Stat 506 Assignment 3

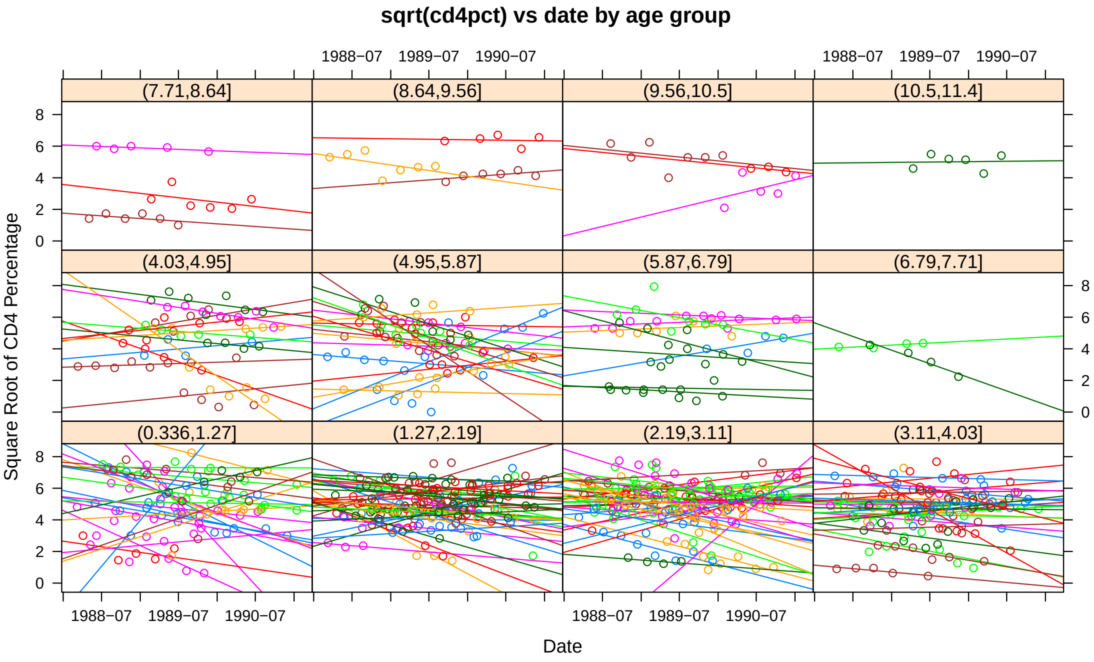
Leslie Gains-Germain

February 6, 2015

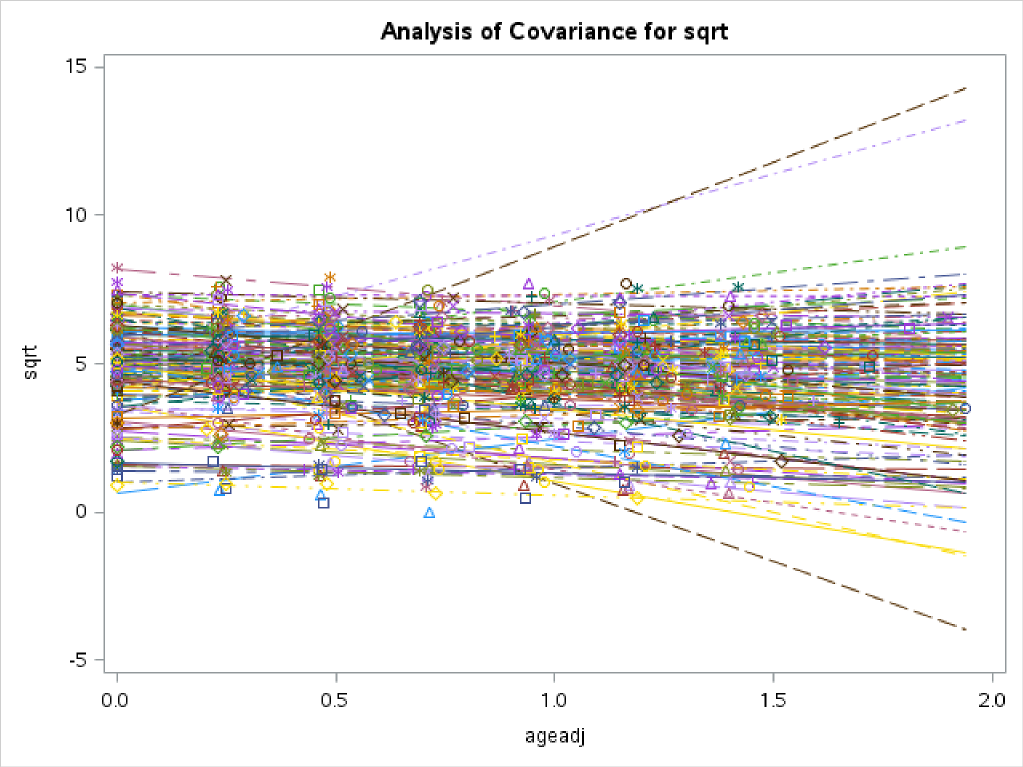
1. Exercise 11.4 p 249, but use these data (a subset with missing values and singletons removed.)

