

Cystic Fibrosis and Pulmonary Function

Cystic fibrosis is a genetic disease that leads to pulmonary complications and ultimately death.

These data represent a subsample of measurements available in a registry database.

Several specific aspects are of interest:

- What is the rate of decline in FEV1?
- Is the time course different for males and females?
- Is the time course different for F508 homozygous subjects?

Davis P.B. (1997) *Journal of Pediatrics*

(Borrowed from Patrick)

Data

```
> data <- read.table( "NewCFkids.dat", header=F )

> cfkids <- data.frame(
  id = data[,1],
  fev1 = data[,2],
  age = data[,3],
  female = as.integer( data[,4]==2 ),
  pseudoA = as.integer( data[,5]==3 ),
  f508 = 3-data[,6],
  pancreat = as.integer( data[,7]==2 ))

> cfkids[1:10,]
   id  fev1   age female pseudoA f508 pancreat
1 100073 113.80  8.452     1      1      2        1
2 100073  98.18  8.783     1      1      2        1
3 100073  98.73  9.785     1      1      2        1
4 100073 101.79 10.538     1      1      2        1
5 100073  98.04 12.329     1      1      2        1
6 100073  94.32 13.306     1      1      2        1
7 100073  95.48 14.418     1      1      2        1
8 100111  96.85 12.515     1      0      0        0
9 100111 101.05 13.103     1      0      0        1
10 100111 100.33 15.105     1      0      0        1
```

- ID = patient id
- FEV1 = percent-predicted forced expiratory volume in 1 second
- AGE = age (years)
- GENDER = sex (1=male, 2=female)
- PSEUDOA = infection with Pseudo Aeruginosa (0=no, 3=yes)
- F508 = genotype (1=homozygous, 2=heterozygous, 3=none)
- PANCREAT = pancreatic enzyme supplementation (0,1=no, 2=yes)

Data

```
> attach(cfkids)

> length(id)
[1] 1513

> length(table(id))
[1] 200

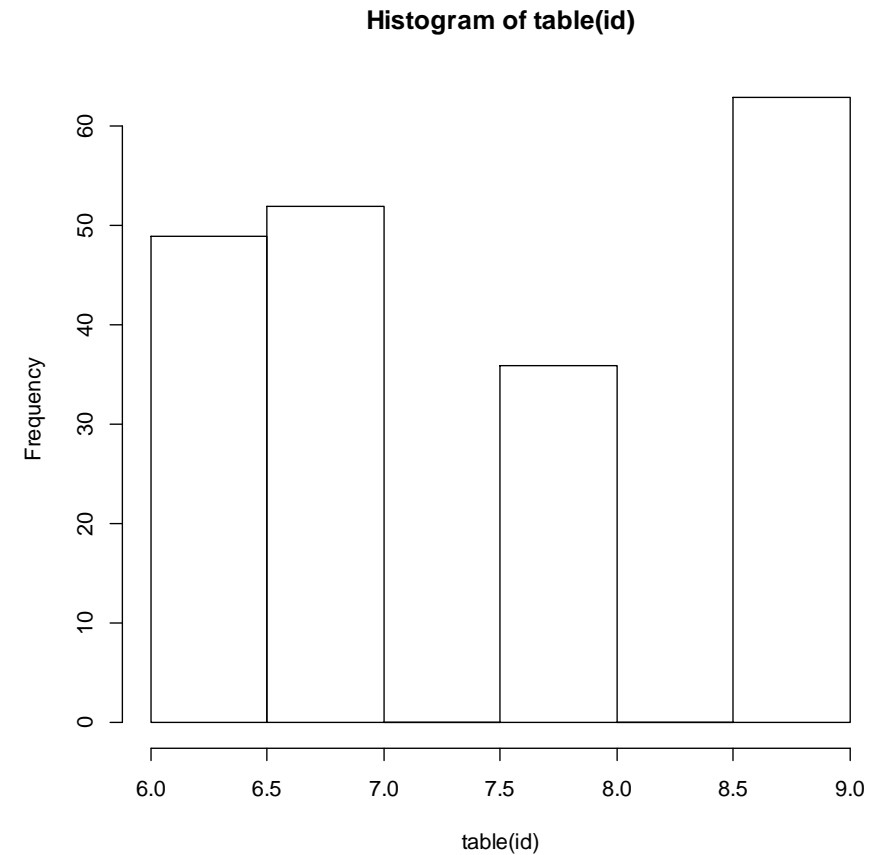
> hist(table(id))

> table(female)
female
 0    1
773 740

> summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 5.002  11.617  15.255  15.776  19.677  29.906

> table(f508)
f508
 0    1    2
165 660 688
```

What's wrong with these?



