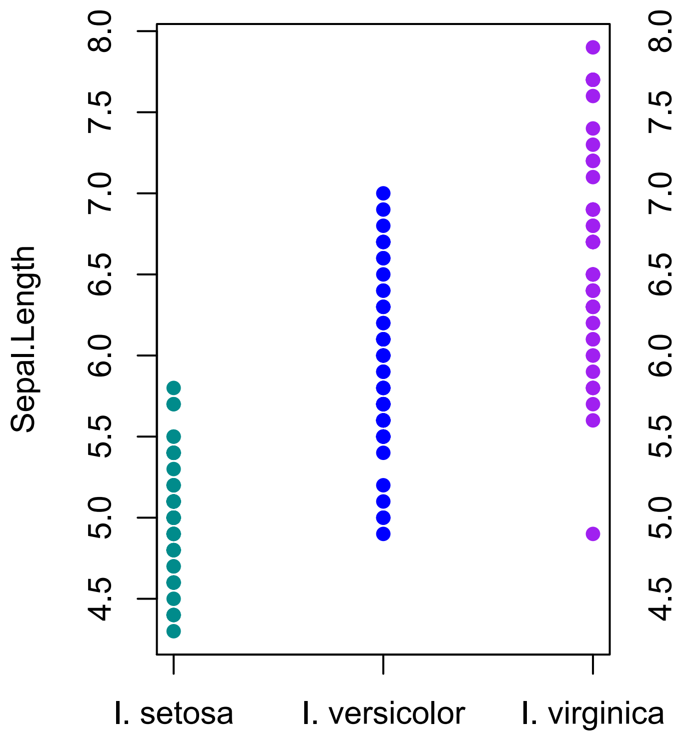
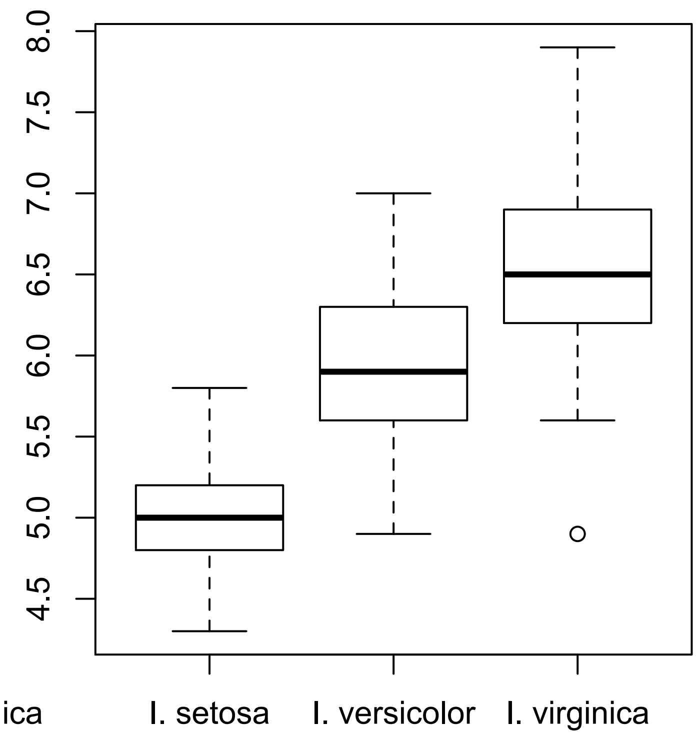
data visualization and regression






```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.0060     0.0728  68.762  < 2e-16
SpeciesI. versicolor  0.9300     0.1030   9.033 8.77e-16
SpeciesI. virginica   1.5820     0.1030  15.366  < 2e-16
```

data visualization: pros and cons

"There is no statistical tool that is as powerful as a well-chosen graph." (William Cleveland)

- dimensionality:
- 2-D, maybe 3-D in 2-D
- type of data we often work with
- makes visualization harder
- "univariate" visualization is still a good tool
- if assumptions are met, regression very useful
- can use vis. to check assumptions are met

what is our kind of data?

