# BTRY/STSCI 4030/5030: Linear Models with Matrices

Giles Hooker

Fall 2018
MW 2:55 - 4:10
101 Phillips Hall

# Instructor

- Professor: Giles Hooker, BSCB
- Office: 1186 Comstock Hall
- Email: gjh27@cornell.edu
- Office Hours: Thursday 2:00 - 3:00, Comstock 1181/1186
- Webpage:

  www.bscb.cornell.edu/∼hooker/

- Labs, homework at

  cmsx.cs.cornell.edu

  discussion boards via

  piazza.com

also

campuswire.com

# TA and Labs

**Zhengze Zhou** (zz433@cornell.edu)

Office hours : Thursday 10:00-11:00, Comstock 1187

**Samriddha Lahiri** (sl2938@cornell.edu)

Office hours : Friday 1:00-2:00, Comstock 1187

**Elly Kipkogei** (ek492@cornell.edu)

Office hours : Wednesday 11:30 - 12:30, Comstock 1187

Labs (All Mann B30A): M/W/Th 7:30-9:25, W 12:20-2:15

Labs approx every two weeks, starting Aug 27 (announcements on Piazza/CMS)

# Learning Outcomes – Or What's The Point?

Learning Outcomes listed as

1. Formulate linear and linear mixed models for data analysis using matrix algebra,
2. Use a statistical computing package to analyze data using linear mixed models, and
3. Derive the repeated sampling properties of estimators and test-statistics obtained from linear mixed models.

My goal:
*Achieve fluency in using matrix algebra to describe linear and mixed models, and the calculations involved in estimating and performing inference with them.*

# Syllabus

- Review of matrix algebra
- Simple and multiple linear regression as matrix equations
- Estimation and inference for linear regression
- Diagnostic tools
- Distribution of quadratic forms
- Model selection methods.
- Mixed effects models and Restricted Maximum Likelihood
- Balanced factorial designs

# Notes and Software

- Lectures will be "chalk-and-talk": on the board (please ask if my writing isn't clear, and stop me if I'm going too fast).
- No official text but references (next slides)
  - Complement lecture material
  - Are available in electronic version through Cornell library
  - See list of specific material on website
  - Ask if you would like suggestions for further reading.
- Typed summary of material from previous years is available on course website; some topics will be different.

**Software**: R in Labs and Rmarkdown for Homework.

# Useful References

Background/Introductory

- Dalgaard (2002). "Introductory Statistics with R". Springer.
- Brown (2014). "Linear Models in Matrix Form", Springer.
- Harville (2008) "Matrix Algebra From a Statistician's Perspective". Springer.

About right (topics covered may vary)

- Christensen (2011) "Plane Answers to Complex Questions: The Theory of Linear Models", Springer.
- Moser (1996) "Linear Models: A Mean Model Approach", Academic Press.
- Draper and Smith (1998). "Applied Regression Analysis", Wiley.
- Renchler and Schaalje (2008). "Linear Models in Statistics", Springer.

# Useful References

More Technical and Mixed Models

- Seber and Lee (2003). “Linear Regression Analysis”, Wiley.
- Verbeke and Molenberghs (2000). “Linear Mixed Models for Longitudinal Data”. Springer.
- Searle, Casella and McCulloch (1992). “Variance Components”. Wiley.
- McCullogh, Searle and Neuhaus (2008). “Generalized Linear and Mixed Models", Wiley.

All available electronically through Cornell Library (links on Piazza).
Some pointers to material on course websites.

# Homework and What to Expect

- Class focus on *how* and *why* statistical procedures work rather than carrying out calculations.
- Presentation and assessment more theoretical than applied.
- Some questions on "derive the following" (e.g We can remove the mean from $Y$ and $X$, ignore the intercept and the slope doesn't change).
- Pen-and-paper calculations can be hand-written and scanned.
- Some applied questions to say "Look, it really does work!"
- Submit these as Rmarkdown and PDF.
- Exams: more tending towards theory, some "What would be the right model?" questions.