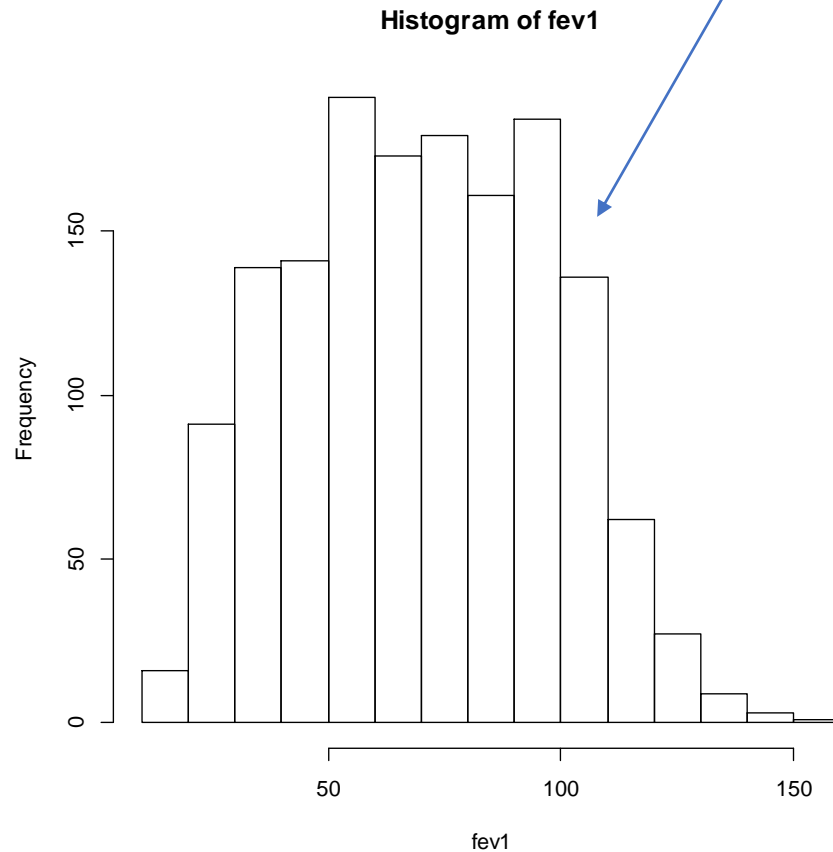
Age at Entry

```
> age_entry = (cfkids %>% group_by(id) %>% top_n(-1, age))$age ## this uses the dplyr package
> cfkids$age0 <- age0 <- rep(age_entry, times = table(id)) ## age at entry
> cfkids$ageL <- ageL <- cfkids$age - cfkids$age0 ## age since entry

> stem(age_entry)
The decimal point is at the |
 5 | 002355666788889
 6 | 0111222334555567789999
 7 | 0001234446778889
 8 | 011223345566899
 9 | 00011244788
10 | 0111113349
11 | 2223446678
12 | 0011122234445557788888999
13 | 01234456
14 | 111245555779
15 | 001223357
16 | 0012347899
17 | 1223567779
18 | 4899
19 | 4
20 | 0123778
21 | 15577
22 | 2459
23 | 001128
```

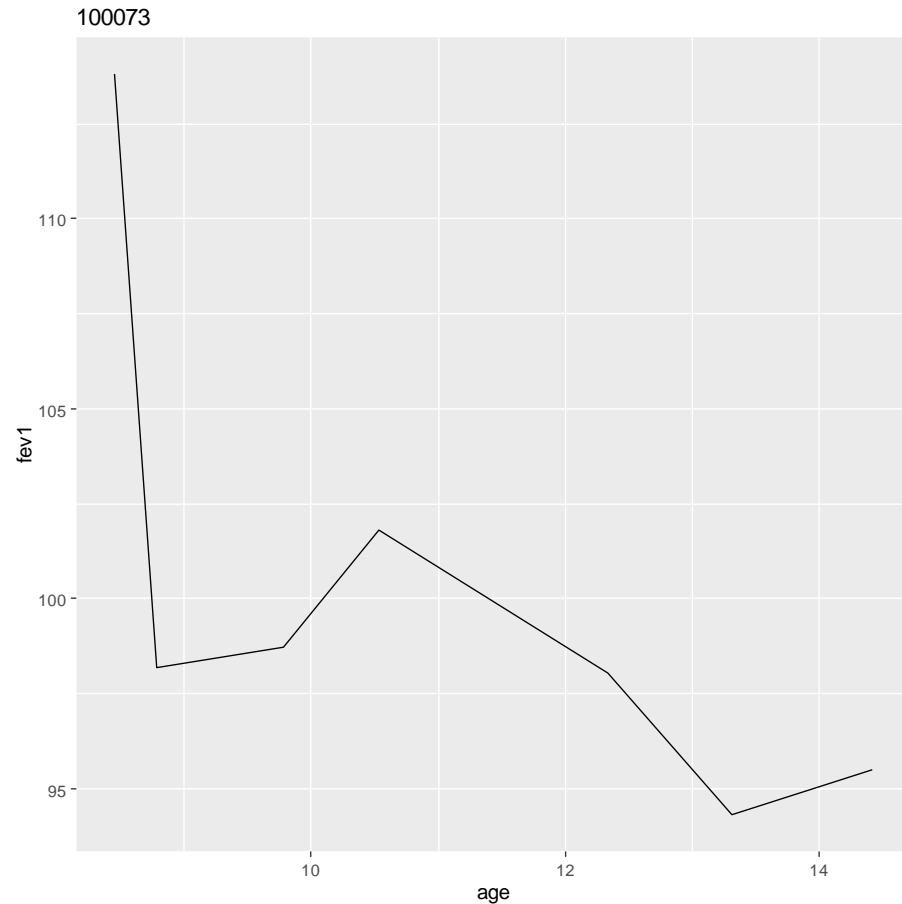
FEV1

Consider transformation *IF* ugly distribution: log, sqrt, square, arctan;
here, it looks ok