- sociolinguistic (response) variables usually binary
- predictor variables often categorical (factors)
  - in part because of limitations of RBRUL software
- usu. 100s of observations from 10s of speakers
- often interested in predictors on two levels
  - social or external: gender, age, social class, etc.
  - linguistic or internal: phon. context, gram. categories
- traditionally analyzed with ordinary regression

what is regression? what's a model?

- regression is descriptive stats: size of effects
- regression is inferential stats: are effects > 0, are two categories equal… (p-values!)
- demonstration using R – always use a script
- most basic function is lm( ) for linear regression
- simple linear regression: one predictor
- lm(y ~ x)
- plot(y ~ x)

regression terminology

y	x
Dependent Variable	Independent Variable
Explained Variable	Explanatory Variable
Response Variable	Control Variable
Predicted Variable	Predictor Variable
Regressand	Regressor

- distinction between predictors of interest and control predictors
- I prefer "response" and "predictors"
- errors or residuals( )

regression assumptions

- independence (of residuals)
- linearity
- normality (of residuals)
- omitted variable bias

- logistic regression (with a binary response) has fewer assumptions

goodness of fit: R^2

- regression is an attempt to account for the variability in a data set
- with linear regression, you can calculate how much of the variation has been accounted for
- this is called R^2
- it ranges from 0 to 1

extensions of linear regression

- GLM (generalized linear models)
- logistic regression
- log-odds of the response: $\ln(p / (1 - p))$
- Poisson regression: responses that are counts
- etc.

- all these can be called "fixed-effects models"
- meaning: not mixed-effects models

logistic regression

- the general norm in quantitative sciences is linear regression with continuous predictors
- in sociolinguistics, the norm is logistic regression with categorical predictors
- in logistic regression, the predictors still have linear effects and combinations of effects
- but the effect is not on the 0's and 1's directly but on the log-odds: $\ln(p / (1 - p))$
- residuals work differently – because of 0 or 1

basics of R

- what is R?
- command-line interface, but don't use it
- use scripts and execute one part at a time (how)
- we assign models to objects (give them names)
- we can then examine the models
- and compare the models, find the "best model"
- best data format
  - rows are observations, columns are variables
  - easy in Excel, save as .csv, then in R, use read.csv( )

