2021

Advanced Statistical Modelling

Module number: DALT7009

Student Number: 19132761

MSc Course: Data Analytics

Word count: 1252

Table of Contents

1. Describe the statistical model and produce a profile plot

The statistical model

Medical researchers are interested in growth measurements for children. The growth measurements included dental measurements from the pituitary gland's center to the pterygomaxillary fissure for 11 girls and 16 boys at ages 8, 10, 12, and 14 years. The subjects are individual children, and there are four repeated measures on each. The data is stored in the SAS Growth data set. The variables included in the data set are people, Gender, Age, and growth rate.

Person	27 persons (1 to 27)
Gender	2 genders (boy and girl)
Age	4 age groups (8, 10, 12, 14 years old)
Growth	measure from the centre of the pituitary gland to the pterygomaxillary fissure

The data has a nested classification because Age is nested within Gender. Age and Gender are considered fixed effects because only four ages (8, 10, 12, and 14) and two genders (boy and girl) are used in the study, and we are only interested in making inference about these four age groups in girl and boy. Person is considered a random effect because they are randomly selected from a population.

Nested Classifications

Gender	Girl				Boy			
Age	8	10	12	14	8	10	12	14
Person	1.11	1..11	1..11	1..11	12..27	12..27	12..27	12..27

The purpose of the study is to:

- Estimate and compare the growth means over the entire population of children.
- Account for the variability in the response variable (growth) due to the Person variance.

The Model:

$$y_{ijk} = \mu + \alpha_i + \gamma_k + (\alpha\gamma)_{ik} + \varepsilon_{ijk}$$

$\varepsilon_{ijk} \sim N(0,\ R)$

y_{ijk} the growth measurement at the i^{th} Gender on the j^{th} Person and k^{th} Age

μ overall mean and an unknown fixed effect

α_i the fixed intercept effect of the i^{th} Gender

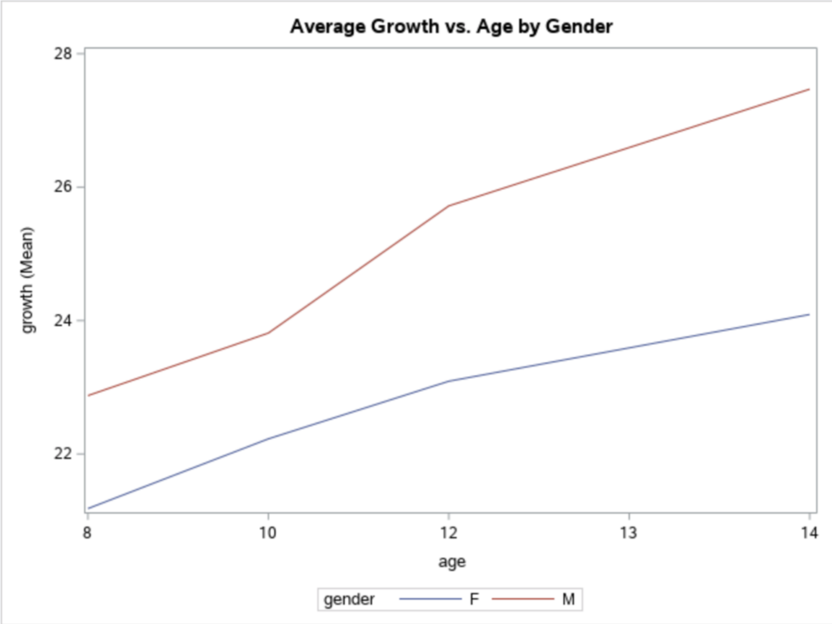
γ_k the fixed effect of the k^{th} Age

$(\alpha\gamma)_{ik}$ the fixed effect of the interaction between the i^{th} Gender and the k^{th} Age

ε_{ijk} the random error associated with the j^{th} Person at i^{th} Gender at Age k.

Produce a profile plot for repeated measures data

```
/* 1 */
proc sgplot data=Growth;
vline age / group=gender stat=mean response=growth;
title 'Average Growth vs. Age by Gender';
run;
title;
```



The average growth measurements for boys and girls are relatively linear. Boys seem to have a more significant growth measure than girls. Boys also have a slightly faster growth rate than girls starting at 10 years old.

