Linear Regression Models
P8111

Lecture 20

Jeff Goldsmith
April 7, 2016

THE DEPARTMENT OF BIOSTATISTICS
Columbia University
MAILMAN SCHOOL
OF PUBLIC HEALTH

Today's Lecture

- Longitudinal data analysis

Focus on covariance

- We've extensively used OLS for the model

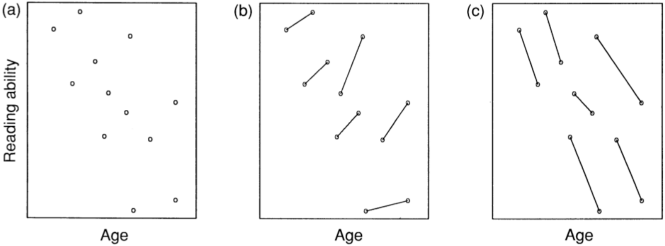
$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

  where $E(\boldsymbol{\epsilon}) = 0$ and $Var(\boldsymbol{\epsilon}) = \sigma^2 I$
- We are now more interested in the case of $Var(\boldsymbol{\epsilon}) = \sigma^2 V$
- WLS and GLS were useful in this setting, but required a known V matrix

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 $i = 1, \ldots I$ at visits $j = 1, \ldots, J_i$
- 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?

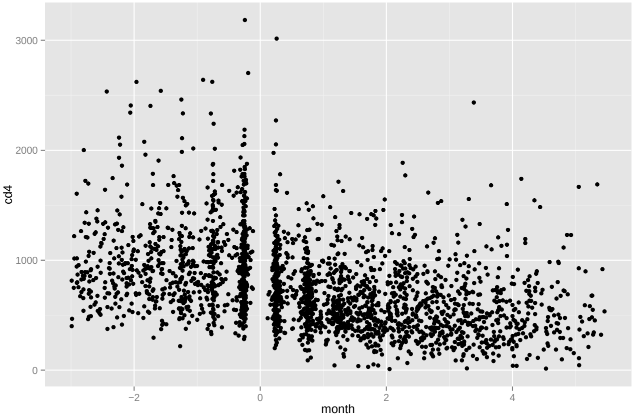
- Correct inference
- More efficient estimation of shared effects
- Estimation of subject-level effects / correlation
- The ability to "borrow strength" – use both subject- and population-level information

Example dataset

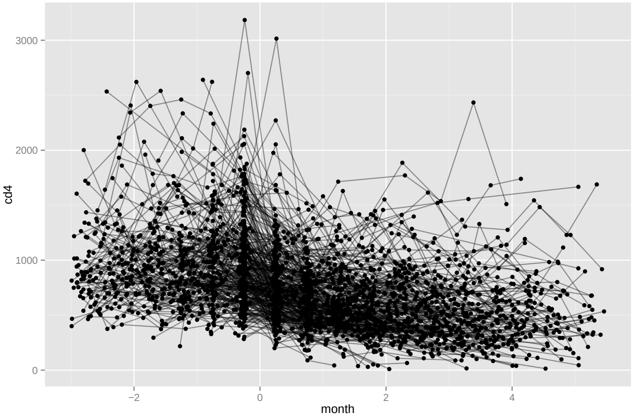
An example dataset comes from the Multicenter AIDS Cohort Study

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

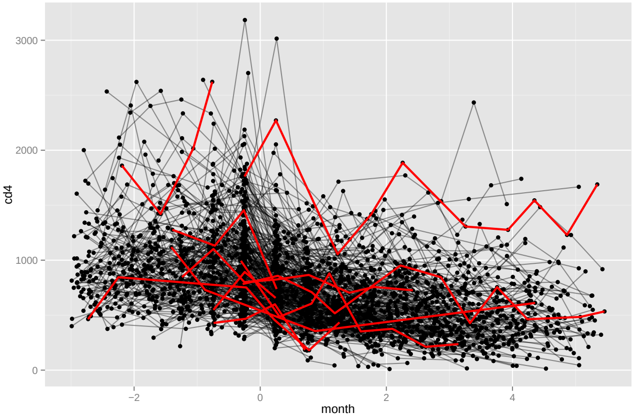
Example dataset



Example dataset



Example dataset



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

Approaches to LDA

We'll consider two main approaches to LDA

- Random effects models, which introduce random subject effects (i.e. effects coming from a distribution, rather than from a "true" parametric model)
- Marginal models, which focus on estimating the main effects and variance matrices but don't introduce subject effects

First problem: uniform 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

If we assume a uniform correlation structure, the marginal model is

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

where

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

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

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

Under this model

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

Relationship between marginal and RI models

The random intercept model implies a correlation structure equivalent to the mixed model, with