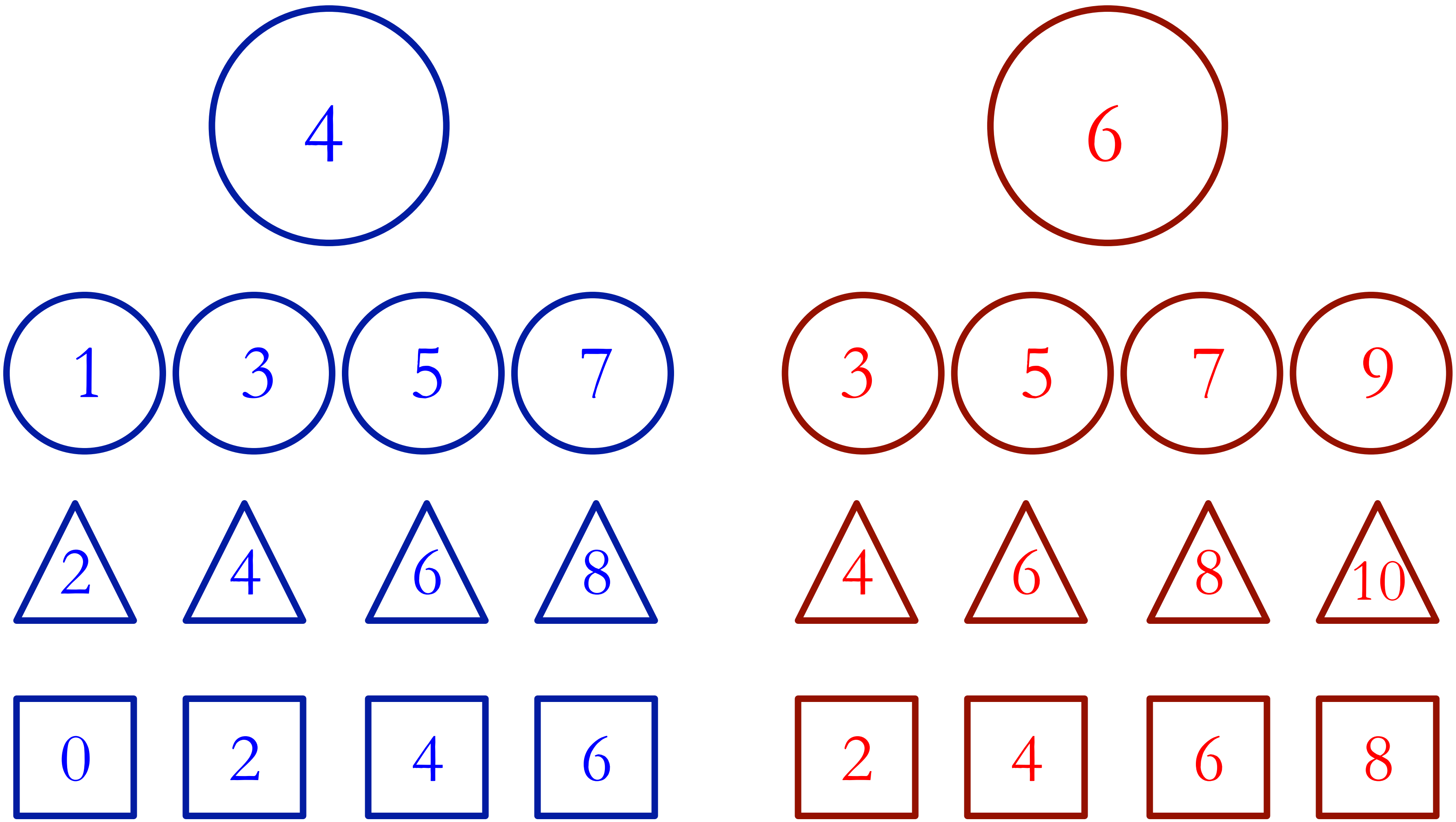
basic fixed-effects regression in R

- the function: lm( ), glm( ), lmer( ), glmer( ), other
- the formula: y ~ x1 + x2 + …
- the family – gaussian (linear), binomial (logistic), poisson (Poisson), others…

