# Longitudinal Data Analysis

Author: Nicholas G Reich, Jeff Goldsmith

*This material is part of the **statsTeachR** project*

# Focus on covariance

- We've extensively used OLS for the model

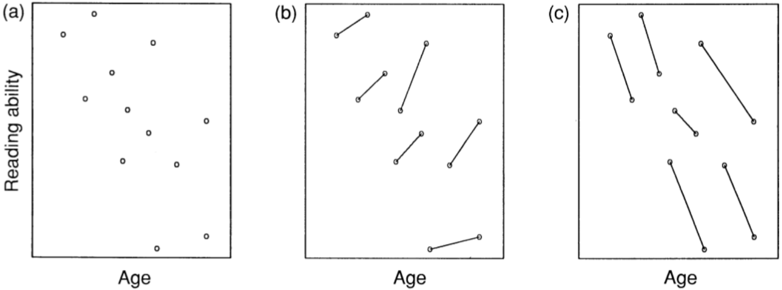
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

  where $E(\boldsymbol{\epsilon}) = 0$ and $Var(\boldsymbol{\epsilon}) = \sigma^2 I$
- We are now more interested in the case of $Var(\boldsymbol{\epsilon}) = \sigma^2 V$

# Longitudinal data

- Data is gathered at multiple time points for each study participant
- Repeated observations / responses
- Longitudinal data regularly violates the "independent errors" assumption of OLS
- LDA allows the examination of changes over time (aging effects) and adjustment for individual differences (subject effects)

# Some hypothetical data



# Notation

- We observe data $y_{ij}, \boldsymbol{x}_{ij}$ for subjects $i = 1, \dots I$ at visits $j = 1, \dots, J_i$
- Vectors $\boldsymbol{y}_i$ and matrices $\boldsymbol{X}_i$ are subject-specific outcomes and design matrices
- Total number of visits is $n = \sum_{i=1}^{I} J_i$
- For subjects $i$, let

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{\epsilon}_i$$

  where $\text{Var}(\boldsymbol{\epsilon}_i) = \sigma^2 V_i$

# Notation

- Overall, we pose the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2 V$ and

$$V = \begin{bmatrix} V_1 & 0 & \dots & 0 \\ 0 & V_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & & V_I \end{bmatrix}$$

# Covariates

The covariates $\boldsymbol{x}_i = x_{ij1} \ldots x_{ijp}$ can be

- Fixed at the subject level – for instance, sex, race, fixed treatment effects
- Time varying – age, BMI, smoking status, treatment in a cross-over design

# Motivation

Why bother with LDA?

- Correct inference
- More efficient estimation of shared effects
- Estimation of subject-level effects / correlation
- The ability to "borrow strength" – use both subject- and population-level information
- Repeated measures is a very common feature of real data!

# Example dataset

An example dataset comes from the Multicenter AIDS Cohort Study (`CD4.txt`).

- 283 HIV+ individuals
- Observation of CD4 cell count (a measure of disease progression)
- Between 1 and 14 observations per subject (1817 total observations)

