STOR 557: ADVANCED METHODS OF DATA ANALYSIS

Instructor: Richard L. Smith

Class Notes:

August 19, 2021

THE UNIVERSITY *of* NORTH CAROLINA *at* CHAPEL HILL

PREREQUISITES FOR THE COURSE

1. An introductory statistics course at the level of STOR 151 or STOR 155 (should be a prerequisite to STOR 455)

2. STOR 455 or equivalent: an undergraduate-level introduction to linear models and regression, including the R statistical package (or RStudio)

3. STOR 435 - probability

4. Linear Algebra is not officially a prerequisite but most students in this class have had a course in this topic. Although the course will not require the mathematical theory of linear algebra (vector spaces etc.), I will find it useful to use matrix algebra for some derivations and to express computational formulas in a compact way.

TOPICS OF THE COURSE

1. Linear regression — assumed as a prerequisite but we will review

2. Generalized linear models

3. Random effects linear models

4. Bayesian statistics

5. Nonparametric linear models, e.g. fitting smooth curves using splines

6. The book covers some more specialized topics such as trees and neural networks — probably won't get to these but we may

OTHER REQUIREMENTS FOR THE COURSE

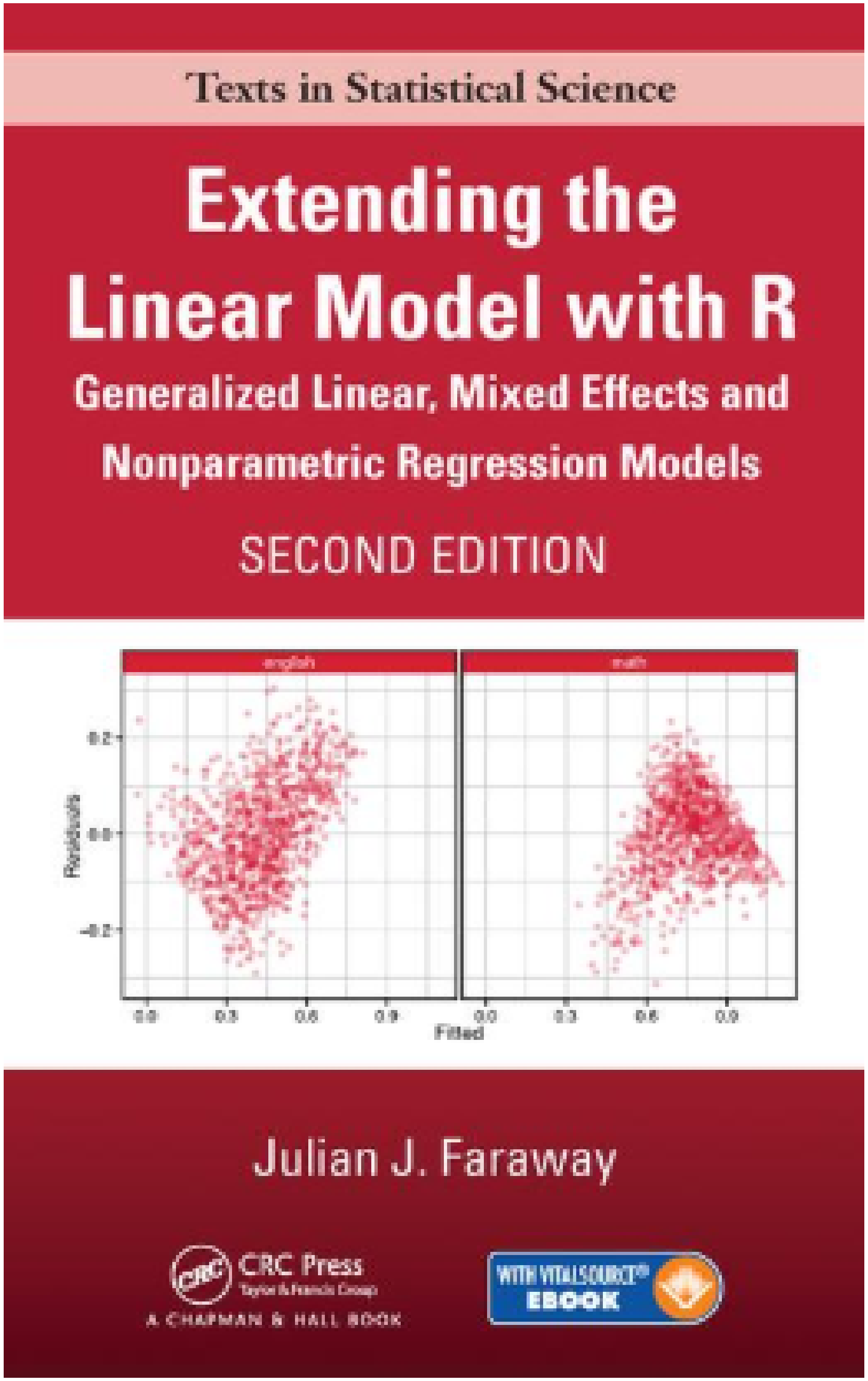
1. Text: *Extending the Linear Model with R* by Julian Faraway. Available through campus store. There is an e-book version and this would also be acceptable.