   (a) Plot square root CD4 percentage over time with regression lines. Use the as.Date function to make visit days into a ”date” object for plotting. A centered time column is also a good idea.

   I first looked at a plot showing the relationship between square root CD4 percentage and time for every kid. Before adding regression lines for each kid, I broke the kids up by age group to make the plot more readable (so it wouldn't just be one plot with 173 lines on it). The result is below. I think I could work more on how to define the age groups so that there is about the same number of kids in each group.



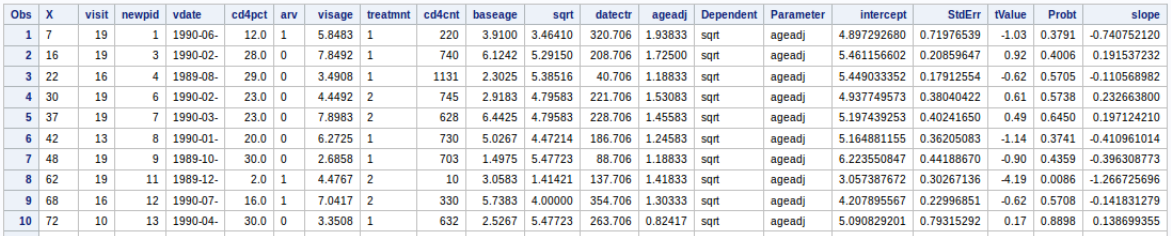
   Here is the SAS version of the plot with a regression line for each kid.



(b) Create a fit for each kid using lmList in nlme package. (The intervals command creates 95% CIs for each kid's intercept and slope.)

The following shows the intercept and slope estimates, with 95% confidence intervals, for the first 6 kids. I also used ODS OUTPUT in SAS to do this. I outputted a dataset with a row for each kid and their estimated intercept and slope. This was a pain to do in SAS. The dataset is below.

```
## , , (Intercept)
##
##   lower est. upper
## 1  3.84 4.90  5.95
## 3  4.57 5.46  6.35
## 4  4.45 5.45  6.44
## 6  3.97 4.94  5.91
## 7  4.27 5.20  6.13
## 8  3.63 5.16  6.70
##
## , , ageadj
##
##    lower   est. upper
## 1 -1.652 -0.741  0.17
## 3 -0.665  0.192  1.05
## 4 -1.485 -0.111  1.26
## 6 -0.898  0.233  1.36
## 7 -0.874  0.197  1.27
## 8 -2.301 -0.411  1.48
```



Obs	X	visit	newpid	vdate	cd4pct	arv	visage	treatmnt	cd4cnt	baseage	sqrt	datectr	ageadj	Dependent	Parameter	intercept	StdErr	tValue	Probt	slope
1	7	19	1	1990-06-	12.0	1	5.8483	1	220	3.9100	3.46410	320.706	1.93833	sqrt	ageadj	4.897292680	0.71976539	-1.03	0.3791	-0.740752120
2	16	19	3	1990-02-	28.0	0	7.8492	1	740	6.1242	5.29150	208.706	1.72500	sqrt	ageadj	5.461156602	0.20859647	0.92	0.4006	0.191537232
3	22	16	4	1989-08-	29.0	0	3.4908	1	1131	2.3025	5.38516	40.706	1.18833	sqrt	ageadj	5.449033352	0.17912554	-0.62	0.5705	-0.110568982
4	30	19	6	1990-02-	23.0	0	4.4492	2	745	2.9183	4.79583	221.706	1.53083	sqrt	ageadj	4.937749573	0.38040422	0.61	0.5738	0.232663800
5	37	19	7	1990-03-	23.0	0	7.8983	2	628	6.4425	4.79583	228.706	1.45583	sqrt	ageadj	5.197439253	0.40241650	0.49	0.6450	0.197124210
6	42	13	8	1990-01-	20.0	0	6.2725	1	730	5.0267	4.47214	186.706	1.24583	sqrt	ageadj	5.164881155	0.36205083	-1.14	0.3741	-0.410961014
7	48	19	9	1989-10-	30.0	0	2.6858	1	703	1.4975	5.47723	88.706	1.18833	sqrt	ageadj	6.223550847	0.44188670	-0.90	0.4359	-0.396308773
8	62	19	11	1989-12-	2.0	1	4.4767	2	10	3.0583	1.41421	137.706	1.41833	sqrt	ageadj	3.057387672	0.30267136	-4.19	0.0086	-1.266725696
9	68	16	12	1990-07-	16.0	1	7.0417	2	330	5.7383	4.00000	354.706	1.30333	sqrt	ageadj	4.207895567	0.22996851	-0.62	0.5708	-0.141831279
10	72	10	13	1990-04-	30.0	0	3.3508	1	632	2.5267	5.47723	263.706	0.82417	sqrt	ageadj	5.090829201	0.79315292	0.17	0.8898	0.138699355