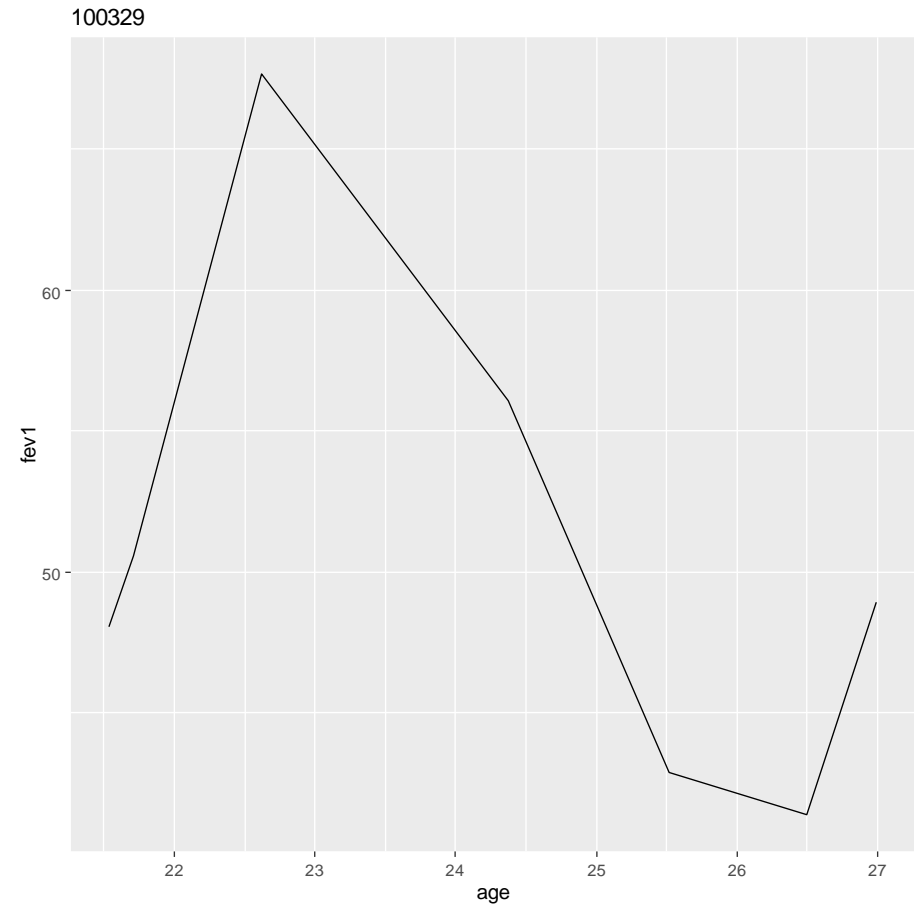


- ID = patient id
 - FEV1 = percent-predicted forced expiratory volume in 1 second
 - AGE = age (years)
 - GENDER = sex (1=male, 2=female)
 - PSEUDOA = infection with Pseudo Aeruginosa (0=no, 3=yes)
 - F508 = genotype (1=homozygous, 2=heterozygous, 3=none)
 - PANCREAT = pancreatic enzyme supplementation (0,1=no, 2=yes)
-
- What is the rate of decline in FEV1?
 - Is the time course different for males and females?
 - Is the time course different for F508 homozygous subjects?

FEV1 over Time

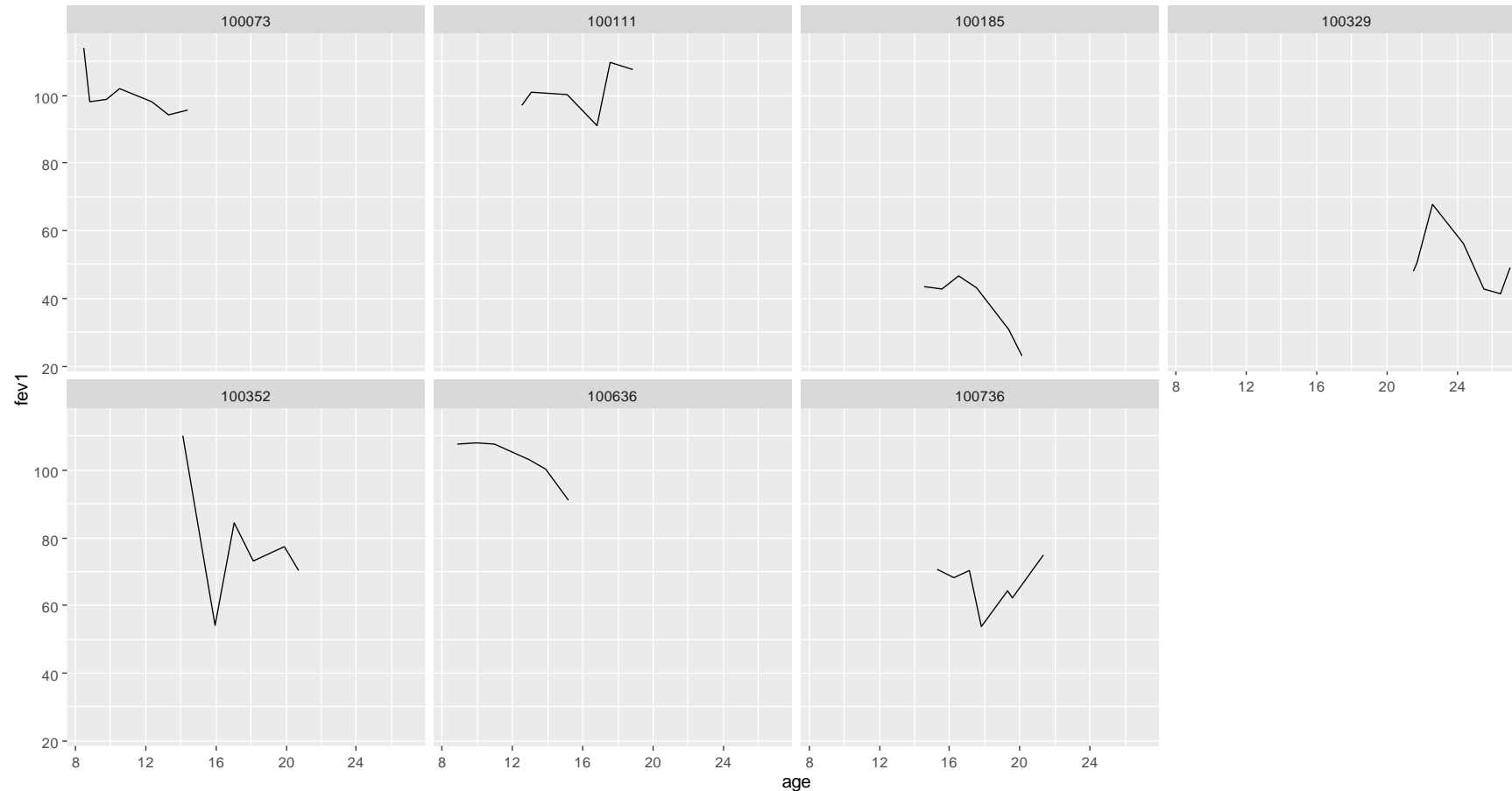


```
> ggplot(data= cfkids %>%filter(id== 100073), aes(x=age,  
y=fev1)) + geom_line()+ggtitle("100073")
```



