Longitudinal Analysis 1

Instructor:

E. John Orav, PhD

Division of General Medicine, Brigham and Women's Hospital

Harvard Medical School, Boston, MA, USA

Additional Reading:

"Applied Longitudinal Analysis"

GM Fitzmaurice, NM Laird, and JH Ware.

2004. Wiley Series in Probability and Statistics.

ISBN 978-0-471-21487-8

Lecture 1: Study Design and Correlation

Outline

- Independent versus dependent outcomes
- Multivariate designs
- Within-subject versus between-subject effects
- An example: CD4 counts
- Effects of correlation on sample size

Terminology

Dependence: Two measurements are said to be dependent if knowing one of them gives you some information about what value the other might be.

- Related terms:
  - Association;
  - Predictability;
  - Correlation (or Covariance)

Types of Dependence

1. Between a Predictor and an Outcome (i.e., the independent and dependent variables in a regression). You wish to assess whether a predictor affects the outcome.
2. Between Competing Predictors. This can be a reflection of confounding or collinearity and must be considered as part of building a regression model.
3. Between Outcomes. **This is the type of dependence that is the focus of this lecture.**

More Terminology

A Regression with Dependent Outcomes =

A Regression with Correlated Outcomes =

Multivariate Analysis

<u>Types of Multivariate Studies:</u>

- Longitudinal data = Repeated measures = Panel data
- Clustered data = Multi-Level data = Hierarchical data
- Classic time series
- Classic multivariate analysis of variance (MANOVA)

Cluster Designs

1. We measure an outcome for each patient in our study (i.e., whether the patient lives or dies). The patient will be our "*unit of analysis*".
2. Two patients seen by the same physician are more likely to have similar outcomes than patients seen by different physicians. Therefore, physicians constitute "clusters" (i.e., patients are "nested" within physicians)
3. Two physicians working in the same clinic are more likely to have similar outcomes than physicians in different clinics: clinics are also clusters (i.e., patients are nested within physicians, who are nested within clinics)

Classic Time Series

Classic Time Series: We have one very long series of many repeated measurements.

- Daily measurements of temperature over 10 years.

- The yearly incidence of tuberculosis, over the last century

Repeated Measures: We have many short series of a few repeated measurements (one series for each patient).

Classical MANOVA

- For each subject, we measure “k” outcomes and compare the outcomes between groups of subjects (between-subject effects only).
- The “k” outcomes do not need to be measuring the same thing (i.e., a heart’s mass, wall thickness, contractility).
- A single test compares all the outcomes between groups, with a single p-value (i.e., do women’s hearts differ from men’s hearts?).
- If the test is significant, it is not possible to say which outcome (or outcomes) were different.

Longitudinal Analyses: Purposes

We have repeated measurements of the outcome over time for each subject in the study, so that we can:

1. Assess how the outcome changes over time (*within-subject effects*)
2. Assess subject characteristics which influence the outcome (*between-subject effects*)
3. Assess subject characteristics which influence the pattern of change over time (*between-subject time-interaction effects*)

