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

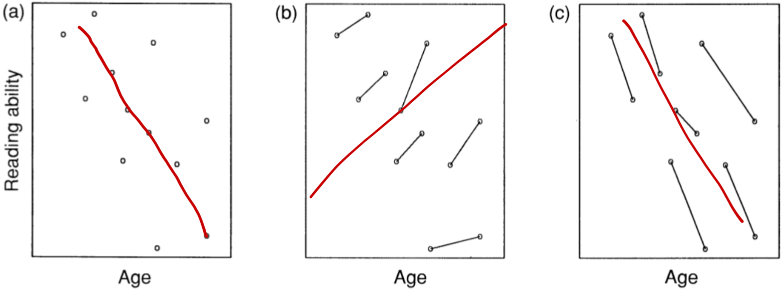
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

  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 $y_{ij}, \mathbf{x}_{ij}$ for subjects $i = 1, \ldots I$ at visits $j = 1, \ldots, J_i$ $\longleftarrow$ balanced if $J_i = J \; \forall i$
- Vectors $\mathbf{y}_i$ and matrices $\mathbf{X}_i$ are subject-specific outcomes and design matrices
- Total number of visits is $n = \sum_{i=1}^{I} J_i$
- For subjects i, let

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

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

Notation

- Overall, we pose the model

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

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

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

$$\begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ 0 & & 1 & \\ & & & \ddots \end{bmatrix}$$

independence

unstructured

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 \\ & 1 & \rho_5 & \rho_6 & \rho_7 \\ & & 1 & \ddots & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}$$

autoregressive

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}$$

Covariates

The covariates $\mathbf{x}_i = x_{ij1} \ldots x_{ijp}$ 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`).

- 366 HIV+ individuals
- Observation of CD4 cell count (a measure of disease progression)
- Between 1 and 11 observations per subject (1888 total observations)

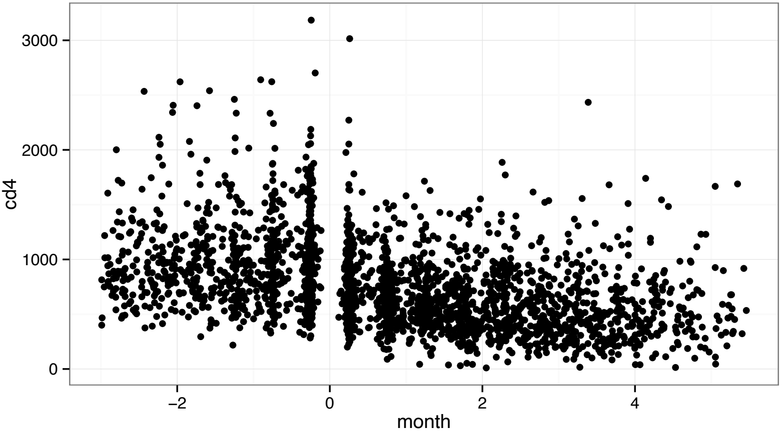
CD4 dataset

```
data <- read.table("CD4.txt", header = TRUE)
head(data, 15)

##      month  cd4  age packs drug part ceased    ID
## 1  -0.7420  548 6.57     0    0    5      8 10002
## 2  -0.2464  893 6.57     0    1    5      2 10002
## 3   0.2437  657 6.57     0    1    5     -1 10002
## 4  -2.7296  464 6.95     0    1    5      4 10005
## 5  -2.2505  845 6.95     0    1    5     -4 10005
## 6  -0.2218  752 6.95     0    1    5     -5 10005
## 7   0.2218  459 6.95     0    1    5      2 10005
## 8   0.7748  181 6.95     0    1    5     -3 10005
## 9   1.2567  434 6.95     0    1    5     -7 10005
## 10 -1.2402  846 2.64     0    1    5     18 10029
## 11 -0.7420 1102 2.64     0    1    5     18 10029
## 12 -0.2519  801 2.64     0    1    5     38 10029
## 13  0.2519  824 2.64     0    1    5      7 10029
## 14  0.7693  866 2.64     0    1    5     15 10029
## 15  1.4127  704 2.64     0    1    5     21 10029
```

