Final Project

- Option #1: Research Paper
  - Feb 12: Indicate choice
  - Feb 12: Choose your paper
  - 10 minute presentation (Feb 26)
  - Watch 2 other videos and comment (Mar 5)
  - max 2 page report (Mar 5)
- Option #2: Simulation Study
  - max 2 page report (Mar 5)
  - Simulation code (Mar 5)
  - Simulation can take a couple of hours to run (don't leave to last minute!)

Informative or Random dropout

- Once a subject's tumor diameter exceeds 1.2cm they are removed from the study MAR
- In a breast feeding study, women who have stopped breast feeding are less likely to fill out follow-up surveys MNAR
- In a smoking cessation study, those that have started smoking are less likely to report their status MNAR

Recap of Day 1

Recap of Day 1

- Joint models are needed when the marker and survival processes are correlated
  - Interested in survival outcome but want to incorporate the effects of a marker measured with error
  - Interested in a longitudinal outcome but have to deal with informative dropout
- Components of a joint model
  - Longitudinal submodel
  - Survival submodel
- Both components are conditional on random effects ("shared random effects" models)

Longitudinal Data Analysis

Day 2

- Review of linear mixed-effects models

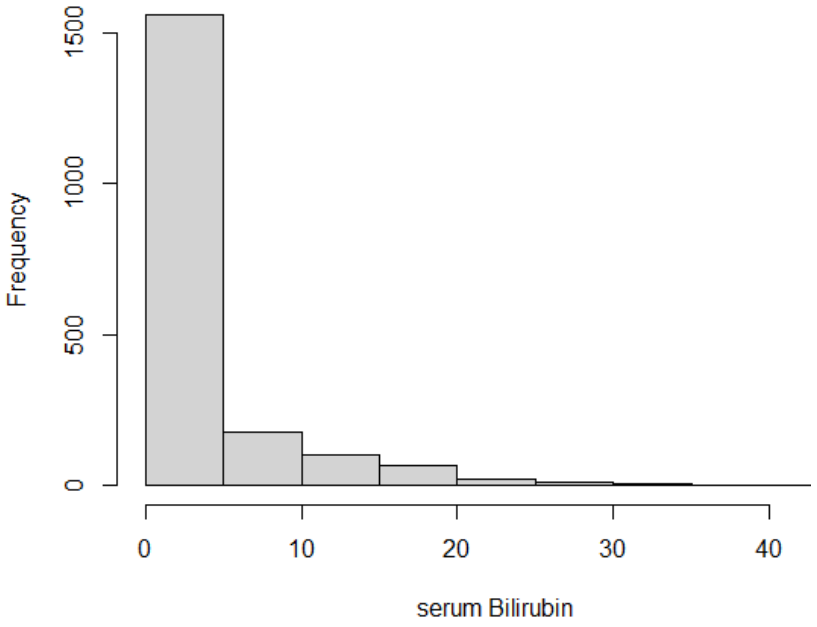
Breakout Session #2 (10 min)

1. What are two advantages of linear mixed effects models for modeling longitudinal data?
2. In what situations would you use REML (restricted maximum likelihood) to estimate parameters for a linear mixed model and why?
3. Why or why should we not consider an unstructured covariance matrix in our PBC data example?
4. What is your current TV show/movie recommendation?

Features of Longitudinal Data

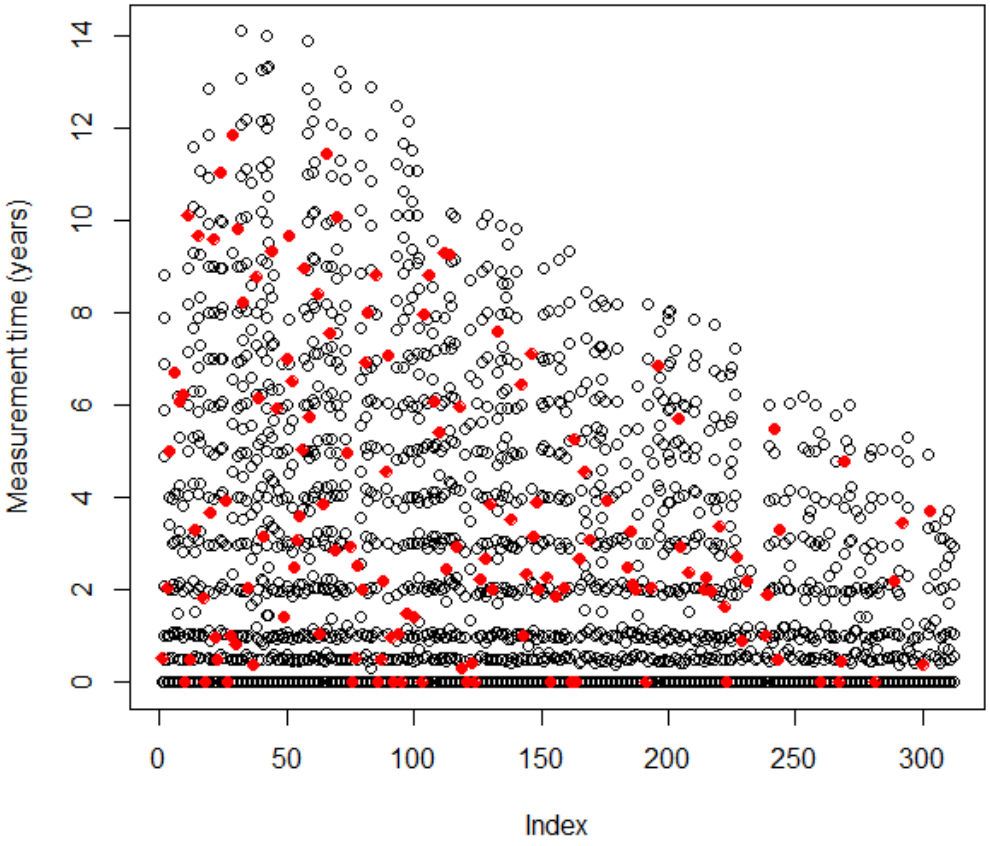
- Repeated evaluations of the same outcome in a subject over time
  - Serum bilirubin in PBC patients
- Interested in investigating:
  1. How treatment means differ at specific time points (**cross-sectional effect**)
  2. How treatment means or differences between treatment means change over time (**longitudinal effect**)
- Measures on the same subject are expected to be (positively) correlated
  - Need to use appropriate statistical methods to account for this

