LMM: Linear Mixed Models and FEV1 Decline

- We can use linear mixed models to assess the evidence for differences in the rate of decline for subgroups defined by covariates.

- **S+** / **R** has a function `lme()`.

- **SAS** has the `MIXED` procedure.

SAS Program:

```
options linesize=80 pagesize=60;

data cfkids;
  infile 'NewCFkids-SAS.data';
  input id fev1 age female pseudoA f508 panc age0 ageL;
run;

data cfkids; set cfkids;
  f508_1 = 0;
  if f508=1 then f508_1 = 1;
  f508_2 = 0;
  if f508=2 then f508_2 = 1;
run;

proc mixed data=cfkids method=reml;
        class id;
        model fev1 = age0 ageL female f508_1 f508_2 female*ageL
                     f508_1*ageL f508_2*ageL / s;
        repeated / type=cs  subject=id;
run;
```

```
proc mixed data=cfkids method=reml;
        class id;
        model fev1 = age0 ageL female f508_1 f508_2 female*ageL
                     f508_1*ageL f508_2*ageL / s;
        random intercept / type=un subject=id g;
run;

proc mixed data=cfkids method=reml;
        class id;
        model fev1 = age0 ageL female f508_1 f508_2 female*ageL
                     f508_1*ageL f508_2*ageL / s;
        random intercept ageL / type=un subject=id g;
run;

proc mixed data=cfkids method=reml;
        class id;
        model fev1 = age0 ageL female f508_1 f508_2 female*ageL
                     f508_1*ageL f508_2*ageL / s;
        random intercept / type=un subject=id g;
        repeated / type=sp(pow)(ageL) subject=id;
run;

proc mixed data=cfkids method=reml;
        class id;
        model fev1 = age0 ageL female f508_1 f508_2 female*ageL
                     f508_1*ageL f508_2*ageL / s;
        random intercept / type=un subject=id g;
        repeated / type=sp(pow)(ageL) subject=id local;
```

```
run;

proc mixed data=cfkids method=reml;
        class id;
        model fev1 = age0 ageL female f508_1 f508_2 female*ageL
                     f508_1*ageL f508_2*ageL / s;
        random intercept ageL / type=un subject=id g;
        repeated / type=sp(pow)(ageL) subject=id;
run;

proc mixed data=cfkids method=reml;
        class id;
        model fev1 = age0 ageL female f508_1 f508_2 female*ageL
                     f508_1*ageL f508_2*ageL / s;
        random intercept ageL / type=un subject=id g;
        repeated / type=sp(pow)(ageL) subject=id local;
run;
```

Comments on Syntax and Model

$\boxed{\textbf{Model}:}$ $Y_{ij} = \mu_{ij} + \overbrace{\boxed{b_{i,0}}}^{\texttt{(a)}} + \overbrace{\boxed{e_{ij}}}^{\texttt{(b)}}$

- SYNTAX: `random intercept / type=un subject=id g;`
- DESCRIPTION: The `random` statement is used to declare random effects. After the forward slash an ID variable must be specified using `subject =` *your-id-variable-name*. The option `type=un` is not necessary here (intercept only). The option g simply asks that the output display the random effects covariance matrix (we've called this $\mathbf{D}$) be written out as a matrix.
- NOTE: The default is to include the errors `(b)` above as independent errors with a constant variance.

Comments on Syntax and Model

$$\boxed{\textbf{Model}:}\ Y_{ij} = \mu_{ij} + \overbrace{\boxed{b_{i,0} + b_{i,1}\texttt{ageL}}}^{\texttt{(a)}} + \overbrace{\boxed{e_{ij}}}^{\texttt{(b)}}$$

- SYNTAX: `random intercept ageL / type=un subject=id g;`
- DESCRIPTION: The `random` statement is used to declare random effects. The option `type=un` asks that the variances and the covariance of random effects be an arbitrary (unstructured) matrix we've called this $\mathbf{D}$. One could specify other options such as asking for independent random effects, but for linear mixed models this isn't usually of interest.
- NOTE: The default is to include the errors `(b)` above as independent errors with a constant variance.

Comments on Syntax and Model

$$\boxed{\textbf{Model}:}\ Y_{ij} = \mu_{ij} + \overbrace{\boxed{b_{i,0}}}^{\texttt{(a)}} + \overbrace{\boxed{W_i(t_{ij})}}^{\texttt{(b)}}$$

- SYNTAX for (a): `random intercept / type=un subject=id g;`
- SYNTAX for (b): `repeated / type=sp(pow)(ageL) subject=id;`
- DESCRIPTION: The use of the `repeated` command allows one to relax the assumption that within-subject errors are independent. To include a component of serial correlation (autocorrelated errors) we can use commands like `type = ar(1)` which assume that observations j and k for a subject have within-subject errors with covariance $\sigma^2\rho^{|j-k|}$. When observations are not equally spaced in time the command `type=sp(pow)(ageL)` allows a covariance $\sigma^2\rho^{d_{jk}}$, where the distance is computed as $|t_{ij} - t_{ik}|$, and the argument ageL is specifying the time variable to compute distance.

