

Example 1: Relationship between SES and Educational Achievement

First a peek at the data:

```
proc print data=BASUG.hsb (obs=50);  
  var school_id student_id math cses sector;  
run;
```

Obs	school_ ID	student_ ID	math	cses	sector
1	1224	1	5.876	-1.09362	0
2	1224	2	19.708	-0.15362	0
3	1224	3	20.349	-0.09362	0
4	1224	4	8.781	-0.23362	0
5	1224	5	17.898	0.27638	0
6	1224	6	4.583	0.45638	0
7	1224	7	-2.832	-0.18362	0
:	:	:	:	:	:
45	1224	45	23.584	1.18638	0
46	1224	46	14.053	0.44638	0
47	1224	47	2.183	0.01638	0
48	1288	48	7.857	-0.90960	0
49	1288	49	10.171	-0.44960	0
50	1288	50	15.699	0.35040	0

Now fitting the model:

```
proc mixed data=BASUG.hsb noclprint;  
  class school_id;  
  model math = sector cses sector*cses / notest solution ddfm=bw  
    outpredm=predfix outpred=predrand;  
  random intercept cses / subject=school_id type=un g;  
run;
```

Notes on Syntax:

- PROC MIXED statement:
 - NOCLPRINT tells SAS not to print the levels of all variables declared on the CLASS statement
 - COVTEST tells SAS to print standard errors and Wald tests for variance-covariance parameter estimates.
- CLASS statement:
 - School_ID is an ID variable indicating the school the student attends. This ID variable need not be declared on the CLASS statement *if* the data is sorted by the school ID in advance of running the MIXED procedure.
- MODEL statement:
 - General form: criterion = predictors with fixed effects. Enter all fixed effects here.
 - NOTEST suppresses F tests, which are redundant with t-tests obtained with SOLUTION option. The latter option also requests tabled estimates for the fixed effects.
 - DDFM indicates the method used to compute degrees of freedom for *F*- and *t*-tests of fixed effects. Given the large sample size, this is largely irrelevant here, so the computationally simple Between and Within method is used, which is based on analogy to df computation for repeated measures ANOVA.

- OUTPREDM generates marginal predicted values for the observations based on the fixed effect estimates only. OUTPRED generates conditional (school-specific) predicted values for the observations based on estimates of both the fixed and random effects.
- RANDOM statement:
 - The random intercept is included by specifying the keyword “intercept” (not a variable in the dataset) and the random effect of cses is included by specifying “cses”.
 - SUBJECT=School_ID indicates that the random effects vary over schools.
 - TYPE=UN indicates that all elements of the covariance matrix of random effects are to be estimated (default is to estimate only variances).
 - G option requests that SAS print the variance-covariance matrix of the random effects (in matrix form, as opposed to list form).

Abridged Results:

Estimated G Matrix					
Row	Effect	school_ID	Col1	Col2	
1	Intercept	1224	6.7288	1.0489	
2	cses	1224	1.0489	0.2657	
Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	school_ID	6.7288	0.8634	7.79	<.0001
UN(2,1)	school_ID	1.0489	0.3416	3.07	0.0021
UN(2,2)	school_ID	0.2657	0.2288	1.16	0.1228
Residual		36.7066	0.6258	58.66	<.0001
Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	11.3939	0.2926	158	38.94	<.0001
sector	2.8075	0.4389	158	6.40	<.0001
cses	2.8028	0.1550	7023	18.09	<.0001
sector*cses	-1.3411	0.2338	7023	-5.74	<.0001

Notes on results...

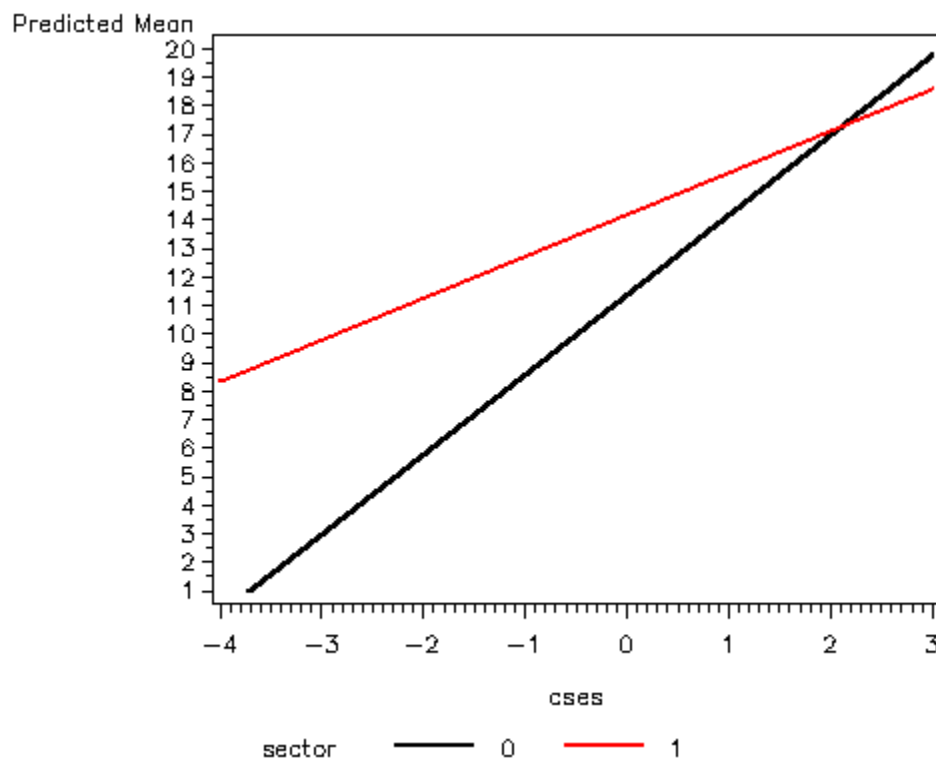
- Fixed Effects:
 - Expected math score for a child of average SES in a public school is 11.39.
 - Expected math score for a child of average SES in a private school is 2.81 units higher.
 - The expected gap in math achievement between two public school students differing by one point in SES is 2.80 units.
 - The expected gap in math achievement between two private school students differing by one point in SES is 1.34 units less than that, or 1.46 (i.e., 2.80 - 1.34)

➤ Random Effects:

- The relatively large variance of the random intercept indicates that, even after controlling for sector, there are differences across schools in the average levels of math achievement of their students.
- The small variance of the random slope indicates that, after controlling for sector differences in the effect of SES, there are few remaining differences in the effect of SES across schools.

To produce a plot of the interaction between sector and cses in the fixed effects, I used OUTPREDM to generate predicted values based solely on the fixed effects for all individuals. These are then plotted here:

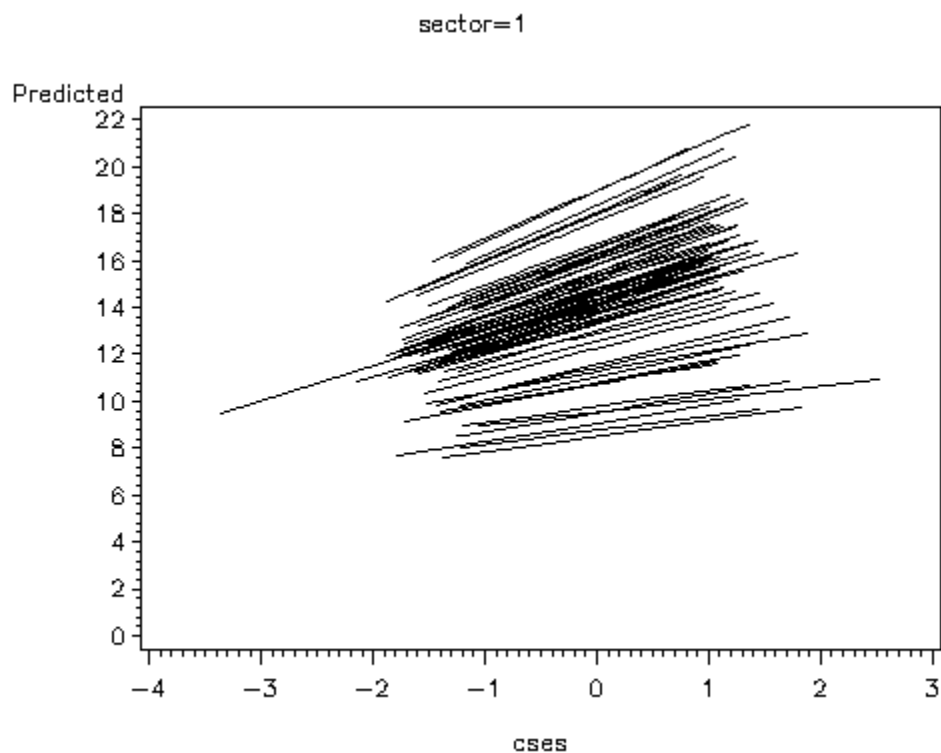
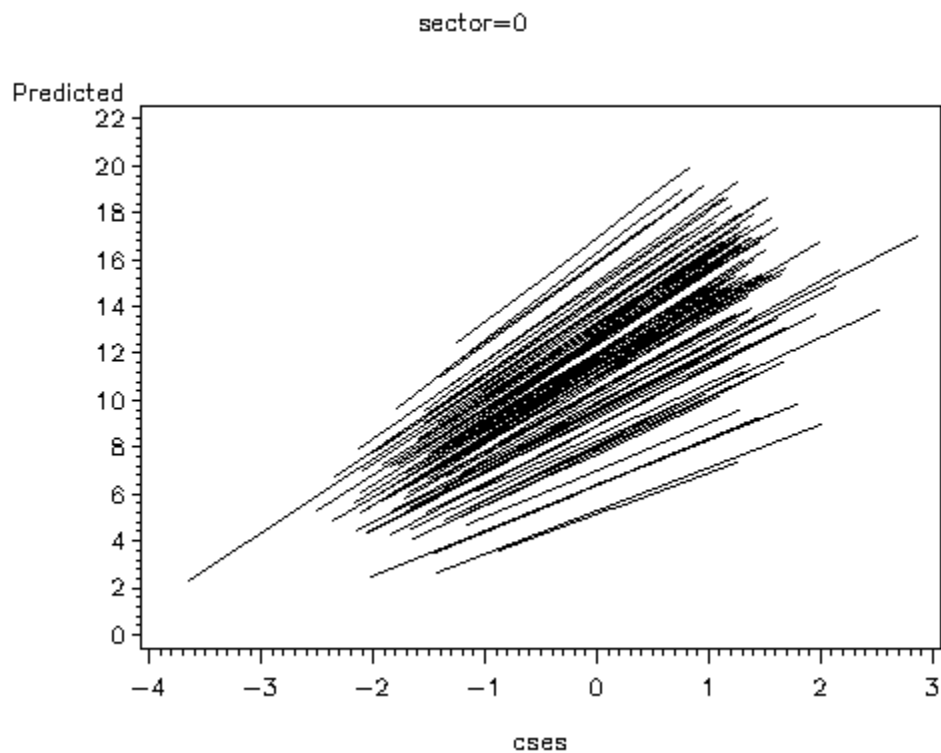
```
symbol value=none i=rl w=2;  
proc gplot data=predfix;  
  plot pred*cses=sector;  
run; quit;
```



Note the larger effect of SES within public schools, producing the interaction.

To clarify the magnitude of the random effects as well, I used OUTPRED to generate predicted values based on *both* the fixed effects and random effects for all individuals (i.e., school-specific predicted values). These are then plotted here:

```
proc sort data=predrand; by sector school_id cses; run;  
symbol value=none i=join color=black w=1 repeat=999;  
axis1 order= 0 to 22 by 2;  
proc gplot data=predrand; by sector;  
  plot pred*cses=school_id /nolegend vaxis=axis1;  
run; quit;
```



Note that in addition to observing the molar sector differences, we also see here the within-sector variability in level produced by the random intercept term. Variability in the regression slopes is less pronounced. The positive covariance between the random intercepts and slopes within sectors is also observed in the steeper slopes of lines within higher intercepts.

Example 2: Race Differences in Cardiovascular Health

First a peek at the data:

```
proc print data=BASUG.HR(obs=100);
  var id time wake timebed sys dia hr black pd bmi male age;
run;
```

	t	i	m	b	M	A
I	t	W	e	S	D	a
D	m	K	e	Y	I	H
e	E	d	d	S	A	R
1.00	8.4833	1	23.1833	108	76	79
1.00	8.7333	1	23.1833	114	66	99
1.00	9.1333	1	23.1833	110	66	89
:	:	:	:	:	:	:
1.00	23.1833	1	23.1833	113	70	74
1.00	24.1000	0	23.1833	91	49	74
1.00	25.1167	0	23.1833	90	40	80
:	:	:	:	:	:	:
1.00	30.1667	0	23.1833	90	60	55
2.00	8.7167	1	13.8000	101	57	80
2.00	9.6833	1	13.8000	110	64	80
2.00	10.7667	1	13.8000	125	56	101
:	:	:	:	:	:	:
2.00	13.8000	1	13.8000	103	54	93
3.00	9.4333	1	23.0667	112	79	81
3.00	9.7833	1	23.0667	108	59	89
3.00	10.0333	1	23.0667	121	73	88
:	:	:	:	:	:	:
3.00	23.0667	1	23.0667	124	76	84
3.00	24.1500	0	23.0667	83	52	68
:	:	:	:	:	:	:
3.00	27.1167	0	23.0667	104	59	77

Note that PD, BMI & Age have been mean-centered.

Given the many repeated measures available per person, the relatively small degree of measurement error, and the modest number of people, a good way to “get to know” the data is by plotting HR readings against time for every case. Here I shall augment these plots by superimposing subject-specific predicted values for a 2-part linear trajectory (obtained by conducting case-wise OLS regression). To generate the predicted values, I must first create two time variables, one for the waking period and another for the sleeping period.

```
data HR; set BASUG.HR;
  if wake = 1 then timewake = time-timebed;
  else timewake = 0;
  if wake = 0 then timesleep = time-timebed;
  else timesleep = 0;
run;

proc print data=HR (obs=100);
  var id time wake timebed timewake timesleep;
run;
```

The new time variables look like this:

ID	time	WAKE	timebed	timewake	timesleep
1.00	8.4833	1	23.1833	-14.7000	0.00000
1.00	8.7333	1	23.1833	-14.4500	0.00000
1.00	9.1333	1	23.1833	-14.0500	0.00000
:	:	:	:	:	:
1.00	23.1833	1	23.1833	0.0000	0.00000
1.00	24.1000	0	23.1833	0.0000	0.91667
1.00	25.1167	0	23.1833	0.0000	1.93333
:	:	:	:	:	:
1.00	30.1667	0	23.1833	0.0000	6.98333
2.00	8.7167	1	13.8000	-5.0833	0.00000
2.00	9.6833	1	13.8000	-4.1167	0.00000
2.00	10.7667	1	13.8000	-3.0333	0.00000
:	:	:	:	:	:
2.00	13.8000	1	13.8000	0.0000	0.00000

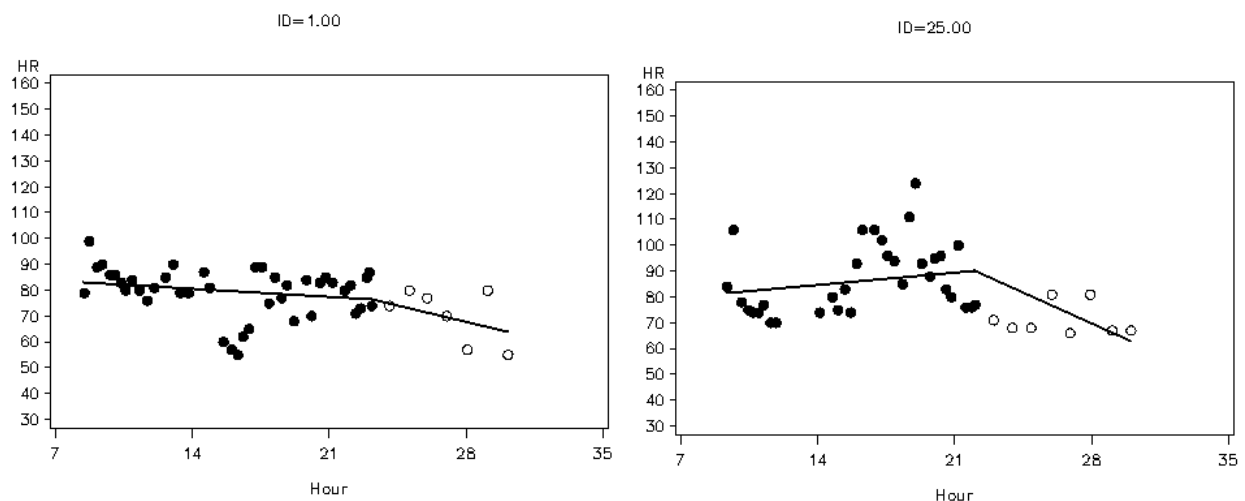
The following code generates the individual trajectory plots:

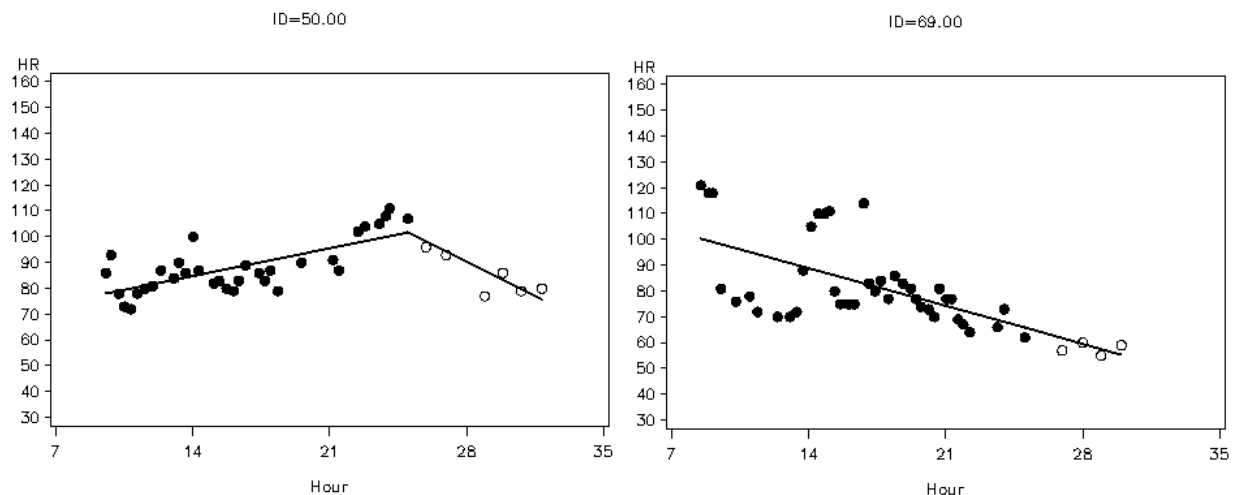
```
proc reg data=HR noprint; by id;
  model HR = timewake timesleep;
  output out=pred p=y_hat;
run;

data pred; set pred;
  if wake = 1 then wakeHR = HR; else wakeHR = .;
  if wake = 0 then sleepHR = HR; else sleepHR = .;
run;

symbol1 value=dot color=black;
symbol2 value=circle color=black;
symbol3 value=none interpol=join color=black h=2 w=2;
axis1 label=("HR") order=(30 to 160 by 10) minor=none;
axis2 label=("Hour") order=(7 to 35 by 7) minor=none;
proc gplot data=pred; by id;
  plot (wakeHR sleepHR y_hat)*time / overlay vaxis=axis1 haxis=axis2
  nolegend;
run; quit;
```

Here are a handful of the plots:





A few observations based on the plots:

- A two-part linear model appears to fit the observed data for most individuals quite well.
- There are apparent individual differences in level and rate of change in both waking and sleeping periods.
 - We will capture these differences via a random intercept and random slopes for TimeWake and TimeSleep.
- There appears to be some oscillation of the residuals about the individual-specific regression lines, suggesting the need for an autoregressive covariance structure on the residuals.

We will begin our formal analysis by fitting an initial model with the proposed random effects and autoregressive residual structure (Model 1).

```
proc mixed data=HR method=reml covtest noclprint;
  class id;
  model HR = timewake timesleep / notest solution ddfm=kr outpred=predicted;
  random intercept timewake timesleep / subject=id type=un gcorr g;
  repeated / type=sp(pow)(time) subject=id rcorr;
run;
```

Notes on syntax:

- MODEL statement:
 - DDFM=KR indicates to use the Kenward-Rodgers method of testing the fixed effects, which is superior to other methods when the number of independent sampling units is small.
- RANDOM statement:
 - GCORR option requests that SAS print the correlation matrix of the random effects
- REPEATED statement is used to impose a covariance structure on the Level 1 residuals.
 - In the absence of this statement, it is assumed that the Level 1 residuals are iid normal.
 - Here, the residuals are allowed to correlate according to a continuous-time autoregressive structure by using TYPE=SP(POW)(TIME) with SUBJECT=ID.
 - RCORR requests that SAS print the correlation matrix of the Level 1 residuals for the first level of the ID variable.

Some abridged results:

Estimated R Correlation Matrix for ID 1.00							
Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7
1	1.0000	0.4722	0.1421	0.06385	0.01922	0.009075	0.003688
2	0.4722	1.0000	0.3010	0.1352	0.04070	0.01922	0.007811
3	0.1421	0.3010	1.0000	0.4492	0.1352	0.06385	0.02595
4	0.06385	0.1352	0.4492	1.0000	0.3010	0.1421	0.05777
5	0.01922	0.04070	0.1352	0.3010	1.0000	0.4722	0.1919
6	0.009075	0.01922	0.06385	0.1421	0.4722	1.0000	0.4064
7	0.003688	0.007811	0.02595	0.05777	0.1919	0.4064	1.0000
8	0.001742	0.003688	0.01225	0.02728	0.09062	0.1919	0.4722
9	0.000640	0.001356	0.004505	0.01003	0.03332	0.07056	0.1736
10	0.000213	0.000451	0.001499	0.003337	0.01109	0.02348	0.05777
11	0.000064	0.000136	0.000451	0.001005	0.003337	0.007067	0.01739
12	0.000021	0.000045	0.000150	0.000334	0.001110	0.002351	0.005785
13	3.705E-6	7.846E-6	0.000026	0.000058	0.000193	0.000408	0.001005
14	1.172E-6	2.483E-6	8.249E-6	0.000018	0.000061	0.000129	0.000318

Estimated G Matrix					
Row	Effect	ID	Col1	Col2	Col3
1	Intercept	1.00	108.90	3.7943	-6.9791
2	timewake	1.00	3.7943	0.4442	-0.2715
3	timesleep	1.00	-6.9791	-0.2715	1.3235

Estimated G Correlation Matrix					
Row	Effect	ID	Col1	Col2	Col3
1	Intercept	1.00	1.0000	0.5455	-0.5813
2	timewake	1.00	0.5455	1.0000	-0.3541
3	timesleep	1.00	-0.5813	-0.3541	1.0000

Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	ID	108.90	24.6758	4.41	<.0001
UN(2,1)	ID	3.7943	1.5794	2.40	0.0163
UN(2,2)	ID	0.4442	0.1432	3.10	0.0010
UN(3,1)	ID	-6.9791	3.5477	-1.97	0.0492
UN(3,2)	ID	-0.2715	0.2397	-1.13	0.2573
UN(3,3)	ID	1.3235	0.7334	1.80	0.0356
SP(POW)	ID	0.04972	0.009577	5.19	<.0001
Residual		138.66	4.8951	28.33	<.0001

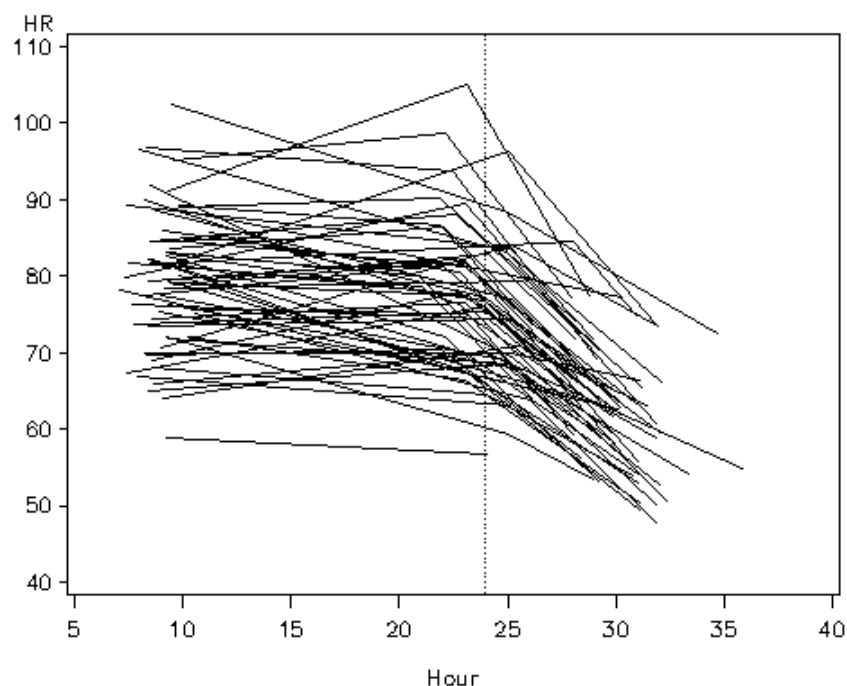
Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	76.9416	1.4751	57.9	52.16	<.0001
timewake	-0.1697	0.1134	56.1	-1.50	0.1401
timesleep	-2.3735	0.2586	40.4	-9.18	<.0001

Notes on results...

