Longitudinal Analysis 2

Instructor:

E. John Orav, PhD

Division of General Medicine, Brigham and Women's Hospital
Harvard Medical School, Boston, MA, USA

Lecture 2: Analysis of Longitudinal Data

Outline

- Incorporating correlation into the regression model.
- Marginal models: Correlation through the error terms.
- Specifying the correlation structure
- An example : CD4 counts over time
- Conditional models: Correlation through random effects.

Review: Independent Outcomes

Linear Regression Model:

$$CD4_i = \alpha_0 + \beta_1 Time_i + \beta_2 Gender_i + \varepsilon_i$$

Assumptions:

- The ε_i are independent random variables. Therefore, the CD4 counts are independent random variables.
- The coefficients, β_1 and β_2, are (unknown) numbers which capture the population-wide relationship between Time, Gender, and the CD4 counts.

Independent Outcomes

Analysis Process: There is only one optimal way to approach the fitting process for normal data (maximum likelihood, or, equivalently, least squares).

All software programs approach the problem the same way and come up with the same answers.

Correlated Outcomes

Idea: Somehow, the model must allow for correlation between some of the outcomes.

- The most obvious way would be to allow the ε to be correlated, instead of independent: Marginal Models

- Alternatively, the coefficients in the model could be random variables specific to subjects (while the errors remain independent): Conditional Models

Correlated Outcomes

Making the Choice: You must choose the modeling approach that you prefer.

- Partly, the choice is philosophical. Marginal models are often used for longitudinal data. Conditional models are often used for clustered data.
- Partly, the choice is driven by software. Can your software fit marginal models (usually, yes)? Can your software fit conditional models (usually yes for linear regression; not always yes for logistic regression).

Making the Choice

- Partly, the choice is driven by precedent. What have other authors used?
  - Often, a marginal model, since the software became available earlier and is more widespread.

The Good News

Usually, the answers you get through the different model choices are very similar.

Marginal Models for Normal Data

Marginal Model Equation:

$$Y_{ij} = \alpha_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \varepsilon_{ij}$$

- where “i” identifies the subject

- and “j” identifies the time point for a specific subject

Assumptions about correlation:

- The ε_{ij} are independent over i (outcomes from different subjects are independent)

- For each subject i, the ε_{ij} are correlated over j (outcomes are correlated over time within subjects)

Marginal Model for Normal Data

Marginal Model Equation:

$$Y_{ij} = \alpha_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \varepsilon_{ij}$$

Interpretation of the Beta's: Each beta coefficient represents the population-average effect of the predictor on the outcome.

Every subject has the same relationship between a predictor and the outcome (i.e., the same value of beta).