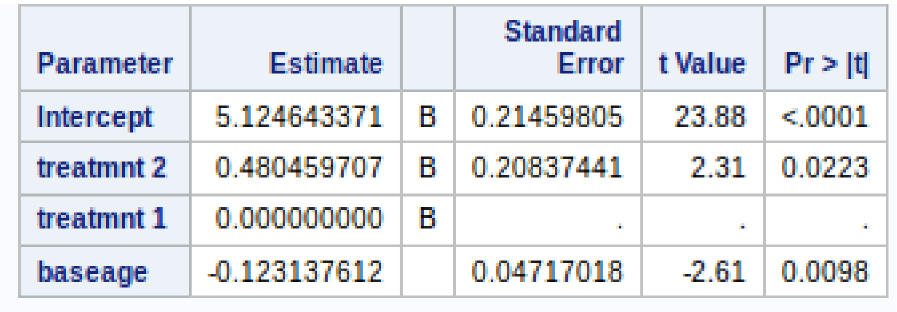
(c) Extract coefficient vectors for each kid and obtain the two coefficient estimates.

The average of the estimated intercepts for each kid is 4.93*. The average of the estimated slopes for each kid is* −0.302*. These are the estimates we would get if we left* `kid` *out of the model.*

(d) Note that these children contracted HIV from their mothers at birth, so their base age tells us how long the virus has been in their systems. Use lm to obtain estimates for the whole population by fitting a model for intercept as a function of initial age and treatment, then fitting a similar model for slopes. Plot these fits and their "data points" (in quotes because they are estimates, not data).

*The first table below shows a summary of the model for intercept as a function of initial age and treatment. There is strong evidence that a kid's intercept is related to baseage (p-value=*0.00985*) after accounting for treatment. For a kid coming in to his first hospital visit, his sqrt(cd4pct) is estimated to be* 0.123 *lower than a kid one year younger. There is moderate evidence that a kid's intercept is related to treatment after accounting for baseage (p-value=* 0.02233*). At his first hospital visit, a kid on treatment* 2 *is estimated to have a sqrt(cd4pct)* 0.48 *higher than a kid on treatment* 1*.*

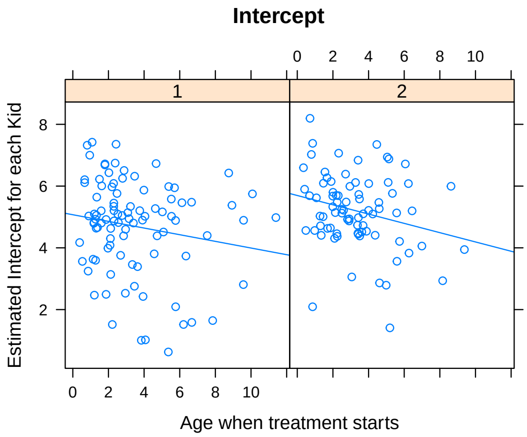
The SAS output is also below. I changed the reference level to treatment 1*, so we can easily see that the output is identical.*