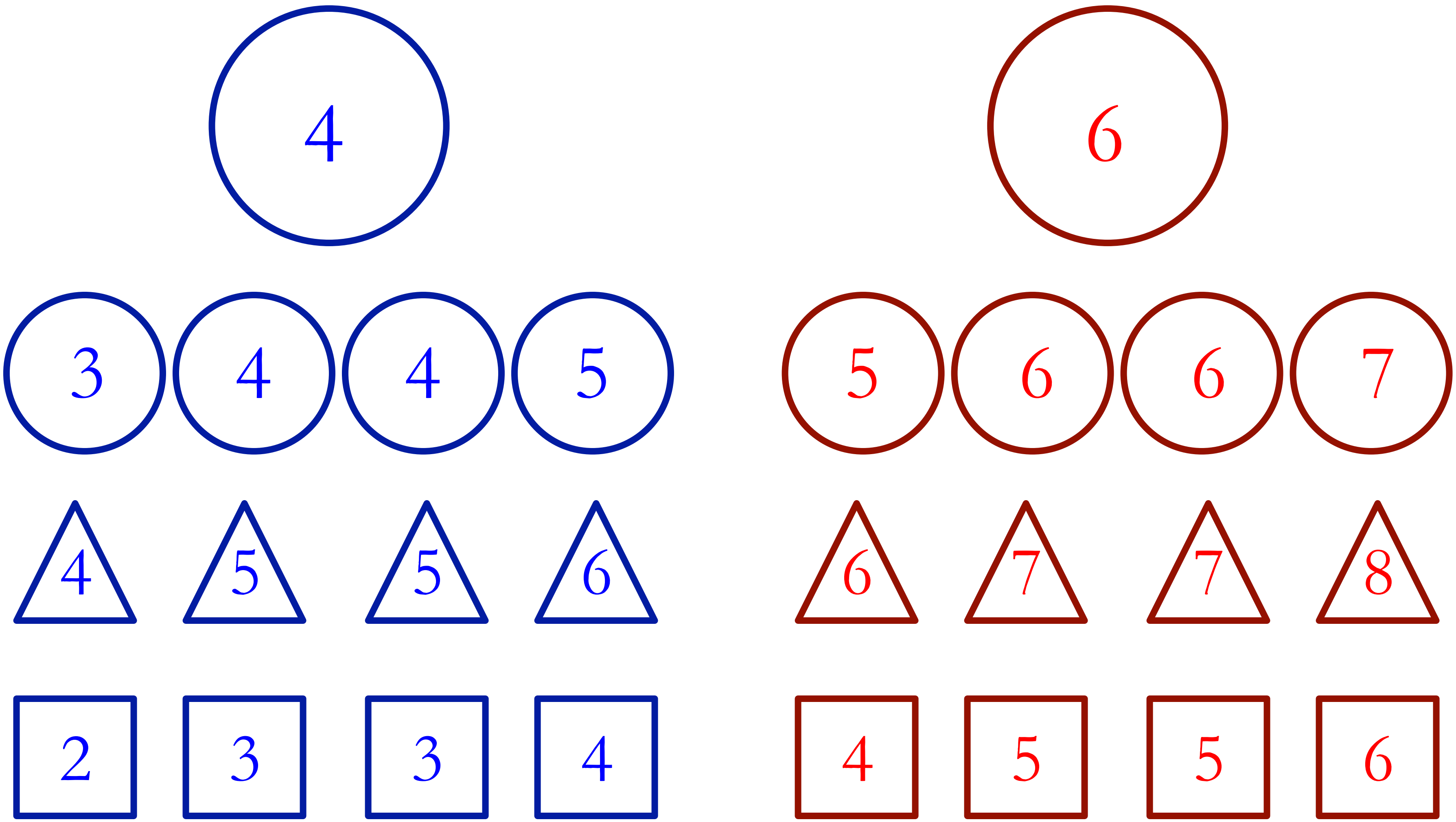
```
> m1 <- function(formula, data=…, family=…)
```

- print methods: > print(m1) or just > m1
- summary methods: > summary(m1)
- ‘anova’ methods: > anova(m1) or
  > anova(m1, m2)

mixed-effects models: why? what?



mixed-effects models: why? what?



why a different kind of model?

- if we leave out the speaker (or similar) level
- and there is any variation at that level:
- independence assumption is violated
- omitted variable bias may be occurring
- if we try to include the speaker (or similar):
- collinearity problem
- impossible to divide effect between speaker and between-speaker variables

four ways fixed effects can fail

1) they overestimate the significance of between-speaker predictors

2) if speakers have different amounts of data, size of between-speaker predictor effects can be ‘wrong’

3) if speakers have different balances of the other predictors, size of within-speaker effects ‘wrong’

4) in logistic regression, general shrinking of effects

how mixed effects do better

- they account for the speaker (etc.) level by estimating the population variance of speakers

- the inference (p-values) now reflects the real hierarchical structure of the data

- they have the same familiar fixed-effects part

random-effect estimates

- are not quite the same as fixed-effect estimates
- are called BLUPs (best linear unbiased predictors)
- or conditional modes
- they are not true parameters of the model
- rather, the group variances are the parameters

- but, we can inspect the BLUPs as if they were part of the model

goodness of fit: a problem

- one drawback to mixed models:
- no obvious analog of R^2
- harder to say how much has been explained