Comments on Syntax and Model

$$\boxed{\textbf{Model}:}\ Y_{ij} = \mu_{ij} + \overbrace{\boxed{b_{i,0}}}^{\texttt{(a)}} + \overbrace{\boxed{W_i(t_{ij})}}^{\texttt{(b)}} + \overbrace{\boxed{e_{ij}}}^{\texttt{(c)}}$$

- SYNTAX for (a): `random intercept / type=un subject=id g;`
- SYNTAX for (b) and (c):

  `repeated / type=sp(pow)(ageL) subject=id local;`
- DESCRIPTION: This is similar to the previous model, but now the option `local` asks for the inclusion of the measurement errors, e_{ij}, which are assumed to be independent. Thus the within-subject errors for this model have both a serial component, and a pure noise component.

SAS Fit 1 **Random Intercepts + Slopes**

```
                The MIXED Procedure

             Class Level Information

   Class      Levels  Values

   ID            200  100073 100111 100185 100329
                      100352 100636 100736 100815
                        .      .      .     .
                        .      .      .     .
                        .      .      .     .

          REML Estimation Iteration History

  Iteration  Evaluations     Objective     Criterion

          0            1  11288.083105
          1            2  9625.5053208    0.00009459
          2            1  9625.0138752    0.00000165
          3            1  9625.0058252    0.00000000

              Convergence criteria met.
```

G Matrix

Effect	ID	Row	COL1	COL2
INTERCEPT	100073	1	512.41416519	-7.62541488
AGEL	100073	2	-7.62541488	4.51229421

Covariance Parameter Estimates (REML)

Cov Parm	Subject	Estimate
UN(1,1)	ID	512.41416519
UN(2,1)	ID	-7.62541488
UN(2,2)	ID	4.51229421
Residual		118.02545375

Model Fitting Information for FEV1

Description	Value
Observations	1513.000
Res Log Likelihood	-6194.59
Akaike's Information Criterion	-6198.59
Schwarz's Bayesian Criterion	-6209.22
-2 Res Log Likelihood	12389.17
Null Model LRT Chi-Square	1663.077
Null Model LRT DF	3.0000
Null Model LRT P-Value	0.0000

Solution for Fixed Effects

Effect	Estimate	Std Error	DF	t	Pr > |t|
INTERCEPT	104.51880927	6.64425792	195	15.73	0.0001
AGE0	-1.91050551	0.33103925	1113	-5.77	0.0001
AGEL	-0.60278138	0.59026446	196	-1.02	0.3084
FEMALE	-1.30051066	3.37009028	1113	-0.39	0.6996
F508_1	-4.23810256	5.56362771	1113	-0.76	0.4464
F508_2	-6.65228283	5.59443908	1113	-1.19	0.2347
AGEL*FEMALE	-0.76242845	0.38119258	1113	-2.00	0.0457
AGEL*F508_1	-0.50010405	0.63571767	1113	-0.79	0.4316
AGEL*F508_2	-0.74589389	0.63445750	1113	-1.18	0.2400

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
AGE0	1	1113	33.31	0.0001
AGEL	1	196	1.04	0.3084
FEMALE	1	1113	0.15	0.6996
F508_1	1	1113	0.58	0.4464
F508_2	1	1113	1.41	0.2347
AGEL*FEMALE	1	1113	4.00	0.0457
AGEL*F508_1	1	1113	0.62	0.4316
AGEL*F508_2	1	1113	1.38	0.2400

SAS Fit 2 Random Intercepts + AR errors

```
                    The MIXED Procedure

            REML Estimation Iteration History

 Iteration  Evaluations      Objective     Criterion

         0            1   11288.083105
         1            2   9842.5776314  152.37610437
         2            2   9666.7258252    0.01014002
         3            1   9612.7982174    0.00047006
         4            2   9610.8212742    0.00002116
         5            1   9610.7155276    0.00000009
         6            1   9610.7151109    0.00000000

                Convergence criteria met.
```

G Matrix

Effect	ID	Row	COL1
INTERCEPT	100073	1	483.46131336

Covariance Parameter Estimates (REML)

Cov Parm	Subject	Estimate
UN(1,1)	ID	483.46131336
SP(POW)	ID	0.33940770
Residual		172.66675040