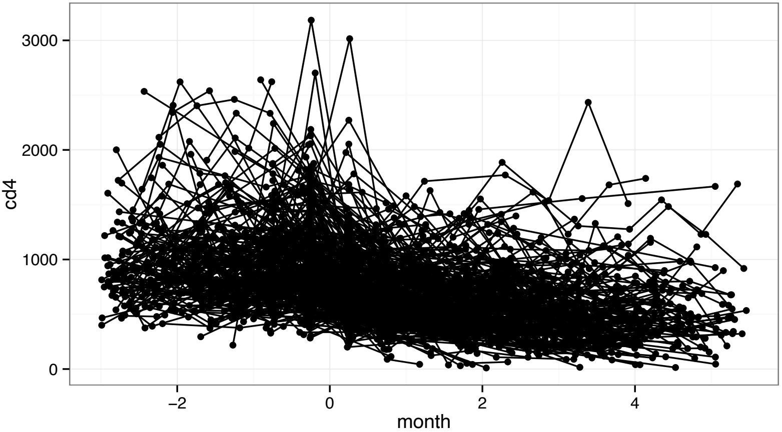
CD4 dataset

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



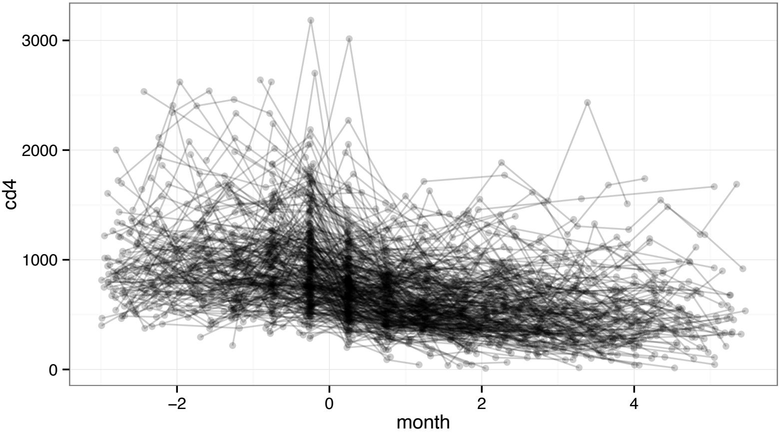
CD4 dataset

```r
qplot(month, cd4, data=data, geom=c("point", "line"),
      group=ID)
```



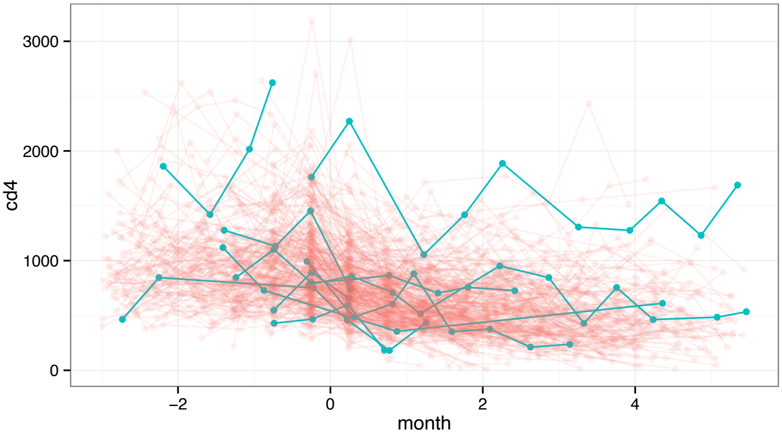
CD4 dataset

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



CD4 dataset

```
IDS <- unique(data$ID)
data$highlight <- as.factor(data$ID %in% IDS[1:10])
qplot(month, cd4, data=data, 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

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

  where $\mathrm{Var}(\boldsymbol{\epsilon}_i) = \sigma^2 V_i$, 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-sectinal 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

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

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.

```
require(gee)

## Warning:  package 'gee' was built under R version 3.1.2

linmod <- lm(cd4~month, data=data)
geemod <- gee(cd4~month, data=data, 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)   838.82      8.136  103.11  0.000e+00
## month         -88.95      3.964  -22.44 2.628e-101

summary(geemod)$coef

##             Estimate Naive S.E. Naive z Robust S.E. Robust z
## (Intercept)   836.93     14.794   56.57      15.261    54.84
## month         -99.73      3.429  -29.08       5.056   -19.73
```

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

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

Fitting a random effects model

subject specific
intercept

```
require(lme4)
memod <- lmer(cd4 ~ (1 | ID) + month, data = data)
summary(memod)$coef

##              Estimate Std. Error t value
## (Intercept)    836.96     14.652   57.12
## month          -99.66      3.448  -28.90

summary(geemod)$coef

##              Estimate Naive S.E. Naive z Robust S.E. Robust z
## (Intercept)    836.93     14.794   56.57      15.261    54.84
## month          -99.73      3.429  -29.08       5.056   -19.73
```

Conclusion

Today we have..

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