- for example, if speakers are being controlled for
- we can test if e.g. age, sex, class is significant
- but the more those fixed effects explain, the less the speaker random effect explains…

fitting mixed-effects models in R

```
> lm(y ~ 1 + x, data)
> glm(y ~ 1 + x, data, family = gaussian)
> glm(y ~ 1 + x, data, family = binomial)

> lmer(y ~ 1 + x + (1|s), data)
> glmer(y ~ 1 + x + (1|s), data, family = binomial)

> glmer(y ~ 1 + x + (1+x|s), data, family = binomial)
```

the formula: fixed-effects part

- same as in a fixed-effects model!
- everything you did, you do the same way

- ideally there is a parallel between the fixed and random effect specifications
- “maximal” random-effect structure means:
- every term in the fixed effects has its place(s) in the random effects, and mostly vice versa

the formula: random-effects part

- identify 'grouping factors' (goes after | symbol)
- if more than one, can be 'nested' or 'crossed'
- simplest random effects are random intercepts

~ 1 + gender + (1|speaker) speaker is a group!

~ 1 + gender + (1|speaker) + (1|small.group)

~ 1 + gender + freq. + (1|speaker) + (1|word)

- between-spkr. variables 'need' spkr. random int.
- between-word variables 'need' word random int.

the formula: random-effects part

- the intercept can usually vary between groups
- if the effects might too, you need random slopes

~ 1 + gender + freq. + (1|speaker) + (1|word)

- gender can't vary by speaker, freq. can't by word!
- gender could vary by word, freq. could by spkr.

~ 1 + gender + freq. + (1 + freq.|speaker) + (1 + gender|word)

- random slopes can cause slow/bad model fitting
- tip: center any continuous predictors
- tip: drop slopes for predictors 'not of interest'

the formula: shorthand

- 1 means intercept and is optional

  ~ 1 + x   is the same as   ~ x

- 0 means no intercept (rarely needed)

  ~ 0 + x

- * is for interactions

  ~ x1 * x2 is the same as   y ~ x1 + x2 + x1:x2

- ^ is for more than one interaction

  ~ (x1 + x2 + x3) ^ 2 equals ~ x1*x2 + x1*x3 + x2*x3

- transformations: log(x), I(x^2), anything else!

categorical predictors: contrasts

- the estimate for a continuous predictor is always:
  - what is the change in y for a one-unit increase in x?
  - y could be the response itself, the log-odds of it, etc.
- for a categorical predictor with k levels:
  - there are k-1 coefficients to be estimated
  - binary: one coefficient – easy: difference between
  - if $k > 2$, several systems of 'contrasts' are used
- 'treatment': levels compared to one baseline (0)
- 'sum': levels are deviations from mean of all (0)

more about contrasts

- changing contrasts does not change the model
- changing contrasts does affect the model output
- with interactions, contrasts become complicated
- can change the baseline with relevel( )
- in treatment contrasts, the missing level is 0
- in sum contrasts, it is 0 - the sum of the others
- missing levels frustrating – Rbrul shows all levels
- treatment: (0), 1, 2   sum: -1, 0, (1)

working with mixed-effects models

Fixed effects:	Estimate	Std. Error	t value
(Intercept)	651.742	39.464	16.515
Antpp	-2.888	30.711	-0.094
Verbosg	34.142	31.164	1.096
Antpp:Verbosg	-88.265	43.620	-2.023

Correlation of Fixed Effects:	(Intr)	Antpp	Verbsg
Antpp	-0.389		
Verbosg	-0.383	0.493	
Antpp:Vrbsg	0.274	-0.705	-0.716

anatomy of the (g)lmer output

```
> lmer(y ~ shape * color + (1 | speaker), d)

Linear mixed model fit by REML ['lmerMod']
Formula: y ~ shape * color + (1 | speaker)
   Data: d
REML criterion at convergence: 6.9096
Random effects:
 Groups   Name        Std.Dev.
 speaker  (Intercept) 0.81074
 Residual             0.05714
Number of obs: 16, groups: speaker, 8
Fixed Effects:
             (Intercept)               shape_triangle
                 2.99198                      2.01150
               color_red  shape_triangle:color_red
                 2.00804                     -0.04212
```