- Fixed Effects:
 - Across individuals, the expected HR reading at the last waking assessment is 77 BPM.
 - Across individuals, the mean trend for the waking period is flat.
 - Across individuals, the mean trend for the sleeping period indicates a decrease in systolic BP of 2.4 BPM per hour.
- Random Effects:
 - The relatively large variance of the random intercept indicates that there are individual differences in overall HR levels.
 - The variances of the random effects of timewake and timesleep also indicate that there are significant individual differences in rates of change in HR across waking and sleeping periods.
 - GCORR output indicates that those individuals with high HR levels on the last waking assessment (high intercepts) experienced the greatest increases in HR over the day (high timewake slopes) and show the greatest decreases in HR over the night (low timesleep slopes).
- Residuals:
 - The autoregressive (spatial power) structure implies that observations close together in time will have positively correlated residuals and that this correlation will diminish with the interval of time between observations. This is seen in the RCORR matrix.

The results for the fixed and random effects can also be observed graphically by plotting the subject-specific predicted values from the model:

```
symbol1 value=none interpol=join color=black repeat=80;  
axis1 label=("HR") order=(40 to 110 by 10) minor=none;  
axis2 label=("Hour") order=(5 to 40 by 5) minor=none;  
proc gplot data=predicted; where ~missing(HR);  
  plot pred*time=id / nolegend vaxis=axis1 haxis=axis2 href=24 lhref=33;  
run; quit;
```



Model 2 examines possible racial differences in cardiovascular health by adding the variable Black as a predictor of each of the random effects, controlling for age, sex, and BMI.

```
proc mixed data=HR method=reml noclprint covtest;
class id;
model hr = timewake timesleep black timewake*black timesleep*black
        bmi male age/ notest solution ddfm=kr;
random intercept timewake timesleep / subject=id type=un gcorr g;
repeated / type=sp(pow)(time) subject=id;
ods output solutionf=fixedeffects;
run;
```

Notes on syntax:

- ODS statement:
 - OUTPUT SOLUTIONF=fixedeffects tells SAS to put the table of fixed effect estimates generated by the SOLUTION option of the MODEL statement into a data file called “fixedeffects” for subsequent processing.

Abridged output:

Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	ID	89.0133	21.8227	4.08	<.0001
UN(2,1)	ID	3.3799	1.4831	2.28	0.0227
UN(2,2)	ID	0.4349	0.1433	3.03	0.0012
UN(3,1)	ID	-7.2335	3.4792	-2.08	0.0376
UN(3,2)	ID	-0.2802	0.2428	-1.15	0.2484
UN(3,3)	ID	1.4084	0.7667	1.84	0.0331
SP(POW)	ID	0.04980	0.009582	5.20	<.0001
Residual		138.67	4.8948	28.33	<.0001

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	75.1896	2.2045	63.5	34.11	<.0001
timewake	-0.2935	0.1587	53	-1.85	0.0699
timesleep	-2.4579	0.3608	34.4	-6.81	<.0001
black	5.9753	2.9038	59	2.06	0.0440
timewake*black	0.2234	0.2256	54.7	0.99	0.3264
timesleep*black	0.2846	0.5224	40.1	0.54	0.5889
bmi	0.05699	0.2508	53.3	0.23	0.8211
MALE	-4.2571	2.2982	55.2	-1.85	0.0693
AGE	0.1516	0.1163	56.4	1.30	0.1974

Notes on results...

- Fixed Effects:
 - Black participants have an expected HR that is 6 BPM higher than white participants, controlling for BMI, age and sex.
 - There are no significant differences between black and white participants in rates of change in HR over the day and night.
 - Males have an expected HR that is 4 BPM lower than females, controlling for BMI, age and race.

To examine race differences in greater detail, I will plot the expected trajectories for black and white participants controlling for the covariates. In this case, I cannot use OUTPREDM to generate the predicted values, as this will not hold the other covariates (age, BMI, Male) constant. As a more flexible alternative, I have output the fixed effects using the ODS statement to a data file that looks like this:

```
proc print data=fixedeffects; run;
```

Effect	Estimate	StdErr	DF	tValue	Probt
Intercept	75.1896	2.2045	63.5	34.11	<.0001
timewake	-0.2935	0.1587	53	-1.85	0.0699
timesleep	-2.4579	0.3608	34.4	-6.81	<.0001
black	5.9753	2.9038	59	2.06	0.0440
timewake*black	0.2234	0.2256	54.7	0.99	0.3264
timesleep*black	0.2846	0.5224	40.1	0.54	0.5889
bmi	0.05699	0.2508	53.3	0.23	0.8211
MALE	-4.2571	2.2982	55.2	-1.85	0.0693
AGE	0.1516	0.1163	56.4	1.30	0.1974

To make this data easier to work with, I will flip it to one row as follows:

```
proc transpose data=fixedeffects (keep=estimate effect) out=fixedT;
  var estimate;
  id effect;
run;
proc print data=fixedT; run;
```

NAME	Intercept	timewake	timesleep	black
Estimate	75.1896	-0.2935	-2.4579	5.9753
timewake_	timesleep_			
black	black	bmi	MALE	AGE
0.2234	0.2846	0.05699	-4.2571	0.1516

I can now generate the predicted values within a data step, holding age, BMI and Male at 0 (= average age, average BMI, female), and plot them as follows:

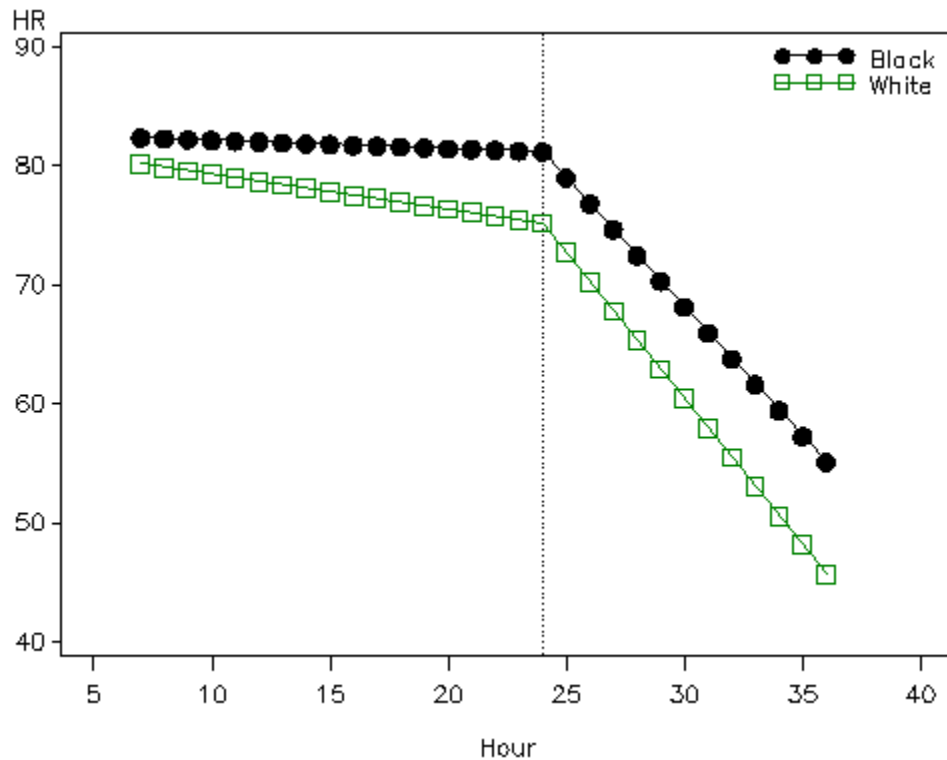
```
data effects; set fixedT;
do time = 7 to 36 by 1;
  if time < 24 then do; *average time to bed ~ midnight;
    twake = time - 24;
    tsleep = 0;
  end;
  else do;
    twake = 0;
    tsleep = time - 24;
  end;
  whitepred = intercept + timewake*twake + timesleep*tsleep;
  blackpred = whitepred + black + timewake_black*twake +
    timesleep_black*tsleep;

  output;
end;
run;
```

```

symbol1 value=dot i=join color=black w=1.25 h=1.25 repeat=1;
symbol2 value=square i=join color=green w=1.25 h=1.25 repeat=1;
axis1 label=('HR') minor=none order=40 to 90 by 10;
axis2 label=('Hour') order=5 to 40 by 5 offset=(2) minor=none;
legend1 position=(inside top right) down=3 label=none
      value=('Black' 'White');
proc gplot data=effects;
  plot (blackpred whitepred)*time /overlay legend=legend1 vaxis=axis1
      haxis=axis2 href=24 lhref=33;
run;

```



Recall that from the tests of fixed effects, only the difference in level apparent in this plot is statistically significant, not the differences in the slopes.