Example: CD4 Counts

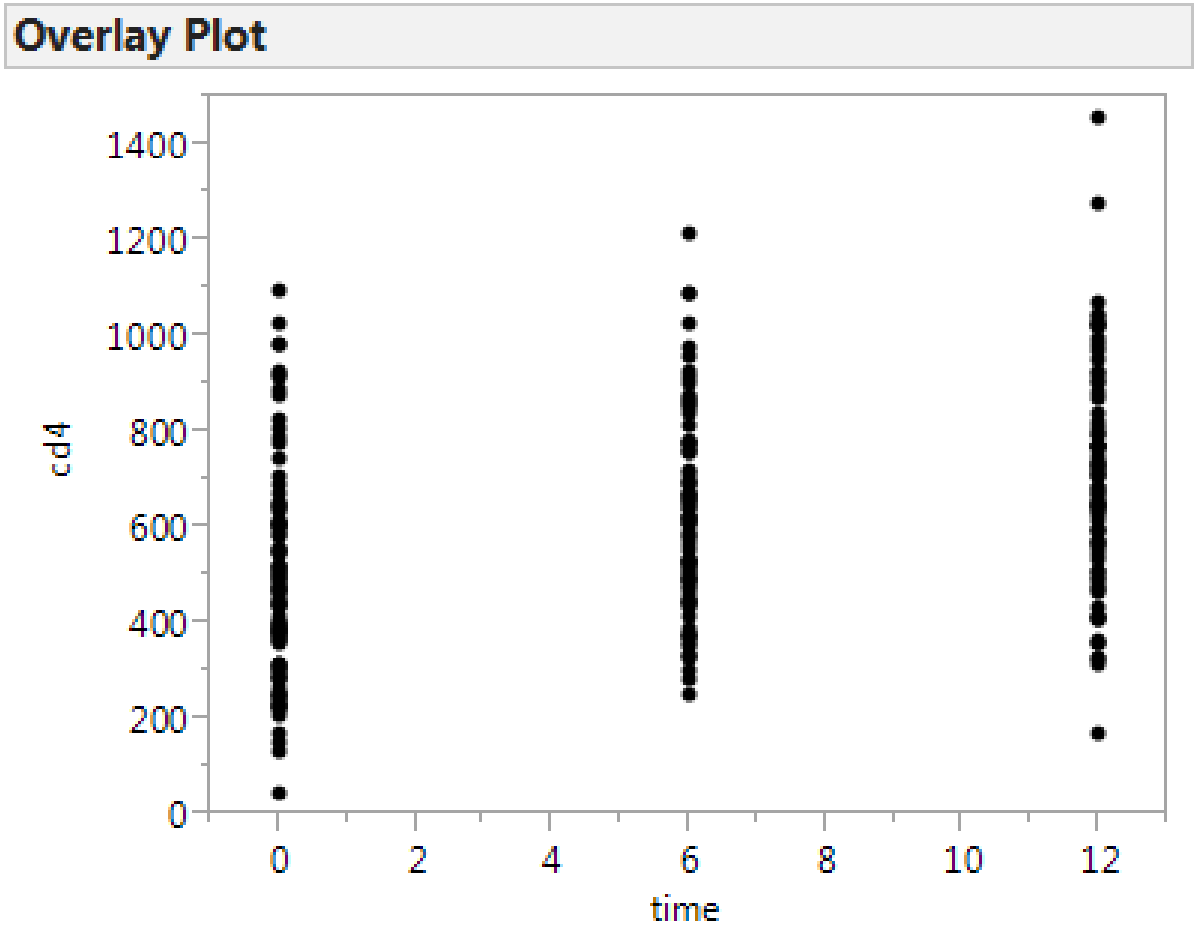
We have a cohort of subjects who have CD4 counts measured. We want to know:

1. Have the CD4 counts changed over time?
2. Are the CD4 counts different between men and women?
3. Is the pattern of change over time different between men and women?