Example: PBC data



id	years	status2	drug	age	year	logBilir
1	1.095170	1	D-penicil	58.76684	0.0000000	2.67414865
1	1.095170	1	D-penicil	58.76684	0.5256817	3.05870707
2	14.152338	0	D-penicil	56.44782	0.0000000	0.09531018
2	14.152338	0	D-penicil	56.44782	0.4983025	-0.22314355
2	14.152338	0	D-penicil	56.44782	0.9993429	0.00000000
2	14.152338	0	D-penicil	56.44782	2.1027270	0.64185389
2	14.152338	0	D-penicil	56.44782	4.9008871	0.95551145
2	14.152338	0	D-penicil	56.44782	5.8892783	1.28093385
2	14.152338	0	D-penicil	56.44782	6.8858833	1.43508453
2	14.152338	0	D-penicil	56.44782	7.8907020	1.28093385
2	14.152338	0	D-penicil	56.44782	8.8325485	1.52605630
3	2.770781	1	D-penicil	70.07447	0.0000000	0.33647224
3	2.770781	1	D-penicil	70.07447	0.4818749	0.09531018
3	2.770781	1	D-penicil	70.07447	0.9966050	0.40546511
3	2.770781	1	D-penicil	70.07447	2.0342788	0.58778666
4	5.270507	1	D-penicil	54.74209	0.0000000	0.58778666
4	5.270507	1	D-penicil	54.74209	0.5147300	0.47000363
4	5.270507	1	D-penicil	54.74209	1.0185084	0.53062825
4	5.270507	1	D-penicil	54.74209	1.9959479	1.16315081
4	5.270507	1	D-penicil	54.74209	3.4333589	1.30833282