2. Determine the appropriate covariance structure

Visualising the Correlation Structure

Step 1: Model the mean structure

$$\mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij}$$

$$\mu + gender + age + (age*gender)$$

μ overall mean

α_i the fixed intercept effect of the i^{th} Gender

γ_j the fixed effect of the j^{th} Age

$(\alpha\gamma)_{ij}$ the interaction between the i^{th} Gender and the j^{th} Age

Step 2: Specify the Covariance Structure

```
/* 2.b */
proc mixed data=Growth;
class gender age;
model growth=gender age gender*age / ddfm=kr2;
repeated age / type=un subject=person r rcorr;
ods output covparms=cov rcorr=corr;
run;
```

The variances and covariances appear to be pretty constant.

Estimated R Matrix for Subject 1				
Row	Col1	Col2	Col3	Col4
1	5.4155	2.7168	3.9102	2.7102
2	2.7168	4.1848	2.9272	3.3172
3	3.9102	2.9272	6.4557	4.1307
4	2.7102	3.3172	4.1307	4.9857

It shows the covariance matrix for the first block (person 1) and every block because every block has an identical covariance structure. The diagonal elements show the variances of repeated measures at each time point; the off-diagonal elements represent the covariance of repeated measures taken at different time points. It might be easier to see the patterns of the variances and covariances from the plot produced later.

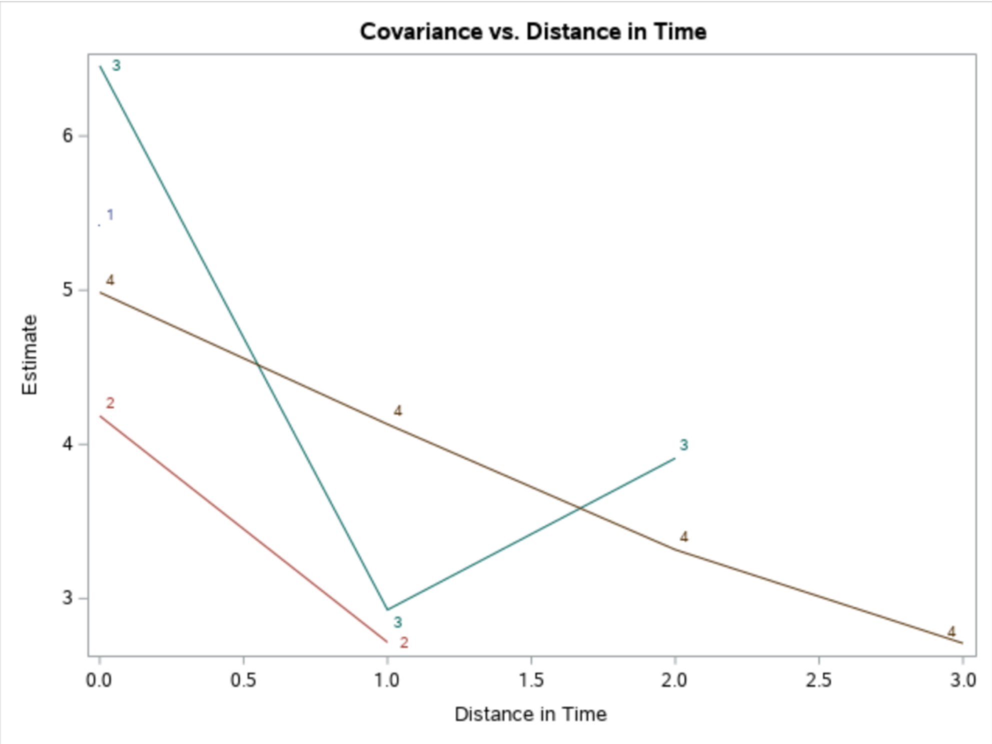
Estimated R Correlation Matrix for Subject 1				
Row	Col1	Col2	Col3	Col4
1	1.0000	0.5707	0.6613	0.5216
2	0.5707	1.0000	0.5632	0.7262
3	0.6613	0.5632	1.0000	0.7281
4	0.5216	0.7262	0.7281	1.0000

The correlation matrix is showed for the first block (person 1) and every block because every block has an identical correlation structure. The diagonal elements are always equal to one; the off-diagonal elements show the correlation of repeated measures taken at different time points.