Might the overall higher HR observed for black participants be explainable on the basis of higher psychosocial stress due to perceptions of discrimination? Model 3 adds Perceived Discrimination as a predictor of each of the random effects:

```

proc mixed data=HR method=reml noclprint covtest;
  class id;
  model hr = timewake timesleep black timewake*black timesleep*black
           bmi male age pd timewake*pd timesleep*pd /
           notest solution ddfm=kr;
  random intercept timewake timesleep / subject=id type=un gcorr g;
  repeated / type=sp(pow)(time) subject=id;
  ods output solutionf=fixedeffects2;
run;

```

Abridged results:

Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	ID	89.0348	21.9124	4.06	<.0001
UN(2,1)	ID	3.4031	1.5028	2.26	0.0235
UN(2,2)	ID	0.4452	0.1468	3.03	0.0012
UN(3,1)	ID	-7.6078	3.3437	-2.28	0.0229
UN(3,2)	ID	-0.2677	0.2351	-1.14	0.2547
UN(3,3)	ID	1.0414	0.6883	1.51	0.0651
SP(POW)	ID	0.04984	0.009588	5.20	<.0001
Residual		138.78	4.8965	28.34	<.0001

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	76.8738	2.4078	65.6	31.93	<.0001
timewake	-0.2727	0.1693	51.4	-1.61	0.1134
timesleep	-2.2151	0.3506	25.7	-6.32	<.0001
black	3.3206	3.4125	62.4	0.97	0.3343
timewake*black	0.1813	0.2589	56.2	0.70	0.4865
timesleep*black	-0.2943	0.5568	36	-0.53	0.6003
bmi	0.2202	0.2484	53.7	0.89	0.3793
MALE	-5.4682	2.2825	54	-2.40	0.0201
AGE	0.2148	0.1171	58	1.83	0.0718
pd	3.0879	2.7910	59.7	1.11	0.2730
timewake*pd	0.07006	0.2189	56.2	0.32	0.7501
timesleep*pd	0.9659	0.4731	34.3	2.04	0.0489

Notes on results...

➤ Fixed Effects:

- The main effect of Black is diminished and no longer statistically significant after controlling for Perceived Discrimination, suggesting that race differences in cardiovascular health may well be of psychosocial origin.
- There is a statistically significant timesleep*pd interaction, indicating that higher levels of PD are associated with *less steep decreases* in HR during sleep.

We will again graph these results via the estimates saved with the ODS statement:

```
proc transpose data=fixedeffects2(keep=estimate effect) out=fixedT2;
  var estimate;
  id effect;
run;

*generating predicted values for black and for pd given fixed effect
estimates holding all other covariates constant;
data effects2; set fixedT2;
do time = 7 to 36 by 1;
  if time < 24 then do; *average time to bed ~ midnight;
    twake = time - 24;
    tsleep = 0;
  end;
  else do;
    twake = 0;
    tsleep = time - 24;
  end;
end;
```

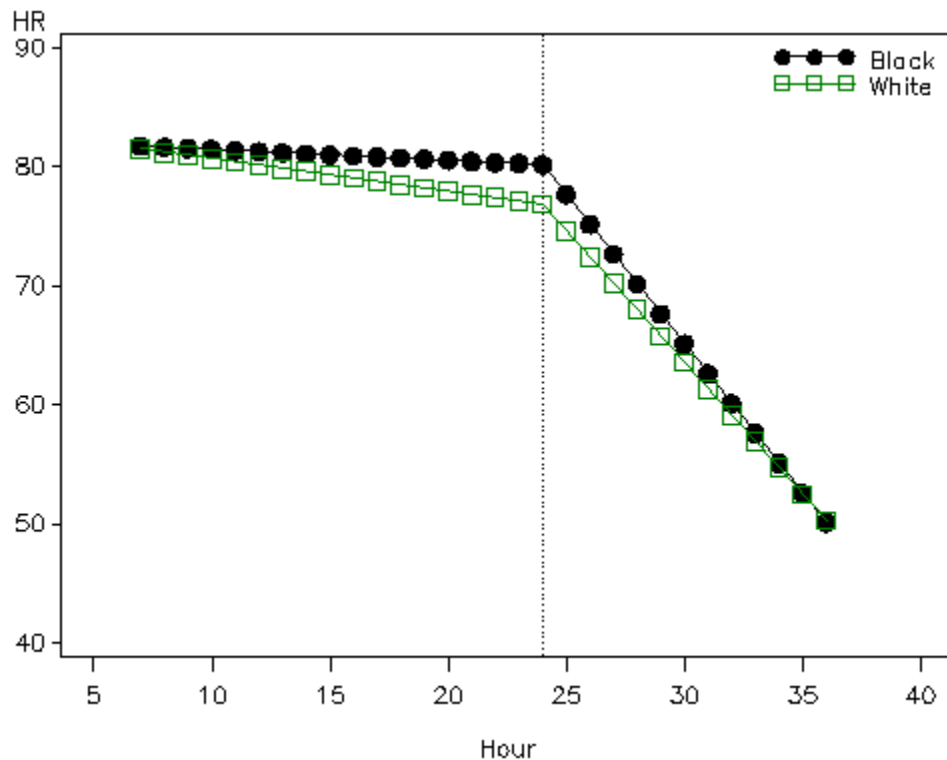
```

whitepred = intercept + timewake*twake + timesleep*tsleep;
blackpred = whitepred + black + timewake_black*twake +
            timesleep_black*tsleep;
* .59 is SD of PD, computing predicted values at +/- 1 SD PD for black
  participants;
pdhipred = blackpred + pd*.59 + timewake_pd*twake*.59 +
            timesleep_pd*tsleep*.59;
pdlopred = blackpred - pd*.59 - timewake_pd*twake*.59 -
            timesleep_pd*tsleep*.59;

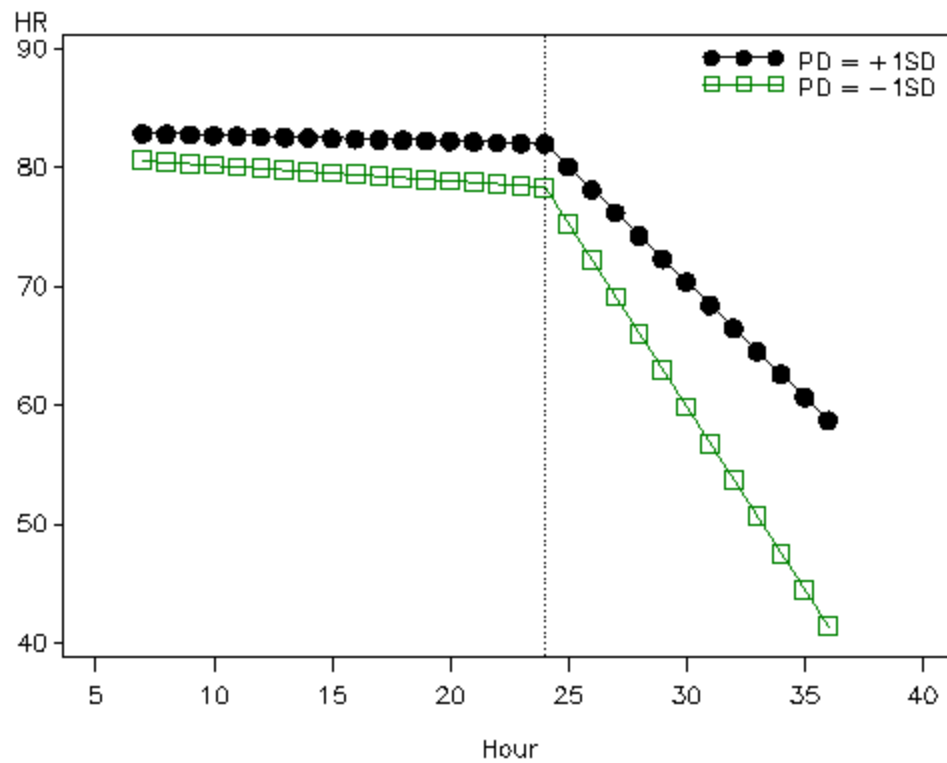
output;
end;
run;

symbol1 value=dot i=join color=black w=1.25 h=1.25;
symbol2 value=square i=join color=green w=1.25 h=1.25;
axis1 label=('HR') minor=none order=40 to 90 by 10;
axis2 label=('Hour') order=5 to 40 by 5 offset=(2) minor=none;
legend1 position=(inside top right) down=2 label=none
      value=('Black' 'White');
legend2 position=(inside top right) down=2 label=none
      value=('PD = +1SD' 'PD = -1SD');
proc gplot data=effects2;
  plot (blackpred whitepred)*time /overlay legend=legend1 vaxis=axis1
      haxis=axis2 href=24 lhref=33;
run;
proc gplot data=effects2;
  plot (pdhipred pdlopred)*time /overlay legend=legend2 vaxis=axis1
      haxis=axis2 href=24 lhref=33;
run;

```



Notice that the Black/White difference is nearly zero (and *ns*), when controlling for PD



Notice that individuals with high levels of perceived discrimination experience a less steep drop in HR during the night, as compared to individuals with low levels of perceived discrimination.

