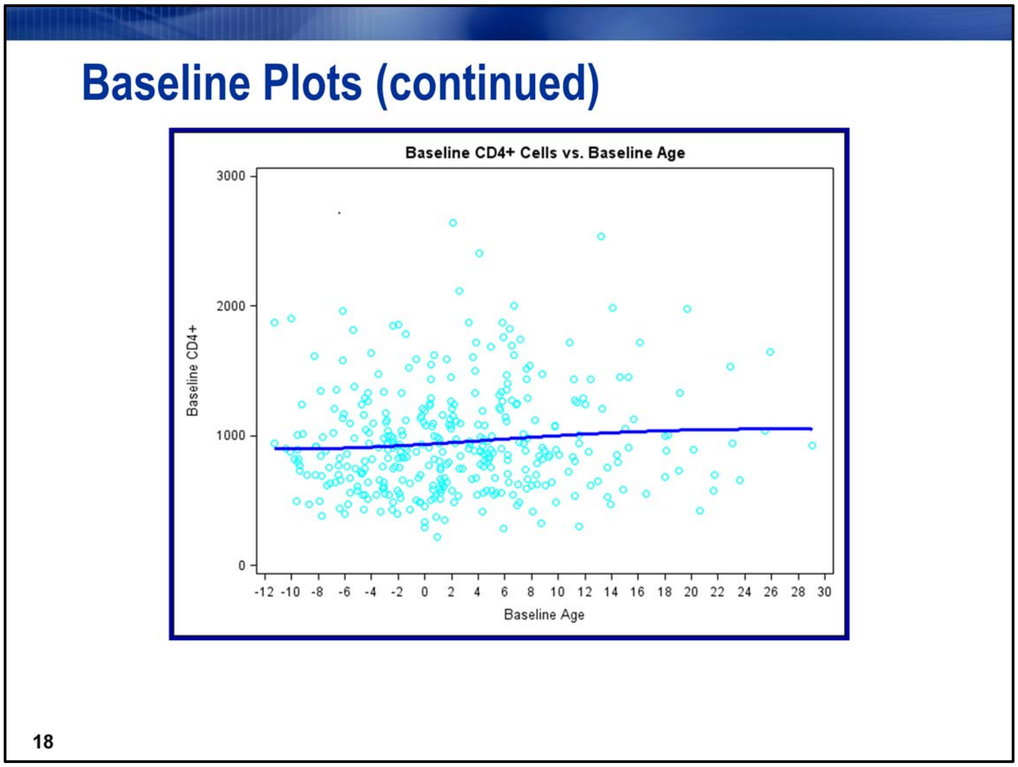
Baseline Plots

17

General coding idea: First, create the baseline scores; the baseline score will be observed values for the outcome variable (CD4) and the predictor variables at the first observation for each patient

Create plot using SGPLOT with a PBSPLINE statement.

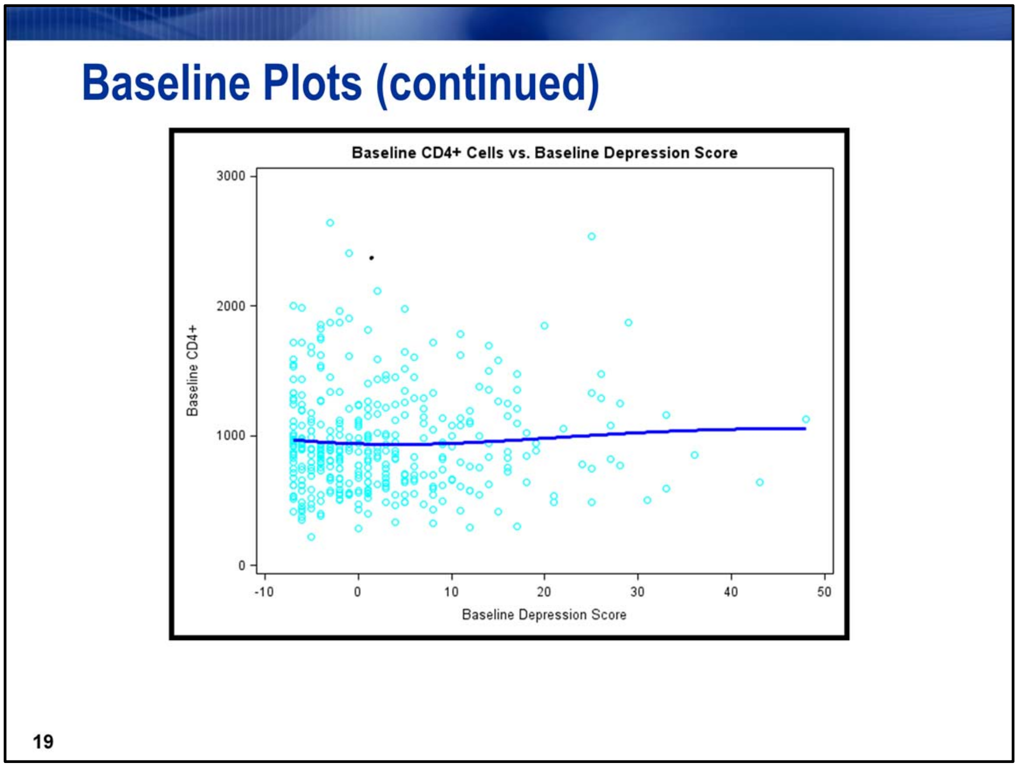
Results: At baseline, CD4 cell counts increase as cigarette use increases, up to 3 packs of cigarettes per day.



Baseline Plots (continued)

18

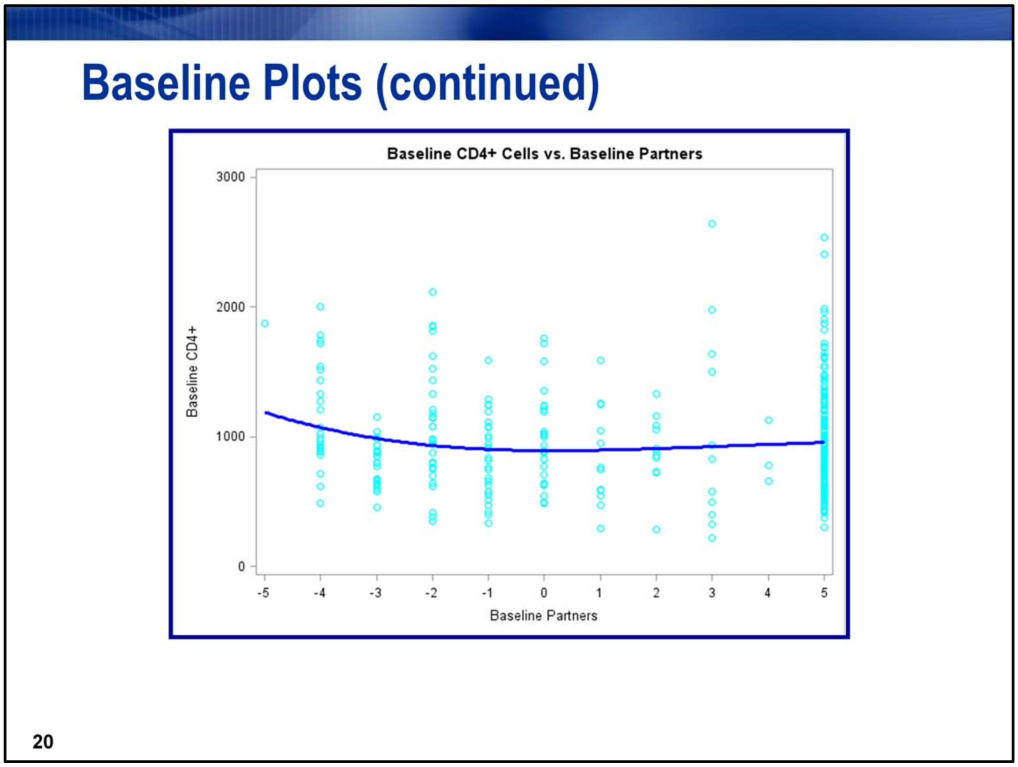
At baseline **age** appears to have no effect.



Baseline Plots (continued)

19

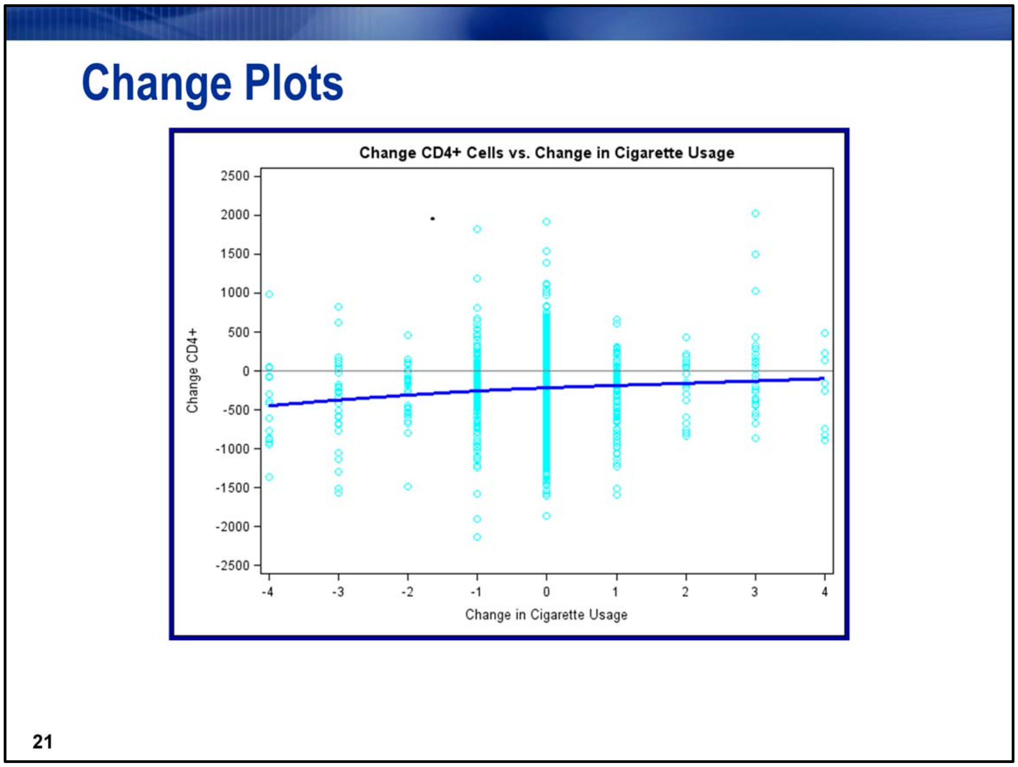
No effect of depression score at baseline



Baseline Plots (continued)

20

The uptick at the first time point is due to one observation. If you remove that observation, the effect of **partners** at baseline appears negligible.

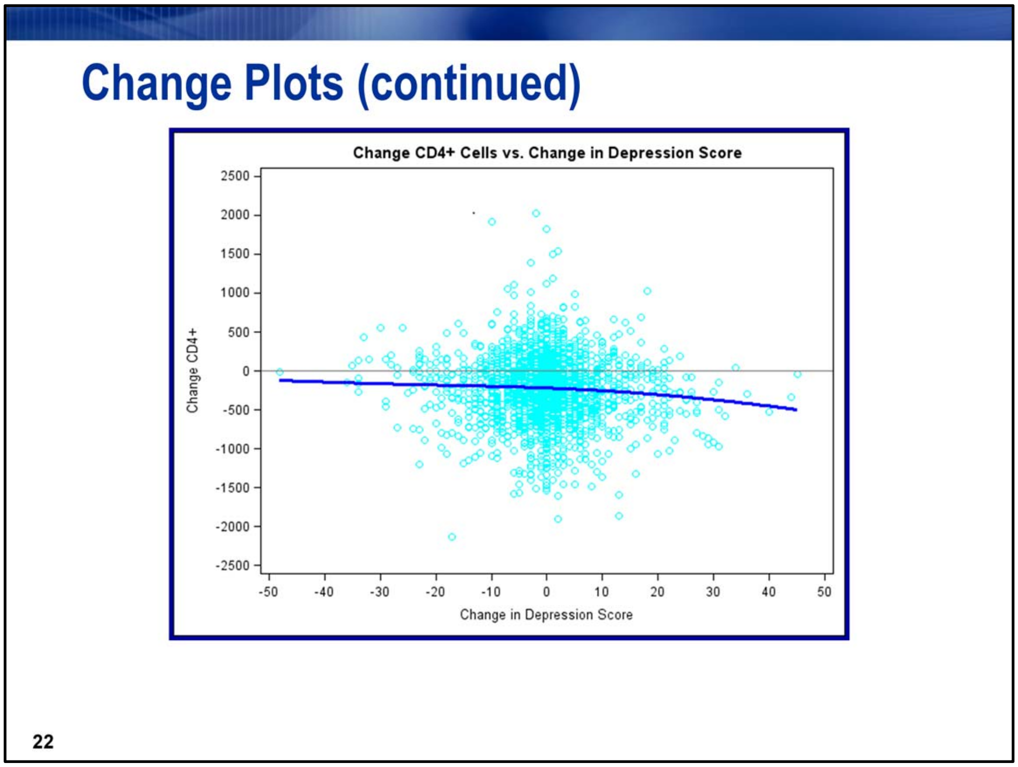


Change Plots

21

General coding idea: Create the variables; the change scores will be calculated within patients. For CD4, this will be the difference between the baseline CD4 value and the CD4 value at each subsequent visit. Change scores for each of the predictor variables are created in the same manner. If a man has 6 visits, he will have 5 change scores.

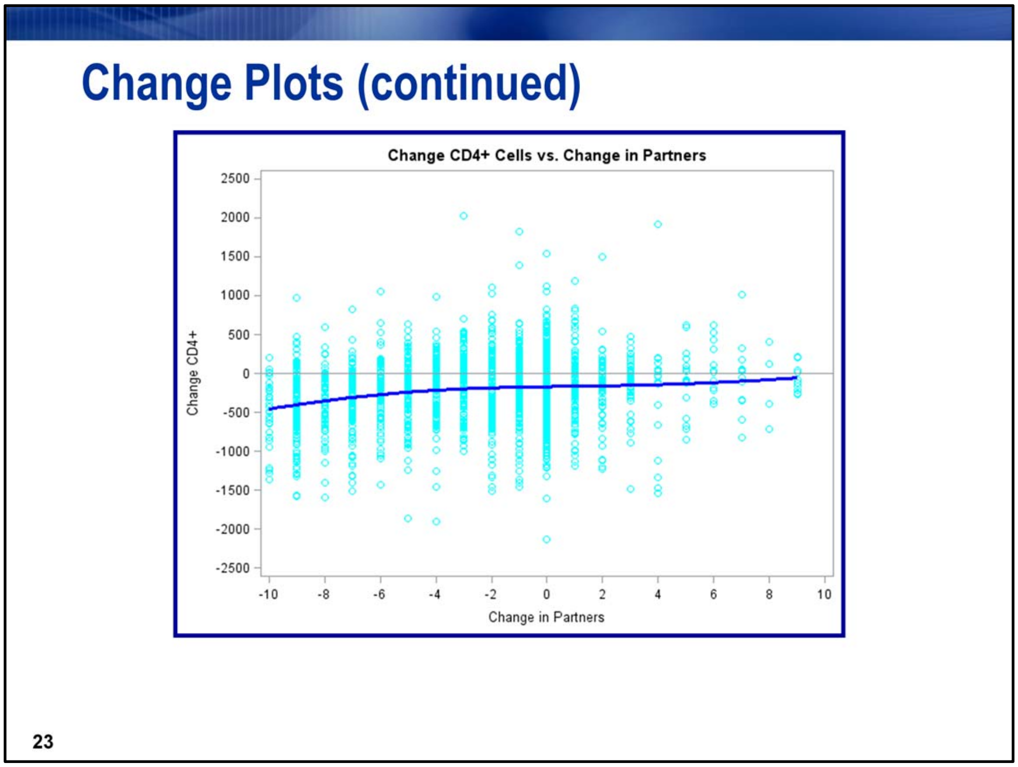
Results: There is an uptrend at baseline. As cigarette use increases, CD4 cell counts increase.