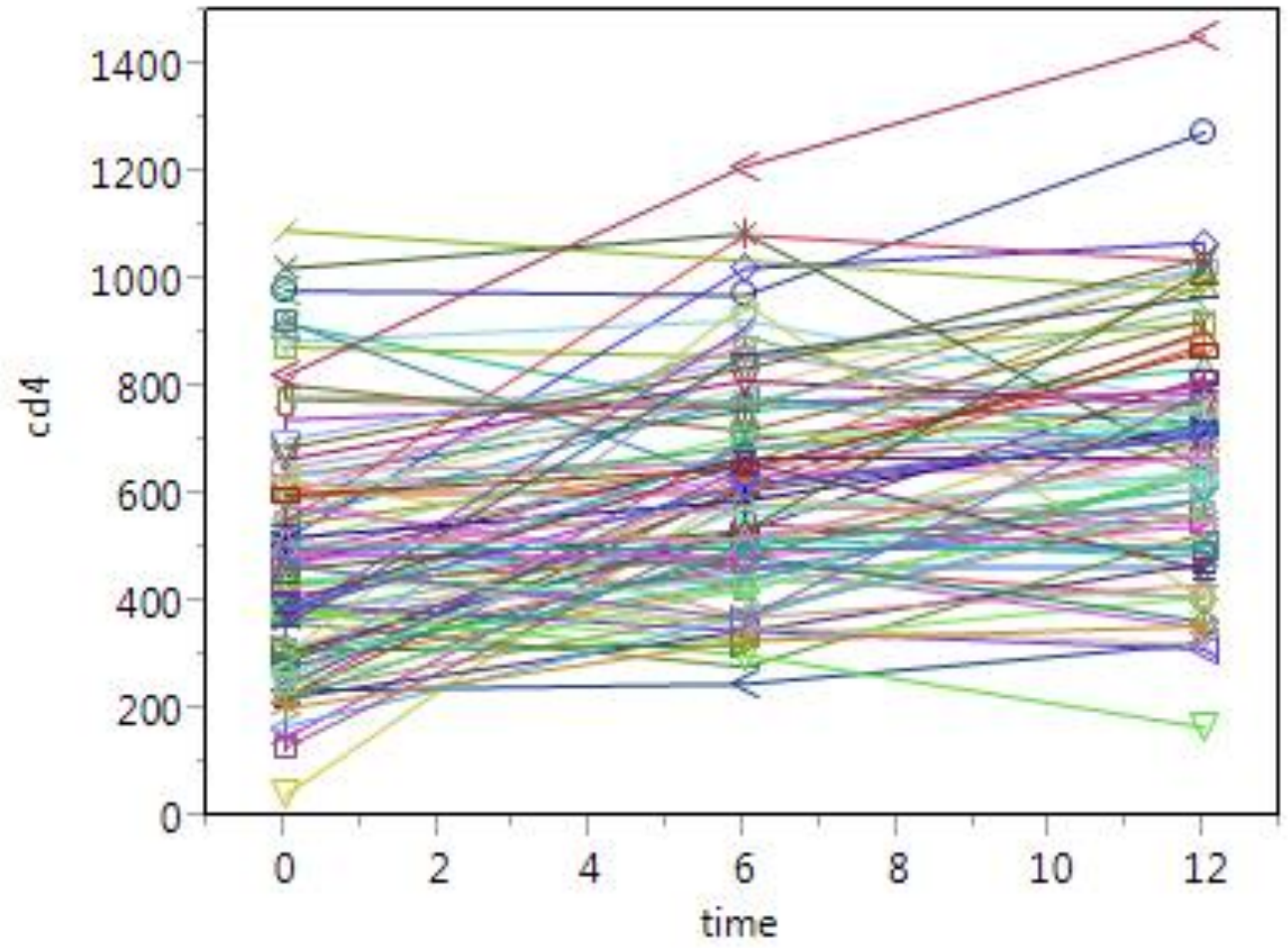
Analysis: Longitudinal Approach

- Enroll a cohort of subjects and record their age and sex
- Evaluate each subject and measure their CD4 count at baseline (time 0), 6 months, and again at 12 months
- Use a longitudinal linear regression model with CD4 count as the outcome (each subject will have multiple outcomes), and:
  - Time and sex as main effects predictors
  - Then, a model with time, sex and a time-by-sex interaction to assess whether time trends are the same for men and women