working w/ fixed-effects estimates

```
Fixed Effects:
           (Intercept)               shape_triangle
               2.99198                      2.01150
             color_red  shape_triangle:color_red
               2.00804                     -0.04212
```

working w/ random-effects estimates

```
Random effects:
 Groups   Name        Std.Dev.
 speaker  (Intercept) 0.81074
 Residual             0.05714
```

p-values from within a model

```
> summary(model)

Fixed effects:
                             Estimate Std. Error t value
(Intercept)                   2.99198    0.40637    7.36
shape_triangle                2.01150    0.04041   49.78
color_red                     2.00804    0.57470    3.49
shape_triangle:color_red    -0.04212    0.05714   -0.74

install.packages("lmerTest") !
```

p-values from comparing models

```
> anova(m, mm)
Models:
mm: y ~ shape + color + (1 | speaker)
m: y ~ shape * color + (1 | speaker)
   Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
mm  5 7.9098 11.773 1.0451  -2.0902
m   6 9.2163 13.852 1.3918  -2.7837 0.6935      1      0.405
```

- test entire predictors (or interactions)
- test contrasts w/in predictor, combining levels
- test the random effects themselves
- some argue that this is not necessary
- larger questions over what belongs in a model

more mixed-effects models in R

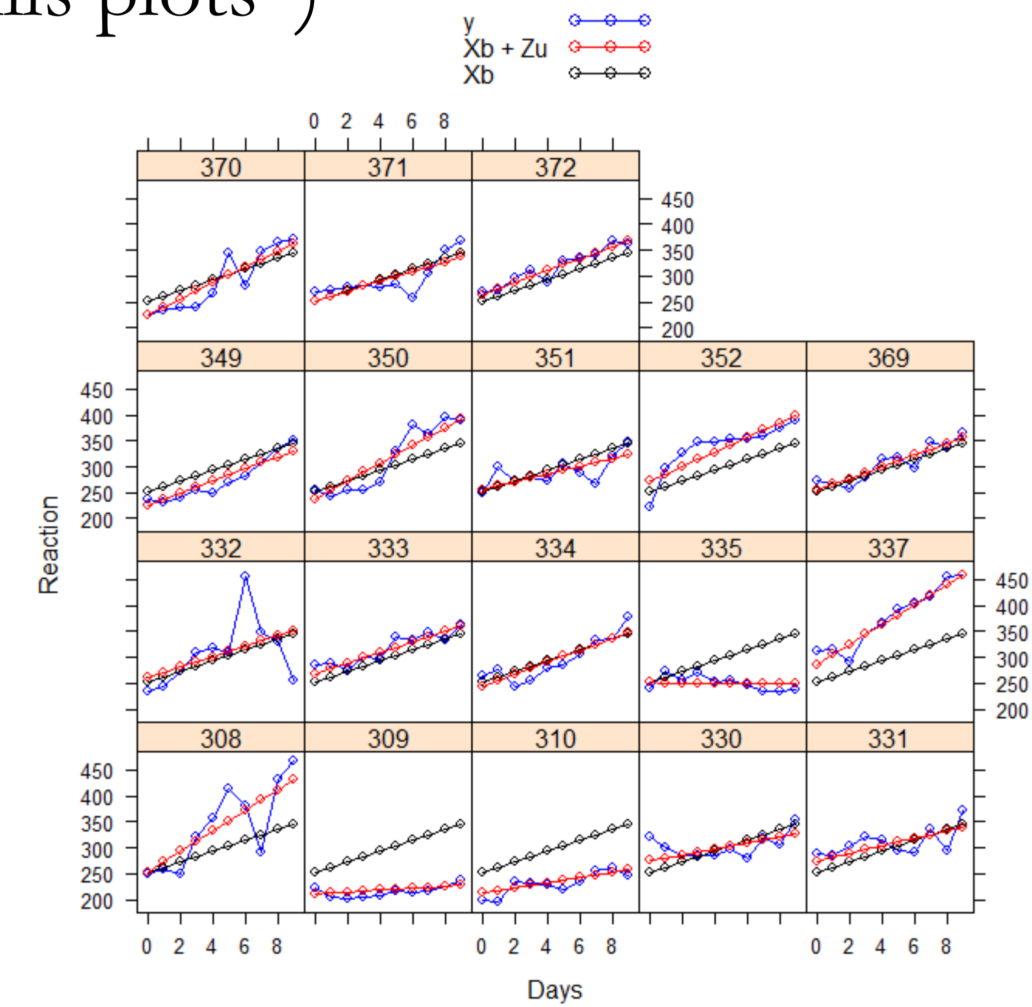
- other R packages besides lme4
- ordinal
- mgcv - GAM(M)s
- MCMCglmm