Example 3: Individual Differences in Vocabulary Development

First a peek at the data:

```
proc print data=BASUG.Vocab (obs=20);
run;
```

id	age	vocab	momspeak	female	group	lnmomspeak
0001	12	1	694	0	0	-1.45753
0001	16	13	694	0	0	-1.45753
0001	18	34	694	0	0	-1.45753
0001	20	47	694	0	0	-1.45753
0001	22	90	694	0	0	-1.45753
0001	24	139	694	0	0	-1.45753
0001	26	205	694	0	0	-1.45753
0002	12	1	4485	0	0	0.40849
0002	14	6	4485	0	0	0.40849
0002	16	46	4485	0	0	0.40849
0002	18	133	4485	0	0	0.40849
0002	20	188	4485	0	0	0.40849
0002	22	355	4485	0	0	0.40849
0002	24	449	4485	0	0	0.40849
0002	26	505	4485	0	0	0.40849
0003	12	1	1445	0	0	-0.72414
0003	16	10	1445	0	0	-0.72414
0003	18	26	1445	0	0	-0.72414
0003	20	43	1445	0	0	-0.72414
0003	22	77	1445	0	0	-0.72414

Note the lnmomspeak is the natural log of momspeak, mean centered after logging.

Now tabulating available sample size at each age by group (=study):

```
proc freq data=BASUG.Vocab;
table age*group / nocol norow nopercnt;
run;
```

age	group		
Frequency	0	1	Total
12	11	11	22
14	5	0	5
16	11	11	22
18	11	0	11
20	11	11	22
22	11	0	11
24	11	11	22
26	11	0	11
Total	82	44	126

Note the difference in assessment schedules between the two studies and the small number of observations at 14 months of age.

Performing some data manipulation for subsequent graphical and statistical analyses:

```
data Vocab; set BASUG.Vocab;
*coding up offset term;
if group = 0 then hours = 5;
if group = 1 then hours = 3;
lnhours = log(hours);
*centering age at 12 months;
agecent = age-12;
*median split on momspeak -- for plots ONLY;
if momspeak > 3103 then momspeakhi = 1;
else momspeakhi=0;
run;
```

The two studies, coded group=0 and group=1, involved 5 versus 3 hours of observations, respectively. Since we are assuming a Poisson process, this can be considered to be differential 'exposure' across the two studies. This difference can be accommodated by including an offset term, $\log(\text{hours})$, in our models so that we will be modeling word production per hour. In other words, we will model the rate of vocabulary production rather than the frequency.

The following plot is used to evaluate the effectiveness of the offset term in eliminating study differences, and to aid in identifying the best function for modeling change in the Poisson rate parameter over time.

```
proc sort data=Vocab; by group age; run;

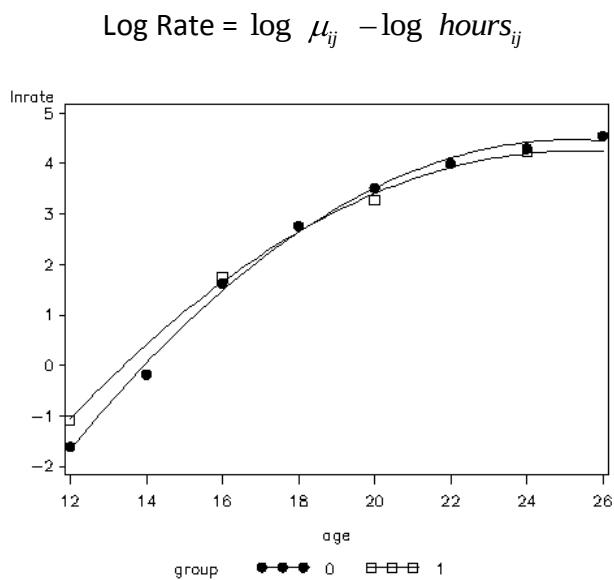
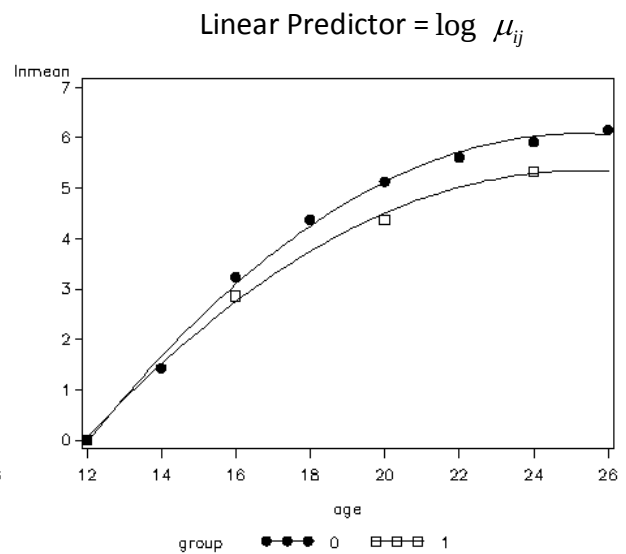
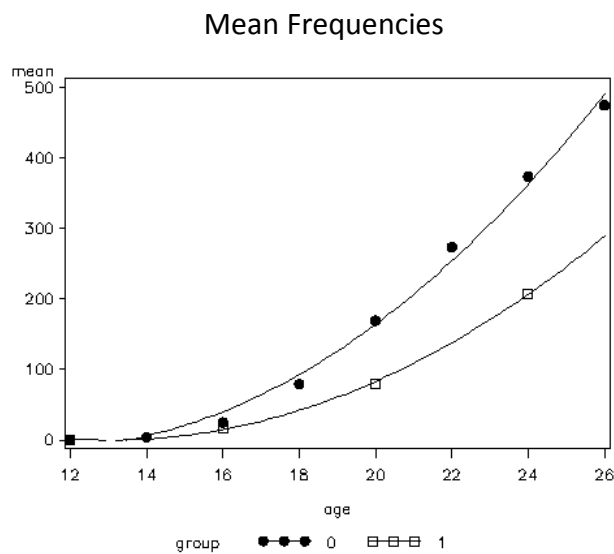
proc means data=Vocab noprint; by group age;
var vocab;
output out=means mean=mean;
run;

proc print data=means; run;

data means; set means;
if group = 0 then hours = 5;
if group = 1 then hours = 3;
lnmean = log(mean);
offset = log(hours);
lnrate = lnmean-offset;
drop _TYPE_ _FREQ_;
run;

symbol1 value=dot color=black i=rq;
symbol2 value=square color=black i=rq;
axis1 minor=none;
axis2 minor=none order=12 to 26 by 2;
proc gplot data=means;
plot (mean lnmean lnrate)*age=group / vaxis=axis1 haxis=axis2;
run; quit;
```

Obs	group	age	_TYPE_	_FREQ_	mean
1	0	12	0	11	1.000
2	0	14	0	5	4.200
3	0	16	0	11	25.455
4	0	18	0	11	79.727
5	0	20	0	11	169.545
6	0	22	0	11	273.545
7	0	24	0	11	373.455
8	0	26	0	11	474.636
9	1	12	0	11	1.000
10	1	16	0	11	17.545
11	1	20	0	11	80.455
12	1	24	0	11	207.455



Note that:

- Mean frequencies are discrepant over studies due to different observation periods and show *accelerating* increases with age
- In contrast, linear predictor shows *decelerating* increases with age.
- Controlling for differential 'exposure' with offset term effectively removes study differences.
- Developmental trends appear roughly quadratic in shape.