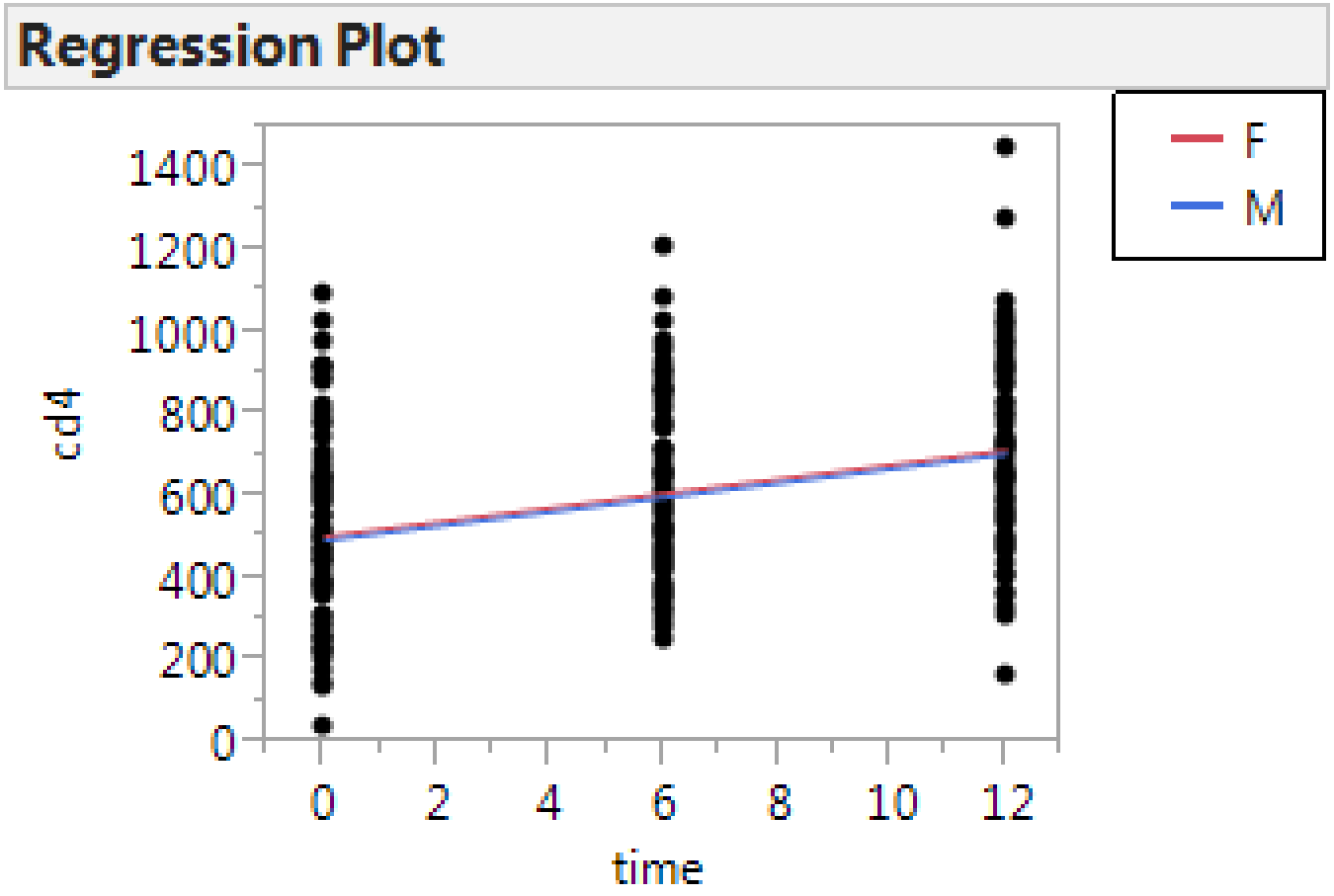
Longitudinal Data



Longitudinal Data



Longitudinal Data: Main Effects



Analysis Results: Assuming Independence

Main Effects Model:

$$Y_i = \alpha_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Or, for our data:

$$CD4_i = \alpha_0 + \beta_1 Time_i + \beta_2 Sex_i$$

Result:

$$CD4 = 494 + 17.6\ Time + 10\ Female$$

Where $\beta_1 = 17.6$ (se=2.7), $p < 0.001$

$\beta_2 = 10$ (se=54), p=.86

Analysis Results: Allowing Dependence

Main Effects Model:

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

Where i: subject and j: time

Or, for our data:

$$CD4_{ij} = \alpha_0 + \beta_1 Time_{ij} + \beta_2 Sex_{ij}$$

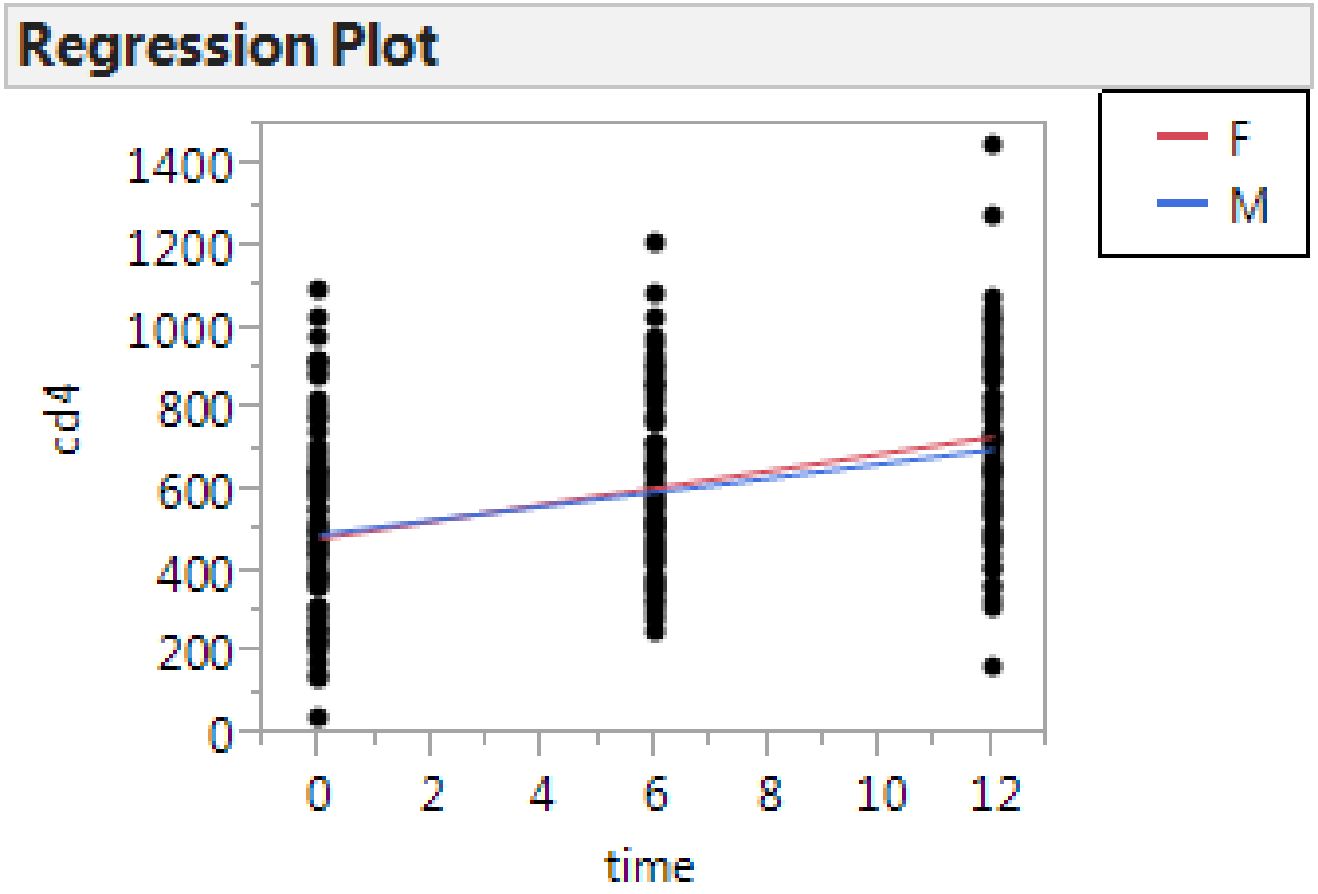
Result:

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

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

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

Longitudinal Data: Interaction Effects



Interaction Results: Assuming Independence

Interaction Model:

$$Y_i = \alpha_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}*X_{2i} + \varepsilon_i$$

Or, for our data:

$$CD4_i = \alpha_0 + \beta_1 Age_i + \beta_2 Sex_i + \beta_3 Age_i*Sex_i$$

Result:

$$CD4 = 495 + 17.4\ Time - 9\ Female + 3.3 Time*Female$$

Where $\beta_1 = 17.4$ (se=2.8), $p < 0.001$

$\beta_2 = -9$ (se=84), p=.91

$\beta_3 = 3.3$ (se=11.4), p=.77

Interaction Results: Allowing Dependence

Interaction Model:

$$Y_{ij} = \alpha_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{1ij}*X_{2ij} + \varepsilon_{ij}$$

Or, for our data:

$$CD4_{ij} = \alpha_0 + \beta_1 Age_{ij} + \beta_2 Sex_{ij} + \beta_3 Age_{ij}*Sex_{ij}$$