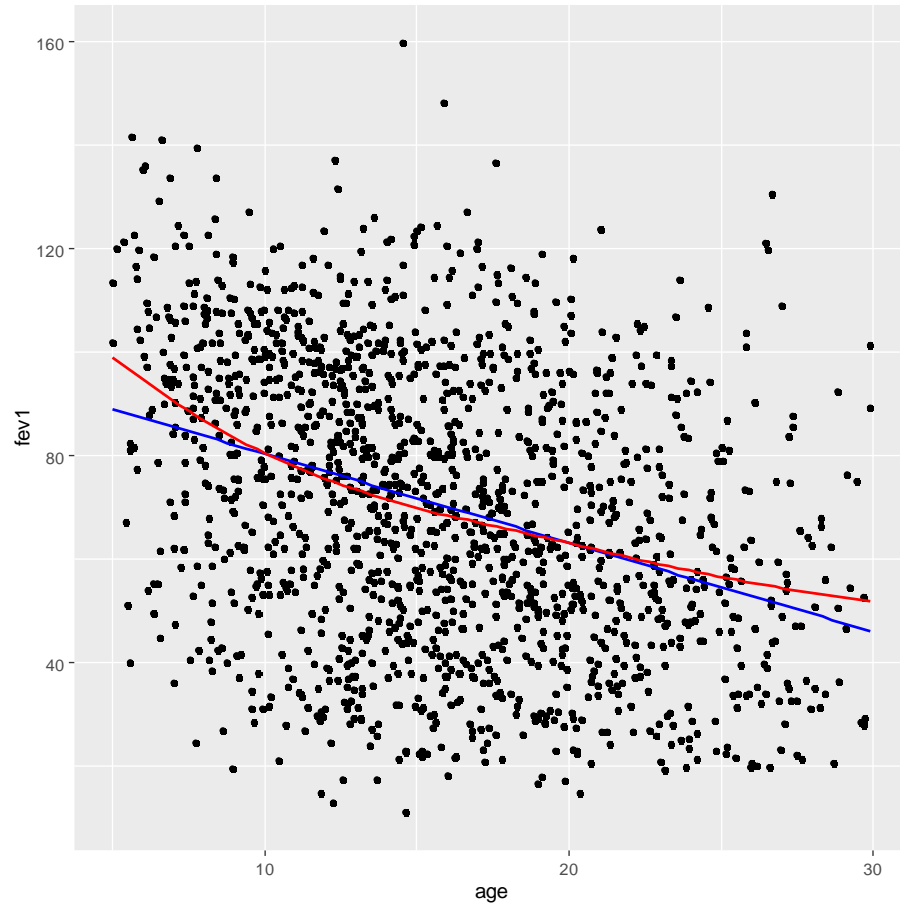
```
ggplot(data= cfkids %>%filter(id== 100329), aes(x=age, y=fev1)) +  
geom_line()+ggtitle("100329")
```

More individual plots

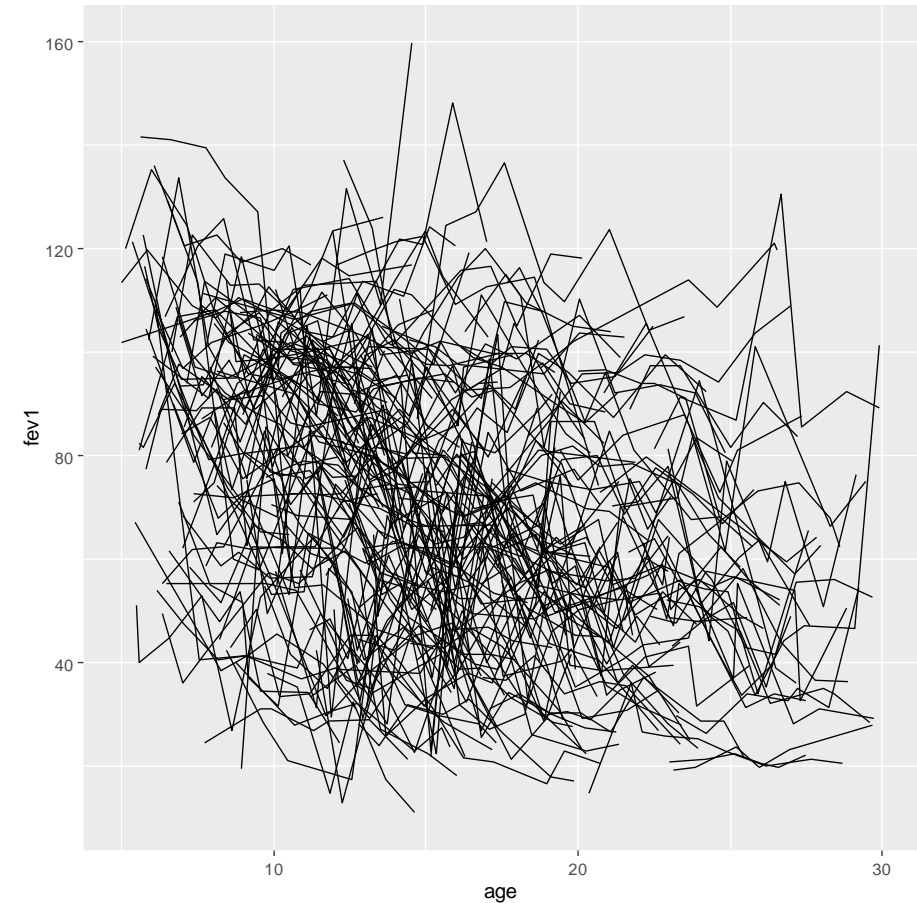


```
> ggplot(data= cfkids[1:45,] , aes(x=age, y=fev1))+geom_line()+facet_wrap(~id, ncol=4)
```