Marginal Model: Normal Longitudinal Data

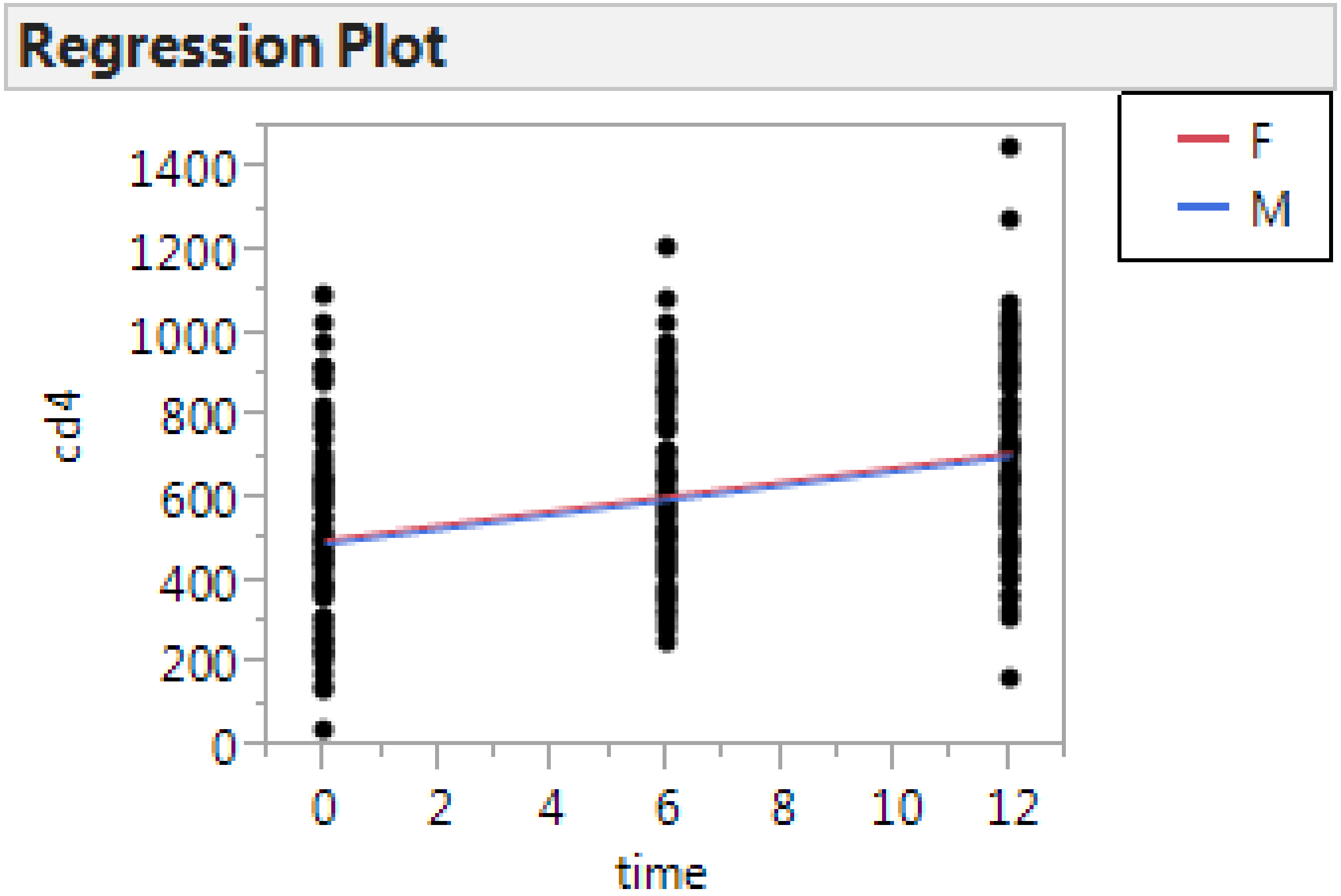
Longitudinal Model Equation:

$$CD4_{ij} = \alpha_0 + \beta_1 \, Time_{ij} + \beta_2 \, Gender_{ij} + \varepsilon_{ij}$$

- i = 1, 2,3, … , 101 (there are 101 people in the study)
- j=1 (baseline); j=2 (6 months); and j=3 (12 months)

- α_0 is the baseline CD4 for men
- β_1 is the slope over time (same for every subject ± noise)
- β_2 is the difference in CD4 between men and women (same for all men and women ± noise)

Longitudinal Data: Main Effects



Regression Plot

Marginal Model: Normal Longitudinal Data

Longitudinal Model Equation:

$$CD4_{ij} = \alpha_0 + \beta_1 Time_{ij} + \beta_2 \, Gender_{ij} + \varepsilon_{ij}$$

Result:

$$CD4 = 500 + 17.3 \, Time + 16 \, Female$$

Where $\beta_1 = 17.3$ (se=2.0), $p < 0.001$

$\beta_2 = 16$ (se=77), $p=.84$

Marginal Model: Correlation Structure

Longitudinal Model Equation:

$$CD4_{ij} = \alpha_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{Gender}_{ij} + \varepsilon_{ij}$$

$i = 1, 2,3, \ldots, 101$ (there are 101 people in the study)

$j=1$ (baseline); $j=2$ (6 months); and $j=3$ (12 months)

Correlation Structure:

- Independence across subjects: $\text{Corr}(\varepsilon_{1j}, \varepsilon_{2j}) = 0$
- Dependence over time, within-subject:

$$\text{Corr}(\varepsilon_{i1}, \varepsilon_{i2}) = \rho_{12}$$

$$= \text{Corr}(CD4_{i,\text{ Baseline}}, CD4_{i,\text{ 6 Months}})$$

Correlations Between Outcomes

Issue: If we only had two variables we could capture their association through a single correlation coefficient.