Result:

$$CD4 = 501 + 17.2Time + 6\ Female + 2.2Time*Female$$

Where $\beta_1 = 17.2$ (se=2.1), $p < 0.001$

$\beta_2 = 6$ (se=86), p=.94

$\beta_3 = 2.2$ (se=8.4), p=.80

Longitudinal Analysis: Allowing Dependence

Statistical Assumptions:

- CD4 counts are normally distributed
- CD4 counts are independent from subject to subject
- CD4 counts are dependent within a subject (over time)

Your Responsibilities:

- Design your study so that you have sufficient power, even after taking account of the dependence
- Use an analysis that takes into account the dependence
- Organize the database so that there is a variable that identifies which outcomes are linked together

Correlation and Sample Size

In Common Language: *Correlation*, *Association* and *Dependence* are used interchangeably.

Statistically: *Correlation* has a formal definition but *Association* and *Dependence* do not. It is possible for two variables to be associated but have zero correlation.

(If both variables are normally distributed, then the sentence above is false. Association and correlation must go together.)

Definition: Correlation

Correlation: A numerical summary which measures the strength of association between two variables.

- It is always ≤ 1 and ≥ -1
- Values close to +1 or -1 indicate a strong relationship between the two variables.
- A value of 0 indicates no association.
- It is possible to test whether the correlation is 0
- It is possible to calculate confidence intervals
- A Pearson correlation is used for normal data and a Spearman correlation for non-normal data.

Correlation

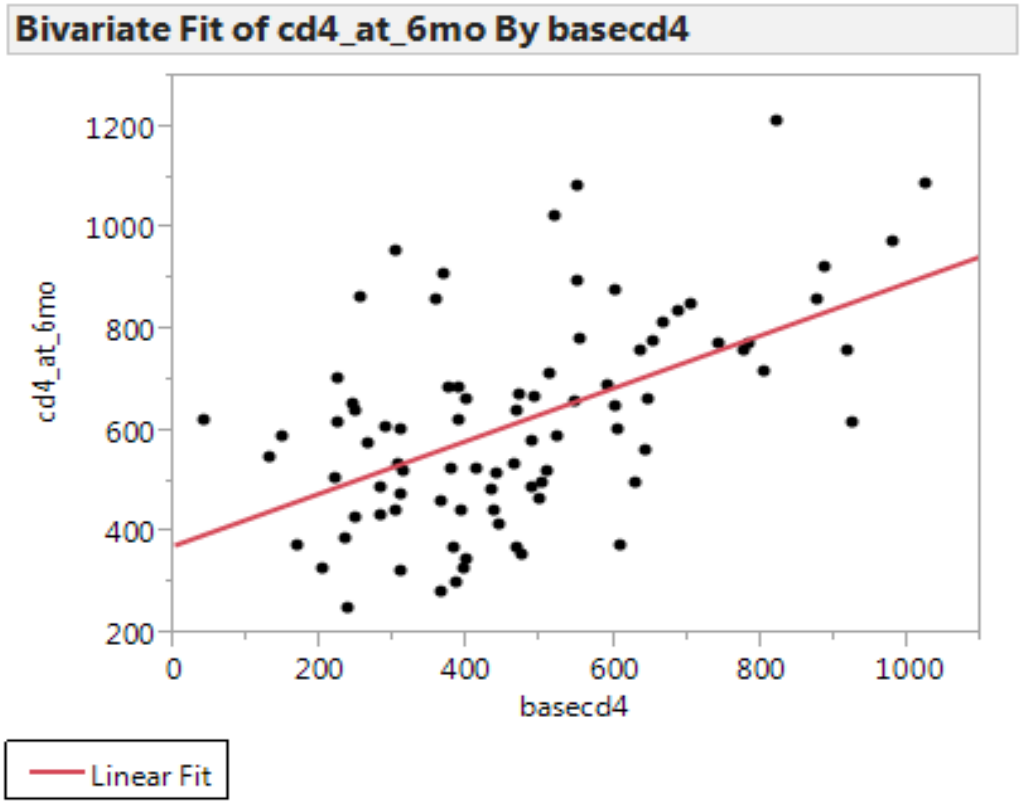
The closer the correlation is to +1 or -1, the more precisely one variable can be predicted by the other variable.