FEV1 over Time

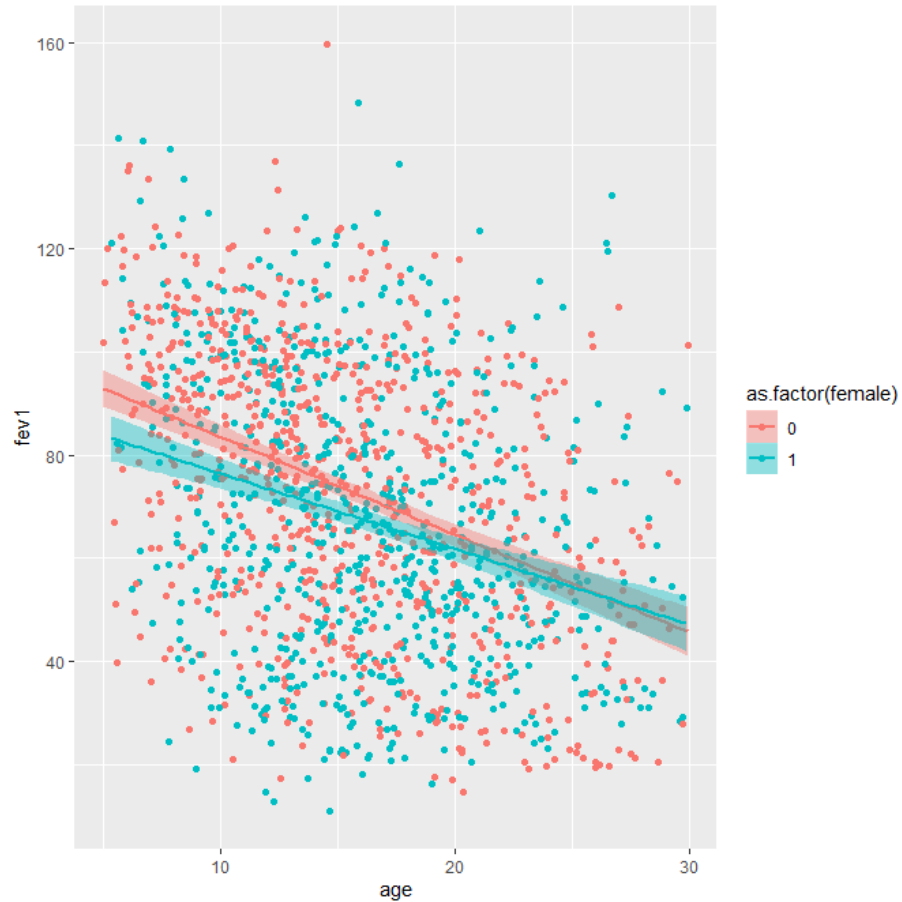


```
ggplot(data=cfkids, aes(x=age, y=fev1)) + geom_point() + geom_smooth(method = lm, color = "blue") + geom_smooth(method = loess, color = "red")
```