Step 3: Produce a plot of covariance versus distance in time.

```
/* 2.c */
data times;
do time1=1 to 4;
do time2=1 to time1;
distance=time1-time2;
output;
end;
end;
run;
data covplot;
merge times cov;
run;
proc print data=covplot;
run;
proc sgplot data=covplot noautolegend;
label distance='Distance in Time';
series y=estimate x=distance / group=time1 datalabel=time1;
title 'Covariance vs. Distance in Time';
run;
title;
```

The times data set produces time pairs, and the hourly distance between them corresponds to the covariance parameter estimates in the data set cov.



As the distance between pairs of observations increases, covariance tends to decrease. The pattern of decreasing covariance with distance is roughly the same for all reference times, indicated by each line's number.

The values plotted at a distance of 0 are the variances among the observations at each of the four-time points. They range from 4.2 to 6.5 and do not seem to be increasing or decreasing in variances with time. The values plotted at distances 1, 2, and 3 represent the covariances between pairs 1, 2, or 3 distances apart. On average, they seem to be pretty constant. Therefore, concluding that a model with constant variance over time is probably adequate for the data.

```
/* 2.d */
%macro forplot(corrdata, dim); proc datasets;
delete corrplot; run;
%do i=1 %to %eval(&dim-1);
data corrplot&i(rename=(col&i=corr row=time1));
set &corrdata(keep=col&i row); distance=row-&i;
time2=&i;
if distance > 0; run;
proc append base=corrplot data=corrplot&i; run;
%end;
 %mend;
 %forplot(corr, 4);
proc print data=corrplot; run;
proc sgplot data=corrplot noautolegend;
label distance='Distance in Time';
series y=corr x=distance / group=time1 datalabel=time1;
yaxis min=0 max=1;
xaxis integer;
title 'Correlation vs. Distance in Time';
run;title;
```

The correlations between pairs of observations at distances 1, 2, and 3 apart in time seem to be relatively stable. A compound symmetry (CS) covariance structure might be appropriate.

Fitting a few covariance structures to the data and comparing the fit statistics.

```
/* 2.e */
ods listing close;
proc mixed data=Growth;
class gender age person;
model growth=gender age gender*age / ddfm=kr2;
repeated age / type=un subject=person;
ods output FitStatistics=FitUn(rename=(value=UN)); run;
proc mixed data=Growth;
class gender age person;
model growth=gender age gender*age / ddfm=kr2;
repeated age / type=ar(1) subject=person;
ods output FitStatistics=FitAR1(rename=(value=AR1)); run;
proc mixed data=Growth;
class gender age person;
model growth=gender age gender*age / ddfm=kr2;
repeated age / type=toep subject=person;
ods output FitStatistics=FitToep(rename=(value=Toep));run;
proc mixed data=Growth;
class gender age person;
model growth=gender age gender*age / ddfm=kr2;
repeated age / type=cs subject=person;
ods output FitStatistics=FitCS(rename=(value=CS));run;
data fits; merge FitUN FitAR1 FitToep FitCS; by descr; run;
ods listing; proc print data=fits; run;
```

Obs	Descr	UN	AR1	Toep	CS
1	-2 Res Log Likelihood	414.0	434.5	418.9	423.4
2	AIC (Smaller is Better)	434.0	438.5	426.9	427.4
3	AICC (Smaller is Better)	436.5	438.7	427.4	427.5
4	BIC (Smaller is Better)	447.0	441.1	432.1	430.0

The CS model and TOEP model seem to provide similar fits to the data.

3. Make statistical inference

CS model:

```
**CS Model;
proc mixed data=Growth;
class gender age person;
model growth = gender age gender*age / ddfm=kr2;
repeated age / type=cs subject=person;
lsmeans gender*age / slice=gender slice=age;
run;
```

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
CS	person	3.2854
Residual		1.9750

Fit Statistics	
-2 Res Log Likelihood	423.4
AIC (Smaller is Better)	427.4
AICC (Smaller is Better)	427.5
BIC (Smaller is Better)	430.0

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
gender	1	25	9.29	0.0054
age	3	75	35.35	<.0001
gender*age	3	75	2.36	0.0781

The covariance of the growth measurements between any pair of the repeated measures for a given subject is estimated to be 3.29. The residual variance is estimated to be 1.98. The gender*age interaction has a p-value of 0.078, which is marginally significant.