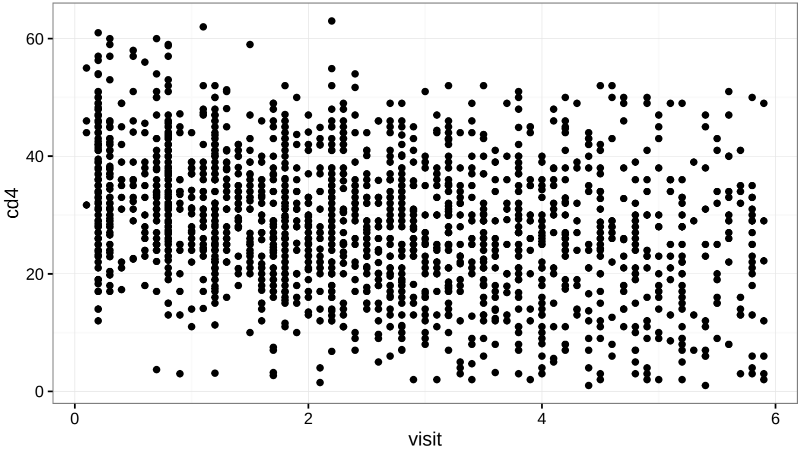
# CD4 dataset

```r
library(timereg)
data(cd4)
head(cd4, 15)
```

```
##     obs   id visit smoke    age precd4 cd4  lt  rt cd4.prev
## 1     1 1022   0.2     0 26.250   38.0  17 0.0 0.2     38.0
## 2     2 1022   0.8     0 26.250   38.0  30 0.2 0.8     17.0
## 3     3 1022   1.2     0 26.250   38.0  23 0.8 1.2     30.0
## 4     4 1022   1.6     0 26.250   38.0  15 1.2 1.6     23.0
## 5     5 1022   2.5     0 26.250   38.0  21 1.6 2.5     15.0
## 6     6 1022   3.0     0 26.250   38.0  12 2.5 3.0     21.0
## 7     7 1022   4.1     0 26.250   38.0   5 3.0 4.1     12.0
## 8     8 1049   0.3     0 32.375   44.5  37 0.0 0.3     44.5
## 9     9 1049   0.6     0 32.375   44.5  44 0.3 0.6     37.0
## 10   10 1049   1.0     0 32.375   44.5  37 0.6 1.0     44.0
## 11   11 1049   1.5     0 32.375   44.5  35 1.0 1.5     37.0
## 12   12 1049   2.0     0 32.375   44.5  25 1.5 2.0     35.0
## 13   13 1049   2.5     0 32.375   44.5  21 2.0 2.5     25.0
## 14   14 1049   3.0     0 32.375   44.5  22 2.5 3.0     21.0
## 15   15 1049   3.5     0 32.375   44.5  21 3.0 3.5     22.0
```

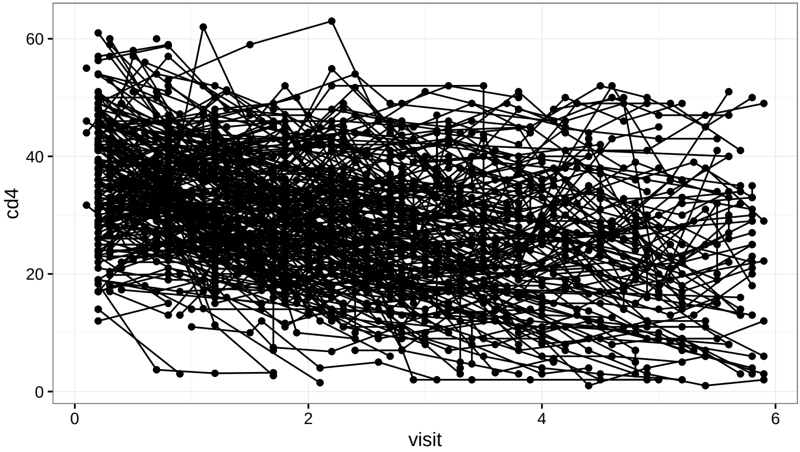
# CD4 dataset

```
qplot(visit, cd4, data=cd4)
```



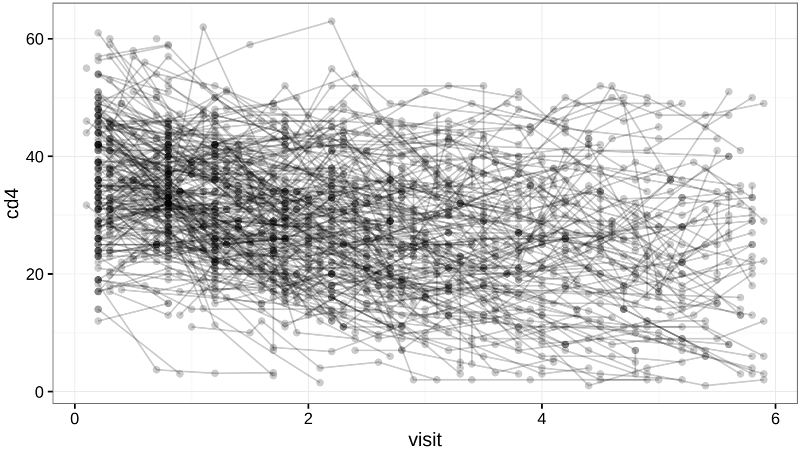
# CD4 dataset

```
qplot(visit, cd4, data=cd4, geom=c("point", "line"),
      group=id)
```