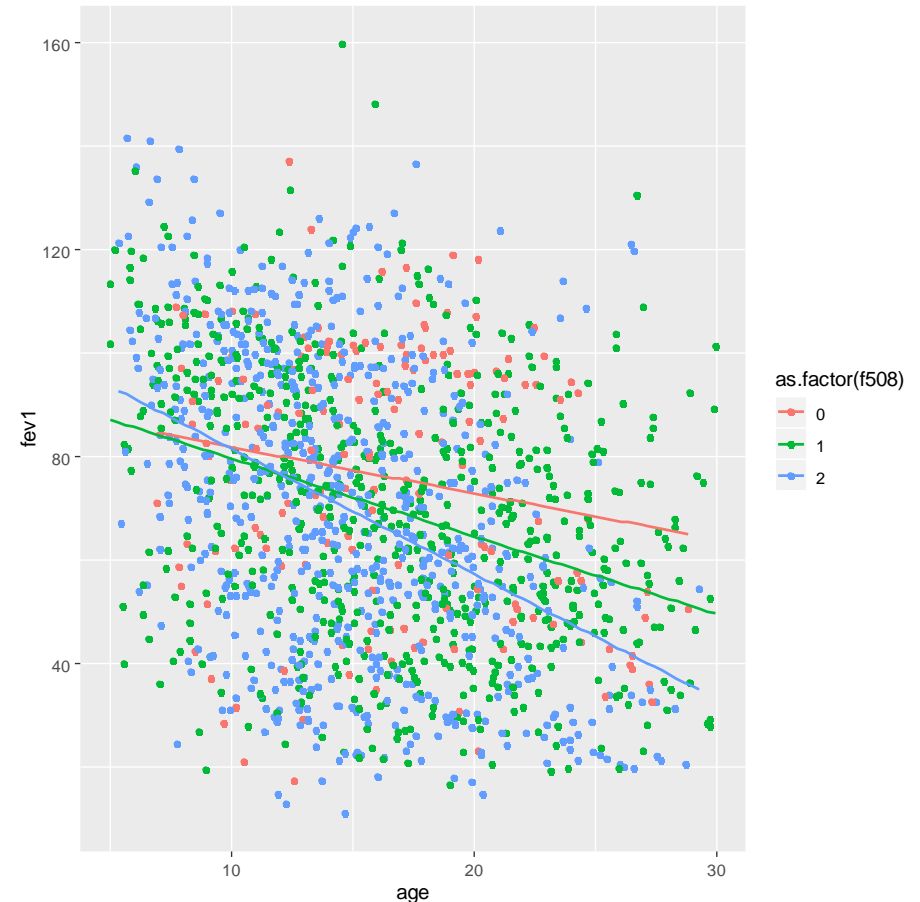


```
ggplot(data=cfkids, aes(x=age, y=fev1, group=id)) + geom_line()
```


Sex and Mutations



```
ggplot(data= cfkids, aes(x=age, y=fev1, color =  
as.factor(female)))+geom_point()+geom_smooth(method=lm,  
aes(fill=as.factor(female)))
```



```
ggplot(data= cfkids, aes(x=age, y=fev1, color =  
as.factor(f508)))+geom_point()+geom_smooth(method=lm,  
aes(fill=as.factor(f508)))
```

General Observations

- Systematic trends in the data:
 - Time
 - Sex
 - F508
- Two different time scales
- How to estimate effects and test?
 - What is the rate of decline in FEV1? (slope of time)
 - Is the time course different for males and females?
 - Is the time course different for F508 homozygous subjects?

Fitting LMMs in R

- Many different software packages for fitting LMMs
 - lmm ← avoid
 - nlme (lme function)
 - lme4 (lmer function)
 - Others...
- SAS is probably better overall

Random Intercept – lme4

```
> summary(lmer(fev1~ age0+ageL+female+(f508==1) + (f508==2)+female*ageL + (f508==1)*ageL + (f508==2)*ageL+ (1|id), REML = T))
Linear mixed model fit by REML ['lmerMod']
Formula: fev1 ~ age0 + ageL + female + (f508 == 1) + (f508 == 2) + female * ageL + (f508 == 1) * ageL + (f508 == 2) * ageL + (1 | id)
```

REML criterion at convergence: 12498.1

Scaled residuals:

Min	1Q	Median	3Q	Max
-4.9430	-0.5282	-0.0037	0.5168	4.2899

Random effects:

Groups	Name	Variance	Std.Dev.
id	(Intercept)	510.7	22.6
Residual		148.8	12.2

Number of obs: 1513, groups: id, 200

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	103.8063	6.7060	15.480
age0	-1.8553	0.3343	-5.550
ageL	-0.5878	0.3931	-1.496
female	-1.1620	3.3979	-0.342
f508 == 1TRUE	-4.2810	5.6110	-0.763
f508 == 2TRUE	-6.7404	5.6426	-1.195
ageL:female	-0.8257	0.2494	-3.311
ageL:f508 == 1TRUE	-0.4877	0.4225	-1.154
ageL:f508 == 2TRUE	-0.6575	0.4214	-1.560

Correlation of Fixed Effects:

	(Intr)	age0	ageL	female	f508=1	f508=2	agL:fm	aL:5=1
age0		-0.630						
ageL	-0.212		0.006					
female	-0.164	-0.088	0.073					
f508==1TRUE	-0.657	-0.002	0.229	-0.021				
f508==2TRUE	-0.725	0.123	0.226	-0.062	0.788			
ageL:female	0.062	-0.005	-0.272	-0.262	0.004	0.011		
aL:508==1TR	0.181	-0.004	-0.856	0.005	-0.267	-0.214	-0.022	
aL:508==2TR	0.179	-0.003	-0.853	0.011	-0.215	-0.266	-0.041	0.805

Random Intercept - nlme

```
> summary(lme(fev1~ age0+ageL+female+(f508==1) + (f508==2)+female*ageL + (f508==1)*ageL + (f508==2)*ageL, random = ~1|id))
```

Linear mixed-effects model fit by REML

Data: NULL

	AIC	BIC	logLik
	12520.08	12578.56	-6249.041

Random effects:

Formula: ~1 | id

(Intercept) Residual

StdDev: 22.59935 12.19734

Fixed effects: fev1 ~ age0 + ageL + female + (f508 == 1) + (f508 == 2) + female * ageL + (f508 == 1) * ageL + (f508 == 2) * ageL

	Value	Std.Error	DF	t-value	p-value
(Intercept)	103.80627	6.706026	1309	15.479550	0.0000
age0	-1.85532	0.334312	195	-5.549663	0.0000
ageL	-0.58782	0.393060	1309	-1.495504	0.1350
female	-1.16203	3.397872	195	-0.341987	0.7327
f508 == 1TRUE	-4.28096	5.611017	195	-0.762956	0.4464
f508 == 2TRUE	-6.74044	5.642572	195	-1.194569	0.2337
ageL:female	-0.82574	0.249427	1309	-3.310541	0.0010
ageL:f508 == 1TRUE	-0.48769	0.422469	1309	-1.154391	0.2486
ageL:f508 == 2TRUE	-0.65745	0.421359	1309	-1.560310	0.1189