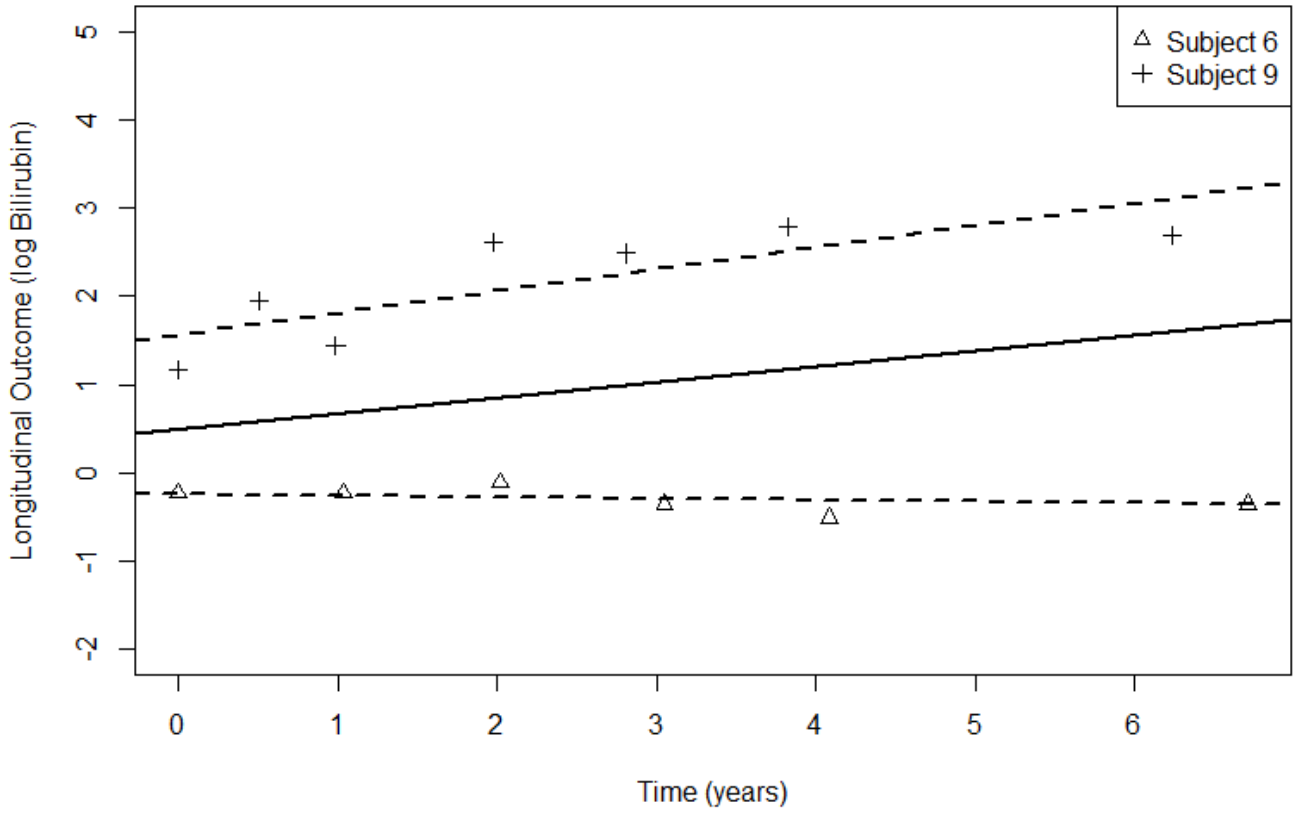
Example: PBC data




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   5.000   6.234   9.000  16.000
```

Linear Mixed Model

- Each individual in the population has their own subject-specific mean response profile over time



Linear Mixed Model

- The evolution of each subject's response in time can be described by a linear model

$$y_{ij} = \tilde{\beta}_{i0} + \tilde{\beta}_{i1} t_{ij} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

  - y_{ij} is the jth response for the ith subject
  - $\tilde{\beta}_{i0}$ is the intercept and $\tilde{\beta}_{i1}$ is the slope for subject i

- **Assume** that since subjects are sampled from a population, subject-specific coefficients are sampled from a population of regression coefficients

$$\tilde{\beta}_i \sim \mathcal{N}(\beta, D)$$

where $\beta = (\beta_0, \beta_1)'$ and D is the variance-covariance matrix

Linear Mixed Model

- We can reformulate the model as

$$y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t_{ij} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

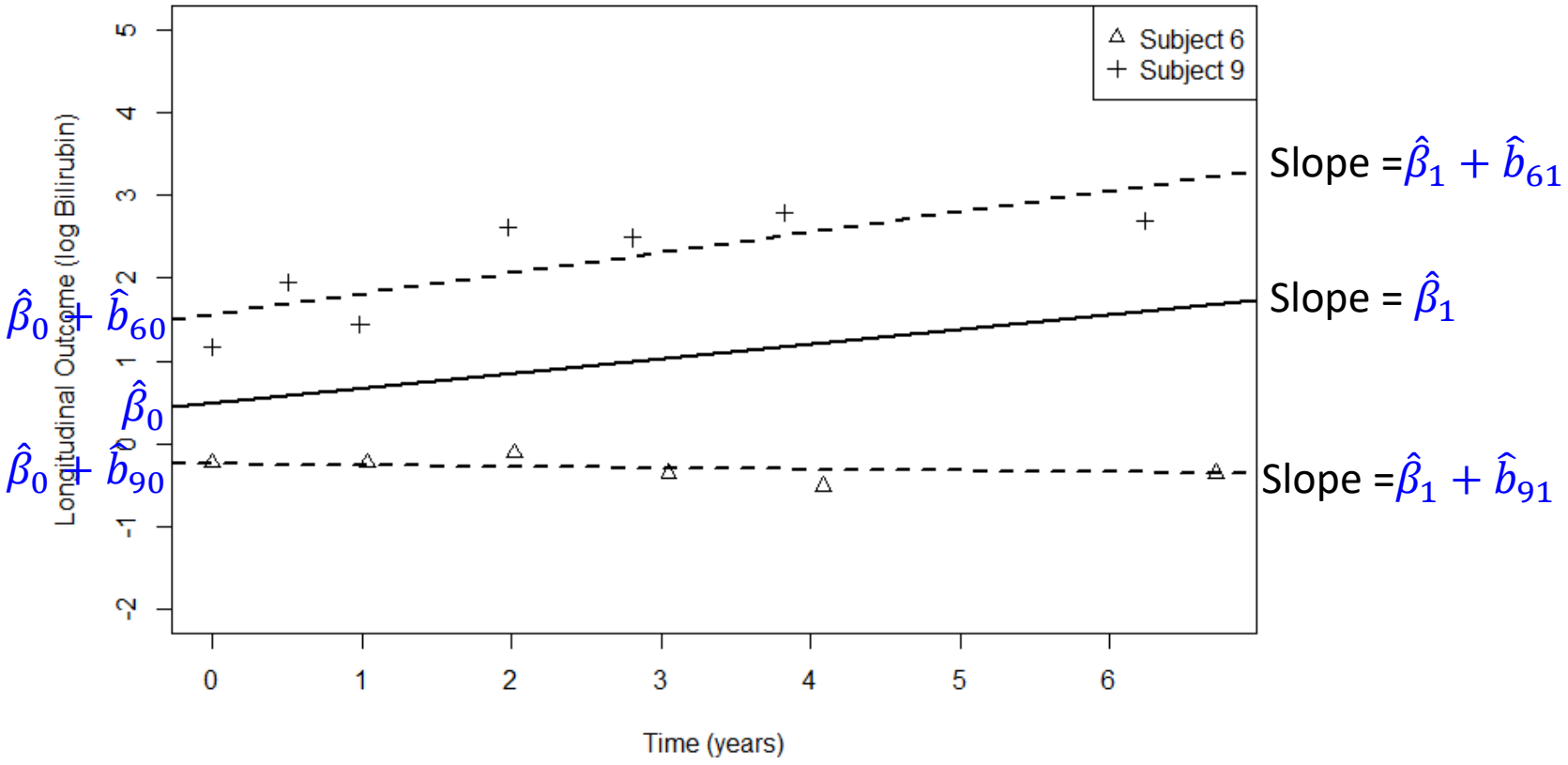
  - βs are the **fixed effects**
  - b_is are the **random effects**

- For the random effects, we assume

$$b_i = \begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} \sim \mathcal{N}(0, D) \qquad D = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}$$

Linear Mixed Model

- Each individual in the population has their own subject-specific mean response profile over time



Linear Mixed Model

- In general form

$$\begin{cases} y_i = X_i\beta + Z_i b_i + \epsilon_i, \\ b_i \sim \mathcal{N}(0, D), \\ \epsilon_i \sim \mathcal{N}(0, \sigma^2 I_{n_i}) \end{cases}$$

  - X is the design matrix for the fixed effects β
  - Z is the design matrix for the random effects b_i
  - $b_i \perp\!\!\!\perp \epsilon_i$

Linear Mixed Model

- **Interpretation:**
- β_j is the change in the average y_i when x_j is increased by one unit
- b_i represents how a subset of the regression parameters for the ith subject deviates from those in the population
- **Advantage: Population + Subject-specific predictions**
- β is the mean response changes in the population
- $\beta + b_i$ describes individual response trajectories

Linear Mixed Model

Benefits for Joint modeling:

- Can reconstruct the complete path of an individual's time-dependent process
- Accommodates unbalanced data
- Accommodates correlation between repeated measurements (parsimonious)

Random effects

- **How do the random effects capture correlation?**
- (**Conditional independence assumption**) Given the random effects, the measurements of each subject are independent

$$p(y_i|b_i;\theta) = \prod_{j=1}^{n_i} p(y_{ij}|b_i;\theta)$$

- Marginally (integrating out the random effects), the measurements of each subject are correlated

$$p(y_i) = \int p(y_i|b_i)p(b_i)db_i$$

$$\Rightarrow y_i \sim \mathcal{N}(X_i\beta, Z_iDZ_i' + \sigma^2 I_{n_i})$$

Random effects

- Can consider more general covariances matrices for the subject-specific errors $\epsilon_i \sim N(0, \Sigma_i)$
- The corresponding marginal model is of the form

$$y_i \sim \mathcal{N}(X_i\beta, Z_i D Z_i' + \Sigma_i)$$

- This model does not assume conditional independence
- Both b_i and Σ_i try to capture the correlation in the observed responses y_i
- Philosophical choice: Random effects vs. Serial correlation

Random effects

- With random effects we model the correlations in the repeated measurements of each subject
- In using random effects for modeling the covariance matrix
  - The more random effects we include the more flexibly we capture the correlations