However, in multivariate data, each subject has a number of outcome variables which may be correlated with each other:

$$\text{Corr}(\text{CD4}_{\text{Baseline}}, \text{CD4}_{\text{Month 6}}) = \rho_{12}$$

$$\text{Corr}(\text{CD4}_{\text{Baseline}}, \text{CD4}_{\text{Month 12}}) = \rho_{13}$$

$$\text{Corr}(\text{CD4}_{\text{Month 6}}, \text{CD4}_{\text{Month 12}}) = \rho_{23}$$

The Correlation Matrix

$$\begin{array}{c|ccc|}
 & \text{Baseline} & \text{Month 6} & \text{Month 12} \\ \hline
\text{Baseline} & \rho_{11} & \rho_{12} & \rho_{13} \\
\text{Month 6} & \rho_{21} & \rho_{22} & \rho_{23} \\
\text{Month 12} & \rho_{31} & \rho_{32} & \rho_{33} \\ \hline
\end{array}$$

where $\rho_{11} = \rho_{22} = \rho_{33} = 1$

Note: $\rho_{12} = \rho_{21}$ and $\rho_{13} = \rho_{31}$ and $\rho_{23} = \rho_{32}$

The Correlation Matrix: CD4 Data

$$\begin{array}{c|ccc|}
 & \text{Baseline} & \text{Month 6} & \text{Month 12} \\ \hline
\text{Baseline} & 1.0 & .53 & .51 \\
\text{Month 6} & .53 & 1.0 & .69 \\
\text{Month 12} & .51 & .69 & 1.0 \\ \hline
\end{array}$$

Types of Correlation Matrices

Unstructured:

$$\begin{array}{l|ccccc|}
 & \text{Baseline} & \text{Month 6} & \text{Month 12} & \text{Month 18} & \text{Month 24} \\
\hline
\text{Baseline} & 1 & \rho_{12} & \rho_{13} & \rho_{14} & \rho_{15} \\
\text{Month 6} & \rho_{21} & 1 & \rho_{23} & \rho_{24} & \rho_{25} \\
\text{Month 12} & \rho_{31} & \rho_{32} & 1 & \rho_{34} & \rho_{35} \\
\text{Month 18} & \rho_{41} & \rho_{42} & \rho_{43} & 1 & \rho_{45} \\
\text{Month 24} & \rho_{51} & \rho_{52} & \rho_{53} & \rho_{54} & 1 \\
\hline
\end{array}$$

Note: Your software will need to estimate 10 correlations in addition to predictor coefficients.

Types of Correlation Matrices

Toeplitz:

$$\begin{array}{c|ccccc|}
 & \text{Baseline} & \text{Month 6} & \text{Month 12} & \text{Month 18} & \text{Month 24} \\
\hline
\text{Baseline} & 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 \\
\text{Month 6} & \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 \\
\text{Month 12} & \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\
\text{Month 18} & \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 \\
\text{Month 24} & \rho_4 & \rho_3 & \rho_2 & \rho_1 & 1 \\
\hline
\end{array}$$

Note: Your software only needs to estimate 4 correlations, but points must be equally spaced.

Types of Correlation Matrices

Autoregressive:

$$
\begin{array}{l|ccccc|}
 & \text{Baseline} & \text{Month 6} & \text{Month 12} & \text{Month 18} & \text{Month 24} \\
\hline
\text{Baseline} & 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\
\text{Month 6} & \rho & 1 & \rho & \rho^2 & \rho^3 \\
\text{Month 12} & \rho^2 & \rho & 1 & \rho & \rho^2 \\
\text{Month 18} & \rho^3 & \rho^2 & \rho & 1 & \rho \\
\text{Month 24} & \rho^4 & \rho^3 & \rho^2 & \rho & 1 \\
\hline
\end{array}
$$

Note: Your software only needs to estimate 1 correlation, but points must be equally spaced.