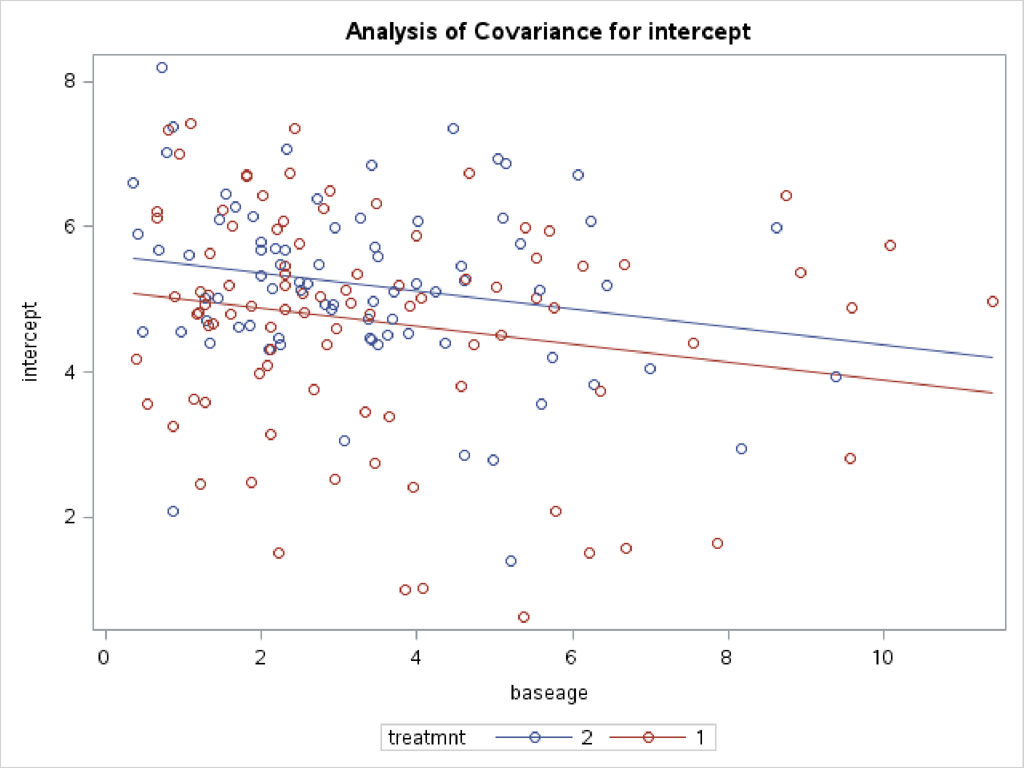


Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	5.124643371	B	0.21459805	23.88	<.0001
treatmnt 2	0.480459707	B	0.20837441	2.31	0.0223
treatmnt 1	0.000000000	B	.	.	.
baseage	-0.123137612		0.04717018	-2.61	0.0098

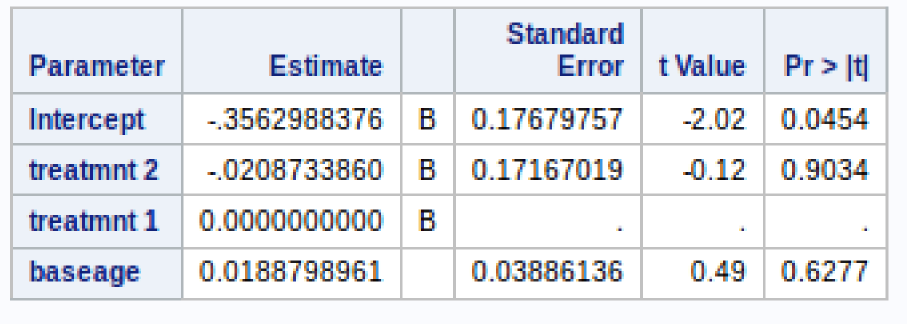
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.1246	0.2146	23.88	0.0000
cd4.treatmnt2	0.4805	0.2084	2.31	0.0223
cd4.baseage	-0.1231	0.0472	-2.61	0.0098

In the plots below, it is pretty clear that the estimated intercept decreases with baseline age for both treatments 1 *and* 2*.*





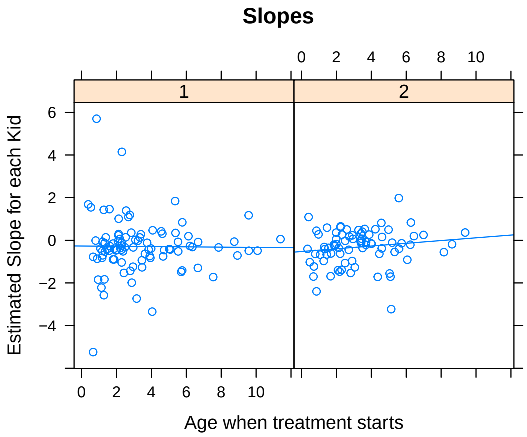
The following shows the output (both R and SAS) for the model for slope as a fucntion of baseage and treatment. There is no evidence that the slope (relationship between sqrt(cd4pct) and age) depends on either baseage or treatment contingent upon the other variable being in the model.