  - By using random effects (other than random intercept alone) we also directly allow for heteroscedasticity (i.e., non-constant variances over time)
  - We do assume a particular type of structure for the correlations and the variances (they are not allowed completely free)
  - Random effects work equally well with balanced or unbalanced data

Estimation

- Assuming independence across subjects, the log-likelihood of the linear mixed model is given by

$$l(\theta) = \sum_{i=1}^{n} \log p(y_i; \theta)$$

$$= \sum_{i=1}^{n} \log \int p(y_i | b_i; \beta, \sigma^2) p(b_i; \theta_b) db_i$$

where $\theta' = (\beta', \sigma^2, \theta_b')$ and $\theta_b = \text{vech}(D)$

Estimation

- Estimation is based on Maximum Likelihood (ML)

$$p(y_i) = \int p(y_i|b_i)p(b_i)db_i \Rightarrow y_i \sim \mathcal{N}(X_i\beta, V_i)$$

$$V_i = Z_i D Z_i' + \sigma^2 I_{n_i}$$

- Then

$$p(y_i;\theta) = (2\pi)^{-n_i/2}|V_i|^{-1/2}\exp\left\{-\frac{1}{2}(y_i - X_i\beta)'V_i^{-1}(y_i - X_i\beta)\right\}$$

where $\theta' = (\beta', \sigma^2, \theta_b')$ and θ_b=vech(D)

Estimation

- **Fixed effects:** For known marginal covariance matrix $V_i = Z_i D Z_i' + \sigma^2 I_{n_i}$, the fixed effects are estimated using generalized least squares

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i' V_i^{-1} X_i\right)^{-1} \sum_{i=1}^{n} X_i' V_i^{-1} y_i$$

- **Variance components:** The unique parameters in V_i are estimated based on either
  - maximum likelihood (ML) – profile likelihood
  - restricted maximum likelihood (REML)
    - REML provides unbiased estimates for the variance components in small samples

Hypothesis Tests

- **Test of random effects**
  - With REML, must the have same fixed effects to compare models
  - Likelihood ratio tests -> mixture of chi-square distributions
  - Compare with AIC
- **Test of fixed effects**
  - Must use ML to compare models with different fixed effects
    - REML depends on the fixed-effects design matrix

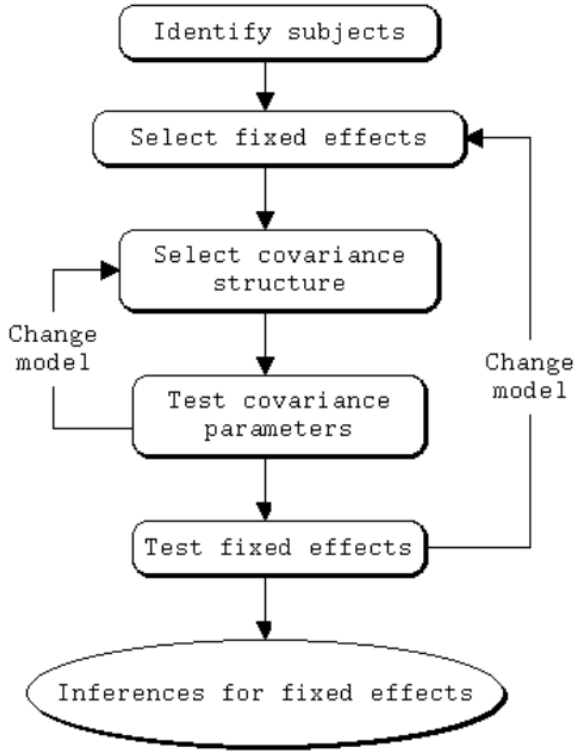


Figure 2. Repeated Measures Analysis in PROC MIXED

Mixed-Effects Models in R

- **nlme**
  - Fits linear and nonlinear mixed effects models, and marginal models for normal data
  - Allows for both random effects and correlated error terms
  - Several options for covariance matrices and variance functions
- **lme4**
  - Fits linear, nonlinear and generalized mixed effects models
  - Uses only random effects
  - Allows for nested and crossed random-effects designs

Mixed-Effects Models in R

- We will only use package **nlme** because the joint modeling package **JM** accepts that as an argument
- The basic function to fit linear mixed models is **lme()** and has three basic arguments
  - **fixed**: a formula specifying the response vector and the fixed-effects structure
  - **random:** a formal specifying the random-effects structure
  - **data:** data frame containing all the variables

Example: PBC data

- We fit a linear mixed model for the PBC dataset to describe the evolution of serum bilirubin
- We will fit the specific model assuming:
  - Different average longitudinal evolutions per treatment group (**fixed part**)
  - Random intercepts and random slopes (**random part**)

Example: PBC data

- Fit a linear mixed effects model for log serum bilirubin, assuming simple linear evolutions in time for each subject (i.e., random intercept, random slope) and different average evolutions per treatment group

$$y_i(t) = \beta_0 + \beta_1 t + \beta_2 \{\text{D-penic}_i \times t\} + b_{i0} + b_{i1} t + \epsilon_i(t)$$

Note: We did not include a main effect for treatment due to randomization

fixed