We shall now consider the potential effects of the covariates:

```
proc sort data=vocab; by group female age; run;
proc means data=vocab noprint; by group female age;
  var vocab;
  output out=means2 mean=mean;
run;

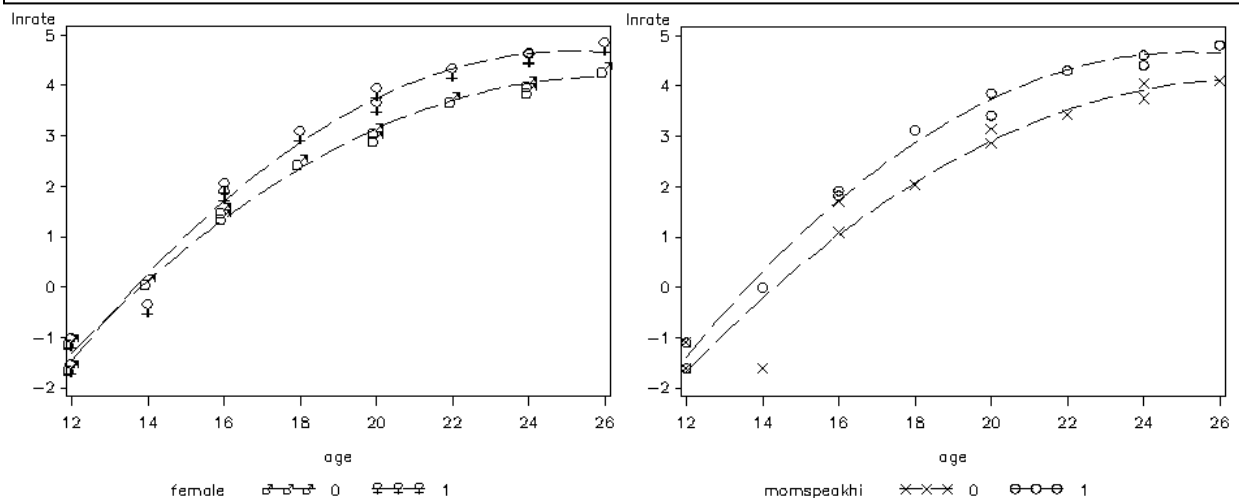
data means2; set means2;
  if group = 0 then hours = 5;
  if group = 1 then hours = 3;
  lnmean = log(mean);
  offset = log(hours);
  lnrate = lnmean-offset;
  drop _TYPE_ _FREQ_;
run;

symbol1 value=> color=black i=rq l=2 w=1.5 h=1.5;
symbol2 value=* color=black i=rq l=2 w=1.5 h=1.5;
axis1 minor=none;
axis2 minor=none order=12 to 26 by 2;
proc gplot data=means2;
  plot lnrate*age=female / vaxis=axis1 haxis=axis2;
run; quit;

proc sort data=vocab; by group momspeakhi age; run;
proc means data=vocab noprint; by group momspeakhi age;
  var vocab;
  output out=means3 mean=mean;
run;

data means3; set means3;
  if group = 0 then hours = 5;
  if group = 1 then hours = 3;
  lnmean = log(mean);
  offset = log(hours);
  lnrate = lnmean-offset;
  drop _TYPE_ _FREQ_;
run;

symbol1 value=x color=black i=rq l=2 w=1 h=1;
symbol2 value=circle color=black i=rq l=2 w=1 h=1;
axis1 minor=none;
axis2 minor=none order=12 to 26 by 2;
proc gplot data=means3;
  plot lnrate*age=momspeakhi / vaxis=axis1 haxis=axis2;
run; quit;
```



These plots suggest that main effects may be sufficient to model sex differences in the effect of language input. Additionally, these plots reinforce the earlier conclusions that the offset term is sufficient to remove study differences and support the idea of using a quadratic trajectory model. The discrepant observations at 14 months are based on a total N of 5, so may be discounted.

The first model fit to the data is designed only to evaluate the effectiveness of a quadratic trajectory model with random intercepts and linear slopes in reproducing the observed data.

```
proc glimmix data=Vocab;
class id;
model vocab = agecent agecent*agecent /
      solution offset=lnhours dist=poisson link=log ddfm=kr;
random intercept agecent /subject=id type=un;
random _residual_ *overdispersion;
output out=predicted pred(blup ilink)=predcount;
run;
```

Notes on syntax:

- The default method of estimation is RSPL, which is equivalent to PQL with REML. This can be changed with the METHOD option on the PROC GLIMMIX statement.
- Relative to the MIXED procedure, we have two new options in the MODEL statement, DIST and LINK, used here to indicate the Poisson distribution and Log link.
- The GLIMMIX procedure does not include a REPEATED statement. The same functionality is provided by the RANDOM statement with keyword `_RESIDUAL_`. Here this statement is used to tell SAS to estimate a multiplicative scale parameter to allow for overdispersion (conditional response variance > nominal variance of Poisson).
- In GLIMMIX, there are no OUTPRED or OUTPREDM options to the MODEL statement. These are subsumed by the OUTPUT statement. Among other things, the OUTPUT statement can be used to generate predicted values with (BLUP) or without (NOBLUP) using the random effects in the scale of linear predictor (default) or the mean (ILINK).

Abridged output:

Covariance Parameter Estimates				
Cov Parm	Subject	Estimate	Standard Error	
UN(1,1)	id	0.9722	0.3527	
UN(2,1)	id	-0.03345	0.01553	
UN(2,2)	id	0.001304	0.000774	
Residual (VC)		3.5685	0.5889	

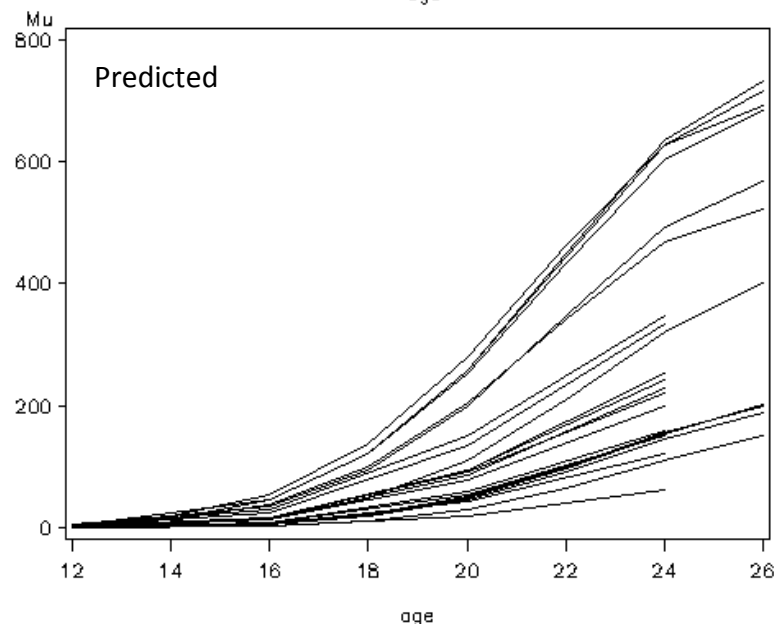
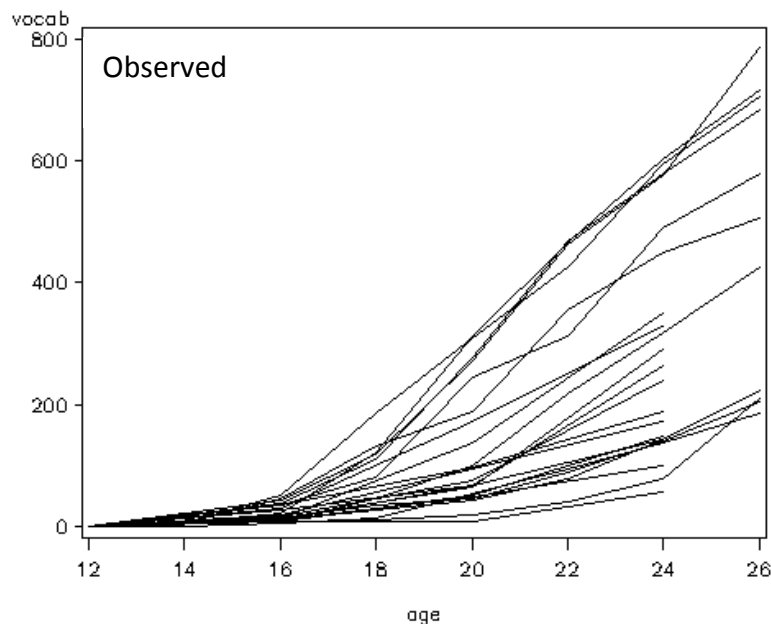
Solutions for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	-1.2621	0.2626	34.93	-4.81	<.0001
agecent	0.7613	0.03207	95.21	23.74	<.0001
agecent*agecent	-0.02575	0.001563	82.82	-16.48	<.0001

Notes on output:

- The random intercept variance estimate [UN(1,1)] indicates significant individual variability in the level of the linear predictor across individuals. The random slope variance estimate [UN(2,2)] is smaller, relative to its SE. Intercepts and slopes are negatively associated [UN(2,1)].
- The fixed effects indicates that the linear predictor increases quickly initially (large positive linear effect), with deceleration over time (negative quadratic effect).