Model Fitting Information for FEV1

Description	Value
Observations	1513.000
Res Log Likelihood	-6187.44
Akaike's Information Criterion	-6190.44
Schwarz's Bayesian Criterion	-6198.41
-2 Res Log Likelihood	12374.88
Null Model LRT Chi-Square	1677.368
Null Model LRT DF	2.0000
Null Model LRT P-Value	0.0000

Solution for Fixed Effects

Effect	Estimate	Std Error	DF	t	Pr > |t|
INTERCEPT	104.19222115	6.76123647	195	15.41	0.0001
AGE0	-1.85599115	0.33226611	1309	-5.59	0.0001
AGEL	-0.58899693	0.50099307	1309	-1.18	0.2399
FEMALE	-1.25755036	3.46074067	1309	-0.36	0.7164
F508_1	-4.70904830	5.71244977	1309	-0.82	0.4099
F508_2	-6.76205358	5.74144814	1309	-1.18	0.2391
AGEL*FEMALE	-0.85649279	0.32240863	1309	-2.66	0.0080
AGEL*F508_1	-0.42254215	0.54066747	1309	-0.78	0.4346
AGEL*F508_2	-0.70366210	0.53899040	1309	-1.31	0.1919

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
AGE0	1	1309	31.20	0.0001
AGEL	1	1309	1.38	0.2399
FEMALE	1	1309	0.13	0.7164
F508_1	1	1309	0.68	0.4099
F508_2	1	1309	1.39	0.2391
AGEL*FEMALE	1	1309	7.06	0.0080
AGEL*F508_1	1	1309	0.61	0.4346
AGEL*F508_2	1	1309	1.70	0.1919

$\boxed{\text{SAS Fit 3}}$ Random Intercepts + AR errors + meas. error

```
                  The MIXED Procedure

           REML Estimation Iteration History

 Iteration  Evaluations     Objective     Criterion

         0            1  11288.083105
         1            2  9777.1891608   83.13453204
         2            2  9632.0064884    8.05381107
         .
        19            1  9579.7150151    0.00000001

             Convergence criteria met.
```

G Matrix

Effect	ID	Row	COL1
INTERCEPT	100073	1	390.13143397

Covariance Parameter Estimates (REML)

Cov Parm	Subject	Estimate
UN(1,1)	ID	390.13143397
Variance	ID	191.19638318
SP(POW)	ID	0.83528160
Residual		72.87412247

Model Fitting Information for FEV1

Description	Value
Observations	1513.000
Res Log Likelihood	-6171.94
Akaike's Information Criterion	-6175.94
Schwarz's Bayesian Criterion	-6186.57
-2 Res Log Likelihood	12343.88
Null Model LRT Chi-Square	1708.368
Null Model LRT DF	3.0000
Null Model LRT P-Value	0.0000

Solution for Fixed Effects

Effect	Estimate	Std Error	DF	t	Pr > |t|
INTERCEPT	104.77553777	6.82547939	195	15.35	0.0001
AGE0	-1.87512100	0.33131471	1309	-5.66	0.0001
AGEL	-0.71223608	0.58209781	1309	-1.22	0.2213
FEMALE	-1.20683124	3.52025659	1309	-0.34	0.7318
F508_1	-5.02332453	5.80876288	1309	-0.86	0.3873
F508_2	-7.10525357	5.83759113	1309	-1.22	0.2238
AGEL*FEMALE	-0.82492646	0.37579975	1309	-2.20	0.0283
AGEL*F508_1	-0.31356545	0.62802526	1309	-0.50	0.6177
AGEL*F508_2	-0.58034315	0.62637981	1309	-0.93	0.3544

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
AGE0	1	1309	32.03	0.0001
AGEL	1	1309	1.50	0.2213
FEMALE	1	1309	0.12	0.7318
F508_1	1	1309	0.75	0.3873
F508_2	1	1309	1.48	0.2238
AGEL*FEMALE	1	1309	4.82	0.0283
AGEL*F508_1	1	1309	0.25	0.6177
AGEL*F508_2	1	1309	0.86	0.3544

Likelihood Summaries

Model	q	$\log L$	AIC	BIC
int + e	2	-6249.0	-6251.0	-6256.4
int+slope + e	4	-6194.6	-6198.6	-6209.2
int + AR	3	-6187.4	-6190.4	-6198.4
int + AR + e	4	-6171.9	-6175.9	-6186.6
int+slope + AR	5	-6174.2	-6179.2	-6192.5
int+slope + AR + e	6	-6170.0	-6176.0	-6192.0

$$\text{AIC} = \log L - q$$

$$\text{BIC} = \log L - q \cdot \log(\sum n_i)/2$$

GEE: Analysis of Decline by Gender & Genotype

- GEE (generalized estimating equations) is a regression method that provides fitted regression coefficients <u>and</u> standard error estimates that account for the correlation in the repeated measurements.