2. Make sure you get the *Second Edition.*

3. Software: use R statistical package available from `https://www.r-project.org/`.

4. Install the `faraway` package (also free — click "Install" and then "Load" from within R.)

5. RStudio also acceptable — `https://www.rstudio.com/`


Texts in Statistical Science
Extending the Linear Model with R
Generalized Linear, Mixed Effects and Nonparametric Regression Models
SECOND EDITION
Julian J. Faraway
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK
WITH VITALSOURCE® EBOOK

CLASS POLICIES

1. This is an *in-person* class and attendance at all classes is expected. That said, I recognize that individual students may have special needs and will accommodate them where possible.

2. The class begins at 9:30 and ends at 10:45. Please do not expect me to end before 10:45.

3. If you know in advance that you will not be present or will not attend the full class, I will appreciate receiving a note about it (personal email to me).

4. *Masks must be worn at all times!!*

EXAMS

1. Midterm 1, *tentatively* — take-home exam, posted online at 6:00 pm Thursday, October 7 and due (via gradescope) at 6:00 pm Friday, October 8.

2. If you have a conflict for those dates please let me know as soon as possible. If many students have a conflict, I may reschedule the exam.

3. The final exam is set by the registrar for *8:00-11:00 am, Tuesday, December 7*. I may switch to a take-home for that but assume it is an in-class exam unless announced otherwise.

BASICS OF LINEAR REGRESSION

$$y_i \;=\; x_{i0}\beta_0 + x_{i1}\beta_1 + ... + x_{ip}\beta_p + \epsilon_i, \;\; i = 1, ..., n$$

where y_i is ith value of the observation of interest, $x_{i0}, ..., x_{ip}$ are the associated covariates, and $\epsilon_1, ..., \epsilon_n$ are random errors. Here $\beta_0, ..., \beta_p$ are the unknown parameters, or regression coefficients. Usually we assume $x_{i0} = 1$ and in that case we call β_0 the intercept. Matrix form:

$$y \;=\; X\beta + \epsilon.$$

Principle of least squares: Find $\beta_0, ..., \beta_p$ to minimize

$$L \;=\; \sum_i \left(y_i - \sum_j x_{ij}\beta_j \right)^2.$$

Solve by calculus.

$$
\begin{aligned}
\frac{\partial L}{\partial \beta_0} &= -2\sum_i \left(y_i - \sum_j x_{ij}\beta_j \right) x_{i0}, \\
\frac{\partial L}{\partial \beta_1} &= -2\sum_i \left(y_i - \sum_j x_{ij}\beta_j \right) x_{i1}, \\
&\cdots \\
\frac{\partial L}{\partial \beta_p} &= -2\sum_i \left(y_i - \sum_j x_{ij}\beta_j \right) x_{ip}.
\end{aligned}
$$

We find the minimizing $\widehat{\beta}_0, ..., \widehat{\beta}_p$ by setting all the partial derivatives to 0, hence

$$
\sum_i \left(y_i - \sum_j x_{ij}\widehat{\beta}_j \right) x_{ik} = 0, \; k = 0, ..., p.
$$

Matrix notation:

$$
X^T y - X^T X \widehat{\beta} = 0.
$$

The Normal Equations

Predicted values, R^2 and R_a^2

$$\widehat{y}_i = \sum_k x_{ik}\widehat{\beta}_k$$

or in matrix notation

$$\widehat{y} = X\widehat{\beta} = X(X^T X)^{-1} X^T y.$$

We define (in case $x_{i0} \equiv 1$)

$$\begin{aligned} RSS &= \sum_i (\widehat{y}_i - y_i)^2, \\ TSS &= \sum_i (y_i - \bar{y})^2, \\ R^2 &= 1 - \frac{RSS}{TSS}. \end{aligned}$$

An alternative is the *adjusted* R^2 given by

$$R_a^2 = 1 - \frac{RSS/(n-p)}{TSS/(n-1)}.$$

Summary Tables in R

The `summary` command in R produces a table of values that includes information about

1. *The residuals* — values $r_i = y_i - \widehat{y}_i$,
2. The standard errors, t-statistics and p-values of each of the parameter estimates.

For a parameter estimate $\widehat{\beta}_k$, R will give us a standard error s_k, then

$$t_k = \frac{\widehat{\beta}_k}{s_k}$$

is called the kth t statistic, so called because it has a t_{n-p} distribution under the null hypothesis that $\beta_k = 0$.