Change Plots (continued)

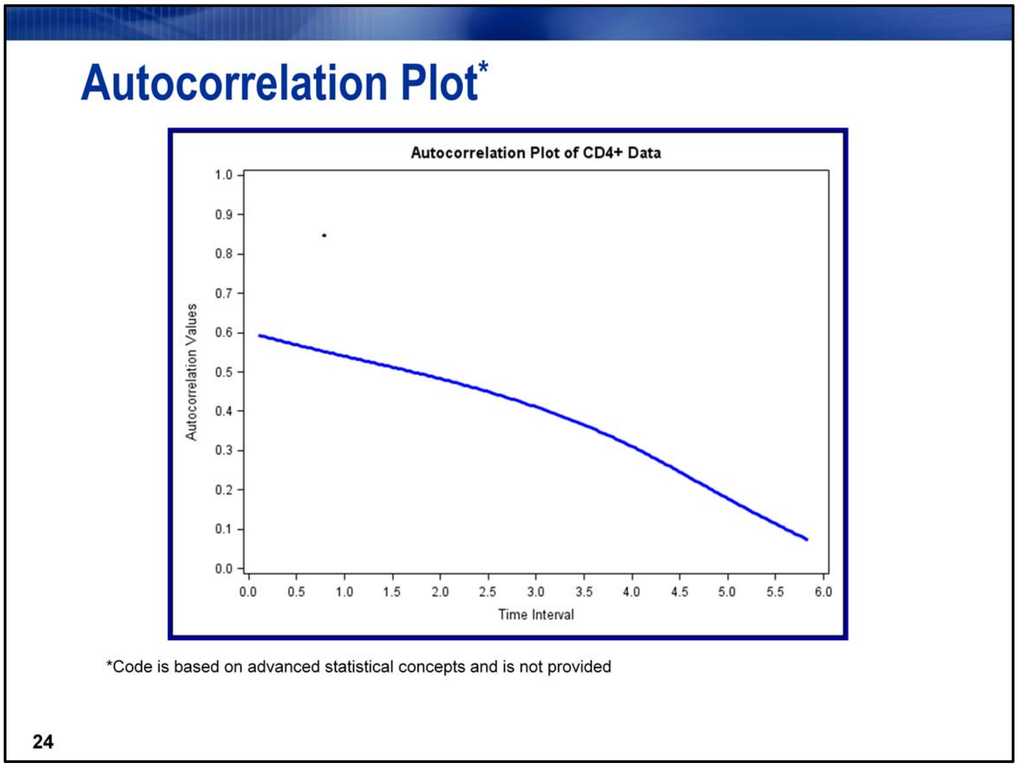
22

Results: over time, an increase in depression scores (meaning more depressed) is accompanied by a decrease in CD4 cell count.



Change Plots (continued)

RESULTS: over time, an increase in partners is associated with an increase in CD4 cell count



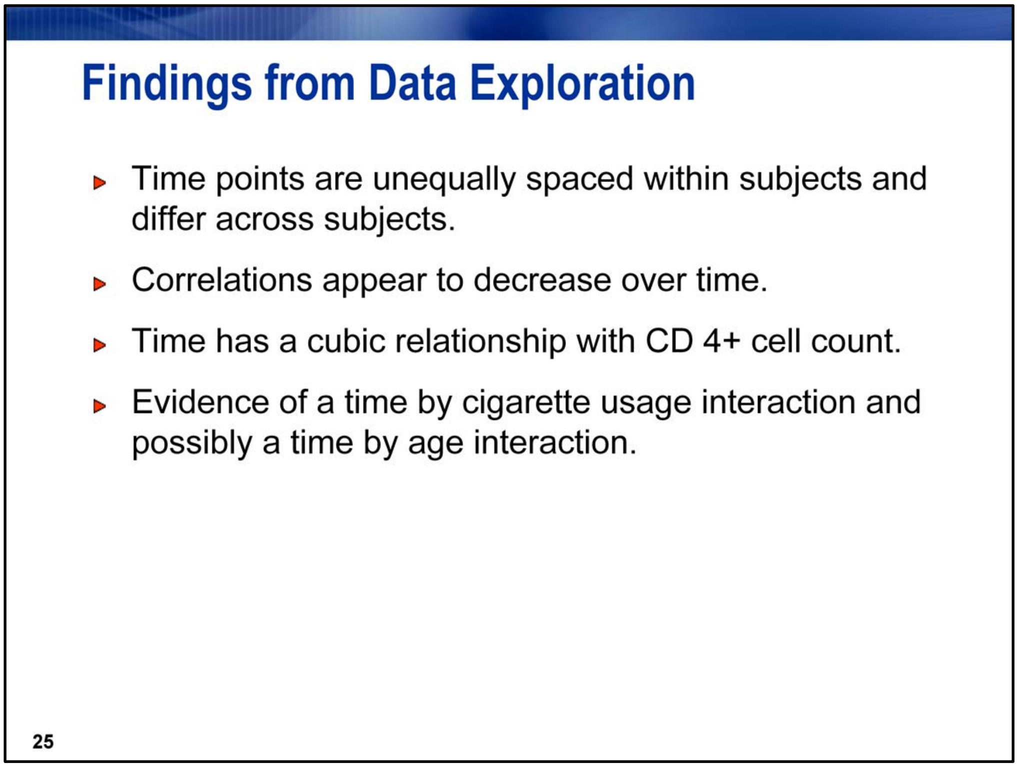
Autocorrelation Plot*

*Code is based on advanced statistical concepts and is not provided

24

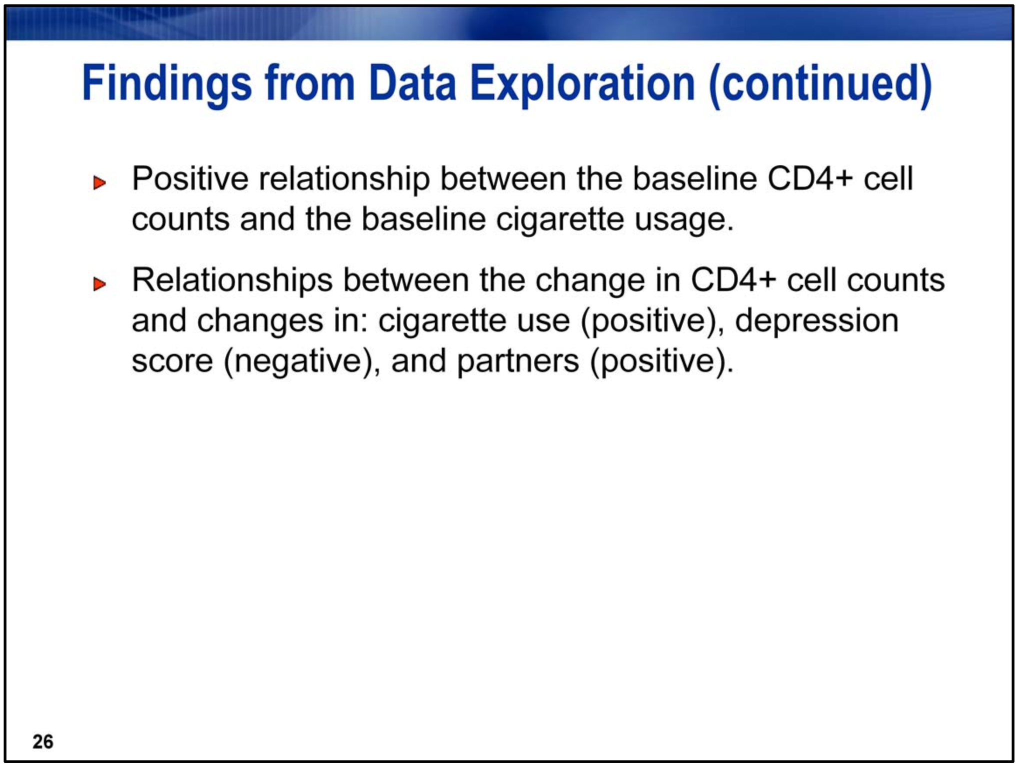
Results: correlations decrease over time, so a covariance structure should be selected that allows for this. Additionally, due to the unbalanced nature of the time points, only a few structures are appropriate; any of the SPATIAL structures would work well.

Autocorrelation plots from PROC CORR require balanced time points across all subjects. This plot is based on a sample variogram developed by Dr. Diggle and does not require balance. Times can be unequally spaced within patients and can differ across subjects. The programming involves advanced statistical subjects and is covered in our LONG93 class. The code is not provided here.



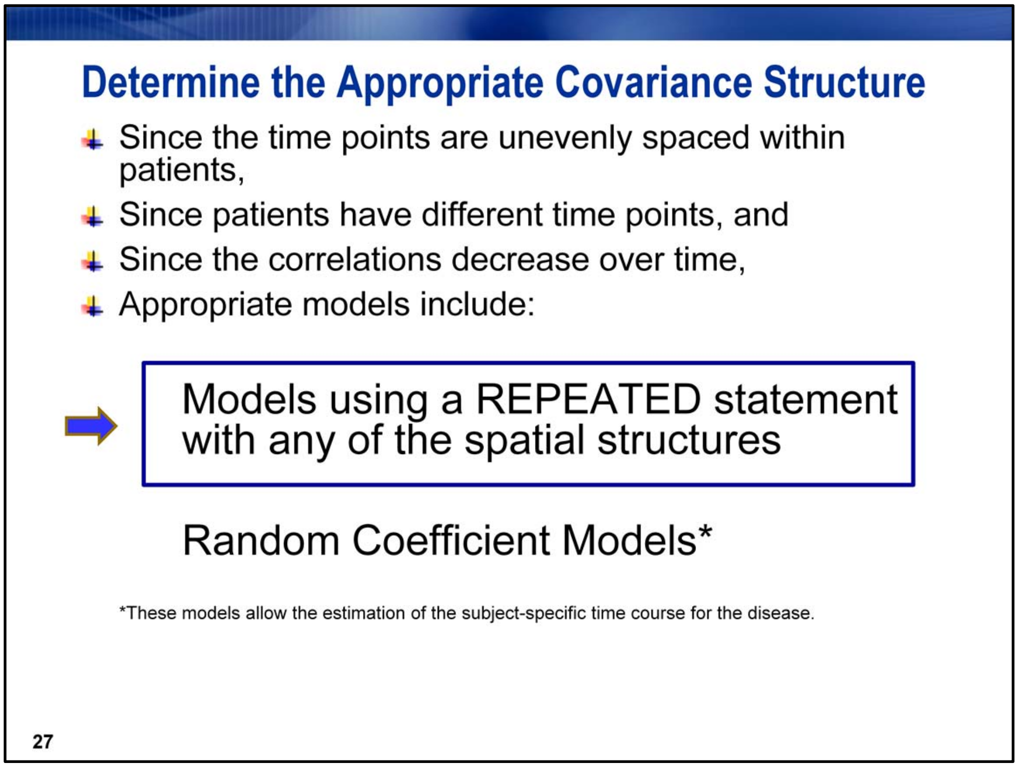
Findings from Data Exploration

- Time points are unequally spaced within subjects and differ across subjects.
- Correlations appear to decrease over time.
- Time has a cubic relationship with CD 4+ cell count.
- Evidence of a time by cigarette usage interaction and possibly a time by age interaction.



Findings from Data Exploration (continued)

- Positive relationship between the baseline CD4+ cell counts and the baseline cigarette usage.
- Relationships between the change in CD4+ cell counts and changes in: cigarette use (positive), depression score (negative), and partners (positive).



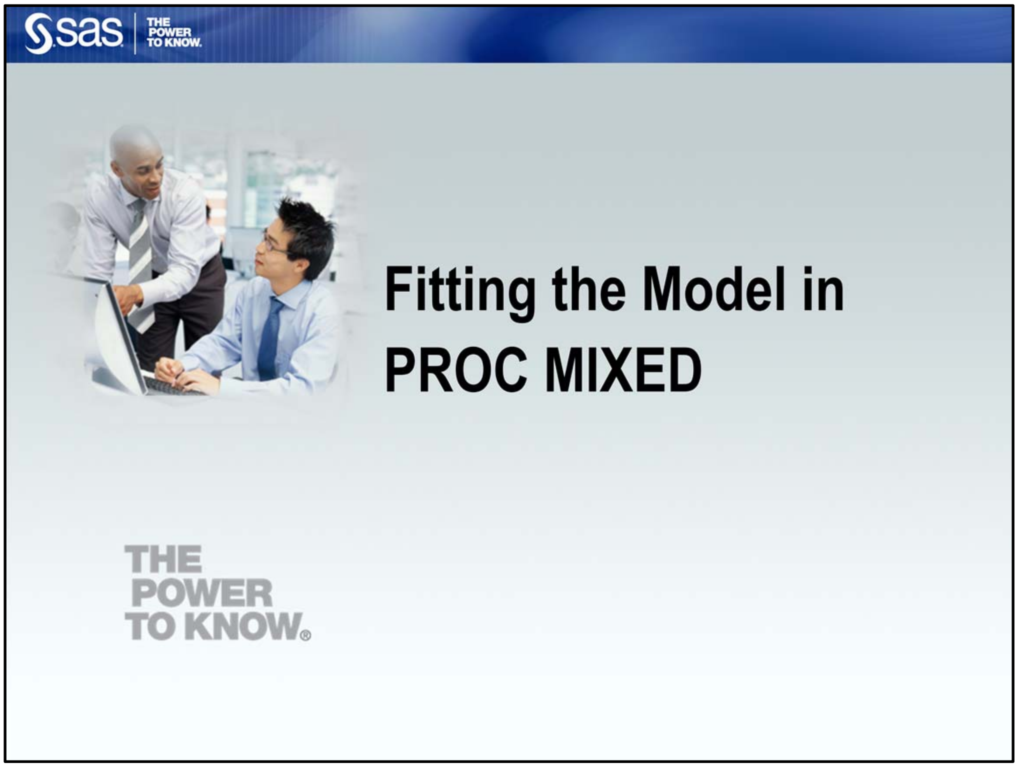
Determine the Appropriate Covariance Structure

- Since the time points are unevenly spaced within patients,
- Since patients have different time points, and
- Since the correlations decrease over time,
- Appropriate models include:

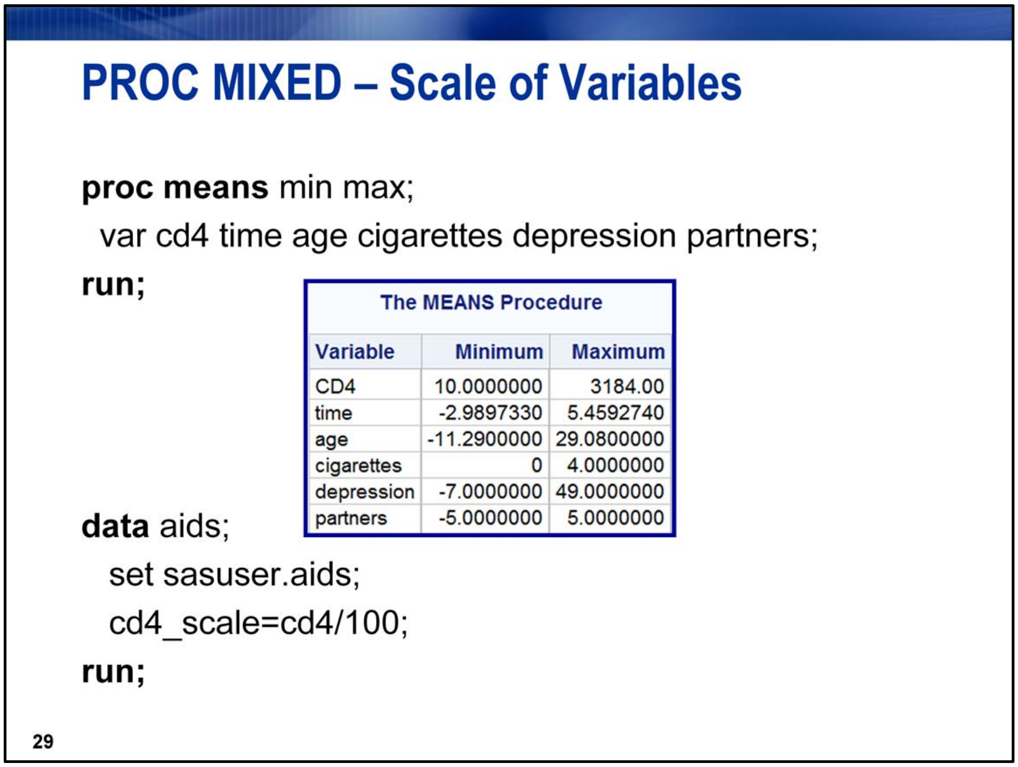
➡ Models using a REPEATED statement with any of the spatial structures

Random Coefficient Models*

*These models allow the estimation of the subject-specific time course for the disease.