Types of Correlation Matrices

Exchangeable (Compound Symmetry):

$$\begin{array}{l|ccccc|}
 & \text{Baseline} & \text{Month 6} & \text{Month 12} & \text{Month 18} & \text{Month 24} \\
\hline
\text{Baseline} & 1 & \rho & \rho & \rho & \rho \\
\text{Month 6} & \rho & 1 & \rho & \rho & \rho \\
\text{Month 12} & \rho & \rho & 1 & \rho & \rho \\
\text{Month 18} & \rho & \rho & \rho & 1 & \rho \\
\text{Month 24} & \rho & \rho & \rho & \rho & 1 \\
\hline
\end{array}$$

Note: Your software only needs to estimate 1 correlations. Often used for clustered data.

Choosing a Correlation Matrix

- If you do not want to make any assumptions: Unstructured
  - However, your software will need to estimate many correlations and this may make the beta’s for the predictors unstable.
- If you choose a specific correlation structure:
  - If you have chosen correctly, you will have better power to estimate the effects of the predictors.
  - If you have chosen incorrectly, you may not get valid results.

Longitudinal Model: CD4 Data

Longitudinal Model Equation:

$$CD4_{ij} = \alpha_0 + \beta_1 \, Time_{ij} + \beta_2 \, Gender_{ij} + \varepsilon_{ij}$$

i = 1, 2,3, … , 101 (there are 101 people in the study)

j=1 (baseline); j=2 (6 months); and j=3 (12 months)

Correlation Structure:

- Independence over subjects: $Corr(\varepsilon_{1j}, \varepsilon_{2j}) = 0$
- Dependence over time, within-subject:

$$Corr(\varepsilon_{11}, \varepsilon_{12}) = \rho_{12} \quad \text{etc.}$$

Correlation Matrix for the CD4 Data

Correlation Matrix (for each person):

$$\begin{array}{c|ccc|}
 & \text{Baseline} & \text{Month 6} & \text{Month 12} \\
\hline
\text{Baseline} & 1 & \rho_{12} & \rho_{13} \\
\text{Month 6} & \rho_{21} & 1 & \rho_{23} \\
\text{Month 12} & \rho_{31} & \rho_{32} & 1 \\
\hline
\end{array}$$

Choices:

- Unstructured? Autoregressive? Toeplitz?