mixed-effects models beyond R

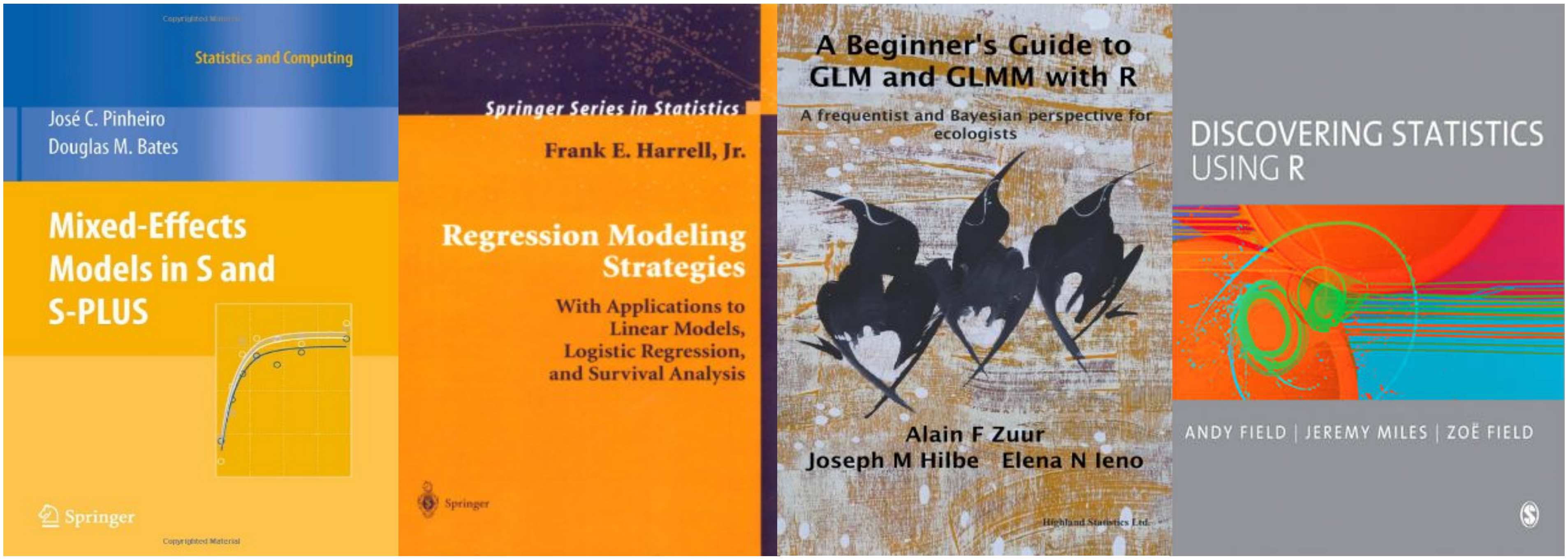
- SAS
- JAGS/BUGS (Bayesian)
- MLwin
- BayesX

visualizing mixed-effects models

- lattice package ("trellis plots")
- effects package



some books I can recommend



- try Rbrul? > source("http://www.danielezrajohnson.com/Rbrul.R")
- email support available at d.e.johnson@lancaster.ac.uk