Fitting the Model in PROC MIXED



PROC MIXED – Scale of Variables

```
proc means min max;
  var cd4 time age cigarettes depression partners;
run;
```

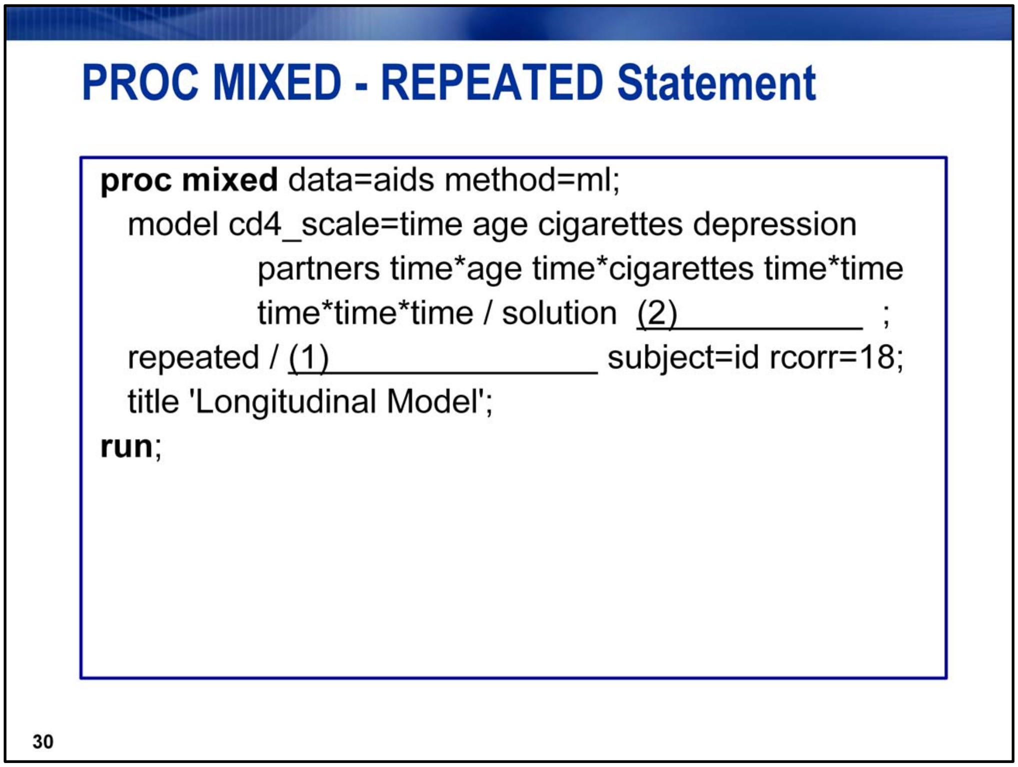
The MEANS Procedure

Variable	Minimum	Maximum
CD4	10.0000000	3184.00
time	-2.9897330	5.4592740
age	-11.2900000	29.0800000
cigarettes	0	4.0000000
depression	-7.0000000	49.0000000
partners	-5.0000000	5.0000000

```
data aids;
   set sasuser.aids;
   cd4_scale=cd4/100;
run;
```

With the exception of the CD4, the variables are on similar scales.

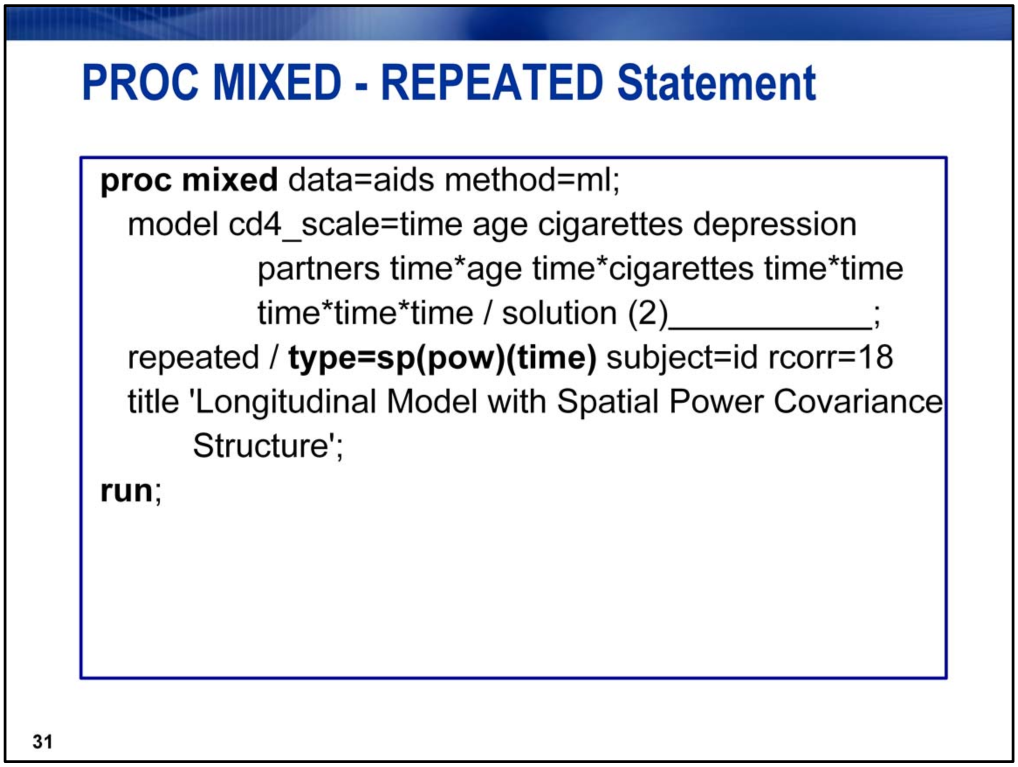
Analytical advice: rescale variables so that they are similar in scale; this helps with model convergence and stability



PROC MIXED - REPEATED Statement

```
proc mixed data=aids method=ml;
  model cd4_scale=time age cigarettes depression
            partners time*age time*cigarettes time*time
            time*time*time / solution  (2)___________  ;
  repeated / (1)_______________ subject=id rcorr=18;
  title 'Longitudinal Model';
run;
```

Cover all code as shown, pointing out that the quadratic and cubic effects of time are in the model, as well as the time*age and time*cigarettes interaction terms. These are from data exploration.



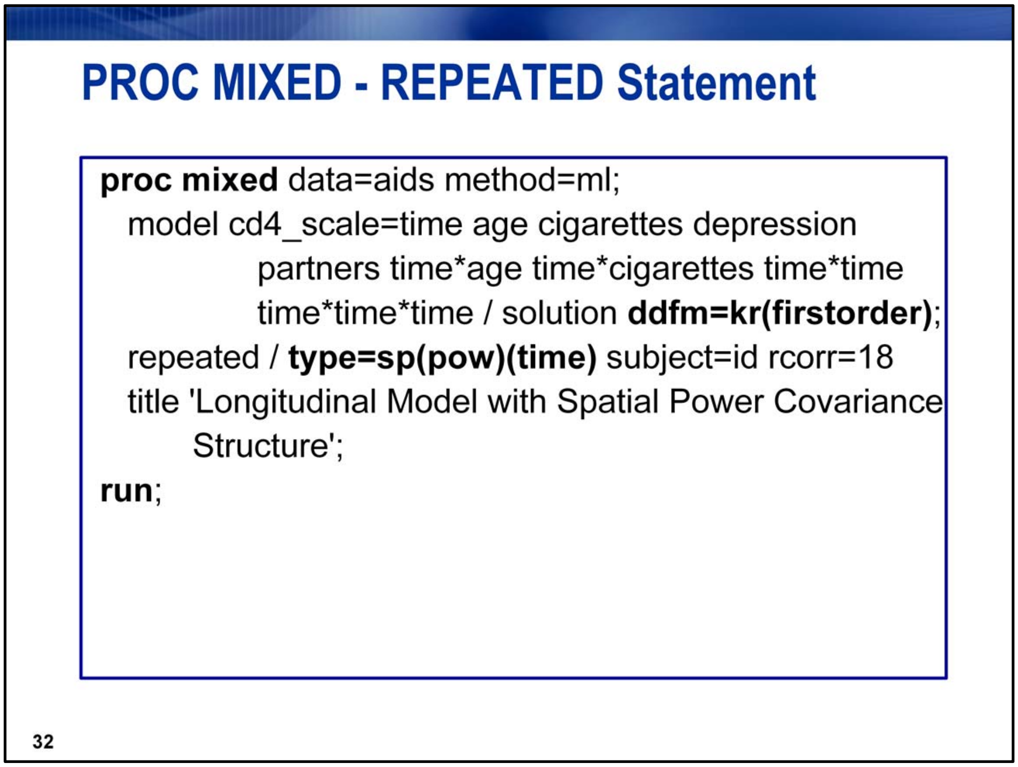
PROC MIXED - REPEATED Statement

```
proc mixed data=aids method=ml;
  model cd4_scale=time age cigarettes depression
            partners time*age time*cigarettes time*time
            time*time*time / solution (2)___________;
  repeated / type=sp(pow)(time) subject=id rcorr=18
  title 'Longitudinal Model with Spatial Power Covariance
       Structure';
run;
```

31

Note: TYPE=SP(POW) (time) Because of the different time points within subjects and different time points across subjects, any of the spatial structures would be appropriate.

Experts indicate that selecting a covariance pattern that fits the structure of the data and allows correlations to decrease over time is more important than fine tuning the structure between compatible but competing structures.

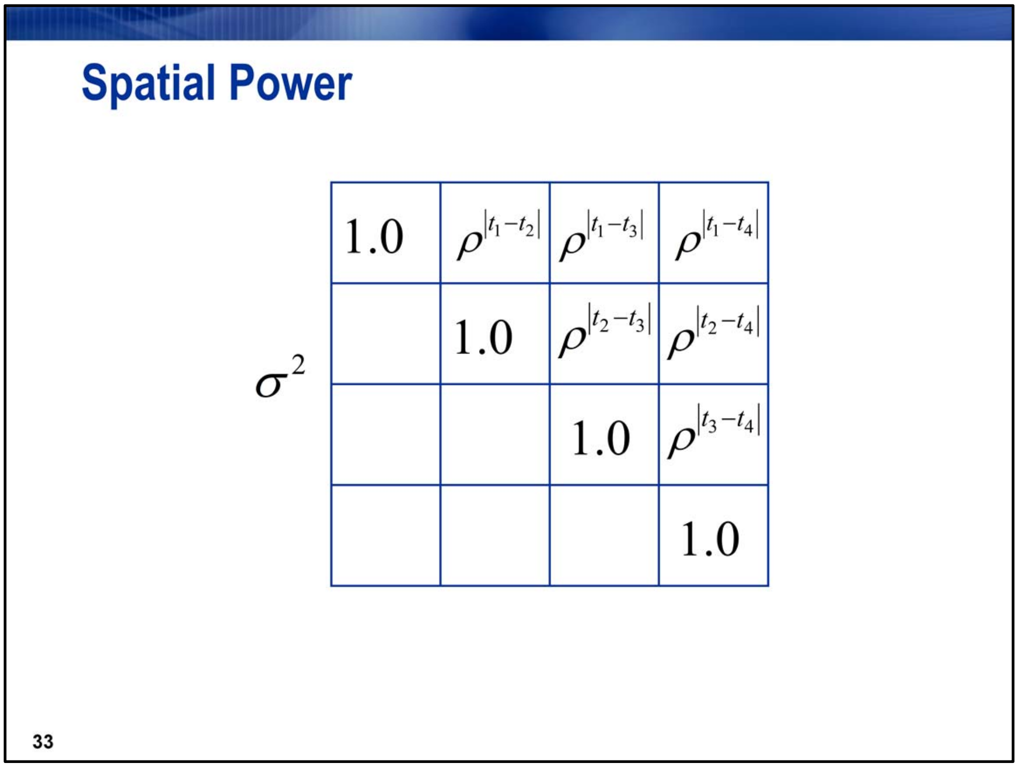


PROC MIXED - REPEATED Statement

```
proc mixed data=aids method=ml;
  model cd4_scale=time age cigarettes depression
            partners time*age time*cigarettes time*time
            time*time*time / solution ddfm=kr(firstorder);
  repeated / type=sp(pow)(time) subject=id rcorr=18
  title 'Longitudinal Model with Spatial Power Covariance
       Structure';
run;
```

32

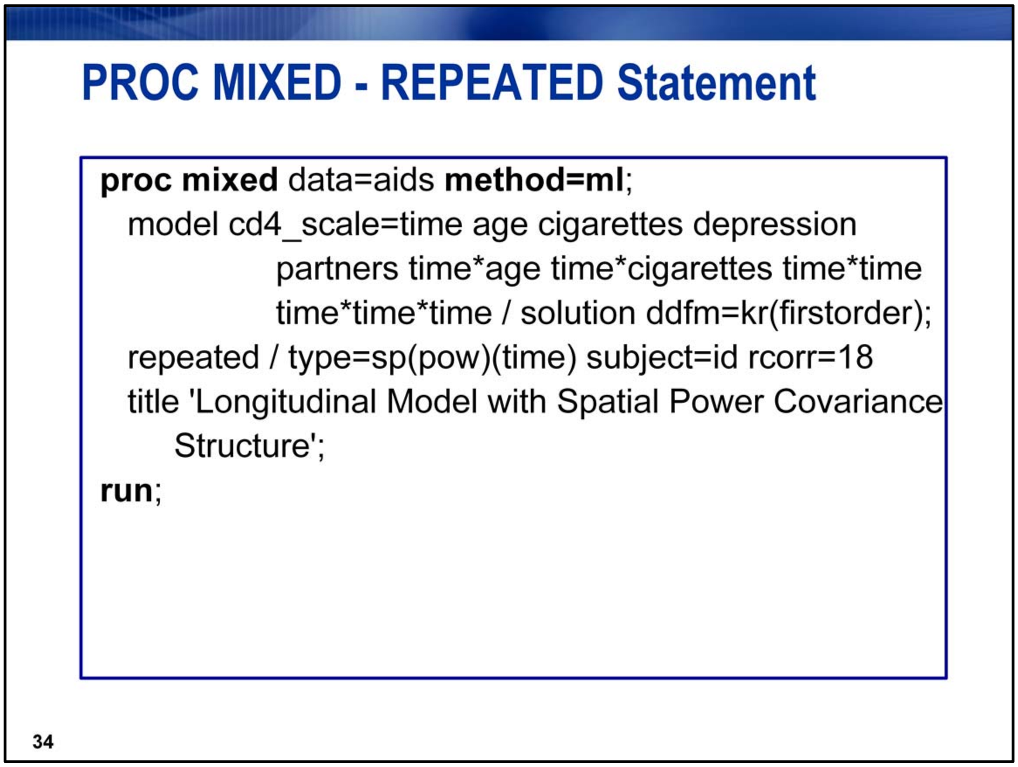
Note: DDFM=KR(firstorder) – recommended. The FIRSTORDER option is available for 9.2 or later. The KR adjustment helps reduce the type1 error rate but can lead to shrinkage of standard errors of fixed effects with complicated correlation structures (i.e., those having nonzero second derivatives), such as the SPATIAL structures. The FIRSTORDER offsets the shrinkage by eliminating second order derivatives from the calculations.