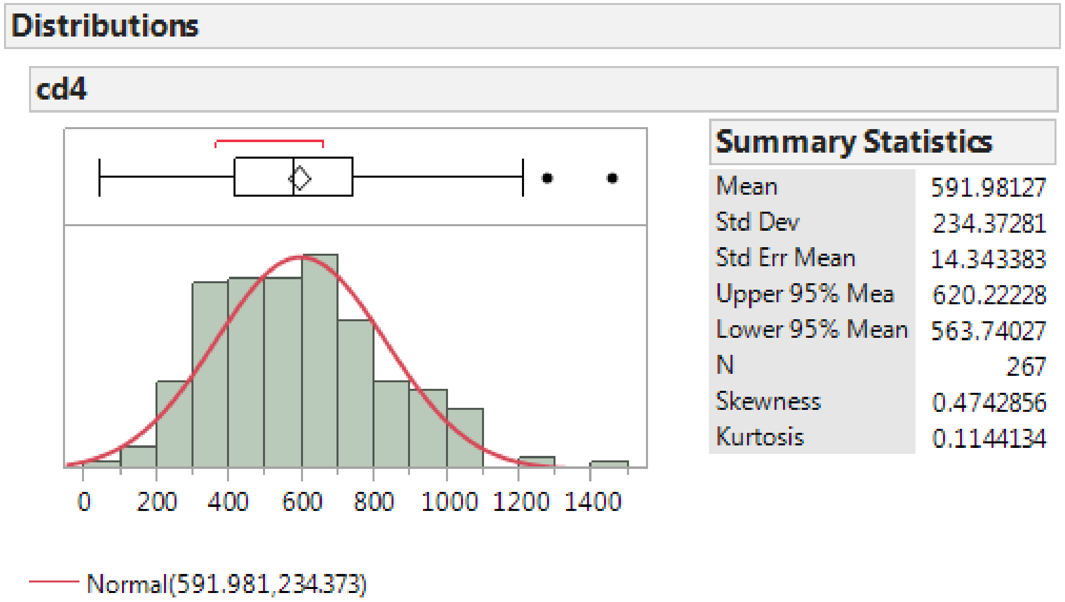
The Modeling Process: Step 1

Step 1: Convert the database into “long” format: Done

Obs	subject	cd4	time	Age	gender
1	JC-01*	1023	0	30.6438	M
2	JC-01*	1087	6	30.6438	M
3	JC-01*	651	12	30.6438	M
4	ND-02*	643	0	42.3507	M
5	ND-02*	561	6	42.3507	M
6	ND-02*	645	12	42.3507	M
7	SJ-03*	463	0	30.3315	M
8	SJ-03*	534	6	30.3315	M
9	SJ-03*	1019	12	30.3315	M

The Modeling Process: Step 2

<u>Step 2:</u> Decide whether your outcomes are normally distributed: Yes, close enough for a class exercise. We will use likelihood based estimation and testing.



Distributions

cd4

Summary Statistics

Mean	591.98127
Std Dev	234.37281
Std Err Mean	14.343383
Upper 95% Mea	620.22228
Lower 95% Mean	563.74027
N	267
Skewness	0.4742856
Kurtosis	0.1144134

The Modeling Process: Step 3

Step 3: Choose a correlation matrix structure:

Thoughts on how?

The Modeling Process: Step 3

Step 3: Choose a correlation matrix structure:

- This is longitudinal, with evenly spaced data.
- We would expect the correlation to decrease as the time separation increased.
- Either unstructured, autoregressive, or Toeplitz might be reasonable. Exchangeable would not be reasonable.
- We have a moderately large sample size, so reduced power if we use unstructured should not be a problem.
- There are many subjects with all 3 time points, so all correlations should be estimable.

The Modeling Process: Step 4

Step 4: Build, run and interpret your model in the usual ways.

- I will run and interpret a variety of models in order to show you how sensitive/insensitive the results are to various choices. You, however, should make one choice and stick with it.

Correlation Structure: Effect on Model Estimates

Data: CD4 counts, measured on 101 patients for up to 3 time points (months 0, 6, and 12). Models were fit using REML and the following correlation structures:

Correlation Structure	Intercept (se)	Slope for Time (se)	Effect of Gender (se)
Independence	494 (20.4)	17.6 (2.72) P<0.001	9.60 (54.8) P=0.86
Unstructured	500 (21.0)	17.3 (2.02) P<0.001	15.7 (77.3) P=0.84
Toeplitz	494 (22.1)	17.8 (1.98) P<0.001	2.02 (79.1) P=0.98
Autoregressive	492 (22.4)	18.0 (2.18) P<0.001	-9.63 (78.0) P=0.90

Conditional Models for Normal Data

General Mixed Model Equation:

$$Y_{ij} = a_i + b_i X_{1ij} + \beta_2 X_{2ij} + \varepsilon_{ij}$$

- where “i” identifies the subject
- and “j” identifies the time point for a specific subject

Assumptions about the ε_{ij} :

- The ε_{ij} are independent over i and j (i.e., the error terms are totally independent)

Conditional Models for Normal Data

General Mixed Model Equation:

$$CD4_{ij} = a_i + b_i Time_{ij} + \beta_2 Gender_{ij} + \varepsilon_{ij}$$

Interpretation of the model coefficients:

- a_i is a random variable representing the baseline CD4 for patient i. We assume that the a_i's follow a normal distribution.
- b_i is a random variable representing the effect of Time for patient i. The b_i's follow a normal distribution.
- β_2 is a fixed number representing the effect of Gender. This effect is assumed to be the same for all subjects.