```
lmeFit.p1 <- lme(log(serBilir) ~ year + drug:year, data = pbc2,
                 random = ~ year | id)
```

random intercept + random slope

```
random = ~ 1 | id)
```

random intercept

Example: PBC data

```
Linear mixed-effects model fit by REML
 Data: pbc2
       AIC      BIC    logLik
  3082.323 3121.323 -1534.161

Random effects:
 Formula: ~year | id
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 0.9990381 (Intr)
year        0.1722826 0.417
Residual    0.3489259

Fixed effects: log(serBilir) ~ year + drug:year
                    Value  Std.Error   DF   t-value p-value
(Intercept)     0.4956686 0.05807539 1631  8.534915  0.0000
year            0.1761726 0.01754759 1631 10.039704  0.0000
year:drugD-penicil 0.0027708 0.02411083 1631  0.114920  0.9085
 Correlation:
                   (Intr) year
year                0.177
year:drugD-penicil  0.002 -0.705

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-4.31683669 -0.49673826 -0.01978638  0.45207679  5.28538333

Number of Observations: 1945
Number of Groups: 312
```

σ_0 σ_1 σ $\frac{\sigma_{01}}{\sigma_0\sigma_1}$

β_0 β_1 β_2

Example: PBC data

- Under the random-intercept, random-slope model, the implied marginal model is

$$y_i = X_i\beta + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, Z_i D Z_i' + \sigma^2 I_{n_i})$$

- And the marginal covariance function for any pair of responses on the same individual is of the form

$$\mathrm{Cov}(y_{ij}, y_{ij'}) = \sigma_1^2 t_{ij} t_{ij'} + \sigma_{01}(t_{ij} + t_{ij'}) + \sigma_0^2$$

$$\mathrm{Var}(y_{ij}) = \sigma_1^2 t_{ij}^2 + 2\sigma_{01} t_{ij} + \sigma_0^2 + \sigma^2$$

Example: PBC data

- The estimated marginal covariance matrix (from a random intercept and slope model) for Patient 6 in our data set is:

id	drug	year	serBilir
6	placebo	0.000000	0.8
6	placebo	1.034936	0.8
6	placebo	2.017851	0.9
6	placebo	3.052787	0.7
6	placebo	4.084985	0.6
6	placebo	6.716132	0.7

```
margCov.p1 <- getVarCov(lmeFit.p1, individuals = 6, type = "marginal")
margCov.p1
```

```
## id 6
## Marginal variance covariance matrix
##        1      2      3      4      5      6
## 1 1.1198 1.0724 1.1430 1.2173 1.2914 1.4803
## 2 1.0724 1.3002 1.2792 1.3853 1.4912 1.7609
## 3 1.1430 1.2792 1.5304 1.5450 1.6809 2.0274
## 4 1.2173 1.3853 1.5450 1.8348 1.8807 2.3080
## 5 1.2914 1.4912 1.6809 1.8807 2.2017 2.5879
## 6 1.4803 1.7609 2.0274 2.3080 2.5879 3.4230
##   Standard Deviations: 1.0582 1.1403 1.2371 1.3545 1.4838 1.8501
```

```
cov2cor(margCov.p1[[1]])
```

```
##           1         2         3         4         5         6
## 1 1.0000000 0.8887190 0.8730637 0.8492049 0.8224226 0.7560682
## 2 0.8887190 1.0000000 0.9068515 0.8969157 0.8813207 0.8346721
## 3 0.8730637 0.9068515 1.0000000 0.9219747 0.9157065 0.8857795
## 4 0.8492049 0.8969157 0.9219747 1.0000000 0.9357149 0.9209516
## 5 0.8224226 0.8813207 0.9157065 0.9357149 1.0000000 0.9426686
## 6 0.7560682 0.8346721 0.8857795 0.9209516 0.9426686 1.0000000
```

Example: PBC data

- The estimated marginal covariance matrix (from a random intercept model) for Patient 6 in our data set is:

```
margCov.p0 <- getVarCov(lmeFit.p0, individuals = 6, type = "marginal")
margCov.p0
```

```
## id 6
## Marginal variance covariance matrix
##        1      2      3      4      5      6
## 1 1.4375 1.1955 1.1955 1.1955 1.1955 1.1955
## 2 1.1955 1.4375 1.1955 1.1955 1.1955 1.1955
## 3 1.1955 1.1955 1.4375 1.1955 1.1955 1.1955
## 4 1.1955 1.1955 1.1955 1.4375 1.1955 1.1955
## 5 1.1955 1.1955 1.1955 1.1955 1.4375 1.1955
## 6 1.1955 1.1955 1.1955 1.1955 1.1955 1.4375
##   Standard Deviations: 1.199 1.199 1.199 1.199 1.199 1.199
```

```
cov2cor(margCov.p0[[1]])
```

```
##           1         2         3         4         5         6
## 1 1.0000000 0.8316326 0.8316326 0.8316326 0.8316326 0.8316326
## 2 0.8316326 1.0000000 0.8316326 0.8316326 0.8316326 0.8316326
## 3 0.8316326 0.8316326 1.0000000 0.8316326 0.8316326 0.8316326
## 4 0.8316326 0.8316326 0.8316326 1.0000000 0.8316326 0.8316326
## 5 0.8316326 0.8316326 0.8316326 0.8316326 1.0000000 0.8316326
## 6 0.8316326 0.8316326 0.8316326 0.8316326 0.8316326 1.0000000
```

Example: PBC data

- We can extract and plot predicted longitudinal evolutions from a mixed model