Spatial Power

$$\sigma^2 \begin{array}{|c|c|c|c|} \hline 1.0 & \rho^{|t_1-t_2|} & \rho^{|t_1-t_3|} & \rho^{|t_1-t_4|} \\ \hline & 1.0 & \rho^{|t_2-t_3|} & \rho^{|t_2-t_4|} \\ \hline & & 1.0 & \rho^{|t_3-t_4|} \\ \hline & & & 1.0 \\ \hline \end{array}$$

Since the actual time points are used to calculate the correlations, no requirements are placed on the structure of the time points within or between subjects.

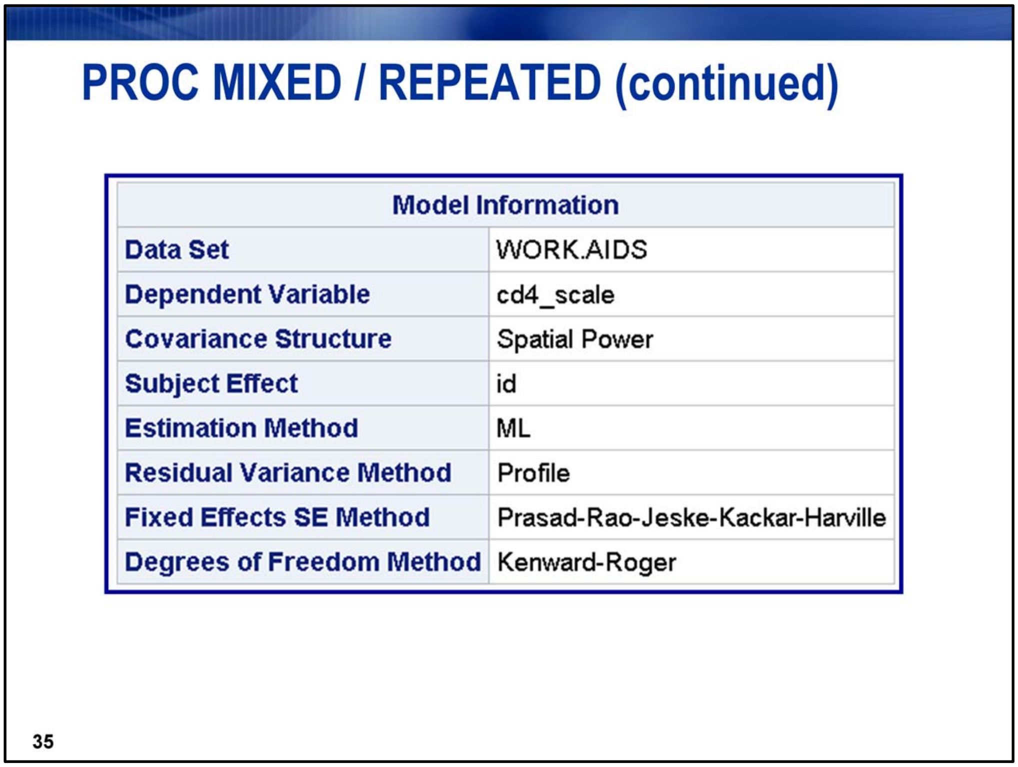


PROC MIXED - REPEATED Statement

```
proc mixed data=aids method=ml;
  model cd4_scale=time age cigarettes depression
              partners time*age time*cigarettes time*time
              time*time*time / solution ddfm=kr(firstorder);
  repeated / type=sp(pow)(time) subject=id rcorr=18
  title 'Longitudinal Model with Spatial Power Covariance
      Structure';
run;
```

34

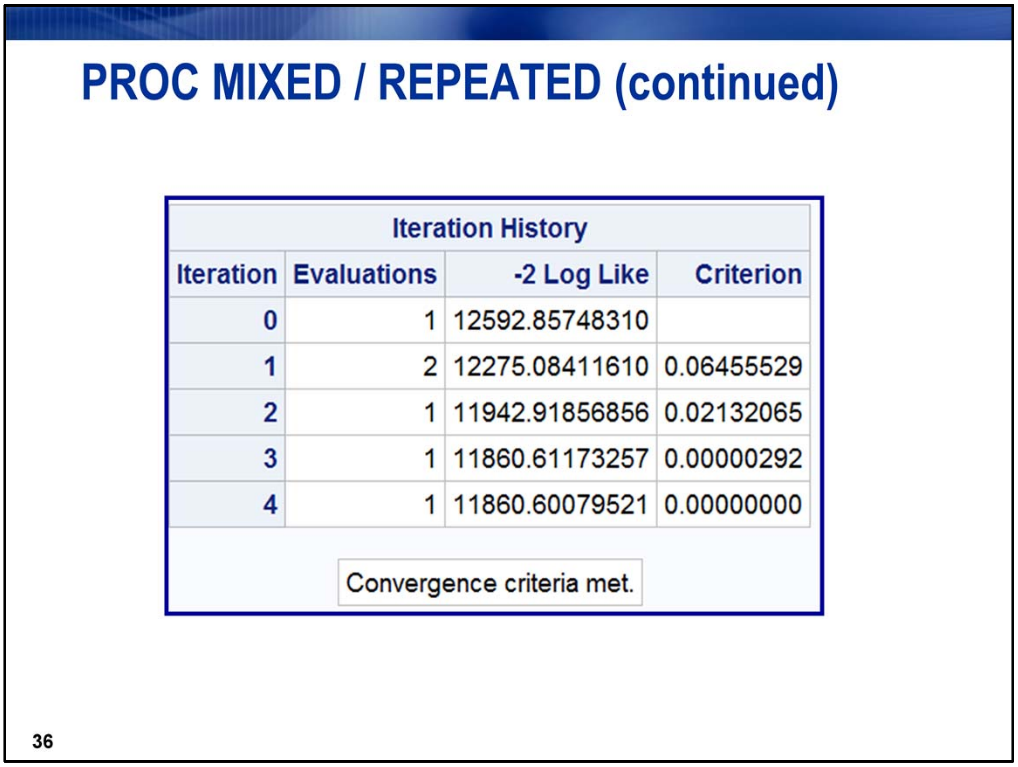
Note the METHOD=ML. We are satisfied with the selected covariance structure because of the unbalanced nature of the data set. So we will turn our attention to the MODEL statement, and will use the ML estimation method, so that we may use the fit statistics to compare models.



PROC MIXED / REPEATED (continued)

Model Information	
Data Set	WORK.AIDS
Dependent Variable	cd4_scale
Covariance Structure	Spatial Power
Subject Effect	id
Estimation Method	ML
Residual Variance Method	Profile
Fixed Effects SE Method	Prasad-Rao-Jeske-Kackar-Harville
Degrees of Freedom Method	Kenward-Roger

Cover all – make note of ML, Prasad-Rao-Jeskie-Kakckac-Harville (from firstorder option) and KR

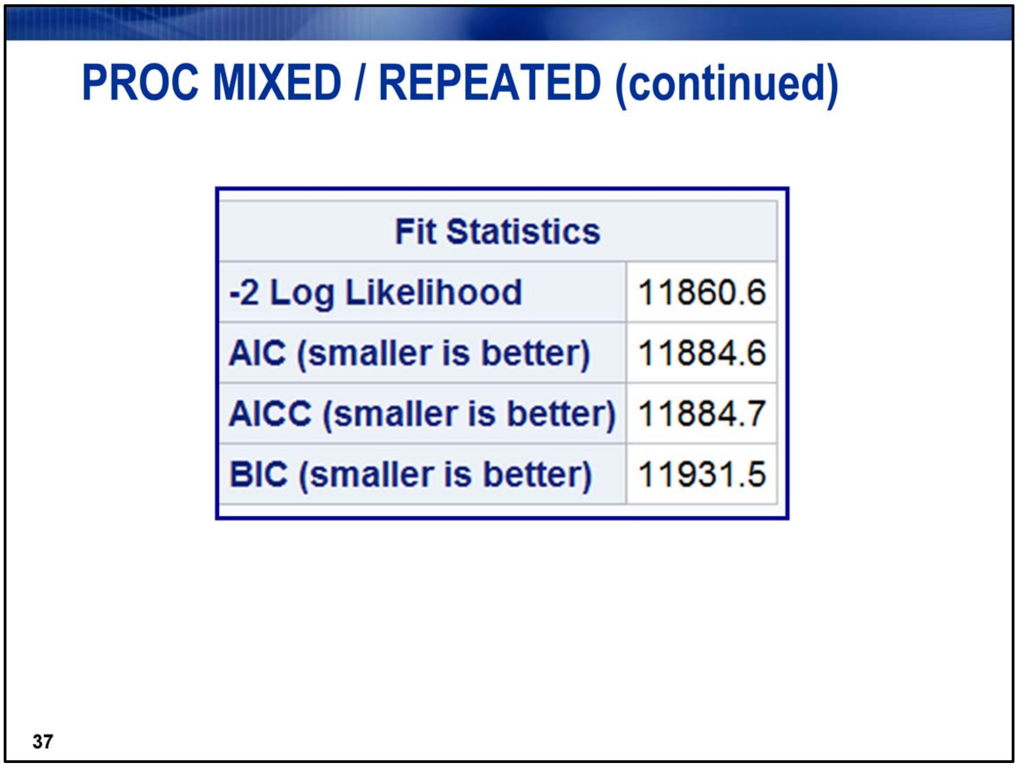


PROC MIXED / REPEATED (continued)

Iteration History			
Iteration	Evaluations	-2 Log Like	Criterion
0	1	12592.85748310	
1	2	12275.08411610	0.06455529
2	1	11942.91856856	0.02132065
3	1	11860.61173257	0.00000292
4	1	11860.60079521	0.00000000

Convergence criteria met.

The model converged quickly in four iterations.

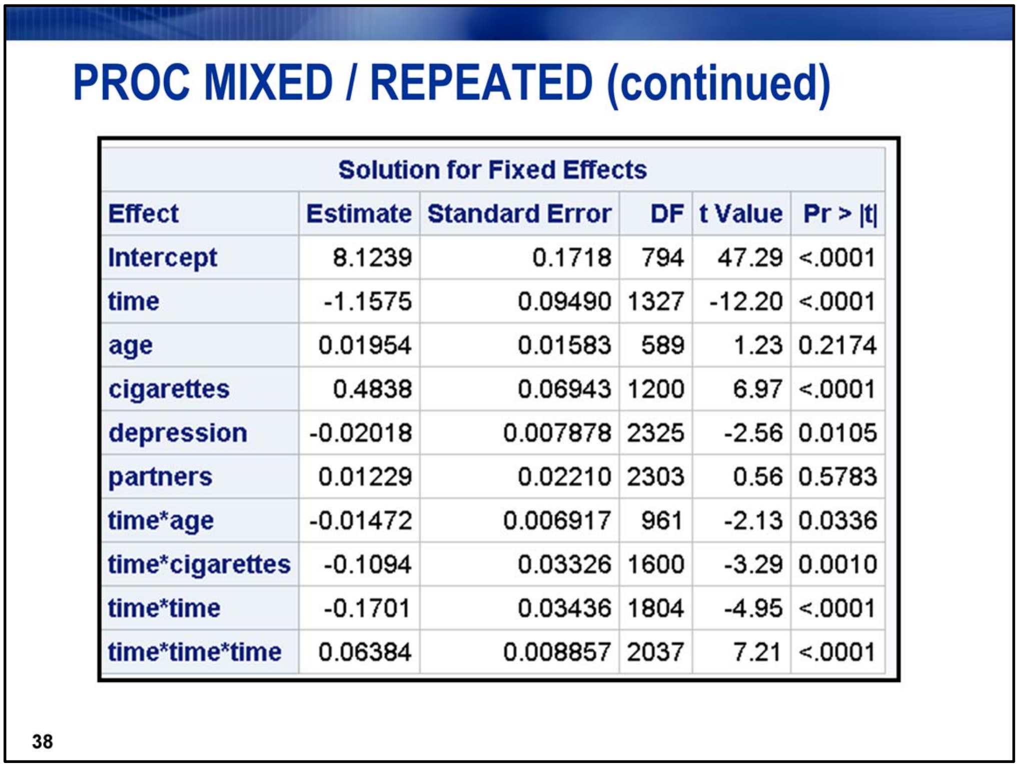


PROC MIXED / REPEATED (continued)

Fit Statistics	
-2 Log Likelihood	11860.6
AIC (smaller is better)	11884.6
AICC (smaller is better)	11884.7
BIC (smaller is better)	11931.5

Since se are deciding on fixed effects we use ML and get -2 Log Likelihood statistics and its variations, rather than -2 res log likelihood and its variations that we would get from the default REML estimation.

The fits statistics under ML are appropriate for comparing models with different fixed effects; first statistics under REML would be appropriate for comparing models with different covariance structures, but with the same mean model.