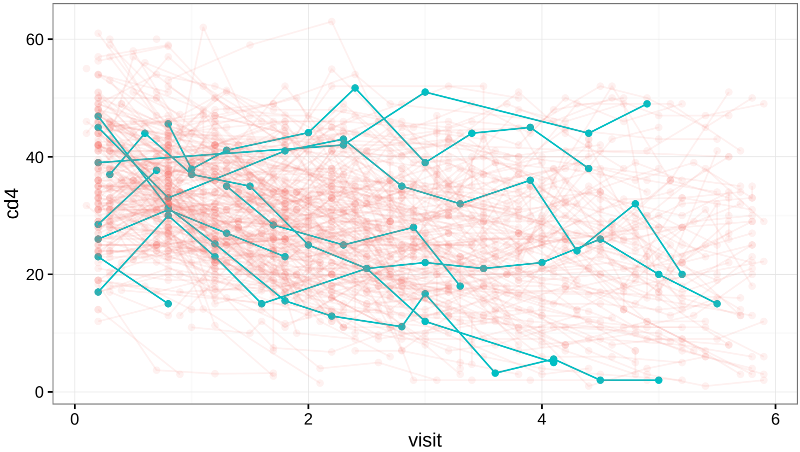
# CD4 dataset

```
qplot(visit, cd4, data=cd4, geom=c("point", "line"),
      group=id, alpha=I(.2))
```



# CD4 dataset

```
ids <- unique(cd4$id)
cd4$highlight <- as.factor(cd4$id %in% ids[1:10])
qplot(visit, cd4, data=cd4, geom=c("point", "line"),
      group=id, color=highlight, alpha=highlight) +
        theme(legend.position="none")
```



# Visualizing covariances

Suppose the data consists of three subjects with four data points each.

- In the model

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{\epsilon}_i$$

  where $\text{Var}(\boldsymbol{\epsilon}_i) = \sigma^2 V_i$, what are some forms for $V_i$?

# Approaches to LDA

We'll consider two main approaches to LDA

- Marginal models, which focus on estimating the main effects and variance matrices but don't introduce subject effects
  - "Simplest" LDA model, just like cross-sectional data
  - Requires new methods, like GEE, to control for variance structure
  - Arguably easier incorporation of different variance structures
- Random effects models, which introduce random subject effects (i.e. effects coming from a distribution, rather than from a "true" parametric model)
  - "Intuitive" model descriptions
  - Explicit estimation of variance components
  - Caveat: can change parameter interpretations

# First problem: exchangeable correlation

Start with the model where

$$V_i = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \\ \rho & \rho & & 1 \end{bmatrix}$$

This implies

- $var(y_{ij}) = \sigma^2$
- $cov(y_{ij}, y_{ij'}) = \sigma^2 \rho$
- $cor(y_{ij}, y_{ij'}) = \rho$

# Marginal model

The marginal model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

- $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2 V$,
- $$V_i = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \\ \rho & \rho & & 1 \end{bmatrix}$$

Tricky part is estimating the variance of the parameter estimates for this new model.

# Fitting a marginal model using GEE

Generalized Estimating Equations provide a semi-parametric method for fitting a marginal model that takes into account the correlation between observations.

$$\mathbb{E}[CD4_{ij}|month] = \beta_0 + \beta_1 \cdot month$$

With GEE, assume $V_i$ is exchangeable.

```
library(geepack)
linmod <- lm(cd4~visit, data=cd4)
geemod <- geeglm(cd4~visit, data=cd4, id=id,
                 corstr="exchangeable")
```

# Fitting a marginal model using GEE

$$\mathbb{E}[CD4_{ij}|month] = \beta_0 + \beta_1 \cdot month$$

With GEE, assume $V_i$ is exchangeable.

```
summary(linmod)$coef

##              Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) 35.010678  0.4585794  76.34595 0.00000e+00
## visit       -2.447625  0.1627810 -15.03630 3.01226e-48

summary(geemod)$coef

##             Estimate   Std.err      Wald Pr(>|W|)
## (Intercept) 35.36883 0.5951037 3532.2872        0
## visit       -2.67221 0.2175556  150.8693        0
```