```
> newdat <- expand.grid(id=1:2, year=c(0.5,1.5), drug=c("D-penicil", "placebo"))
> newdat$pred <- predict(lmeFit.p0, newdata=newdat)
> newdat
  id year      drug      pred
1  1  0.5 D-penicil 2.6815376
2  2  0.5 D-penicil 0.4094144
3  1  1.5 D-penicil 2.7815562
4  2  1.5 D-penicil 0.5094330
5  1  0.5   placebo 2.6764478
6  2  0.5   placebo 0.4043246
7  1  1.5   placebo 2.7662868
8  2  1.5   placebo 0.4941635

> newdat <- expand.grid(year=c(0.5,1.5), drug=c("D-penicil", "placebo"))
> newdat$pred <- predict(lmeFit.p0, newdata=newdat,level=0)
> newdat
  year      drug      pred
1  0.5 D-penicil 0.6207185
2  1.5 D-penicil 0.7207371
3  0.5   placebo 0.6156286
4  1.5   placebo 0.7054676
```

Example: PBC data

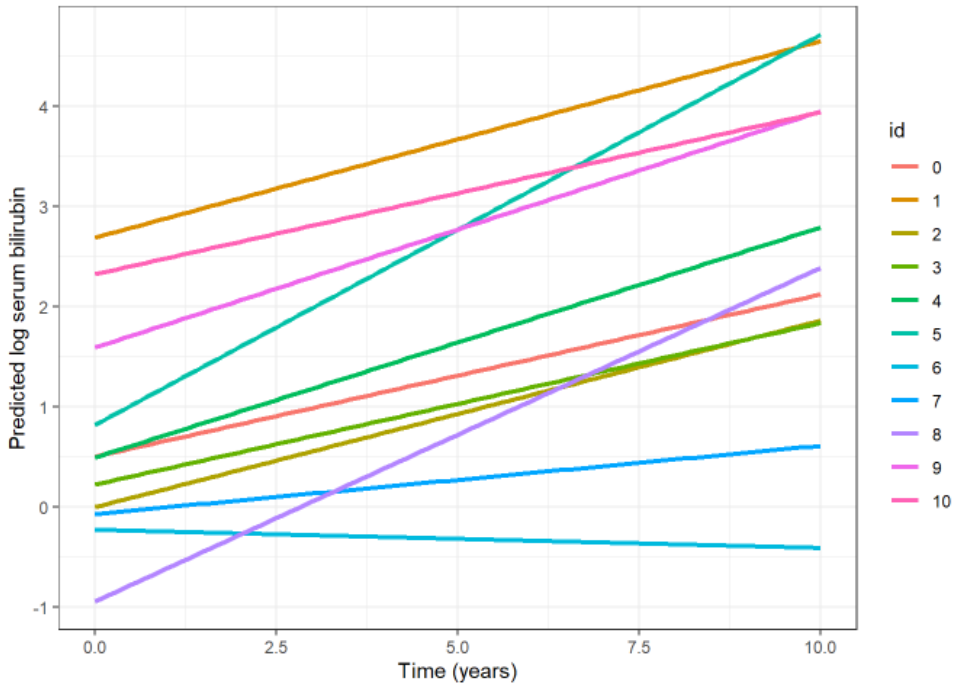
- Random intercepts and random slopes, with a diagonal covariance matrix (using the pdDiag() function)

```
> lmeFit.p2 <- lme(log(serBilir) ~ year+drug:year, data = pbc2,
+                  random = list(id = pdDiag(form = ~ year)))
> lmeFit.p2
Linear mixed-effects model fit by REML
  Data: pbc2
  Log-restricted-likelihood: -1545.702
  Fixed: log(serBilir) ~ year + drug:year
      (Intercept)                year year:drugD-penicil
     0.5018477393       0.1615161046       0.0009132756

Random effects:
 Formula: ~year | id
 Structure: Diagonal
        (Intercept)      year  Residual
StdDev:    1.022807 0.1730064 0.3477198