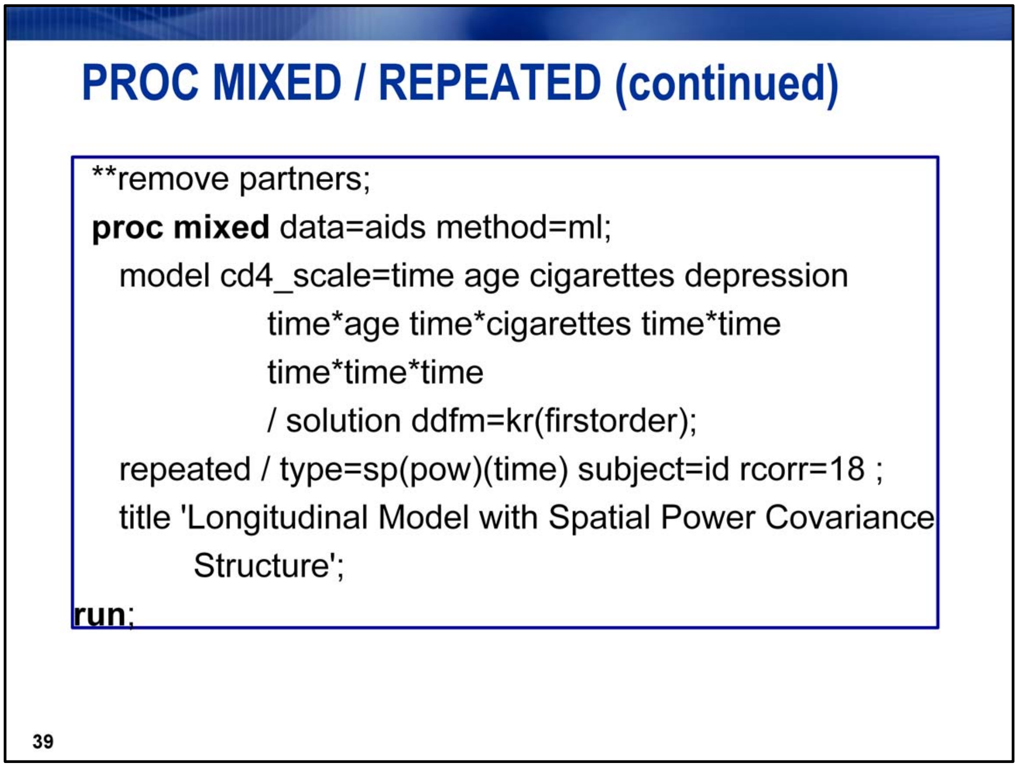


PROC MIXED / REPEATED (continued)

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	8.1239	0.1718	794	47.29	<.0001
time	-1.1575	0.09490	1327	-12.20	<.0001
age	0.01954	0.01583	589	1.23	0.2174
cigarettes	0.4838	0.06943	1200	6.97	<.0001
depression	-0.02018	0.007878	2325	-2.56	0.0105
partners	0.01229	0.02210	2303	0.56	0.5783
time*age	-0.01472	0.006917	961	-2.13	0.0336
time*cigarettes	-0.1094	0.03326	1600	-3.29	0.0010
time*time	-0.1701	0.03436	1804	-4.95	<.0001
time*time*time	0.06384	0.008857	2037	7.21	<.0001

The higher order terms of **time** are significant and **time** is involved in significant interactions with **age** and **cigarettes.** Because of the interaction, **age** will stay in the model as a main effect, even though it is not significant. Of the two variables not involved in interactions, **partners** is not significant, but **depression** is.

Rerun the model without **partners**.

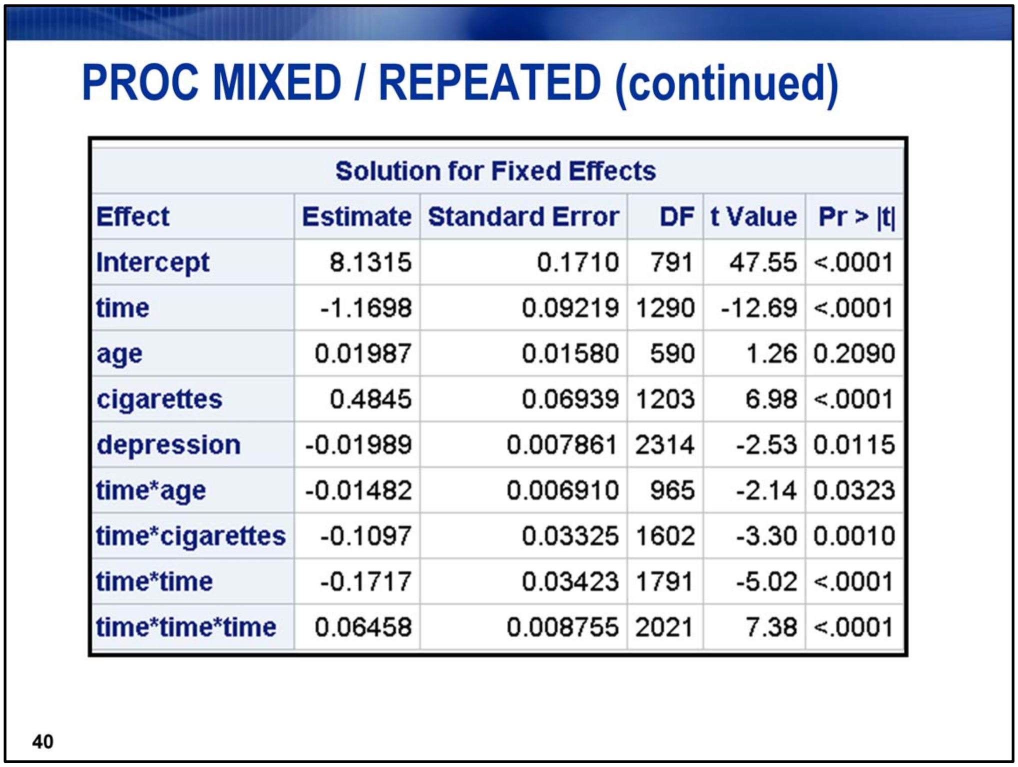


PROC MIXED / REPEATED (continued)

```
 **remove partners;
 proc mixed data=aids method=ml;
   model cd4_scale=time age cigarettes depression
                time*age time*cigarettes time*time
                time*time*time
                / solution ddfm=kr(firstorder);
   repeated / type=sp(pow)(time) subject=id rcorr=18 ;
   title 'Longitudinal Model with Spatial Power Covariance
          Structure';
run;
```

39

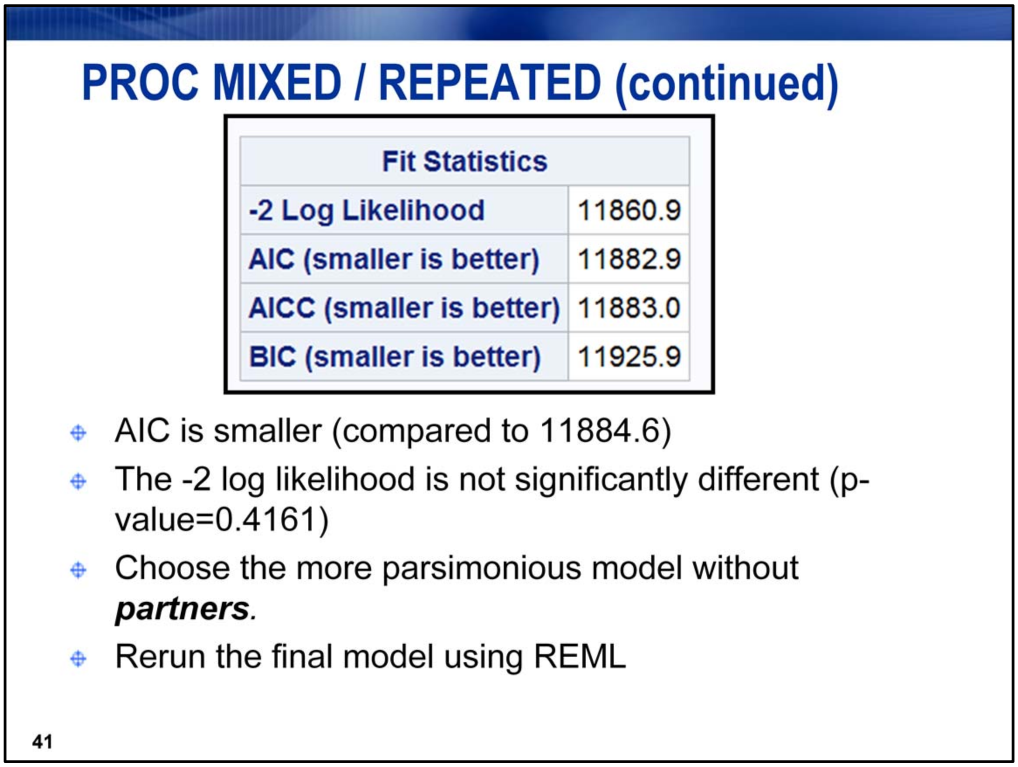
Rerun the model and remove **PARTNERS**, still using ML estimation.



PROC MIXED / REPEATED (continued)

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	8.1315	0.1710	791	47.55	<.0001
time	-1.1698	0.09219	1290	-12.69	<.0001
age	0.01987	0.01580	590	1.26	0.2090
cigarettes	0.4845	0.06939	1203	6.98	<.0001
depression	-0.01989	0.007861	2314	-2.53	0.0115
time*age	-0.01482	0.006910	965	-2.14	0.0323
time*cigarettes	-0.1097	0.03325	1602	-3.30	0.0010
time*time	-0.1717	0.03423	1791	-5.02	<.0001
time*time*time	0.06458	0.008755	2021	7.38	<.0001

Now all terms, except for **AGE** are significant. Remember that these p-values are biased since we are making data-driven decisions.



PROC MIXED / REPEATED (continued)

Fit Statistics	
-2 Log Likelihood	11860.9
AIC (smaller is better)	11882.9
AICC (smaller is better)	11883.0
BIC (smaller is better)	11925.9

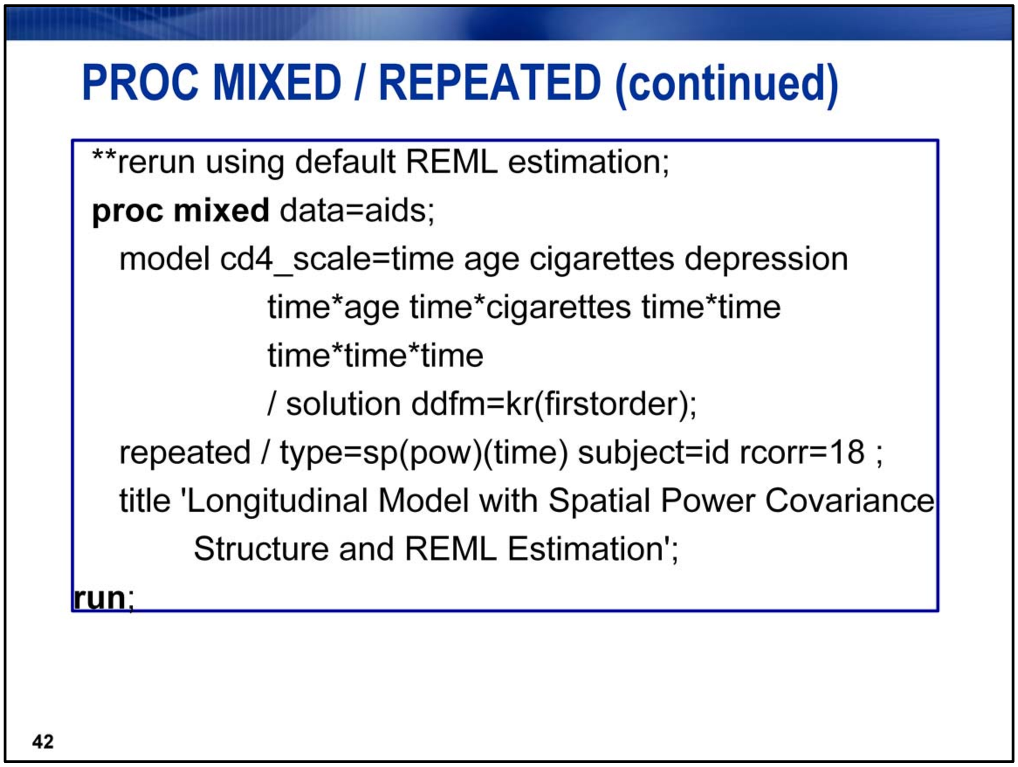
- AIC is smaller (compared to 11884.6)
- The -2 log likelihood is not significantly different (p-value=0.4161)
- Choose the more parsimonious model without ***partners****.*
- Rerun the final model using REML

Fit Statistics from previous model:

-2 Log Likelihood	11860.6
AIC (smaller is better)	11884.6
AICC (smaller is better)	11884.7
BIC (smaller is better)	11931.5

```
*likelihood ratio test;
data _null_;
  dev1=11860.6;
  dev2=11860.9;
  chisq=dev2-dev1;
  pval=probchi(chisq,1);
  put pval=;
run;
```

From the log: pval=0.4161175792

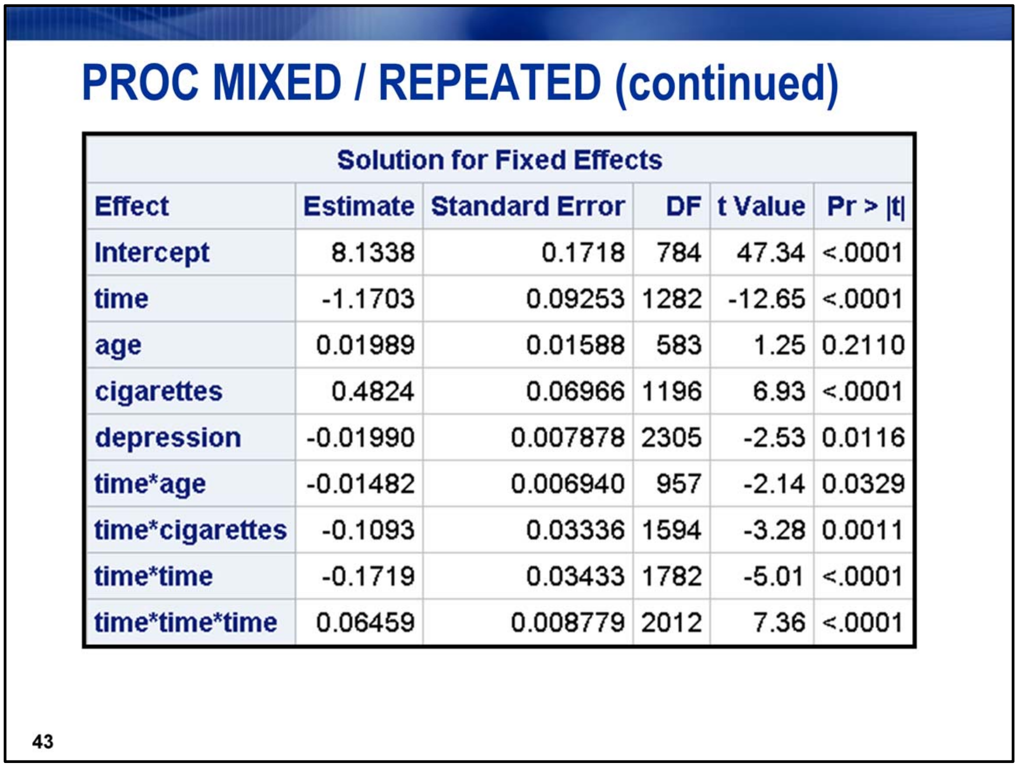


PROC MIXED / REPEATED (continued)

```
 **rerun using default REML estimation;
 proc mixed data=aids;
   model cd4_scale=time age cigarettes depression
              time*age time*cigarettes time*time
              time*time*time
              / solution ddfm=kr(firstorder);
   repeated / type=sp(pow)(time) subject=id rcorr=18 ;
   title 'Longitudinal Model with Spatial Power Covariance
         Structure and REML Estimation';
run;
```

42

Rerun the final model using REML. It adjusts for the downward bias in the covariance parameters.