Least Squares Means							
Effect	gender	age	Estimate	Standard Error	DF	t Value	Pr > |t|
gender*age	F	8	21.1818	0.6915	46.1	30.63	<.0001
gender*age	F	10	22.2273	0.6915	46.1	32.14	<.0001
gender*age	F	12	23.0909	0.6915	46.1	33.39	<.0001
gender*age	F	14	24.0909	0.6915	46.1	34.84	<.0001
gender*age	M	8	22.8750	0.5734	46.1	39.89	<.0001
gender*age	M	10	23.8125	0.5734	46.1	41.53	<.0001
gender*age	M	12	25.7188	0.5734	46.1	44.85	<.0001
gender*age	M	14	27.4688	0.5734	46.1	47.91	<.0001

The output from the SLICE=gender option in the LSMEANS statement suggests that Age is a significant factor for both boys and girls.

Tests of Effect Slices						
Effect	gender	age	Num DF	Den DF	F Value	Pr > F
gender*age	F		3	75	8.55	<.0001
gender*age	M		3	75	33.84	<.0001
gender*age		8	1	46.1	3.55	0.0658
gender*age		10	1	46.1	3.11	0.0843
gender*age		12	1	46.1	8.56	0.0053
gender*age		14	1	46.1	14.14	0.0005

The output from the SLICE=age option indicates that at ages 12 and 14, the p-values are 0.0053 and 0.0005, respectively. Hence, the growth measurements between boys and girls are significantly different. At ages 8 and 10, the p-values are 0.066 and 0.084, respectively, showing that the differences in growth measurements between boys and girls are marginally significant.

TOEP model:

```
**TOEP Model;
proc mixed data=Growth;
class gender age person;
model growth = gender age gender*age / ddfm=kr2;
 repeated age / type=toep subject=person;
lsmeans gender*age / slice=gender slice=age;
run;
```

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
TOEP(2)	person	3.3325
TOEP(3)	person	3.7210
TOEP(4)	person	2.4870
Residual		5.3195

Fit Statistics	
-2 Res Log Likelihood	418.9
AIC (Smaller is Better)	426.9
AICC (Smaller is Better)	427.4
BIC (Smaller is Better)	432.1

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
gender	1	25.3	9.19	0.0055
age	3	40.5	29.03	<.0001
gender*age	3	40.5	2.27	0.0953

Least Squares Means							
Effect	gender	age	Estimate	Standard Error	DF	t Value	Pr > |t|
gender*age	F	8	21.1818	0.6954	45.8	30.46	<.0001
gender*age	F	10	22.2273	0.6954	45.8	31.96	<.0001
gender*age	F	12	23.0909	0.6954	45.8	33.20	<.0001
gender*age	F	14	24.0909	0.6954	45.8	34.64	<.0001
gender*age	M	8	22.8750	0.5766	45.8	39.67	<.0001
gender*age	M	10	23.8125	0.5766	45.8	41.30	<.0001
gender*age	M	12	25.7188	0.5766	45.8	44.60	<.0001
gender*age	M	14	27.4688	0.5766	45.8	47.64	<.0001

Tests of Effect Slices						
Effect	gender	age	Num DF	Den DF	F Value	Pr > F
gender*age	F		3	40.5	6.77	0.0008
gender*age	M		3	40.5	28.56	<.0001
gender*age		8	1	45.8	3.51	0.0673
gender*age		10	1	45.8	3.08	0.0860
gender*age		12	1	45.8	8.46	0.0056
gender*age		14	1	45.8	13.98	0.0005

The results are very similar to the CS model.

4. Fit a linear growth curve treating age as a continuous variable

CS model:

```
**CS Model;
proc mixed data=Growth;
class gender person;
model growth=gender age gender*age/ ddfm=kr2 outp=preddata s;
repeated / type=cs subject=person;
run;
proc sgplot data=preddata;
series y=pred x=age / group=gender;
title1 'Predicted Growth Curve';
title2 'CS Model';
run; title;
```