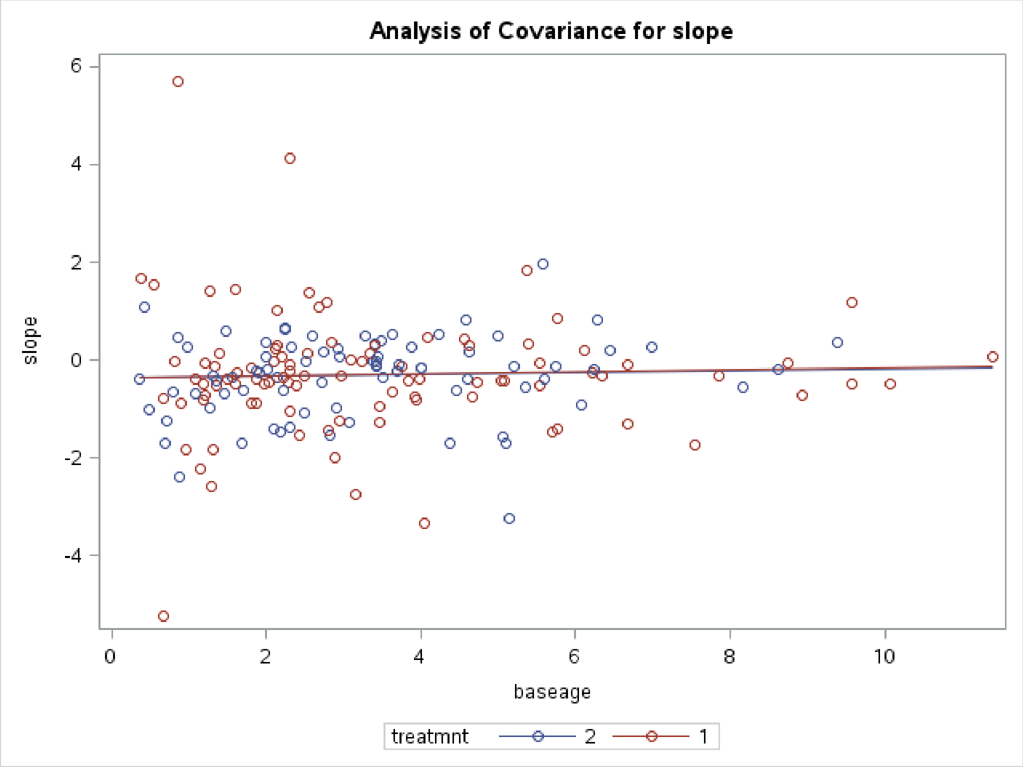


Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-.3562988376	B	0.17679757	-2.02	0.0454
treatmnt 2	-.0208733860	B	0.17167019	-0.12	0.9034
treatmnt 1	0.0000000000	B	.	.	.
baseage	0.0188798961		0.03886136	0.49	0.6277

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.3563	0.1768	-2.02	0.0454
cd4.treatmnt2	-0.0209	0.1717	-0.12	0.9034
cd4.baseage	0.0189	0.0389	0.49	0.6277

We can see in the plots below that the estimated slopes do not appear to change much across baseline age for either treatment.





2. Redo the plot and fit the same models in SAS. Use the plain GLM function without random or repeated statements. Compare output with R.

   I incorporated the SAS output in with the R output.

3. Exercise 12.2 uses the same data, but now fitting with random effects.

(a) Predict CD4% as a function of time using random intercepts and slopes for each child (both R and SAS).

Below is R and SAS output for 1. random intercepts and 2. random intercepts and slopes. For the random intercept model, we calculate $1.323^2 = 1.750$ *and* $0.758^2 = 0.574$*, so the R output matches with the SAS output. For the random intercept and slope model, we calculate* $1.304^2 = 1.700$*,* $(1.304)(0.583)(-0.073) = -0.055$*,* $0.583^2 = 0.340$ *and* $0.698^2 = 0.488$*. The R output matches the SAS output.*

Just intercept random:

```
Linear mixed-effects model fit by maximum likelihood
 Data: cd4
       AIC      BIC    logLik
  2535.376 2554.568 -1263.688

Random effects:
 Formula: ~1 | newpid
        (Intercept)  Residual
StdDev:    1.322925 0.7575567

Fixed effects: sqrt ~ ageadj
                Value  Std.Error  DF  t-value p-value
(Intercept)  4.952422 0.10948883 722 45.23221       0
ageadj      -0.404077 0.05627747 722 -7.18009       0
 Correlation:
       (Intr)
ageadj -0.311

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-4.84075523 -0.47480506  0.00548256  0.46715601  5.09223422

Number of Observations: 896
Number of Groups: 173
```

Both intercept and slope random:

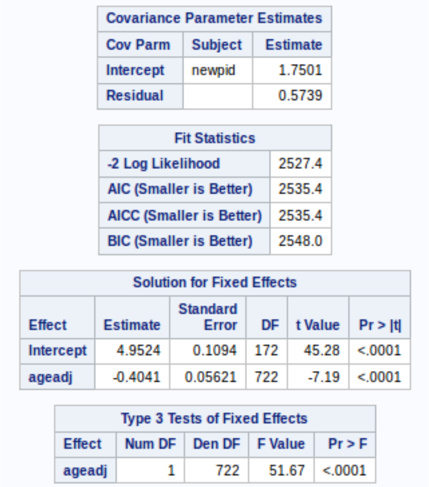
```
Linear mixed-effects model fit by maximum likelihood
 Data: cd4
       AIC      BIC    logLik
  2508.435 2537.222 -1248.217

Random effects:
 Formula: ~1 + ageadj | newpid
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 1.3040801 (Intr)
ageadj      0.5826822 -0.073
Residual    0.6984168

Fixed effects: sqrt ~ ageadj
                Value  Std.Error  DF  t-value p-value
(Intercept)  4.948200 0.10709151 722 46.20535       0
ageadj      -0.387341 0.07096596 722 -5.45812       0
 Correlation:
       (Intr)
ageadj -0.272

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-5.23717060 -0.41696537  0.02376671  0.41749766  5.11482679

Number of Observations: 896
Number of Groups: 173
```



Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	newpid	1.7501
Residual		0.5739

Fit Statistics

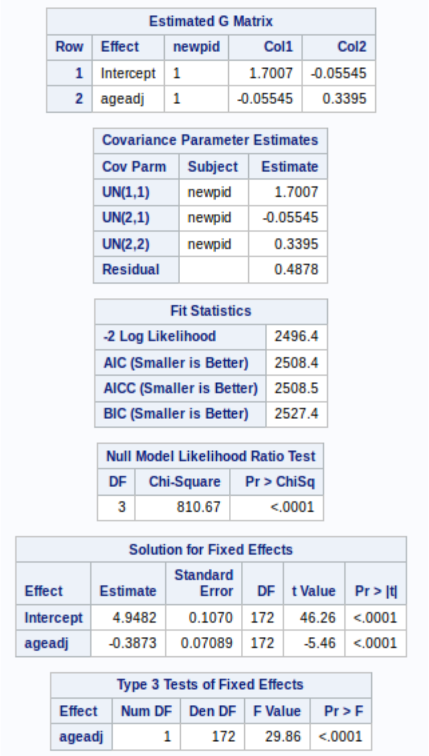
-2 Log Likelihood	2527.4
AIC (Smaller is Better)	2535.4
AICC (Smaller is Better)	2535.4
BIC (Smaller is Better)	2548.0

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	4.9524	0.1094	172	45.28	<.0001
ageadj	-0.4041	0.05621	722	-7.19	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
ageadj	1	722	51.67	<.0001



Estimated G Matrix

Row	Effect	newpid	Col1	Col2
1	Intercept	1	1.7007	-0.05545
2	ageadj	1	-0.05545	0.3395

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
UN(1,1)	newpid	1.7007
UN(2,1)	newpid	-0.05545
UN(2,2)	newpid	0.3395
Residual		0.4878

Fit Statistics

-2 Log Likelihood	2496.4
AIC (Smaller is Better)	2508.4
AICC (Smaller is Better)	2508.5
BIC (Smaller is Better)	2527.4

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
3	810.67	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	4.9482	0.1070	172	46.26	<.0001
ageadj	-0.3873	0.07089	172	-5.46	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
ageadj	1	172	29.86	<.0001

- Interpret the "fixed" intercept and time effect.

  Note that I used a maximum likelihood estimation method here because we are interpreting the fixed effects. There is strong evidence that sqrt(cd4pct) changes as the difference between kids' visit age and baseline age changes (p-value< 0.0001*). For a one year increase in the difference between visit age and baseline age, the sqrt(cd4pct) is estimated to decrease by* 0.3873*, with a* 95% *confidence interval from a* 0.5265 *to a* 0.2482 *decrease.*

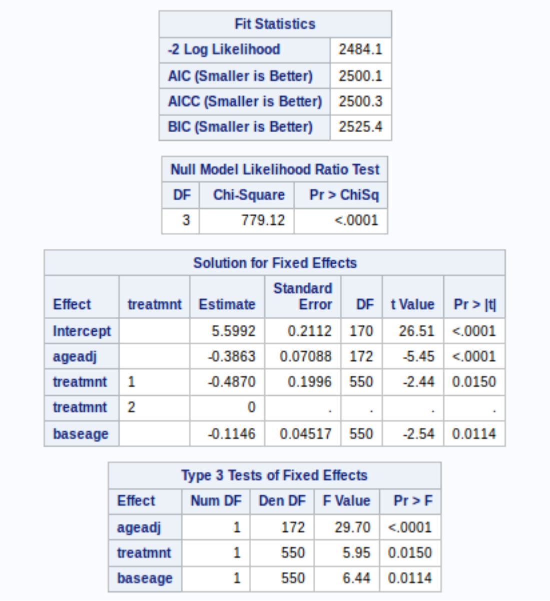
- Is this model an improvement over the simpler "random intercept – fixed slope" model?