PROC MIXED / REPEATED (continued)

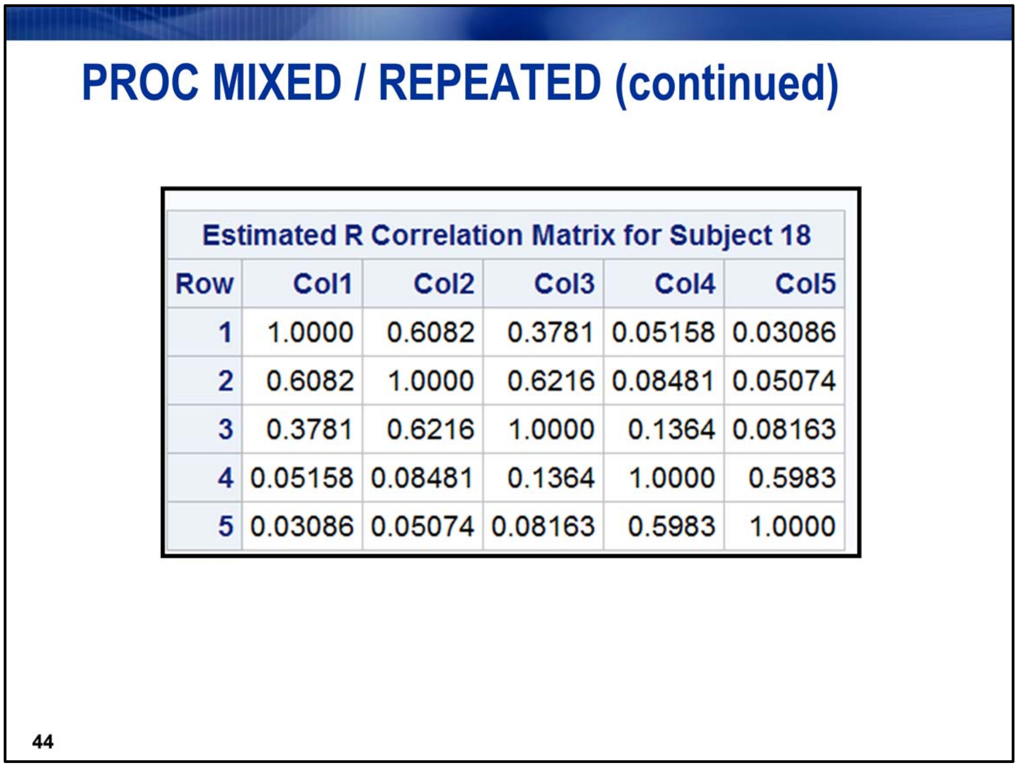
Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	8.1338	0.1718	784	47.34	<.0001
time	-1.1703	0.09253	1282	-12.65	<.0001
age	0.01989	0.01588	583	1.25	0.2110
cigarettes	0.4824	0.06966	1196	6.93	<.0001
depression	-0.01990	0.007878	2305	-2.53	0.0116
time*age	-0.01482	0.006940	957	-2.14	0.0329
time*cigarettes	-0.1093	0.03336	1594	-3.28	0.0011
time*time	-0.1719	0.03433	1782	-5.01	<.0001
time*time*time	0.06459	0.008779	2012	7.36	<.0001

43

None of the inference has changed, but the betas and standard errors have changed slightly.

The only term not involved in an interaction is **DEPRESSION**. It's estimated effect is for each 1-unit in crease in DEPRESSION score, scaled CD4 decreases by 0.01990. On the original scale, that would translate to a decrease of almost 2 CD4 cells. (1.99)

The best way to explain terms involved in interactions or higher-order terms is either the results of ESTIMATE statements or interaction plots.

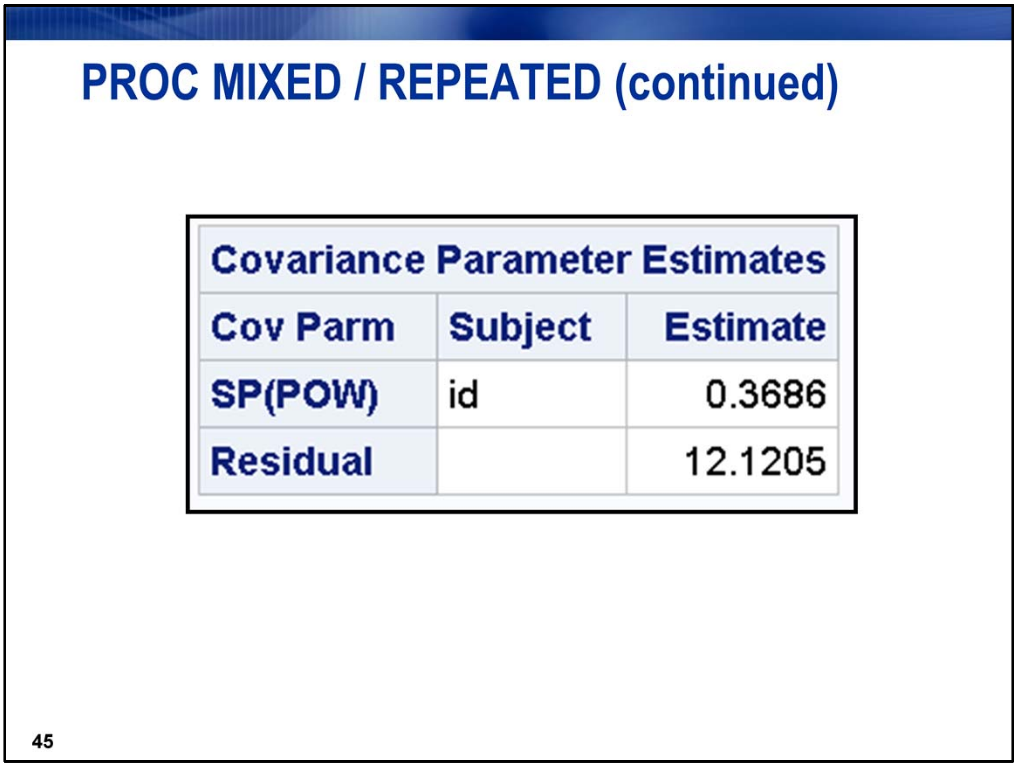


PROC MIXED / REPEATED (continued)

Estimated R Correlation Matrix for Subject 18					
Row	Col1	Col2	Col3	Col4	Col5
1	1.0000	0.6082	0.3781	0.05158	0.03086
2	0.6082	1.0000	0.6216	0.08481	0.05074
3	0.3781	0.6216	1.0000	0.1364	0.08163
4	0.05158	0.08481	0.1364	1.0000	0.5983
5	0.03086	0.05074	0.08163	0.5983	1.0000

44

Note how the correlations decrease over time, but not in a uniform pattern. Also, every correlation is different. The Spatial structures give you the flexibility of an unstructured correlation matrix, but only require the estimation of two covariance parameters, instead of T*(T+1)/2. The two covariance parameters are rho and the residual variance.

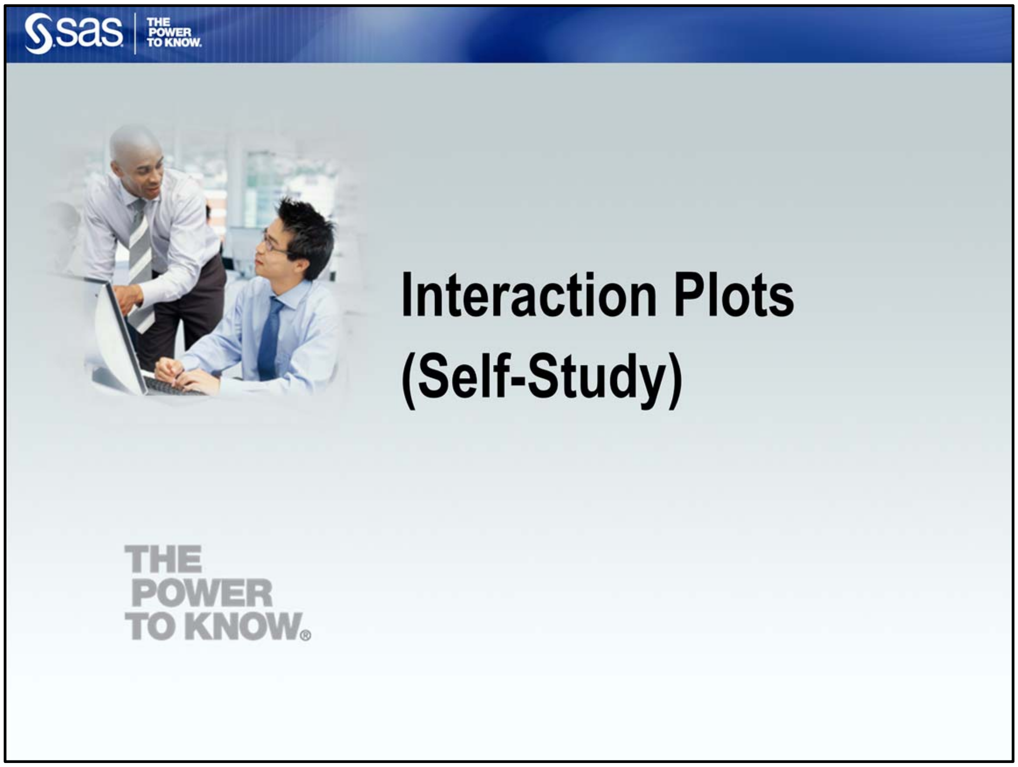


PROC MIXED / REPEATED (continued)

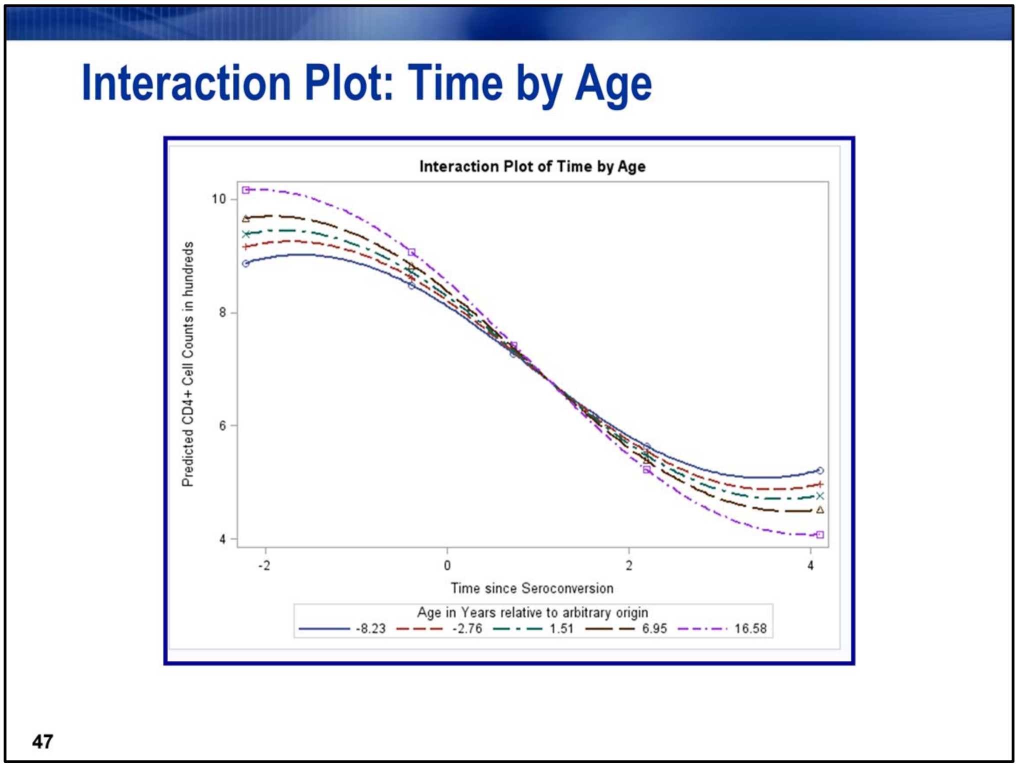
Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
SP(POW)	id	0.3686
Residual		12.1205

45

These are the only two covariance parameters that need to be estimated, even though all correlations have the flexibility to be different(since the time-spacings are unequal).

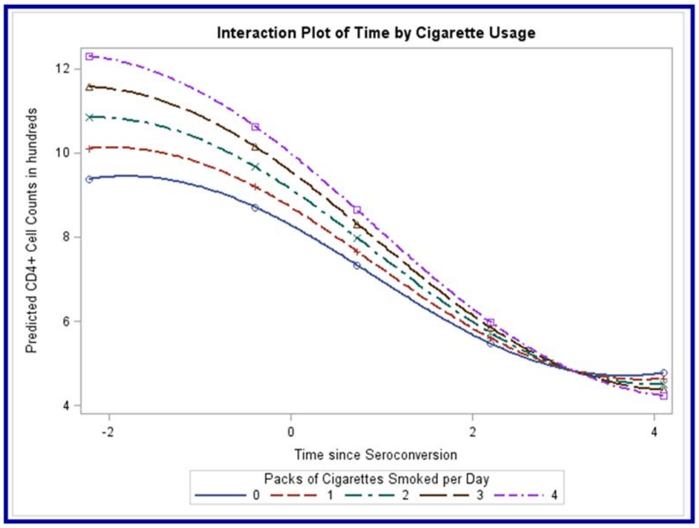


Interaction Plots (Self-Study)