Interpreting a Conditional Model

Random Effects for Longitudinal Data:

$$Y_{ij} = a_i + b_i \text{Time}_{ij} + \varepsilon_{ij}$$

Patient #1: $Y_1 = a_1 + b_1 \text{ Time}_{1j} + \varepsilon_{1j}$

Patient #2: $Y_2 = a_2 + b_2 \text{ Time}_{2j} + \varepsilon_{2j}$

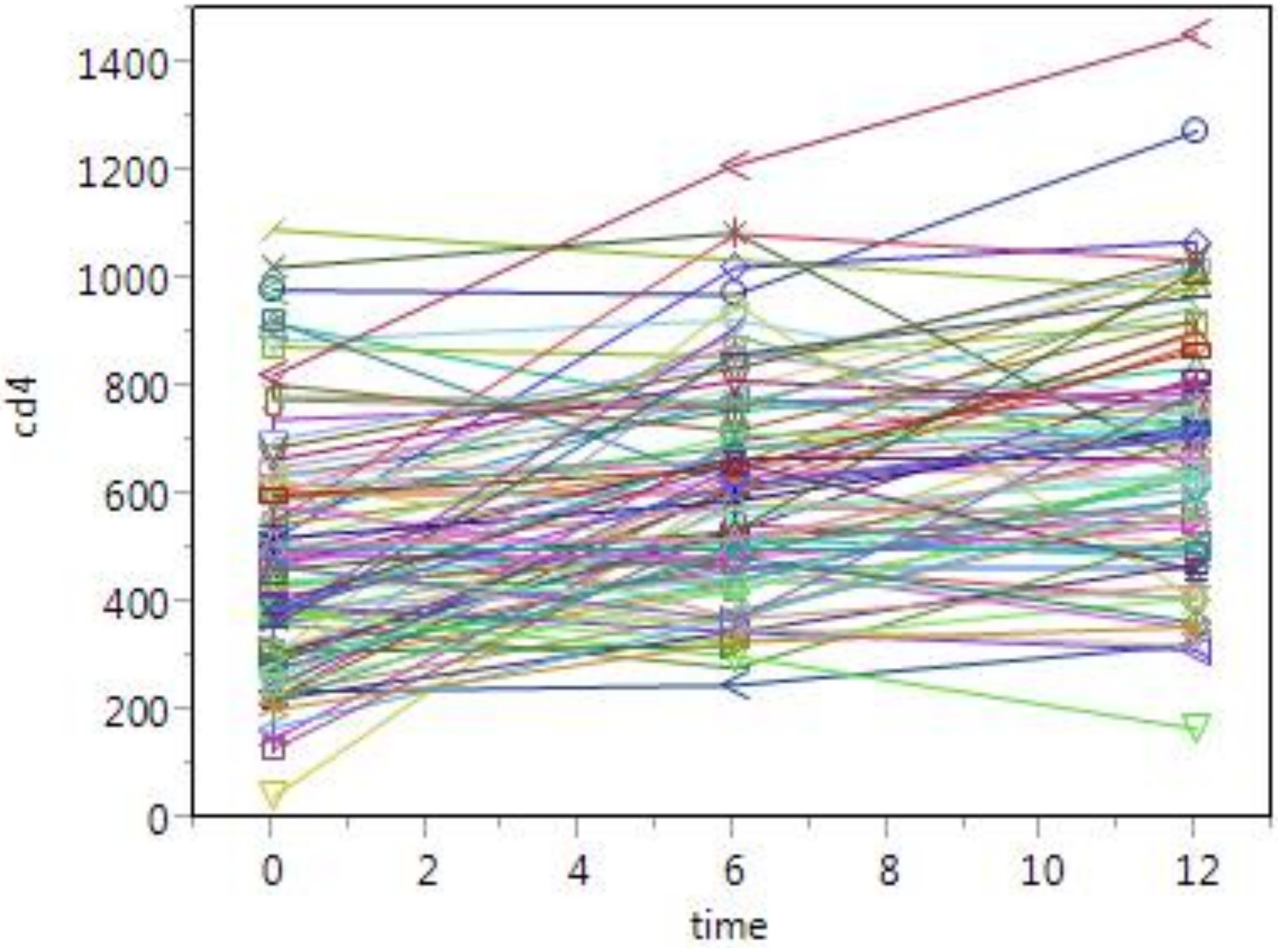
Patient #3: $Y_3 = a_3 + b_3 \text{ Time}_{3j} + \varepsilon_{3j}$

…

Where a_i is the intercept (CD4 count at Time=0) for patient i.