  Looking at the SAS output, we can see that the AIC for the random intercept model is 2535.4*, and the AIC for the random intercept and random slope model is* 2508.4*. So, by AIC comparison, the random intercept random slope model is an improvement over the simple random intercept-fixed slope model.*

  Now, turning our attention to the R output, there is strong evidence that the random intercept-random slope model is an improvement over the simpler model (p-value< 0.0001 *from LRT-statistic=*30.94*).*

```
##              Model df  AIC  BIC logLik   Test L.Ratio p-value
## fit.lme          1  4 2535 2555  -1264
## fit.lme.int      2  6 2508 2537  -1248 1 vs 2    30.9  <.0001
```

(b) Include age at baseline and treatment as fixed effects. Do they improve the model? (Both R and SAS).



Fit Statistics	
-2 Log Likelihood	2484.1
AIC (Smaller is Better)	2500.1
AICC (Smaller is Better)	2500.3
BIC (Smaller is Better)	2525.4

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
3	779.12	<.0001

Solution for Fixed Effects

Effect	treatmnt	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		5.5992	0.2112	170	26.51	<.0001
ageadj		-0.3863	0.07088	172	-5.45	<.0001
treatmnt	1	-0.4870	0.1996	550	-2.44	0.0150
treatmnt	2	0	.	.	.	.
baseage		-0.1146	0.04517	550	-2.54	0.0114

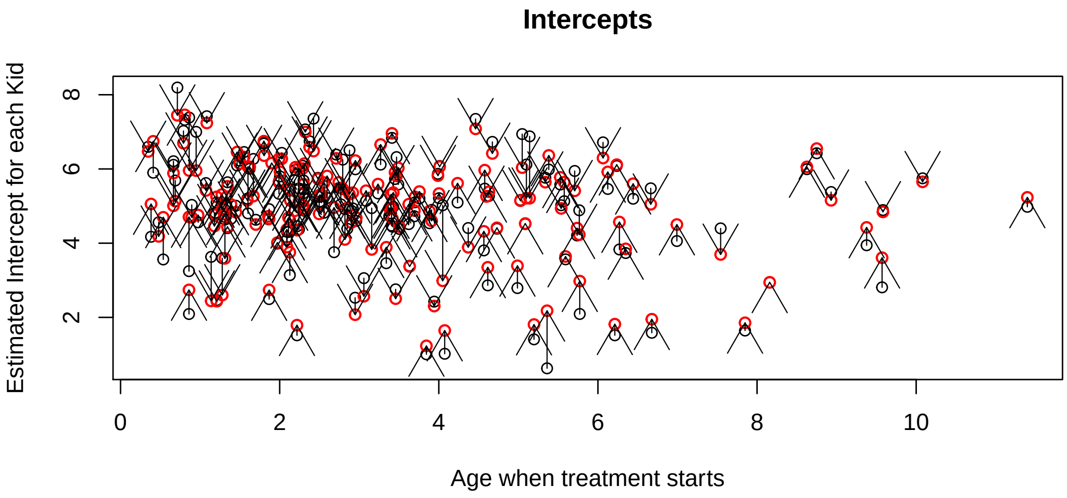
Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
ageadj	1	172	29.70	<.0001
treatmnt	1	550	5.95	0.0150
baseage	1	550	6.44	0.0114

We can see in the SAS output that the AIC for the model with the `baseage` *and* `treatment` *variables is* 2500.1. *So, this model is an improvement to the model with* `ageadj` *as the only fixed effect, but only by* 8 *AIC points. Again, turning to the R output, we see that there is moderate evidence that the model with baseage and treatment is an improvement (p-value=* 0.0021 *from LRT-stat=* 12.29*).*

```
##              Model df  AIC  BIC logLik   Test L.Ratio p-value
## fit.lme.int      1  6 2508 2537  -1248
## fit.more         2  8 2500 2539  -1242 1 vs 2    12.3  0.0021
```

(c) These data are noisy, so we don't get as obvious a picture of shrinkage as in Gelman's radon data, but we'll build the plot anyway. In either R or SAS provide a plot which shows the intercept estimates for each child plotted over initial age. Add an arrow (or can you think of a better way?) to show where they move to (how each one changes) when we fit them as random effects.