Correlation:

	(Intr)	age0	ageL	female	f508=1	f508=2	agL:fm	aL:5=1
age0		-0.630						
ageL		-0.212	0.006					
female		-0.164	-0.088	0.073				
f508 == 1TRUE		-0.657	-0.002	0.229	-0.021			
f508 == 2TRUE		-0.725	0.123	0.226	-0.062	0.788		
ageL:female		0.062	-0.005	-0.272	-0.262	0.004	0.011	
ageL:f508 == 1TRUE		0.181	-0.004	-0.856	0.005	-0.267	-0.214	-0.022
ageL:f508 == 2TRUE		0.179	-0.003	-0.853	0.011	-0.215	-0.266	-0.041

Standardized Within-Group Residuals:

	Min	Q1	Med	Q3	Max
	-4.942954121	-0.528230691	-0.003676248	0.516817787	4.289855016

Number of Observations: 1513

Number of Groups: 200

Random Intercept/Slope – lme4

```
> summary(lmer(fev1~ age0+ageL+female+(f508==1) + (f508==2)+female*ageL + (f508==1)*ageL + (f508==2)*ageL+(1+ageL|id), REML = T))
Linear mixed model fit by REML ['lmerMod']
Formula: fev1 ~ age0 + ageL + female + (f508 == 1) + (f508 == 2) + female * ageL + (f508 == 1) * ageL + (f508 == 2) * ageL + (1 + ageL | id)

REML criterion at convergence: 12389.2

Scaled residuals:
    Min       1Q   Median       3Q      Max
-5.3289 -0.4618  0.0052  0.4784  4.3196

Random effects:
 Groups   Name                Variance Std.Dev. Corr
 id       (Intercept)         512.406   22.636
          ageL                4.513    2.124   -0.16
Residual              118.023   10.864
Number of obs: 1513, groups: id, 200

Fixed effects:
              Estimate Std. Error t value
(Intercept)    104.5192     6.6443   15.731
age0            -1.9105     0.3310    -5.771
ageL            -0.6028     0.5903    -1.021
female          -1.3005     3.3701    -0.386
f508 == 1TRUE   -4.2381     5.5636    -0.762
f508 == 2TRUE   -6.6523     5.5944    -1.189
ageL:female     -0.7624     0.3812    -2.000
ageL:f508 == 1TRUE -0.5001     0.6357    -0.787
ageL:f508 == 2TRUE -0.7459     0.6345    -1.176

Correlation of Fixed Effects:
      (Intr) age0    ageL    female f508=1 f508=2 agL:fm aL:5=1
age0      -0.630
ageL      -0.211  0.002
female    -0.164 -0.088  0.075
f508==1TRUE -0.657 -0.002  0.230 -0.021
f508==2TRUE -0.725  0.122  0.227 -0.062  0.788
ageL:female  0.060 -0.003 -0.279 -0.265  0.005  0.012
aL:508==1TR  0.179 -0.001 -0.850  0.005 -0.269 -0.214 -0.023
aL:508==2TR  0.178  0.000 -0.845  0.012 -0.215 -0.268 -0.046  0.797
```

Random Intercept/Slope - nlme

```
> summary(lme(fev1~ age0+ageL+female+(f508==1) + (f508==2)+female*ageL + (f508==1)*ageL + (f508==2)*ageL, random = ~1+ageL|id))
Linear mixed-effects model fit by REML
Data: NULL
      AIC      BIC    logLik
12415.17 12484.28 -6194.586

Random effects:
Formula: ~1 + ageL | id
Structure: General positive-definite, Log-Cholesky parametrization
      StdDev      Corr
(Intercept) 22.636839 (Intr)
ageL         2.124421 -0.158
Residual     10.863789

Fixed effects: fev1 ~ age0 + ageL + female + (f508 == 1) + (f508 == 2) + female * ageL + (f508 == 1) * ageL + (f508 == 2) * ageL
      Value Std.Error   DF   t-value p-value
(Intercept)  104.51929   6.644377 1309  15.730489  0.0000
age0         -1.91054   0.331049  195  -5.771174  0.0000
ageL         -0.60278   0.590297 1309  -1.021148  0.3074
female       -1.30048   3.370124  195  -0.385886  0.7000
f508 == 1TRUE -4.23809   5.563682  195  -0.761742  0.4471
f508 == 2TRUE -6.65233   5.594495  195  -1.189085  0.2359
ageL:female   -0.76242   0.381214 1309  -1.999991  0.0457
ageL:f508 == 1TRUE -0.50011  0.635753 1309  -0.786642  0.4316
ageL:f508 == 2TRUE -0.74591  0.634493 1309  -1.175599  0.2400

Correlation:
      (Intr) age0    ageL    female f508=1 f508=2 agL:fm aL:5=1
age0      -0.630
ageL      -0.211  0.002
female    -0.164 -0.088  0.075
f508 == 1TRUE -0.657 -0.002  0.230 -0.021
f508 == 2TRUE -0.725  0.122  0.227 -0.062  0.788
ageL:female   0.060 -0.003 -0.279 -0.265  0.005  0.012
ageL:f508 == 1TRUE 0.179 -0.001 -0.850  0.005 -0.269 -0.214 -0.023
ageL:f508 == 2TRUE 0.178  0.000 -0.845  0.012 -0.215 -0.268 -0.046  0.797

Standardized Within-Group Residuals:
      Min      Q1      Med      Q3      Max
-5.32898196 -0.46179233  0.00517294  0.47838439  4.31967651

Number of Observations: 1513
Number of Groups: 200
```