In terms of a scatter plot, the closer the correlation is to +1 or -1, the closer the points are to a line going through the scatter plot.

- The angle of the line does not affect the correlation coefficient. (Unless the line is horizontal, in which case the correlation is 0.)

Example: Scatter Plot and Correlation

Baseline CD4 versus 6 Month CD4: Correlation = 0.53



Between-Subject vs Within-Subject Effects

- In a correlated data analysis, you can estimate the effect of any between-subject predictor, or of any within-subject predictor.
- If you have positive correlation within-subject (the usual expectation), then that correlation will:
  - Decrease the power of your analysis of between-subject predictors.
  - Increase the power of your analysis of within-subject predictors.

Empirical Results

Main Effects Model: $CD4 = \alpha_0 + \beta_1 Time + \beta_2 Sex$

Assuming Independence:

$$CD4 = 494 + 17.6\ Time + 10\ Female$$

Where $\beta_1 = 17.6$ (se=2.7), $p < 0.001$

$\beta_2 = 10$ (se=54), p=.86

Allowing Dependence:

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

Where $\beta_1 = 17.6$ (se=2.0), $p < 0.001$

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

Between-Subject Effects

Longitudinal Study:

- Compare serial CD4 counts between HIV-infected and non-infected subjects (i.e., some subjects are HIV-infected; some are not)

- Compare serial blood pressures between intervention and control patients.

Within-Subject Effects

Longitudinal Studies:

- Estimate how serial CD4 counts change over time.
- Estimate the effect of any predictor which also changes over time (i.e., presence/absence of an infection prior to the CD4 count).

Impact of Correlation: Between-Subject Effects

Basic Idea: If the CD4 counts are correlated within subject and, for example, we want to compare CD4 counts between HIV-infected and uninfected subjects (or, between treated and untreated), then:

- 2 measurements on the same subject are not worth as much as 1 measurement on each of 2 subjects

or

- 2 measurements on the same subject are not worth twice as much as one measurement

Correlation and Between-Subject Effects

For example:

If Correlation=0, then $\sigma^2 = 2\sigma^2_1$

If Correlation=1, then $\sigma^2 = \sigma^2_1$

If 0<Correlation<1, then $\sigma^2 = \sigma^2_1$.?

Specifically, If Corr = .10, then $\sigma^2 = 1.8\sigma^2_1$

If Corr = .25, then $\sigma^2 = 1.6\sigma^2_1$

If Corr = .50, then $\sigma^2 = 1.3\sigma^2_1$

Correlation and Between-Subject Effects

Example:

- Assume that the experiment involves 25 subjects, but each subject is measured 4 times.
- This means that there are $4 \times 25 = 100$ outcome measurements available for analysis.
- However, the "effective sample size" is somewhere between N=25 (the number of subjects) and N=100 (the number of measurements).
- The true number depends on the magnitude of the correlation between the 4 measurements on each subject.

The Design Effect

Definition 1: Intra-Class Correlation Coefficient (ICC)

$$\rho = \sigma^2_{Between} / (\sigma^2_{Between} + \sigma^2_{Within})$$

(This ranges from 0 to 1, where we always presume positive correlation.)

Definition 2: Design Effect (D)

Assume that there are k correlated measures on each "subject":

$$D = 1 + (k-1)\rho$$

The Effective Sample Size

Formula:

$$\text{Effective Sample Size} = \frac{\text{Total Number of Outcomes}}{\text{D}}$$

Implication: The power of a study is based on the Effective Sample Size, not the total number of outcomes measured.

Therefore, power calculations should be based on the Effective Sample Size.