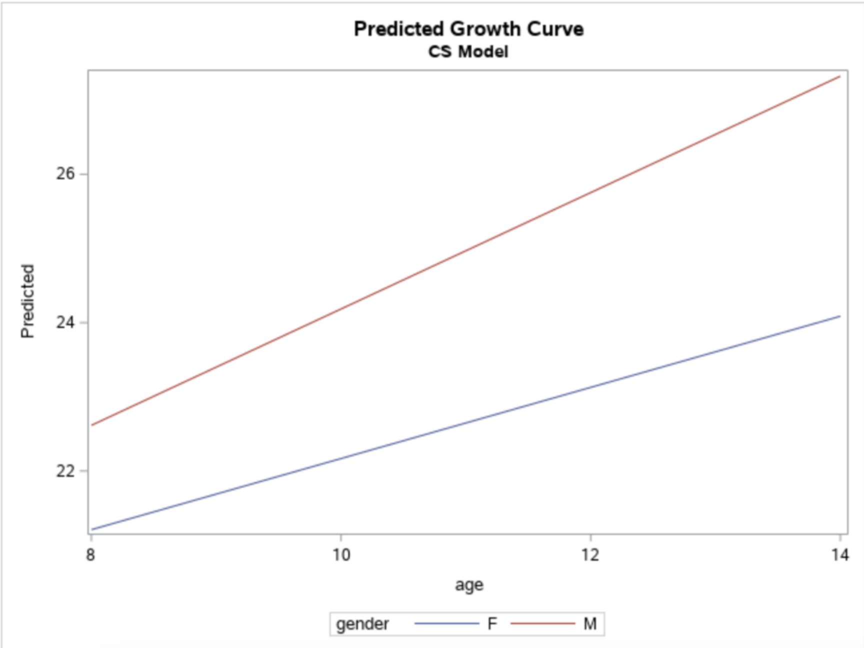
Solution for Fixed Effects						
Effect	gender	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		16.3406	0.9813	104	16.65	<.0001
gender	F	1.0321	1.5374	104	0.67	0.5035
gender	M	0	.	.	.	.
age		0.7844	0.07750	79	10.12	<.0001
age*gender	F	-0.3048	0.1214	79	-2.51	0.0141
age*gender	M	0	.	.	.	.

The linear regression models are:

- boys: growth = 16.3406 + 0.7844 * age
- girls: growth = (16.3406+1.0321) + (0.7844–0.3048)*age = 17.3727+0.4796*age

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
gender	1	104	0.45	0.5035
age	1	79	108.36	<.0001
age*gender	1	79	6.30	0.0141

The age * gender interaction is significant, indicating the need of an unequal slope model. The age effect is also significant, which indicates that the overall slope is nonzero.



TOEP model:

```
**TOEP Model;
proc mixed data=Growth;
class gender person;
model growth=gender age gender*age/ ddfm=kr2 outp=preddata s;
repeated / type=toep subject=person;
run;
proc sgplot data=preddata;
series y=pred x=age / group=gender;
title1 'Predicted Growth Curve';
title2 'TOEP Model';
run; title;
```

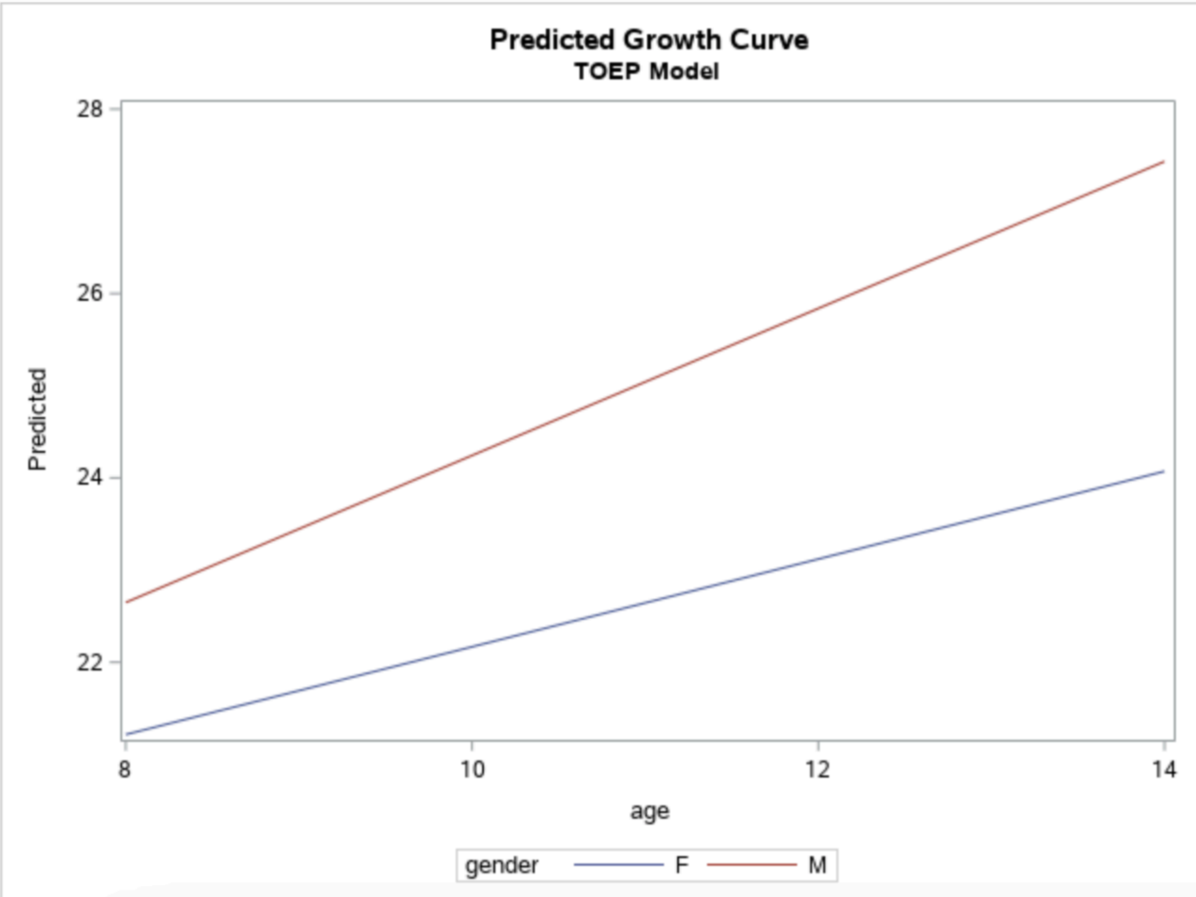
Solution for Fixed Effects						
Effect	gender	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		16.2704	1.0723	41.2	15.17	<.0001
gender	F	1.1385	1.6799	41.2	0.68	0.5017
gender	M	0	.	.	.	.
age		0.7973	0.08674	28.1	9.19	<.0001
age*gender	F	-0.3214	0.1359	28.1	-2.37	0.0252
age*gender	M	0	.	.	.	.