Code Appendix

```
cd4 <- read.csv("~/Documents/Stat506/Homework/HW3/mycd4.csv", head=T)
cd4 <- subset(cd4, baseage!="NA")
cd4 <- subset(cd4, cd4cnt!="NA")
```

```
cd4$vdate <- as.Date(cd4$vdate)
cd4$sqrt <- as.numeric(sqrt(cd4$cd4pct) )
cd4$datectr <- as.numeric(cd4$vdate-mean(cd4$vdate))
cd4$ageadj <- as.numeric(cd4$visage-cd4$baseage)
#require(lattice)
#xyplot(sqrt~vdate|as.factor(treatmnt), cd4, ylab="Square Root of CD4 Percentage",
  #xlab="Date", type=c("p", "r"), strip=strip.custom(factor.levels=c("trt1", "trt2")),
  #main="Relationship between sqrt(cd4pct) and date across treatment")
```

```
age.factor <- cut(cd4$baseage, breaks=12)
require(lattice)
xyplot(sqrt~vdate|age.factor, group=newpid, cd4,
       ylab="Square Root of CD4 Percentage", xlab="Date", type=c("p", "r"), main="sqrt(cd4pct) vs date by age group")
```

```
require(nlme)
fit.age <- lmList(sqrt~ageadj|newpid, cd4)
intervals(fit.age)[1:6, ,]
```

```
mean(coef(fit.age)[[1]])
mean(coef(fit.age)[[2]])
```

```
cd4.new <- data.frame(cd4$newpid, cd4$treatmnt, cd4$baseage)
cd4.new <- unique(cd4.new)
cd4.new$cd4.treatmnt <- as.factor(cd4.new$cd4.treatmnt)
lm.int <-lm(coef(fit.age)[[1]]~cd4.treatmnt+cd4.baseage, data=cd4.new)
require(xtable)
xtable(summary(lm.int))
```

```
xyplot(coef(fit.age)[[1]]~cd4.baseage|cd4.treatmnt, data=cd4.new, type=c("p","r"), xlab
       ="Age when treatment starts", ylab=c("Estimated Intercept for each Kid"), main="Intercept")
```

```
lm.slope <- lm(coef(fit.age)[[2]]~cd4.treatmnt+cd4.baseage, data=cd4.new)
xtable(summary(lm.slope))
```

```
fit.lme <- lme(sqrt~ageadj, data=cd4, random=~1|newpid, method="ML")#random intercept for each kid
fit.lme.int <- lme(sqrt~ageadj, data=cd4, random= ~1+ageadj|newpid, method="ML") #random intercept and slope
```

```
anova(fit.lme, fit.lme.int) #yes improvement
```

```
fit.more <- lme(sqrt~ageadj+baseage+treatmnt, data=cd4, random= ~ ageadj|newpid, method="ML")
##Still random intercept and slope for each kid
anova(fit.lme.int, fit.more)
```

```
with(cd4.new, plot(cd4.baseage, coef(fit.age)[[1]], xlab="Age when treatment starts",
                   ylab=c("Estimated Intercept for each Kid"), main="Intercepts"))
#Shows intercept when kid is treated as fixed effect
points(cd4.new$cd4.baseage, coef(fit.lme)[[1]], col="red", lwd=1.5)
arrows(cd4.new$cd4.baseage, coef(fit.age)[[1]], cd4.new$cd4.baseage, coef(fit.lme)[[1]], lwd=0.8)
#Now add arrows from black points to red ponits
#why are there so fewer red points?
```

SAS Code

```
DATA cd4;
  INFILE "/folders/myfolders/newcd4.csv" firstobs=2 delimiter =',' dsd;
  INPUT X $ visit newpid vdate $ cd4pct arv $ visage treatmnt $ cd4cnt baseage
           sqrt datectr ageadj;
  ;
RUN;

ODS trace on;
PROC GLM DATA=cd4;
CLASS newpid;
MODEL sqrt = ageadj / SOLUTION;
BY newpid;
*OUTPUT OUT=fits P=ints;
ODS OUTPUT ParameterEstimates=ests;
RUN;


/*DATA fits2;
SET fits;
    if ageadj=0;
RUN;*/

DATA fitints;
SET ests;
    if Parameter="Intercept";
RUN;

DATA fitslopes;
SET ests;
if Parameter="ageadj";
RUN;

DATA test;
SET cd4;
BY newpid;
if last.newpid;
RUN;

DATA intslopes;
   MERGE test fitints(rename=(Estimate=intercept)) fitslopes(rename=(Estimate=slope));
   BY newpid;
*PROC PRINT DATA=intslopes;
RUN;

PROC GLM DATA=intslopes PLOTS=all;
```

```
CLASS treatmnt / reference=first;
MODEL intercept = treatmnt baseage / SOLUTION;
RUN;

PROC GLM DATA=intslopes PLOTS=all;
CLASS treatmnt / reference=first;
MODEL slope = treatmnt baseage / SOLUTION;
RUN;

PROC MIXED DATA=cd4 method=ml;
  CLASS newpid;
  MODEL sqrt = ageadj / SOLUTION;
  random intercept / sub=newpid;
RUN;

PROC MIXED DATA=cd4 method=ml;
  CLASS newpid;
  MODEL sqrt = ageadj / s;
  random intercept ageadj / type=un sub=newpid g;
RUN;

PROC MIXED DATA=cd4 method=ml;
  CLASS newpid treatmnt;
  MODEL sqrt = ageadj treatmnt baseage / s;
  random intercept ageadj / type=un sub=newpid g;
RUN;
```