To get a better sense of individual differences in change over time, and the quality of our model, we will plot and compare the observed and predicted individual trajectories:

```
symbol value=none color=black i=join l=1 repeat=22;  
axis1 order=12 to 26 by 2 minor=none;  
axis2 order=0 to 800 by 200 minor=none;  
proc gplot data=predicted;  
  plot (vocab predcount)*age=id/nolegend haxis=axis1 vaxis=axis2;  
run; quit;
```



Overall, the model appears to do a reasonably good job of reproducing the individual trajectories.

In the next model, the covariates are included to predict individual variation not due merely to differences in duration of observation (accounted for by the offset term). (Note that no age*covariate interactions were significant, and hence none are included in the model)

```
proc glimmix data=Vocab;
class id;
model vocab = agecent lnmonspeak female agecent*agecent /
      solution offset=lnhours dist=poisson link=log ddfm=kr;
random intercept agecent / subject=id type=un;
random _residual_;
ods output parameterestimates=fixedeffects;
run;
```

Abridged output:

Covariance Parameter Estimates				
Cov Parm	Subject	Estimate	Standard Error	
UN(1,1)	id	0.7622	0.2982	
UN(2,1)	id	-0.03106	0.01512	
UN(2,2)	id	0.001402	0.000852	
Residual (VC)		3.4973	0.5811	

Solutions for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	-1.3807	0.2571	40.82	-5.37	<.0001
agecent	0.7591	0.03189	96.07	23.80	<.0001
lnmonspeak	0.4456	0.1824	14.34	2.44	0.0280
female	0.3250	0.1997	11.81	1.63	0.1299
agecent*agecent	-0.02573	0.001550	82.21	-16.60	<.0001

Notes on output:

- We observe that the effect of Female is positive but non-significant when controlling for language input. This suggests that the female advantage observed in vocabulary production may be due to differences in the way mothers speak to female versus male infants.
- The effect of language input from the mother (lnmonspeak) is positive and significant indicating that infants who experience high levels of early language input show the greatest gains in vocabulary production over time.

These results can be difficult to interpret, given that they are in the scale of the linear predictor. The results can be understood more effectively via graphs of predicted values implied by the fixed effect estimates (saved via the ODS statement):

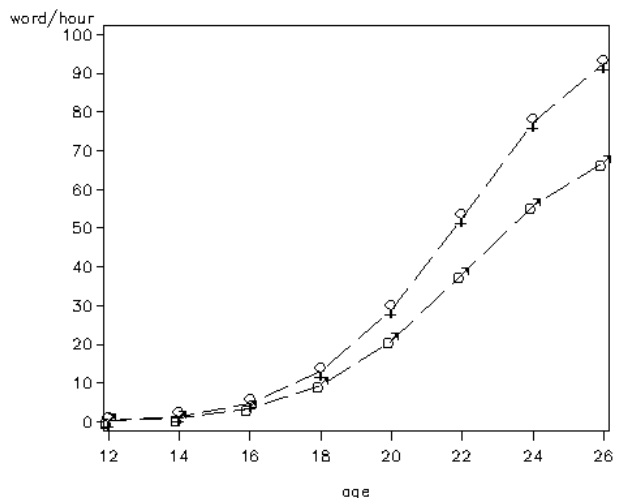
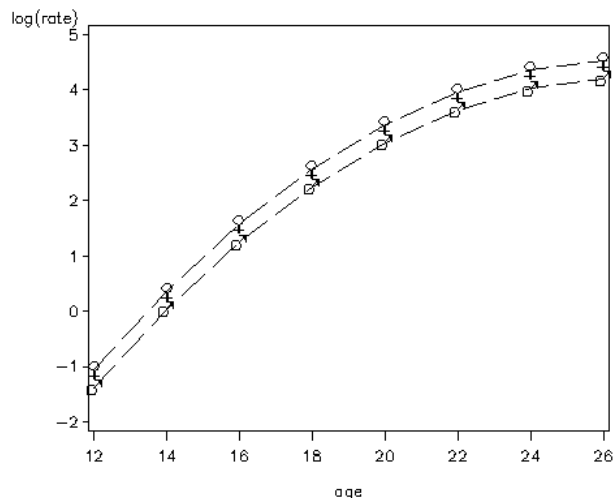
```

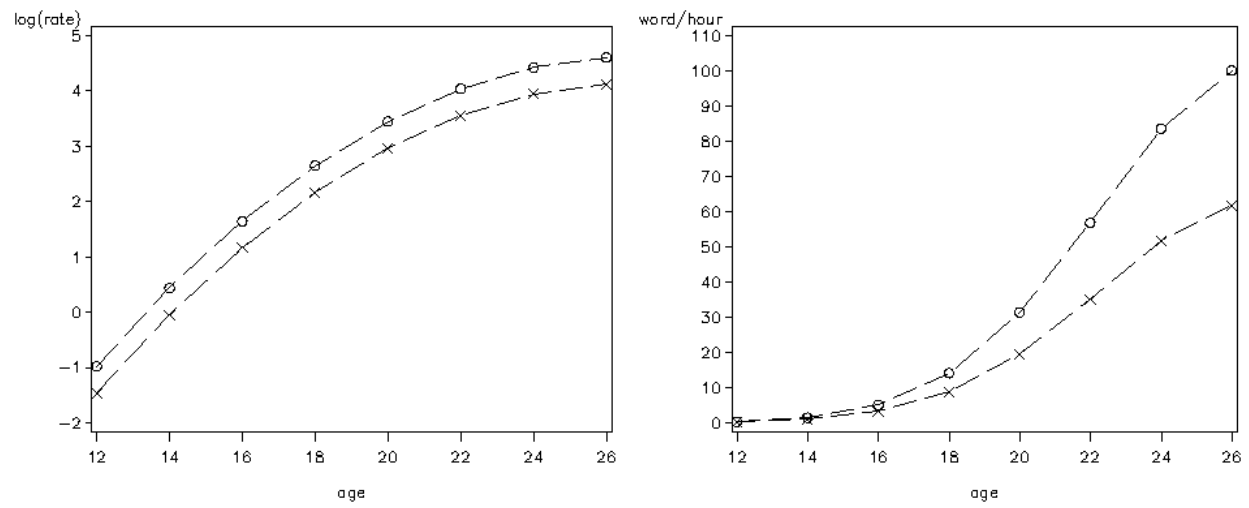
proc transpose data=fixedeffects(keep=estimate effect) out=fixedT;
  var estimate;
  id effect;
run;

data fixedpred; set fixedT;
  do age = 12 to 26 by 2;
    agec = age-12;
    *male and female holding lnmomspeak at mean, 1 hour;
    linpredmale = intercept + agecent*agec + agecent_agecent*(agec**2);
    linpredfemale = linpredmale + female;
    *high and low lnmomspeak (SD=.54), balanced on gender;
    linpredmomhi = linpredmale + .5*female + lnmomspeak*.54;
    linpredmomlo = linpredmale + .5*female - lnmomspeak*.54;
  array linpred[4] linpredmale linpredfemale linpredmomhi linpredmomlo;
  array pred[4] predmale predfemale predmomhi predmomlo;
  do i = 1 to 4;
    pred[i] = exp(linpred[i]); *inverse link;
  end;
  output;
end;
drop i;
run;

symbol1 value=> color=black i=join l=2 w=1.5 h=1.5 repeat=1;
symbol2 value=* color=black i=join l=2 w=1.5 h=1.5 repeat=1;
axis1 minor=none order=12 to 26 by 2;
axis2 minor=none label=("log(rate)");
axis3 minor=none label=("word/hour");
proc gplot data=fixedpred;
  plot (linpredmale linpredfemale)*age / overlay haxis=axis1 vaxis=axis2;
run; quit;
proc gplot data=fixedpred;
  plot (predmale predfemale)*age / overlay haxis=axis1 vaxis=axis3;
run; quit;
symbol1 value=x color=black i=join l=2 w=1 h=1 repeat=1;
symbol2 value=circle color=black i=join l=2 w=1 h=1 repeat=1;
proc gplot data=fixedpred;
  plot (linpredmomlo linpredmomhi)*age / overlay haxis=axis1 vaxis=axis2;
run; quit;
proc gplot data=fixedpred;
  plot (predmomlo predmomhi)*age / overlay haxis=axis1 vaxis=axis3;
run; quit;

```





These plots serve as a useful reminder that main effects in the log scale (left panels) are multiplicative after passing through the inverse link (right panels). Hence, *main effects* are sufficient to reproduce the differences in vocabulary *acceleration* identified in our initial graphical analyses.