These linear regression models are similar to the CS model:

- boys: growth = 16.2704 + 0.7973 * age
- girls: growth = (16.2704+1.1385) + (0.7973–0.3214)*age = 17.4089+0.4759*age

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
gender	1	41.2	0.46	0.5017
age	1	28.1	87.78	<.0001
age*gender	1	28.1	5.59	0.0252

The results are very similar to the CS model.



5. Add the GROUP=gender option

Add the GROUP=gender option

```
**CS Model;
proc mixed data=Growth;
class gender age person;
model growth = gender age gender*age / ddfm=kr2;
repeated age / type=cs subject=person group=gender;
run;
```

GROUP=gender

Fit Statistics	
-2 Res Log Likelihood	406.4
AIC (Smaller is Better)	414.4
AICC (Smaller is Better)	414.8
BIC (Smaller is Better)	419.5

without GROUP

Fit Statistics	
-2 Res Log Likelihood	423.4
AIC (Smaller is Better)	427.4
AICC (Smaller is Better)	427.5
BIC (Smaller is Better)	430.0

Compared with the model without the GROUP = option, both AICC and BIC values from this model are smaller, concluding that the GROUP = option is helpful for the data.

```
**TOEP Model;
proc mixed data=Growth;
class gender age person;
model growth = gender age gender*age / ddfm=kr2;
repeated age / type=toep subject=person group=gender;
run;
```

GROUP=gender

Fit Statistics	
-2 Res Log Likelihood	401.2
AIC (Smaller is Better)	417.2
AICC (Smaller is Better)	418.8
BIC (Smaller is Better)	427.6

without GROUP

Fit Statistics	
-2 Res Log Likelihood	418.9
AIC (Smaller is Better)	426.9
AICC (Smaller is Better)	427.4
BIC (Smaller is Better)	432.1

The CS model seems to be the best fitting model with GROUP = gender option.