# Looking at the correlation structures: exchangeable

```
summary(geemod)
```

```
##
## Call:
## geeglm(formula = cd4 ~ visit, data = cd4, id = id, corstr = "exchangeable")
##
##  Coefficients:
##             Estimate Std.err   Wald Pr(>|W|)
## (Intercept)  35.3688  0.5951 3532.3   <2e-16 ***
## visit        -2.6722  0.2176  150.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Estimated Scale Parameters:
##             Estimate Std.err
## (Intercept)    116.7   7.036
##
## Correlation: Structure = exchangeable  Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.6566  0.0369
## Number of clusters:   283   Maximum cluster size: 14
```

# Looking at the correlation structures: AR(1)

```
geemod1 <- geeglm(cd4~visit, data=cd4, id=id,
                  corstr="ar1")
summary(geemod1)
```

```
##
## Call:
## geeglm(formula = cd4 ~ visit, data = cd4, id = id, corstr = "ar1")
##
##  Coefficients:
##             Estimate Std.err Wald Pr(>|W|)
## (Intercept)   35.644   0.642 3079   <2e-16 ***
## visit         -2.761   0.233  140   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Estimated Scale Parameters:
##             Estimate Std.err
## (Intercept)      117    7.03
##
## Correlation: Structure = ar1  Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.891   0.016
## Number of clusters:   283   Maximum cluster size: 14
```

# Comparing different GEE models

Not a straight-forward way to compare different correlation structures

- Some work on AIC in the context of GEEs (Pan, 2001)
- Not implemented in standard GEE packages
- In practice, knowledge of data structure guides choice.

# Marginal model

The marginal model formulation is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

- $\boldsymbol{\epsilon} \sim \mathrm{N}\left[0, \sigma^2 V\right]$

This approach focuses on the *marginal* distribution of $\boldsymbol{y}$, rather than on a subject-level *conditional* distribution.

# Can use Generalized Least Squares

Given the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 V)$ with $V$ known, we are essentially assuming

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 V)$$

Using MLE, we find that $\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}^T V^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T V^{-1} \boldsymbol{y}$

# Estimation – marginal model

- If we can use MLE when $V$ is known, maybe we can use MLE to estimate $V$ as well
- Our log likelihood function is

$$
\begin{aligned}
l(\boldsymbol{\beta}, \sigma^2, V; \boldsymbol{y}, \boldsymbol{X}) &= -\frac{1}{2}\Big[n\log(\sigma^2) + \log(|V|) \\
&\quad + \frac{1}{\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T V^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\Big]
\end{aligned}
$$

- Using profile likelihood, we find that for any $V_0$

$$\hat{\boldsymbol{\beta}}(V_0) = (\boldsymbol{X}^T V_0^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T V_0^{-1}\boldsymbol{y}$$

# Estimation – marginal model

- Estimation of $V$ and $\sigma$ is done through restricted maximum likelihood
  - Standard MLE produces biased variance estimates; REML adjusts for the number of fixed effects components that are estimated
- Often $V$ is structured parametrically to ease estimation and computation
- We won't worry about how this is done

# Random effects model

A random intercept model with one covariate is given by

$$y_{ij} = \beta_0 + b_i + \beta_1 x_{ij} + \epsilon_{ij}$$

where

- $b_i \sim \mathrm{N}\left[0, \tau^2\right]$
- $\epsilon_{ij} \sim \mathrm{N}\left[0, \nu^2\right]$

**For exchangeable correlation and continuous outcomes**, the random intercept model is equivalent to the marginal model. Under this model

- $var(y_{ij}) =$
- $cov(y_{ij}, y_{ij'}) =$
- $cor(y_{ij}, y_{ij'}) = \rho =$

# Fitting a random effects model

```
library(lme4)
memod <- lmer(cd4 ~ (1 | id) + visit, data = cd4)
summary(memod)$coef

##             Estimate Std. Error t value
## (Intercept)    35.37      0.597    59.3
## visit          -2.67      0.108   -24.8

summary(geemod)$coef

##             Estimate Std.err Wald Pr(>|W|)
## (Intercept)    35.37   0.595 3532        0
## visit          -2.67   0.218  151        0
```

# Conclusion

Today we have..

- introduced longitudinal data analysis.
- defined and fitted Marginal and Random Effects models.