Number of Observations: 1945
Number of Groups: 312
```

Example: PBC data

- Can consider different functions of time for describing the longitudinal trajectory

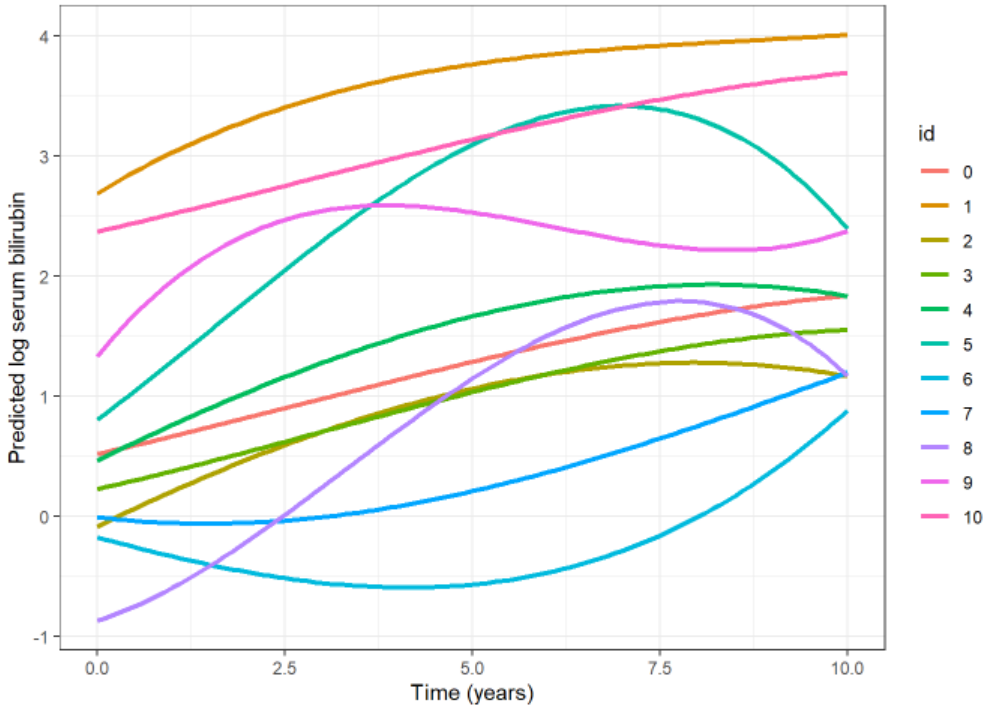
```
lmeFit.y1 <- lme(log(serBilir) ~ year, data = pbc2,
                 random = list(id = pdDiag(form = ~ year)))
```



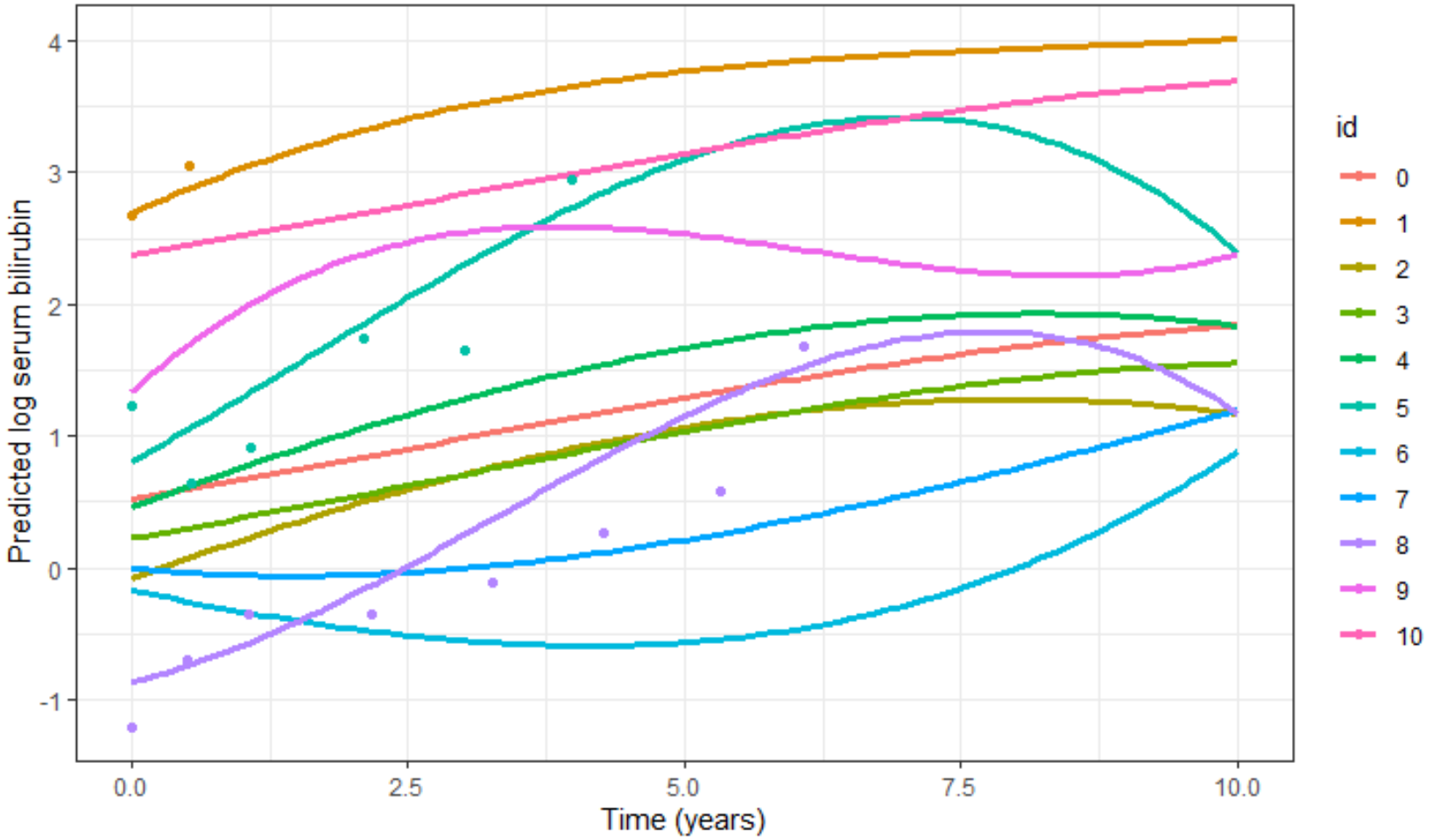
Example: PBC data

- Model trajectory over time more flexibly using splines