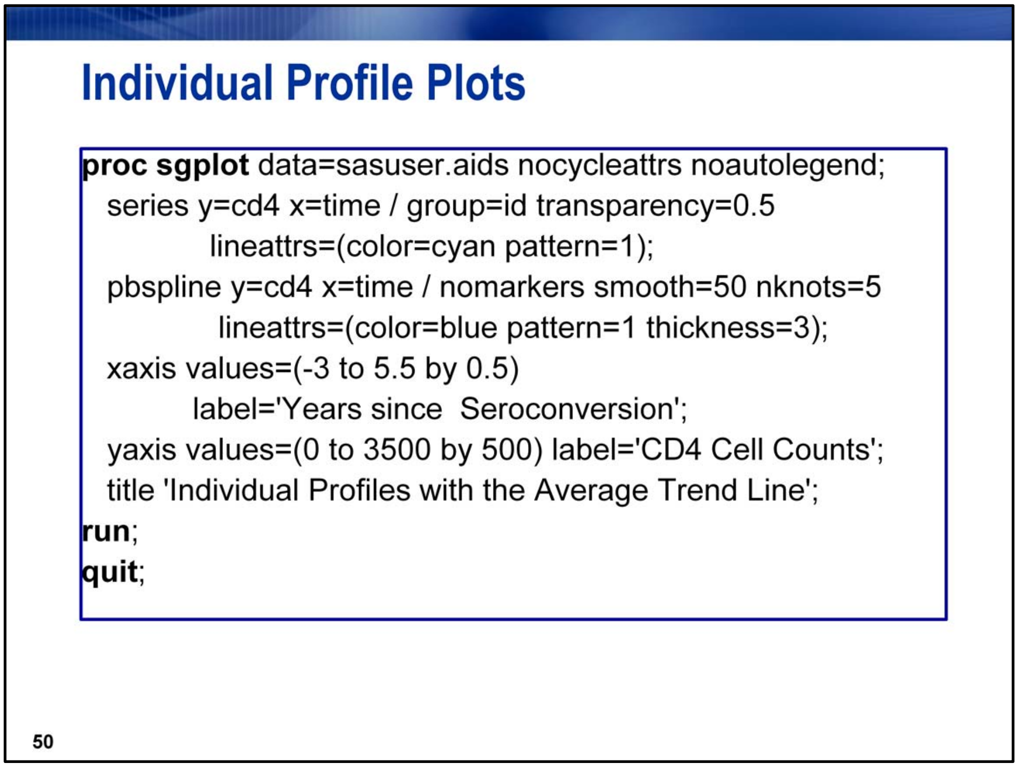
Interaction Plot: Time by Age

Interaction Plot: Time by Cigarettes



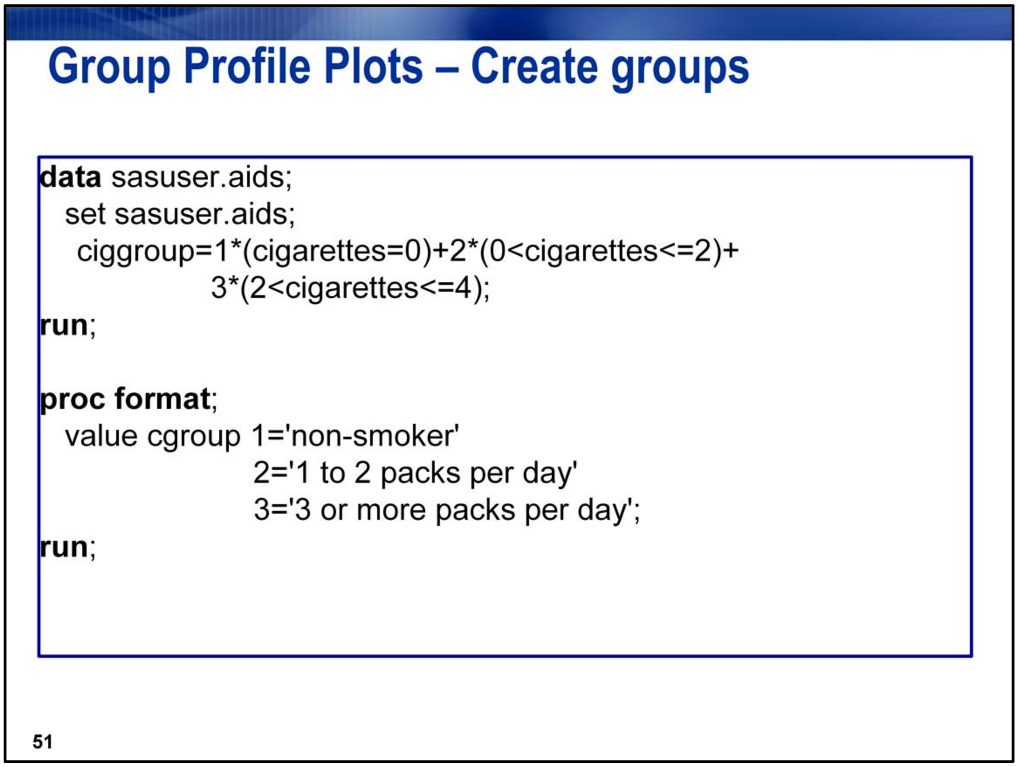
Appendix

SAS Programs



Individual Profile Plots

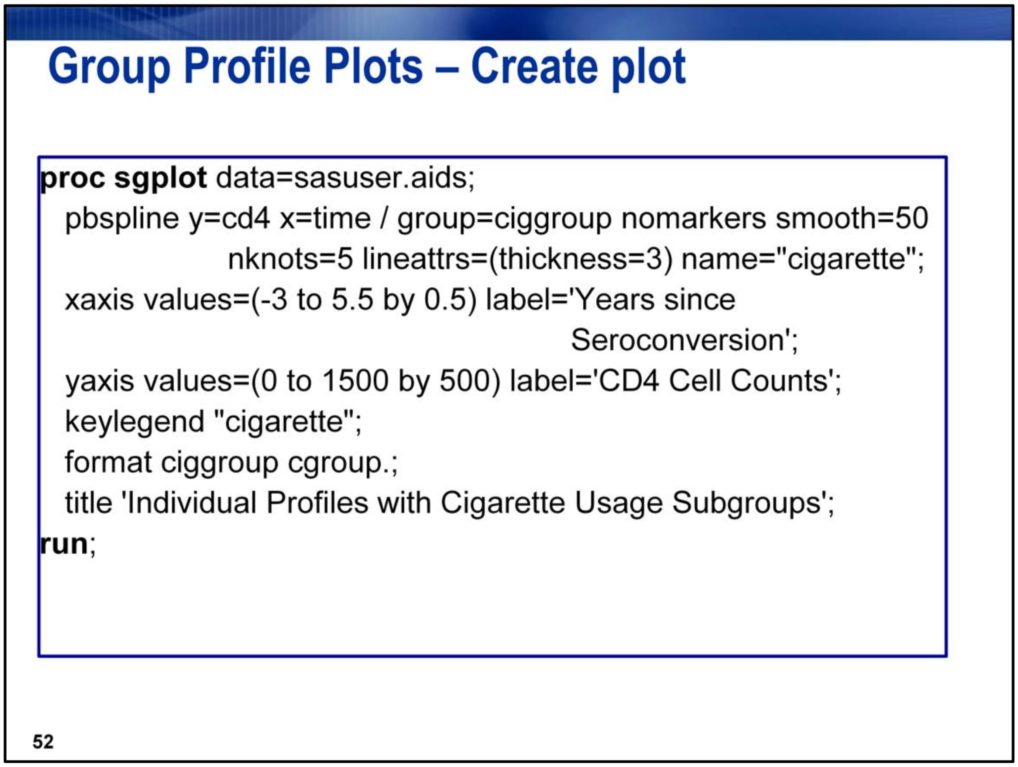
```
proc sgplot data=sasuser.aids nocycleattrs noautolegend;
  series y=cd4 x=time / group=id transparency=0.5
            lineattrs=(color=cyan pattern=1);
  pbspline y=cd4 x=time / nomarkers smooth=50 nknots=5
             lineattrs=(color=blue pattern=1 thickness=3);
  xaxis values=(-3 to 5.5 by 0.5)
          label='Years since  Seroconversion';
  yaxis values=(0 to 3500 by 500) label='CD4 Cell Counts';
  title 'Individual Profiles with the Average Trend Line';
run;
quit;
```



Group Profile Plots – Create groups

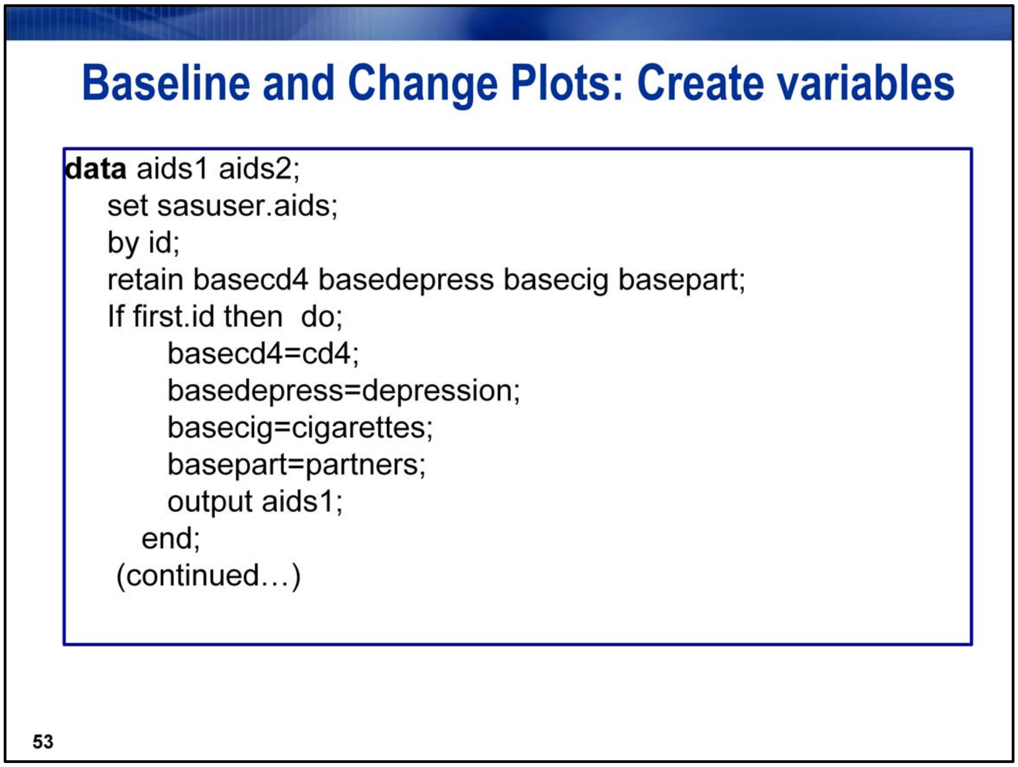
```
data sasuser.aids;
  set sasuser.aids;
   ciggroup=1*(cigarettes=0)+2*(0<cigarettes<=2)+
              3*(2<cigarettes<=4);
run;

proc format;
  value cgroup 1='non-smoker'
                  2='1 to 2 packs per day'
                  3='3 or more packs per day';
run;
```



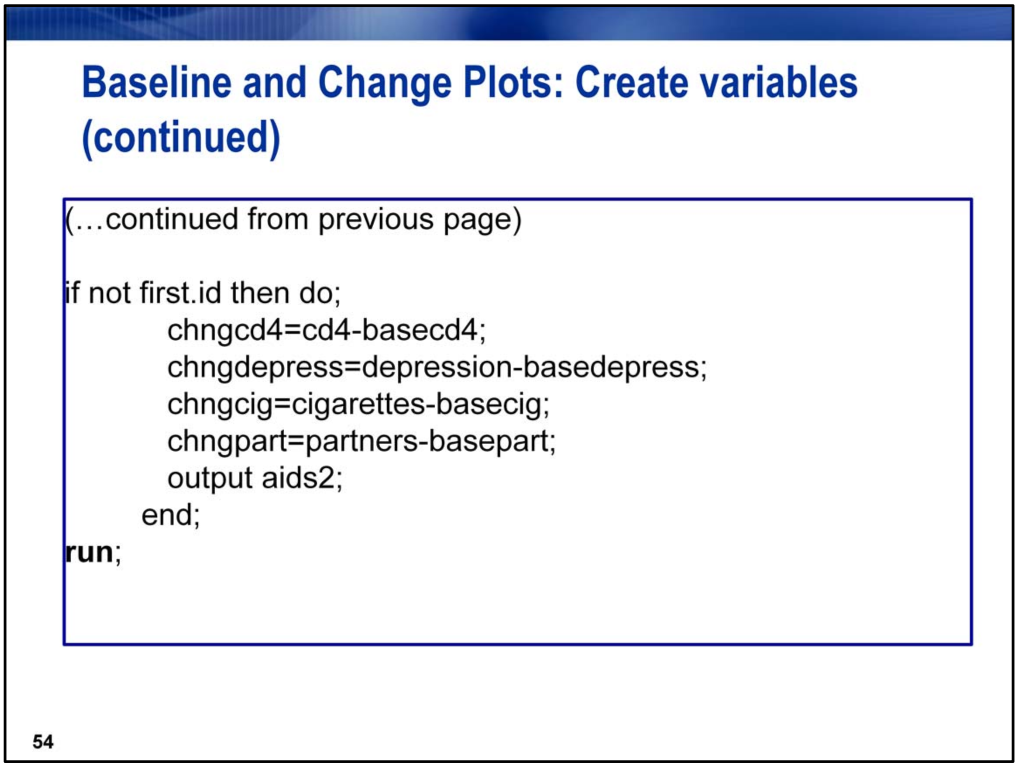
Group Profile Plots – Create plot

```
proc sgplot data=sasuser.aids;
  pbspline y=cd4 x=time / group=ciggroup nomarkers smooth=50
                nknots=5 lineattrs=(thickness=3) name="cigarette";
  xaxis values=(-3 to 5.5 by 0.5) label='Years since
                                        Seroconversion';
  yaxis values=(0 to 1500 by 500) label='CD4 Cell Counts';
  keylegend "cigarette";
  format ciggroup cgroup.;
  title 'Individual Profiles with Cigarette Usage Subgroups';
run;
```



Baseline and Change Plots: Create variables

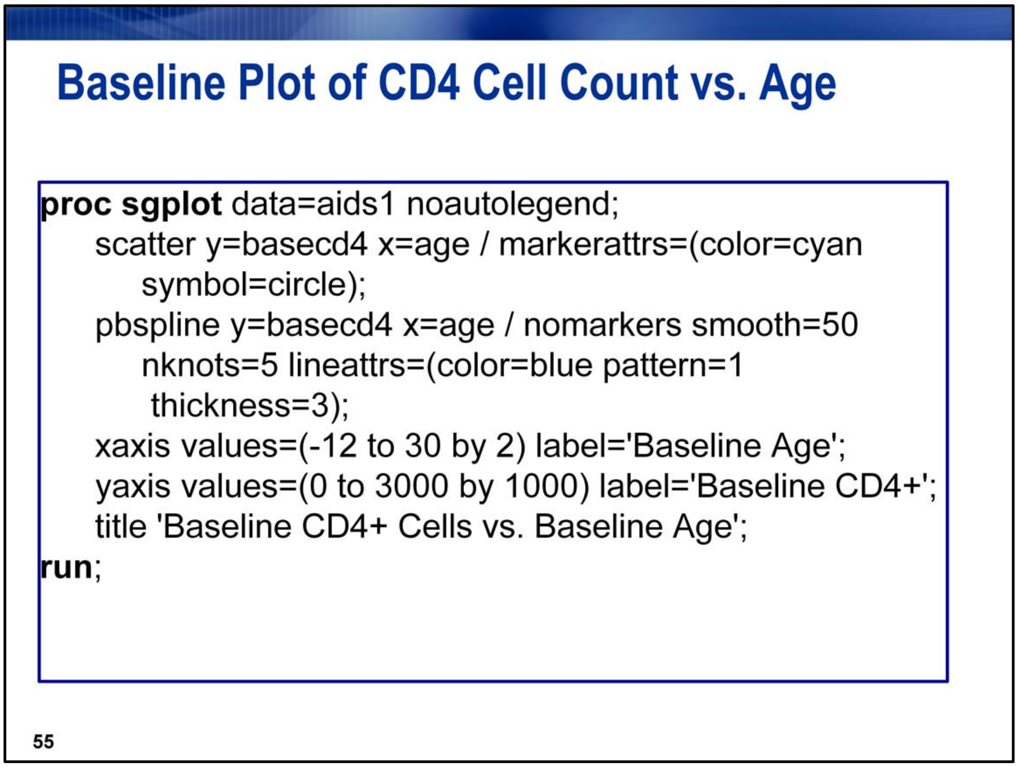
```
data aids1 aids2;
    set sasuser.aids;
    by id;
    retain basecd4 basedepress basecig basepart;
    If first.id then  do;
          basecd4=cd4;
          basedepress=depression;
          basecig=cigarettes;
          basepart=partners;
          output aids1;
       end;
     (continued…)
```



Baseline and Change Plots: Create variables (continued)