Fit a growth curve model with GROUP=gender

```
** for CS Model;
proc mixed data=Growth;
class gender person;
model growth=gender age gender*age/ ddfm=kr2 outp=preddata s;
repeated / type=cs subject=person group=gender;
run;

proc sgplot data=preddata;
series y=pred x=age / group=gender;
title1 'Predicted Growth Curve';
title2 'CS Model with GROUP=gender';
run; title;
```

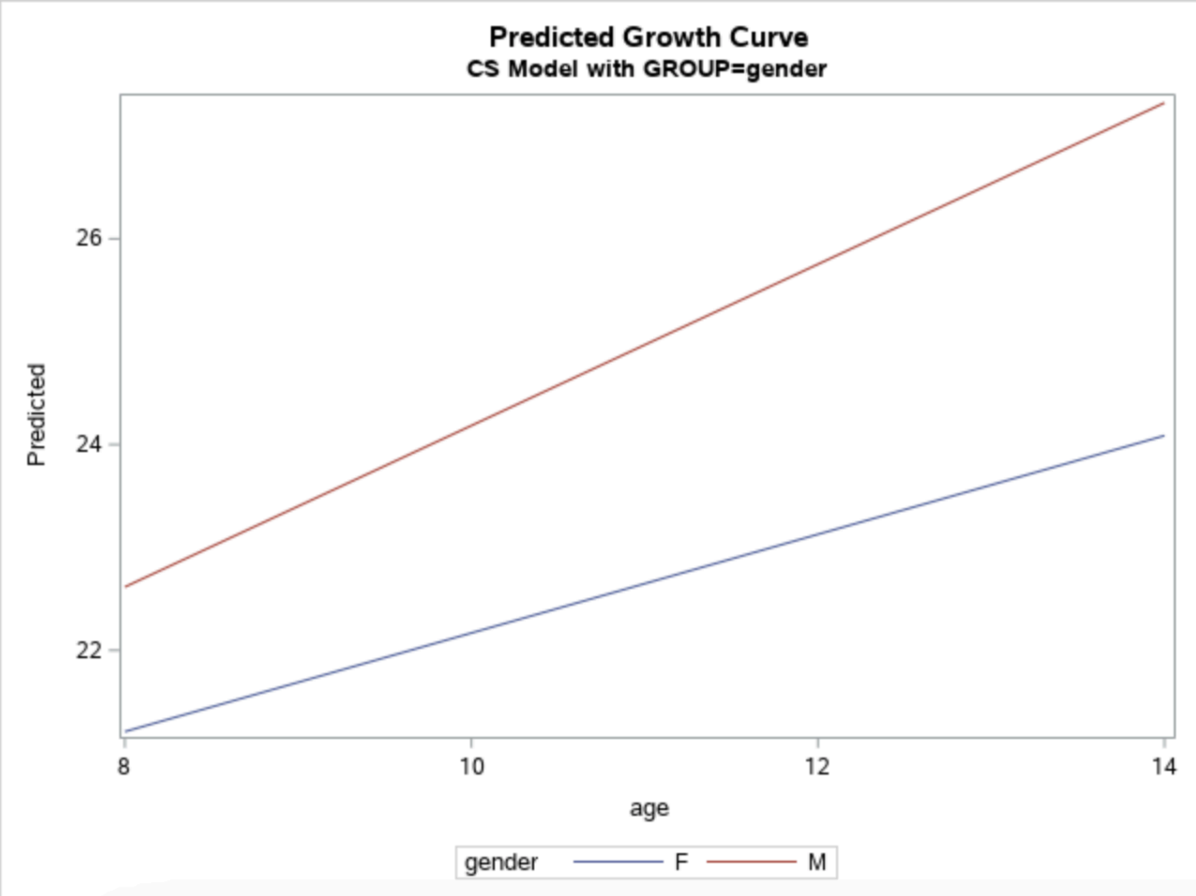
Solution for Fixed Effects						
Effect	gender	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		16.3406	1.1287	60	14.48	<.0001
gender	F	1.0321	1.4183	86.5	0.73	0.4687
gender	M	0	.	.	.	.
age		0.7844	0.09382	47	8.36	<.0001
age*gender	F	-0.3048	0.1076	70.9	-2.83	0.0060
age*gender	M	0	.	.	.	.

The linear regression models are as follows:

- boys: growth = 16.3406 + 0.7844 * age
- girls: growth = (16.3406+1.0321) + (0.7844–0.3048)*age = 17.3727+0.4796*age

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
gender	1	86.5	0.53	0.4687
age	1	70.9	138.11	<.0001
age*gender	1	70.9	8.03	0.0060

The Type 3 tests indicate that Age*gender is a significant factor (p-value is small). The slopes for Age are not the same between the boys and the girls.