- $\sigma^2 = \tau^2 + \nu^2$
- $\rho = \frac{\tau^2}{\tau^2+\nu^2}$

(This works with continuous responses, but be careful with generalized outcomes)

Partitioning variance

- Whether we look at random effects or marginal modeling, we have to partition total variability into subject-level variance and population-level variance
- In a random effects framework, we estimate between and within subject variance components
- In a marginal model framework, we estimate a within subject variance and a covariance matrix

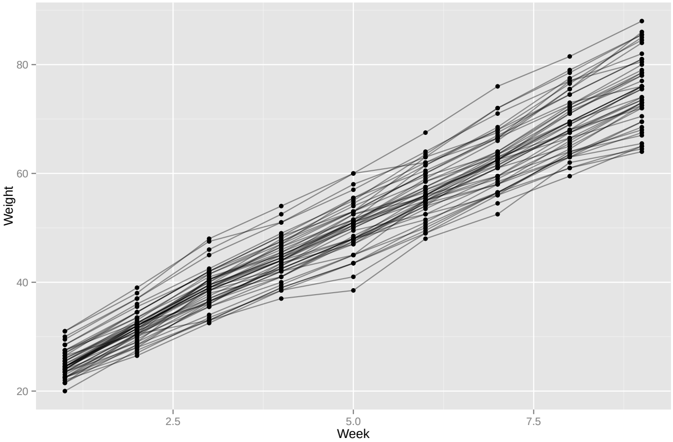
Interpretation of ICC

- The quantity $\rho = \frac{\tau^2}{\tau^2+\nu^2}$ is called the intraclass correlation
- It tells how strongly observations within a subject are correlated relative to the overall population variance
- Alternatively, the ICC tells what proportion of variability is within-subject variability

Pig weight data

- Weight on 48 pigs
- Nine measurements per pig

Pig weight data



Pig weight data

- Apparent linear relationship
- High variance across pigs compared to variance within pigs
- Each pig's "baseline" is very important for future observations

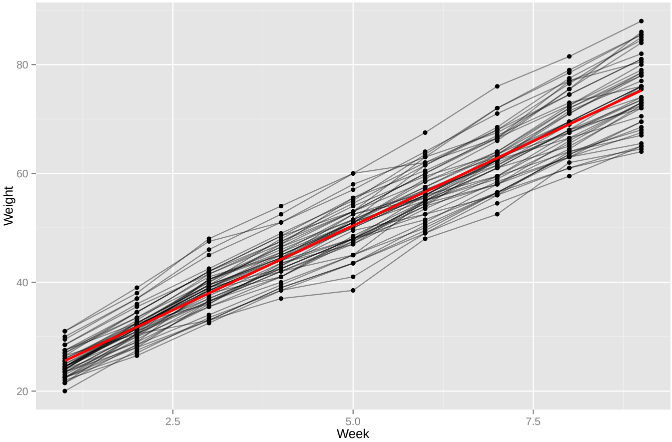
Pig weight data analysis

Using ordinary least squares, we find

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.35561    0.46054   42.03   <2e-16 ***
num.weeks    6.20990    0.08184   75.88   <2e-16 ***

Residual standard error: 4.392 on 430 degrees of freedom
```

Pig weight data analysis



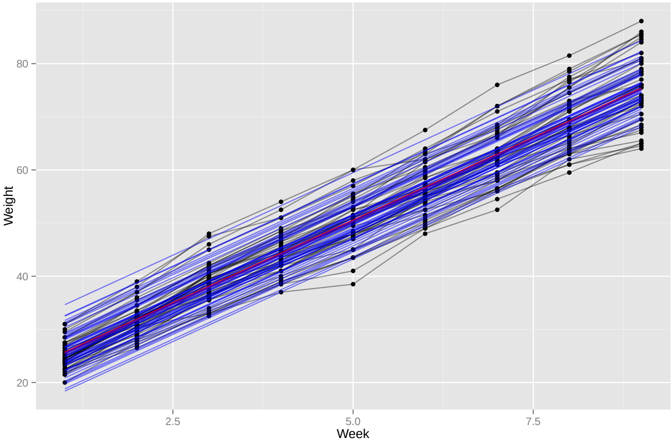
Pig weight data analysis

Using a random intercept model, we find

```
Random effects:
 Groups   Name        Variance Std.Dev.
 id.num   (Intercept) 15.1418  3.8913
 Residual              4.3947  2.0964
Number of obs: 432, groups: id.num, 48

Fixed effects:
            Estimate Std. Error t value
(Intercept) 19.35561    0.60311   32.09
num.weeks    6.20990    0.03906  158.97
```

Pig weight data analysis



Next time

- Why do we use random effects rather than creating subject-level indicator variables and estimating fixed effects?
- Next time we'll talk about estimation of random effect and marginal models

Today's big ideas

- Longitudinal data analysis
- Uniform correlation models