Effective Sample Size: Example

Prior Example:

If we assume that for each of the 25 subjects, the correlation between the k=4 repeated measurements was ρ=0.5,

Then $D = 1 + 3(.5) = 2.5$

$N_{Correlated} = 100$

And Eff Sample Size $= N_{Correlated} / D$

$= 100/2.5 = 40$

So, the "effective sample size" is 40

Example: Longitudinal Design

Sample Size Required, Per Group:

Power=90%, α=5% (2-sided), $(u_1-u_2)/\sigma = .50$

	Number of Patients	Number of Measurements
Independent Data:	84	84
2 Measurements Per Subject		
$\rho = .05$	44	88
$\rho = .10$	46	92
$\rho = .25$	52	105
$\rho = .50$	63	126

Example: Longitudinal Design

Sample Size Required, Per Group:

Power=90%, α=5% (2-sided), $(u_1-u_2)/\sigma = .50$

	Number of Patients	Number of Measurements
Independent Data:	84	84
3 Measurements Per Subject		
$\rho = .05$	31	92
$\rho = .10$	34	101
$\rho = .25$	42	126
$\rho = .50$	56	168

Correlation: Within-Subject Effects

Basic Idea: If the CD4 counts are correlated within subject and, for example, we want to compare how CD4 counts change over time (or analyze a cross-over study), then:

- 2 measurements on the same subject are worth **more** than 1 measurement on each of 2 subjects

or

- 2 measurements on the same subject are worth **more** than twice as much as one measurement

Correlation and Within-Subject Effects

For example:

If Correlation=0, then 2 = 2x1

If Corr = .10, then 2 = 2.2

If Corr = .25, then 2 = 2.7

If Corr = .50, then 2 = 4.0

Note: The usual rule-of-thumb is that you need 4-times as many patients for a parallel trial as for a cross-over trial. This presumes a correlation of 0.5 between serial measurements on the same patient.

Within Subject Effects: Example

Situation: Assume we want to compare CD4 counts at diagnosis to CD4 counts 1-year after diagnosis, where CD4 counts have a standard deviation of 250 and change by 100 over the year.

- If we measure 100 subjects at the time of diagnosis and a different 100 subjects at 1 year, we will have 80% power.

- If we measure the same 100 subjects both at diagnosis and at 1 year, and assume a correlation of 0.30, we will have 92% power.

Practical Issues: Study Design

- In planning a study where you will collect correlated outcomes, you must take the correlation into account in your power/sample size calculation.
  - Estimate the average number of correlated outcomes per "subject" (i.e., k)
  - "Guesstimate" the correlation within subject
  - Calculate the design effect, D. (Use a negative sign in the calculation if you are interested in within-subject effects.)

Practical Issues: Study Design

- If you are doing a sample size calculation: Use your usual software/formula to calculate the necessary sample size (this assumes independent data). Then multiply this sample size by D.

- If you are doing a power calculation: Divide the total sample size (# of subjects times k) by D to get the effective sample size. Use your usual software/formula to calculate the power based on the effective sample size.

Practical Issues: Analysis

- If your study design involves correlated outcomes, then you must use software that recognizes and adjusts for the correlation.
  - You will need to specify the form of the correlation matrix (next lecture).
  - You will need to identify a variable which is the same for every outcome from the same subject.
  - Then, the software will measure and adjust for the correlation. The process will be transparent to you. You do not have to know whether it is a within-person or between-person effect.

Data Format: Wide

Usual Form of a Database:

Obs	subject	basecd4	at_6mo	12_mo	Age	gender
1	JC-01*	1023	1087	651	30.6438	M
2	ND-02*	643	561	645	42.3507	M
3	SJ-03*	463	534	1019	30.3315	M
4	DK-04	652	777	682	34.9671	M
5	KM-05*	289	608	715	37.3753	M
6	KS-06*	551	1085	1037	36.1671	M
7	NF-07**	131	548	496	22.8767	F
8	RM-08*	263	575	631	39.1918	M
9	JF-09	365	459	594	35.6795	M
10	AR-10	919	757	836	30.6959	M

Data Format: Long

In Order to Carry Out Correlated Data Analyses:

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

Conclusions

- In some studies, the outcomes you collect may not be independent of one another.
  - Multiple outcomes over time; Multiple patients from each clinic.
- If you have a study with dependent/correlated outcomes, it will effect your results.
  - The standard errors will change more than the effect estimates.
  - If you are looking at predictors which change within subjects (time), the standard errors will go down.
  - If you are looking at predictors which compare subjects to each other (gender), the standard errors will go up.
- Your power calculations must take account of the correlation.
- Your analysis must take account of the correlation.
- You database must be organized so that correlated observations are evident.