Where b_i is the slope for patient i.

Longitudinal Data



Conditional Models for Normal Data

General Mixed Model Equation:

$$Y_{ij} = a_i + b_i X_{1ij} + \varepsilon_{ij}$$

Correlation Structure of the Y_{ij}:

- For different subjects (i.e., for different values of i), the Y_{ij} are independent: $\mathrm{Corr}(Y_{1j}, Y_{2j}) = 0$
- Within a subject (i.e., over j, for a given value of i), the Y_{ij} are dependent: $\mathrm{Corr}(Y_{11}, Y_{12}) = \rho_{12}$
- Even though the ε_{ij} are independent, the Y_{ij} are dependent over j because they share a_i and b_i

Random Effects Longitudinal Models

Random Slope and Intercept Model:

$$Y_{ij} = a_i + b_i \mathrm{Time}_{ij} + \varepsilon_{ij}$$

- Every patient has their own unique regression line

Random Intercept Model:

$$Y_{ij} = a_i + \beta \mathrm{Time}_{ij} + \varepsilon_{ij}$$

- Every patient has their own unique starting value (intercept)
- But the slope over time is the same for everyone.

Random Effects Longitudinal Models

Random Slope Model:

$$Y_{ij} = \alpha + b_i \, \text{Time}_{ij} + \varepsilon_{ij}$$

- Every patient starts from the same baseline value (intercept).
- But the slope over time is unique for each subject.

Random Effects Models: Assumptions

Mixed Effects Model:

$$Y_{ij} = a_i + b_i \text{Time}_{ij} + \beta_2 \text{Gender} + \varepsilon_{ij}$$

Usual Assumptions:

- The a_i are normally distributed random variables with mean α and variance σ_a^2
- The b_i are normally distributed random variables with mean β and variance σ_b^2
- The ε_{ij} are independent, normally distributed random variables with mean 0 and variance σ_ε^2

Results: Random Slope and Intercept Model for CD4 Counts

Random Effects Model:

$$CD4_{ij} = a_i + b_i Time_{ij} + \beta_2 Gender_{ij} + \varepsilon_{ij}$$

Result:

$$CD4_{ij} = 502 + 17.9\ Time_{ij} + 5.86\ Female_{ij}$$

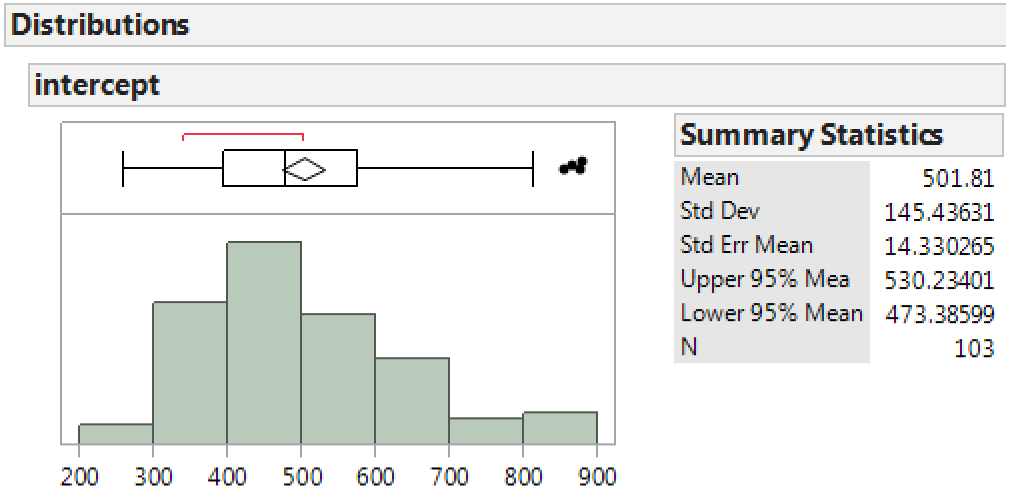
So $\alpha = 502$ (se=76)