```
lmeFit.y2 <- lme(log(serBilir) ~ bs(year), data = pbc2,
                 random = list(id = pdDiag(form = ~ bs(year))))
```



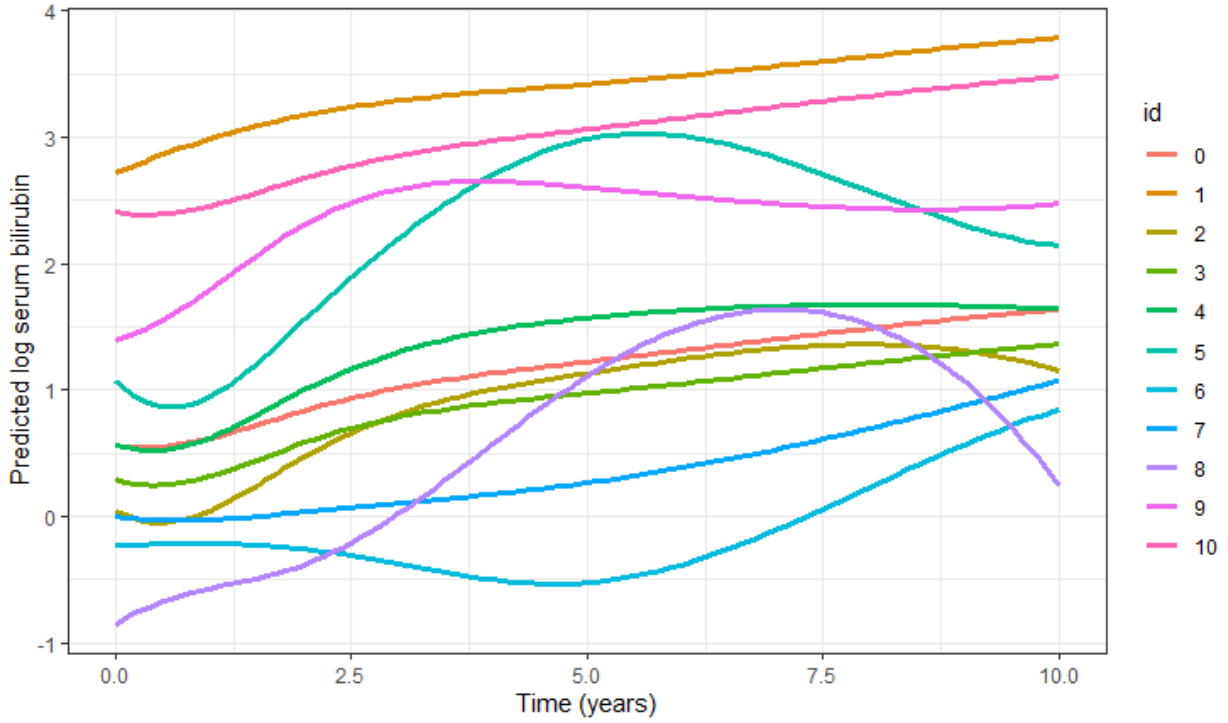
Example: PBC data



Example: PBC data

- Model trajectory over time more flexibly using splines

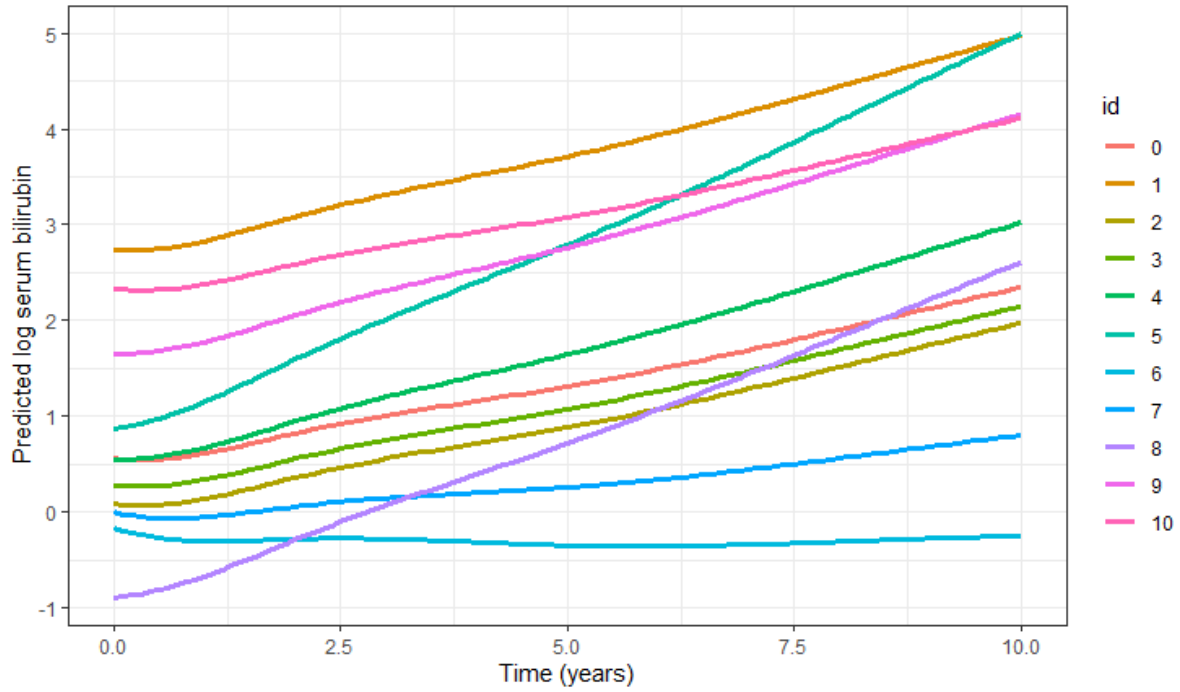
```
lmeFit.y3 <- lme(log(serBilir) ~ bs(year, knots=c(2,5)), data = pbc2,
                 random = list(id = pdDiag(form = ~ bs(year,knots=c(2,5)))))
```



Example: PBC data

- Model trajectory over time more flexibly using splines (basis, natural, etc.)

```
lmeFit.y4 <- lme(log(serBilir) ~ bs(year, knots=c(2,5)), data = pbc2,
                 random = list(id = pdDiag(form = ~ year)))
```



Breakout Session #2 (10 min)

1. What are two advantages of linear mixed effects models for modeling longitudinal data?
2. In what situations would you use REML (restricted maximum likelihood) to estimate parameters for a linear mixed model and why?
3. Why or why should we not consider an unstructured covariance matrix in our PBC data example to have increased flexibility?
4. What is your current TV show/movie recommendation?