6. Fit a random coefficient model with an unstructured covariance

Fit random coefficient model

```
/* 6.a */
proc mixed data=Growth;
class gender person;
model growth=gender age gender*age / s ddfm=kr2;
random int age / type=un subject=person s;
run;
```

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
UN(1,1)	person	5.7864
UN(2,1)	person	-0.2896
UN(2,2)	person	0.03252
Residual		1.7162

Fit Statistics	
-2 Res Log Likelihood	432.6
AIC (Smaller is Better)	440.6
AICC (Smaller is Better)	441.0
BIC (Smaller is Better)	445.8

Solution for Fixed Effects						
Effect	gender	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		16.3406	1.0185	25	16.04	<.0001
gender	F	1.0321	1.5957	25	0.65	0.5237
gender	M	0	.	.	.	.
age		0.7844	0.08600	25	9.12	<.0001
age*gender	F	-0.3048	0.1347	25	-2.26	0.0326
age*gender	M	0	.	.	.	.

Solution for Random Effects						
Effect	person	Estimate	Std Err Pred	DF	t Value	Pr > |t|
Intercept	1	-0.6413	2.3457	3.18	-0.27	0.8014
age	1	-0.04475	0.2043	2.58	-0.22	0.8428
Intercept	2	-0.6602	2.3457	3.18	-0.28	0.7957
age	2	0.09029	0.2043	2.58	0.44	0.6930
Intercept	3	-0.2489	2.3457	3.18	-0.11	0.9218
age	3	0.1136	0.2043	2.58	0.56	0.6230
Intercept	4	1.6611	2.3457	3.18	0.71	0.5273
age	4	0.02821	0.2043	2.58	0.14	0.9003
Intercept	5	0.5710	2.3457	3.18	0.24	0.8226
age	5	-0.05496	0.2043	2.58	-0.27	0.8080
Intercept	6	-0.8263	2.3457	3.18	-0.35	0.7467
age	6	-0.04806	0.2043	2.58	-0.24	0.8315
Intercept	7	0.05820	2.3457	3.18	0.02	0.9817
age	7	0.02348	0.2043	2.58	0.11	0.9169
Intercept	8	1.4133	2.3457	3.18	0.60	0.5871
age	8	-0.07178	0.2043	2.58	-0.35	0.7521
Intercept	9	-0.5389	2.3457	3.18	-0.23	0.8323
age	9	-0.07478	0.2043	2.58	-0.37	0.7423
Intercept	10	-2.9842	2.3457	3.18	-1.27	0.2884
age	10	-0.06270	0.2043	2.58	-0.31	0.7821

Example of equation for person 1 (girl) is:

$$\text{growth} = (16.3406 + 1.0321) + (0.7844 - 0.3048) * \text{age} - 0.6413 - 0.04475 * \text{age}$$

$$= 16.7314 + 0.4349 * \text{age}$$

The predicted growth measurement for the first person at age 13

```
/* 6.b */
data new;
input person gender $ age;
datalines;
1 F 13
; run;
data growth;
set Growth new;
run;
proc mixed data=growth;
class gender person;
model growth=gender age gender*age / ddfm=kr2 outp=pred;
random int age / type=un subject=person;
run;
proc print data=pred;
where age=13;
run;
```

Obs	person	gender	growth	age	Pred	StdErrPred	DF	Alpha	Lower	Upper	Resid
109	1	F	.	13	22.3837	0.77389	69.6779	0.05	20.8401	23.9273	.

The predicted growth measurement for the first person at age 13 is 22.3837.

7. Fit a random coefficient model plus an AR(1) structure with GROUP=gender

```
/* 7 */
proc mixed data=Growth;
class gender person;
model growth = gender age gender*age / s ddfm=kr2;
random int age / type=un subject=person;
repeated / type=ar(1) group=gender subject=person;
run;
```

Covariance Parameter Estimates			
Cov Parm	**Subject**	**Group**	**Estimate**
UN(1,1)	person		5.4638
UN(2,1)	person		-0.2909
UN(2,2)	person		0.03737
Variance	person	gender F	0.3870
AR(1)	person	gender F	-0.1314
Variance	person	gender M	2.2322
AR(1)	person	gender M	-0.2555

Fit Statistics	
-2 Res Log Likelihood	410.5
AIC (Smaller is Better)	424.5
AICC (Smaller is Better)	425.6
BIC (Smaller is Better)	433.6

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
gender	1	24.7	0.73	0.4018
age	1	26.7	118.46	<.0001
age*gender	1	26.7	7.20	0.0124

This model has more minor fit statistics than the model fitted for the previous random coefficient model with independent errors. It seems to fit the data better. The Type 3 Tests of Fixed Effects also remains the same conclusion.