And $\beta = 17.9$ (se=2.05), $p<.001$

Results: Random Intercepts for CD4 Counts

$$CD4_{ij} = a_i + b_i Time_{ij} + \beta_2 Sex_{ij} + \varepsilon_{ij}$$

$$CD4_{ij} = 502 + 17.9\ Time_{ij} + 5.86\ Female_{ij}$$



Distributions

intercept

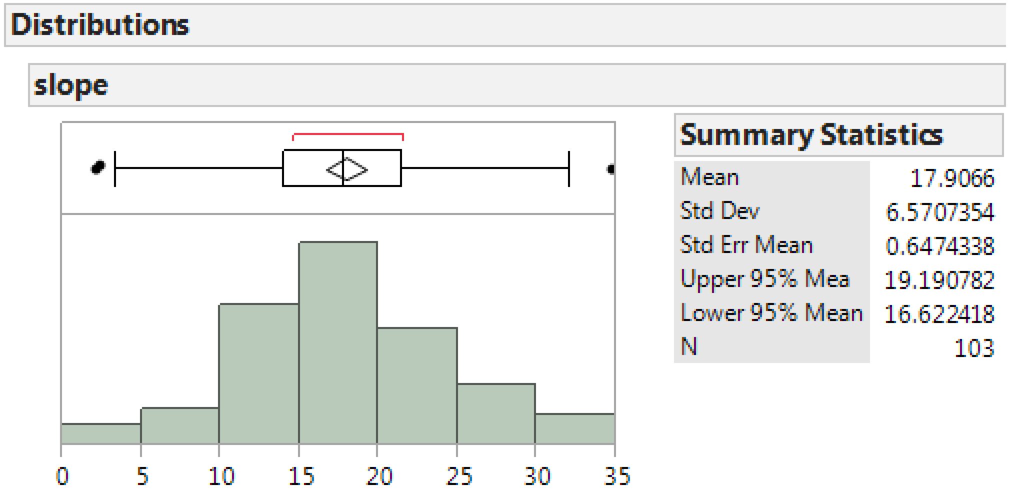
Summary Statistics

Mean	501.81
Std Dev	145.43631
Std Err Mean	14.330265
Upper 95% Mea	530.23401
Lower 95% Mean	473.38599
N	103

Results: Random Slopes for CD4 Counts

$$CD4_{ij} = a_i + b_i Time_{ij} + \beta_2 Sex_{ij} + \varepsilon_{ij}$$

$$CD4_{ij} = 502 + 17.9\ Time_{ij} + 5.86\ Female_{ij}$$



Distributions

slope

Summary Statistics

Mean	17.9066
Std Dev	6.5707354
Std Err Mean	0.6474338
Upper 95% Mea	19.190782
Lower 95% Mean	16.622418
N	103

Random Effects Compared to REML and GEE Estimates

Modeling Approach	Intercept (se)	Slope for Time	Effect of Gender
Marginal Model	494 (21)	17.6 (2.10)	9.6 (71)
Conditional: Random intercept and slope	502 (76)	17.9 (2.05)	5.9 (78)
Conditional: Random intercept, fixed slope	507 (77)	17.6 (1.84)	11.0 (79)
Conditional: Fixed intercept, random slope	494 (66)	17.8 (2.43)	1.6 (64)

Conclusions

- When you have correlated outcomes, there are two main approaches to analysis: marginal models and conditional models.
- Marginal models look the same as ordinary regression models, but allow the error terms to be correlated rather than independent.
  - You will need to specify the form of the correlation matrix.
- Conditional models have coefficients for predictors that are random variables and that are unique to each subject.
  - You will need to specify which coefficients are random and which are fixed.
- The two approaches often produce similar results