Key Aspects of Output

- Random Effects (each has mean zero, so only variance)
- Fixed effects: lmer does not give p-values
- (Standardized) Residuals
- AIC/BIC
- lmer does not by default output p-values: lots of caveats with p-values in mixed models
 - Can use anova function to compare models (LRT)
 - lmerTest also will spit out p-values

anova with lmer

```
> mod0 = lmer(fev1~ age0+ageL+female+(f508==1) + (f508==2)+ (f508==1)*ageL + (f508==2)*ageL+(1+ageL|id), REML = T)
> mod1 = lmer(fev1~ age0+ageL+female+(f508==1) + (f508==2)+female*ageL + (f508==1)*ageL + (f508==2)*ageL+(1+ageL|id), REML = T)
> anova(mod1,mod0)
```

refitting model(s) with ML (instead of REML)

Data: NULL

Models:

mod0: fev1 ~ age0 + ageL + female + (f508 == 1) + (f508 == 2) + (f508 ==

mod0: 1) * ageL + (f508 == 2) * ageL + (1 + ageL | id)

mod1: fev1 ~ age0 + ageL + female + (f508 == 1) + (f508 == 2) + female *

mod1: ageL + (f508 == 1) * ageL + (f508 == 2) * ageL + (1 + ageL |

mod1: id)

	Df	AIC	BIC	logLik	deviance	Chisq	Chi	Df	Pr(>Chisq)
mod0	12	12432	12496	-6204.2	12408				
mod1	13	12430	12500	-6202.2	12404	4.0441		1	0.04433 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Choosing a Model

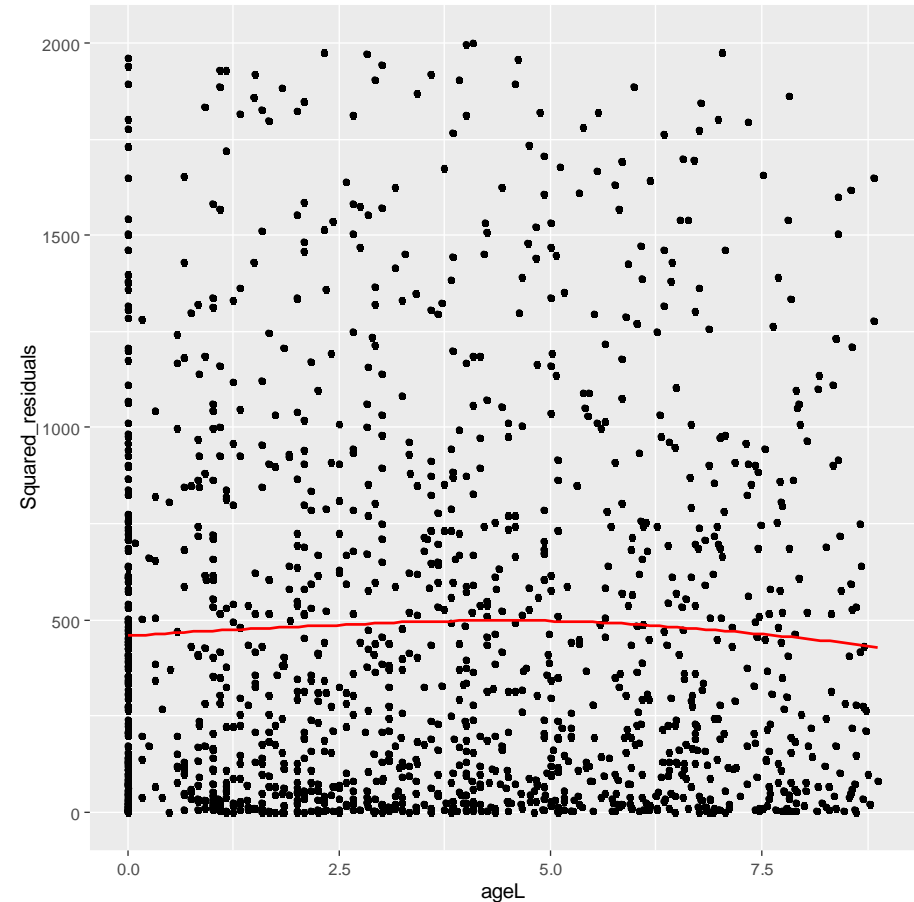
- For LMM, important to get appropriate mean AND covariance structure
- Overly simplistic models lead to invalid inference
- Overly complicated models lead to poor efficiency
- General guidelines:
 - get fixed effects
 - get random effects
 - get covariance structure
 - Reduce the model

Choosing Initial Mean Model

- Start with bigger model than necessary
- Saturated model: kitchen sink
- Not seeing much need for quadratic effects over time
- Possibly some heterogeneity by sex and F508

Choosing Random Effects

- Usually only random effects for variables that change
 - For non time dependent variables, subject specific changes are absorbed into random intercept
- Add variables that are included as fixed effects (require **b** to have mean zero)
- Prefer hierarchical models
- (squared) OLS residuals are helpful
 - Constant variability over time implies stationarity
- Err on side of bigger models
 - But may require estimation of lots of effects which can be unstable



Selection of Residual Covariance Structure

- Generally challenging problem
- Basic idea:
 - Fit range of models
 - Compare models via AIC/BIC (sometimes LRT is possible)
- For unbalanced data with lots of measurements, typically select fairly simplistic models which leads to simple models for \mathbf{V}

Model Reduction

- Backward selection
- Beware of LRT for random effects (not the usual null distributions)
- Can sequentially test for fixed effects
- Inference post model selection?