- GEE allows us to focus on the mean model (regression equations).

- GEE requires us to choose a "working correlation" structure. This will impact how data from different individuals are weighted when combined to estimate regression coefficients.

- GEE returns a model-based variance estimate (naive) and an empirical variance estimate (robust).

GEE Analysis

$\boxed{\star}$ One key idea in GEE is the magic "robust variance". This can be obtained in STATA using the `cluster( )` option with almost any regression routine. (aka Huber/White correction; sandwich variance)

- GEE can be used with general correlated data, including clustered data that arises in health services research. (count and binary too)

- The **Model-based** variance is valid if you have selected an appropriate correlation model.

- The **Empirical** variance is valid no matter which correlation model you have selected! However, it only works well if you have enough independent clusters / subjects (ie. 40+).

SAS Program:

```
options linesize=80 pagesize=60;

data cfkids;
  infile 'NewCFkids-SAS.data';
  input id fev1 age female pseudoA f508 panc age0 ageL;
run;

data cfkids; set cfkids;
  f508_1 = 0;
  if f508=1 then f508_1 = 1;
  f508_2 = 0;
  if f508=2 then f508_2 = 1;
run;

proc genmod data=cfkids;
        class id;
model fev1 = age0 ageL female f508_1 f508_2 female*ageL
                     f508_1*ageL f508_2*ageL /
              link=identity dist=normal;
repeated subject=id / type=exch corrw;
run;
```

SAS Fit 1 **GEE Analysis**

The GENMOD Procedure

Analysis Of Initial Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	104.8027	3.9223	97.1151	112.4903	713.93	<.0001
age0	1	-1.8614	0.1358	-2.1276	-1.5953	187.85	<.0001
ageL	1	-0.8910	0.7908	-2.4409	0.6589	1.27	0.2599
female	1	-1.1330	2.2403	-5.5240	3.2580	0.26	0.6131
f508_1	1	-5.2796	3.7594	-12.6479	2.0887	1.97	0.1602
f508_2	1	-8.7756	3.7700	-16.1647	-1.3866	5.42	0.0199
ageL*female	1	-0.8188	0.5030	-1.8046	0.1670	2.65	0.1035
ageL*f508_1	1	-0.1624	0.8506	-1.8295	1.5048	0.04	0.8486
ageL*f508_2	1	-0.0393	0.8478	-1.7010	1.6224	0.00	0.9630
Scale	1	25.2217	0.4585	24.3389	26.1366		

NOTE: The scale parameter was estimated by maximum likelihood.

GEE Model Information

Correlation Structure	Exchangeable
Subject Effect	id (200 levels)
Number of Clusters	200
Correlation Matrix Dimension	9
Maximum Cluster Size	9
Minimum Cluster Size	6

Working Correlation Matrix

	Col1	Col2	Col3	Col4	Col5
Row1	1.0000	0.7483	0.7483	0.7483	0.7483
Row2	0.7483	1.0000	0.7483	0.7483	0.7483
Row3	0.7483	0.7483	1.0000	0.7483	0.7483
Row4	0.7483	0.7483	0.7483	1.0000	0.7483
Row5	0.7483	0.7483	0.7483	0.7483	1.0000
Row6	0.7483	0.7483	0.7483	0.7483	0.7483
Row7	0.7483	0.7483	0.7483	0.7483	0.7483
Row8	0.7483	0.7483	0.7483	0.7483	0.7483
Row9	0.7483	0.7483	0.7483	0.7483	0.7483

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	103.8135	7.2401	89.6232	118.0039	14.34	<.0001
age0	-1.8553	0.3202	-2.4829	-1.2278	-5.79	<.0001
ageL	-0.5899	0.5232	-1.6153	0.4355	-1.13	0.2595
female	-1.1619	3.3322	-7.6930	5.3691	-0.35	0.7273
f508_1	-4.2885	5.8923	-15.8371	7.2601	-0.73	0.4667
f508_2	-6.7549	5.9833	-18.4820	4.9721	-1.13	0.2589
ageL*female	-0.8257	0.3706	-1.5521	-0.0993	-2.23	0.0259
ageL*f508_1	-0.4854	0.5783	-1.6189	0.6481	-0.84	0.4013
ageL*f508_2	-0.6533	0.5679	-1.7663	0.4598	-1.15	0.2500

GEE Limitations

- Focus on the mean structure – but doesn't give summaries of heterogeneity among subjects.
  - ie: How much variation is there in the rate of decline?

- Requires missing data (drop-out) to not depend on the measurements.
  - ie: Invalid if "sick" kids quit study.

- No likelihood to compare models (LR tests, AIC).

$\boxed{\star}$ There are more sophisticated models that address these issues (for continuous response data).

- Linear Mixed Models (SAS PROC MIXED).