# Grading

- Grades will be based on
  - five homework assignments (15% each, best of four)
  - one midterm exam (15%)
  - final (25%).
- Homework
  - will be posted on blackboard
  - will be due on Fridays at 5pm
  - must be submitted to CMS, separated by question
- Two one-day extension available to everyone. Further extensions only in extremis.

# Assessment Schedule

Subject to change under unforseen circumstances:

Homework 1 Due 5pm, Friday, Sep. 14

Homework 2 Due 5pm, Friday, Sept 28

Homework 3 Due 5pm, Friday, Oct. 12

Miterm Exam Tuesday, Oct. 16, 7:30pm - 9:30pm, PLS233

Homework 4 Due 5pm, Friday, Nov. 9

Homework 5 Due 5pm, Friday, Nov. 30

Final Tuesday, Dec. 11, 7:00pm - 9:00pm

Homework will typically be given out two weeks before the due date.

**Homework 1 is available now**

# Curving and letter grades

- Individual items will not be curved (unless in exceptional circumstances).
- Letter grades will be assigned based on distribution of scores among students.
- Formula not pre-set; aim is for steps of about 5%, median B+/A-; credence given to gaps between students.

# Communication

- Labs, homework, announcements will be posted on

  cmsx.cs.cornell.edu

- Discussion boards also on piazza.com, campuswire.com.
- Labs may sometimes cover material not in class – this is still part of the course.
- Labs are also intended as practice – they are better if you participate.
- Discussion boards are also available for
  - general questions
  - each homework assignment

  We will check them regularly. Please use them!
- Questions can be posted anonymously; we will also post answers to questions that are e-mailed to us or asked in office hours if we think they will be useful to others.
- Communication goes two ways. Please provide feedback.

# Assumed Background

It will be helpful for you to have seen

- Matrix Algebra
  - Addition and Multiplication
  - Matrix inverses and simultaneous equations
  - Eigenvalues and eigenvectors
- Probability
  - Means, Variances
  - Properties of variances
  - Properties of the normal distribution
- Statistics
  - Tests, p-values, confidence intervals
  - linear regression
  - random effects models

Although we will review each, briefly.

# A 4-Slide Overview I
# Linear Models

- **Simple linear regression:** single predictor $x$:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- **Multiple Linear Regession:** predictors $x_1, \ldots, x_p$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

- **Matrix formulation**: place predictors in columns of $n \times (p+1)$ matrix $\boldsymbol{X}$:

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

# A 4-Slide Overview II
# Probability and Inference

- Error distribution: $\boldsymbol{\epsilon} \sim (\mathbf{0}, \boldsymbol{\Sigma})$
- Independent normal errors with homogeneous variances:

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$$

- *Least squares estimates* are linear functions of $\boldsymbol{y}$:

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

- What are the properties of $\hat{\beta}$ in repeated sampling?
- *Fitted values* are the *projection* of $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$:

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

# A 4-Slide Overview III
# Linear Mixed Models

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

$$\boldsymbol{b} \sim (0, \boldsymbol{G}),\ \epsilon \sim (0, \boldsymbol{R})$$

- Fixed and random effects
- *Generalized least squares estimator*

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

$$\boldsymbol{\Sigma} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{R}$$

# A 4-Slide Overview IV
# ANOVA Decomposition

- Variation in $\boldsymbol{y}$ decomposed into components determined by the predictors

$$\boldsymbol{I} = \boldsymbol{A}_1 + \boldsymbol{A}_2 + \cdots + \boldsymbol{A}_k$$

$$\sum y_i^2 = \boldsymbol{y}^T \boldsymbol{I} \boldsymbol{y} = \sum_{j=1}^{k} \boldsymbol{y}^T \boldsymbol{A}_j \boldsymbol{y}$$

- Need the properties of *quadratic forms* $\boldsymbol{y}^T \boldsymbol{A}_j \boldsymbol{y}$ under repeated sampling.
- Will apply to both the linear regression and mixed models cases.

Questions?
Concerns?

# Let's Go!