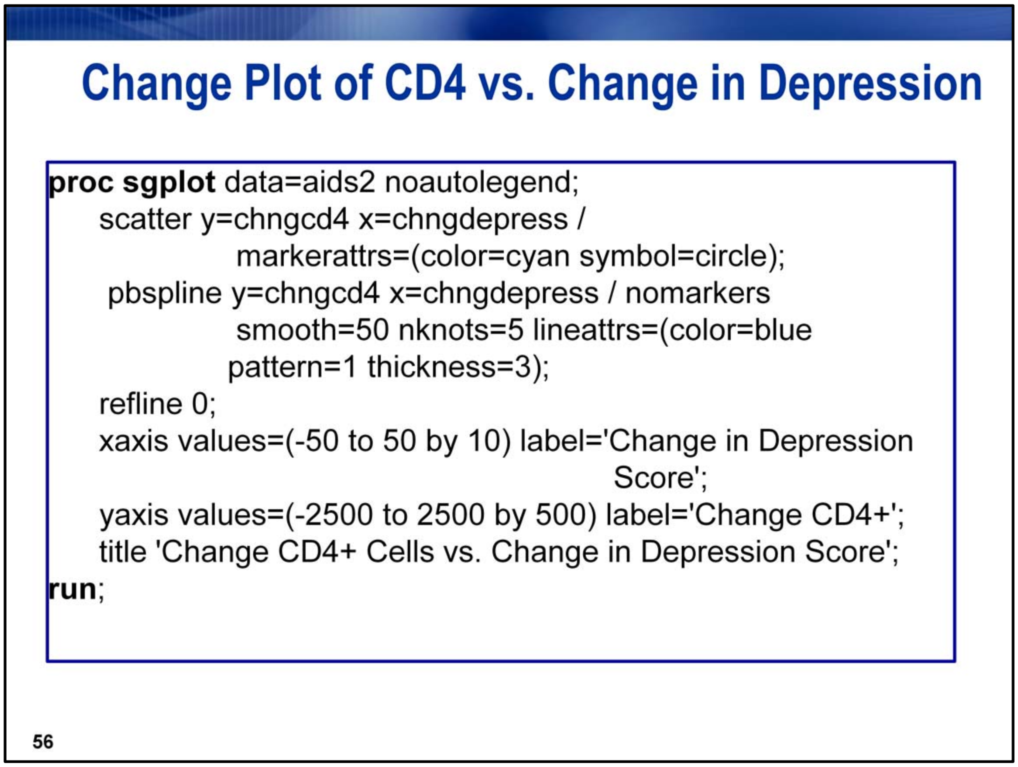
```
(…continued from previous page)

if not first.id then do;
        chngcd4=cd4-basecd4;
        chngdepress=depression-basedepress;
        chngcig=cigarettes-basecig;
        chngpart=partners-basepart;
        output aids2;
      end;
run;
```



Baseline Plot of CD4 Cell Count vs. Age

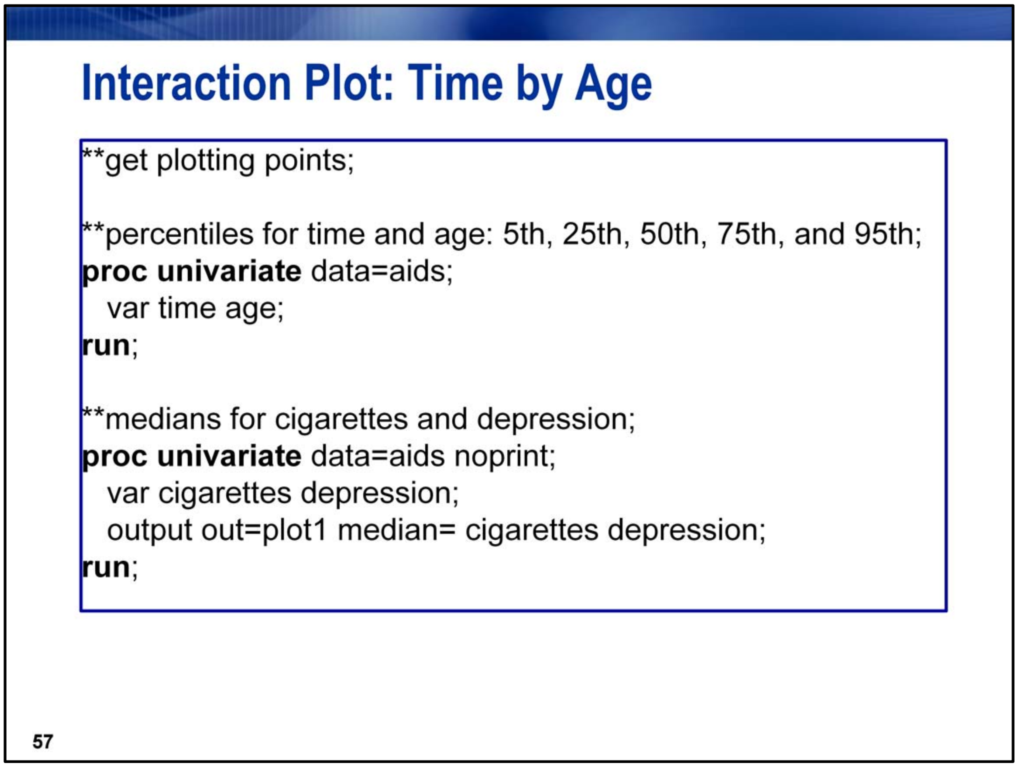
```
proc sgplot data=aids1 noautolegend;
     scatter y=basecd4 x=age / markerattrs=(color=cyan
         symbol=circle);
     pbspline y=basecd4 x=age / nomarkers smooth=50
         nknots=5 lineattrs=(color=blue pattern=1
          thickness=3);
     xaxis values=(-12 to 30 by 2) label='Baseline Age';
     yaxis values=(0 to 3000 by 1000) label='Baseline CD4+';
     title 'Baseline CD4+ Cells vs. Baseline Age';
run;
```



Change Plot of CD4 vs. Change in Depression

```
proc sgplot data=aids2 noautolegend;
     scatter y=chngcd4 x=chngdepress /
                   markerattrs=(color=cyan symbol=circle);
      pbspline y=chngcd4 x=chngdepress / nomarkers
                   smooth=50 nknots=5 lineattrs=(color=blue
                  pattern=1 thickness=3);
     refline 0;
     xaxis values=(-50 to 50 by 10) label='Change in Depression
                                                      Score';
     yaxis values=(-2500 to 2500 by 500) label='Change CD4+';
     title 'Change CD4+ Cells vs. Change in Depression Score';
run;
```

56

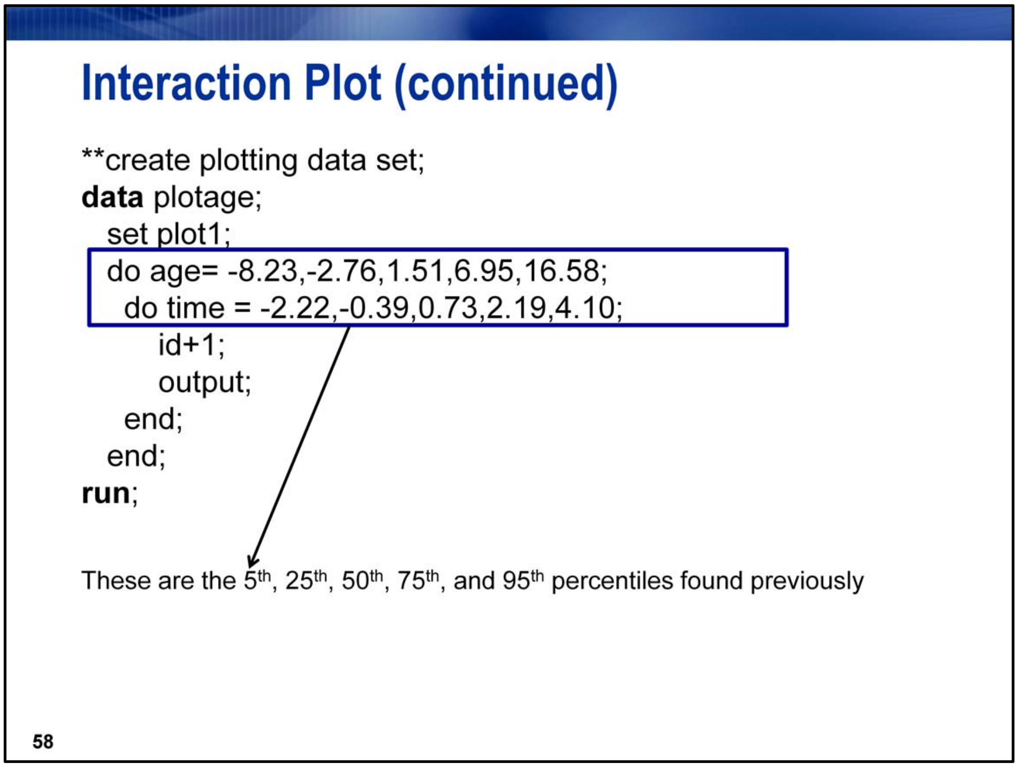


Interaction Plot: Time by Age

```
**get plotting points;

**percentiles for time and age: 5th, 25th, 50th, 75th, and 95th;
proc univariate data=aids;
  var time age;
run;

**medians for cigarettes and depression;
proc univariate data=aids noprint;
  var cigarettes depression;
  output out=plot1 median= cigarettes depression;
run;
```

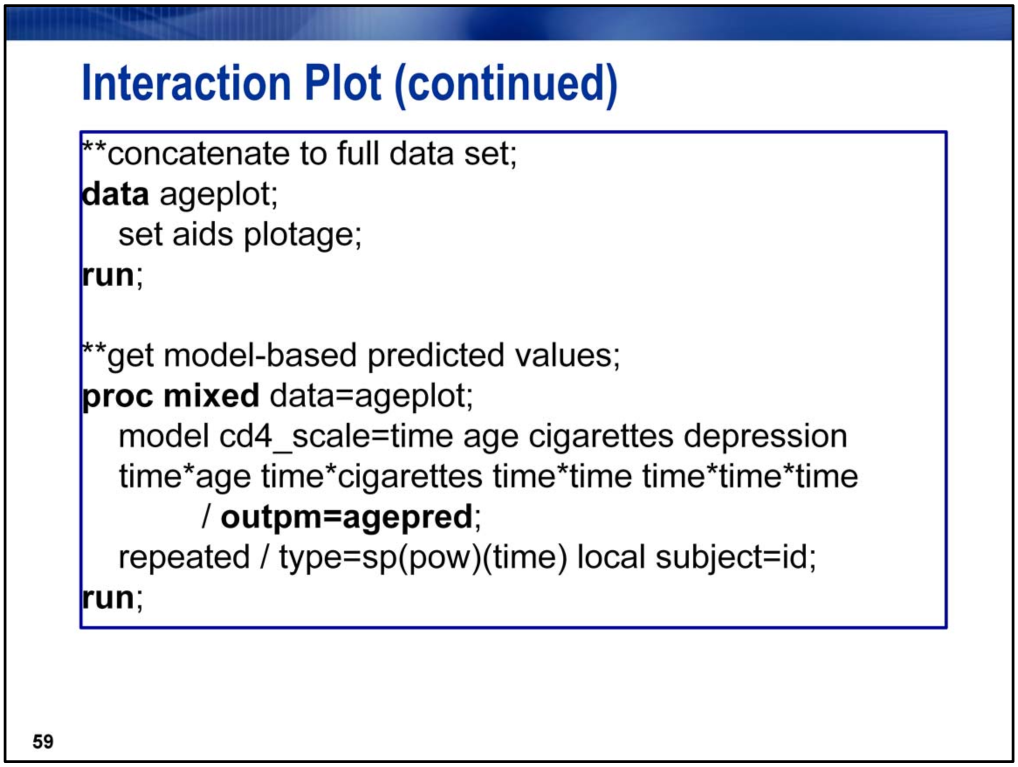


Interaction Plot (continued)

```
**create plotting data set;
data plotage;
  set plot1;
  do age= -8.23,-2.76,1.51,6.95,16.58;
    do time = -2.22,-0.39,0.73,2.19,4.10;
        id+1;
        output;
    end;
  end;
run;
```

These are the 5th, 25th, 50th, 75th, and 95th percentiles found previously

58



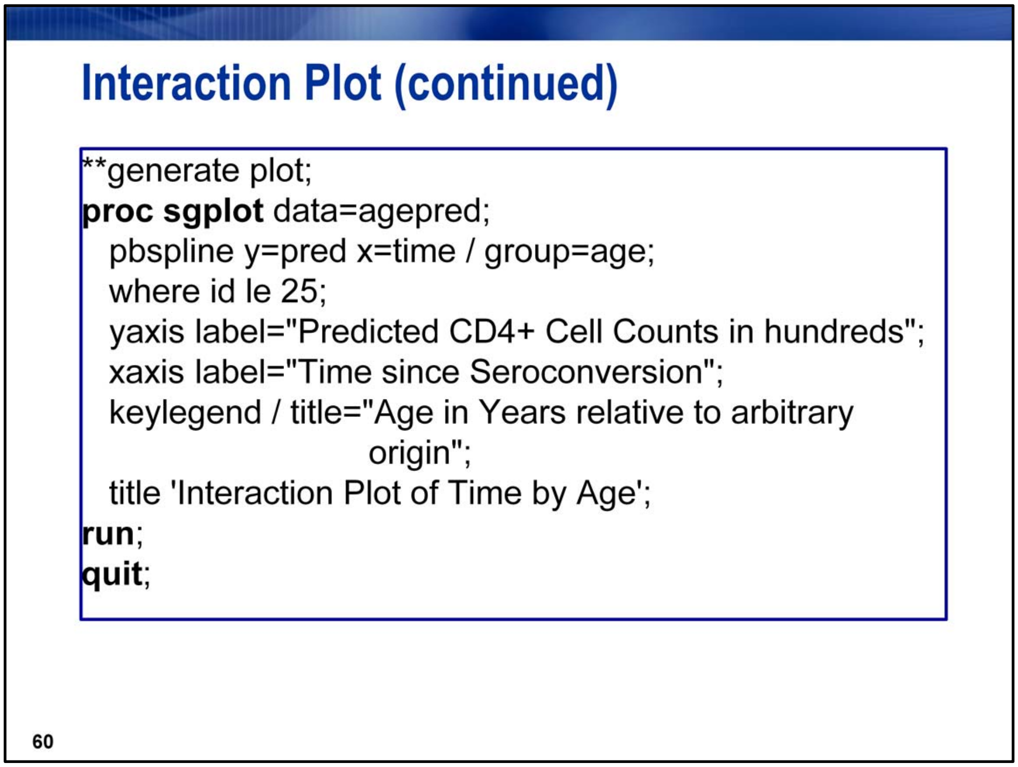
Interaction Plot (continued)

```
**concatenate to full data set;
data ageplot;
   set aids plotage;
run;

**get model-based predicted values;
proc mixed data=ageplot;
   model cd4_scale=time age cigarettes depression
   time*age time*cigarettes time*time time*time*time
          / outpm=agepred;
   repeated / type=sp(pow)(time) local subject=id;
run;
```

59

The OUTPM option creates population-level estimates. Because we are only interested in predicted values and not their standard errors, omitting the DDFM=KR option reduces the run time for the program.



Interaction Plot (continued)

```
**generate plot;
proc sgplot data=agepred;
   pbspline y=pred x=time / group=age;
   where id le 25;
   yaxis label="Predicted CD4+ Cell Counts in hundreds";
   xaxis label="Time since Seroconversion";
   keylegend / title="Age in Years relative to arbitrary
                              origin";
   title 'Interaction Plot of Time by Age';
run;
quit;
```

The WHERE statement restricts the plot to the 25 plotting points.