Today

- HW 3 due April 1
- Project due April 15
- non-specific effects — Cox & Donelly, Ch. 7.2
- generalized linear mixed models and GEEs
- In the News: Election Polling UK

- Project Guidelines
  - report: 3-5 pages: non-technical, no code – Intro, source of data, problem of interest, conclusions, a few tables, a few plots
  - statistical appendix: main statistical methods used, summary of results, code excerpts permitted
  - further plots and tables as needed
  - executable code

Recap: random and mixed effects models

- random effects are useful in a variety of models
- randomized block designs, Latin squares, etc. blocks as random effects
- split plot designs: two levels of randomization ELM §8.5
- nested designs: students within classes within schools; technical replicates within samples within laboratories; ELM §8.6, MASS §10.2
- longitudinal data: PSID, rat growth ELM §9.1, SM Ex.9.18
- multi-level models: combination of crossed (fixed) and nested (random) effects ELM §8.8, 9.3
  – in §9.3, two responses are considered (English and math scores), but a single response is used (English score) with math score as a covariate
- repeated measures: acuity of vision ELM §9.2

Non-specific effects

C&D §7.2

- example: a clinical trial involves several or many centres
- an agricultural field trial repeated at a number of different farms, and over a number of different growing seasons
- a sociological study repeated in broadly similar form in a number of countries
- laboratory study uses different sets of analytical apparatus, imperfectly calibrated
- such factors are non-specific
- how do we account for them
  - on an appropriate scale, a parameter represents a shift in outcome
  - more complicated: the primary contrasts of concern vary across centres
  - i.e. treatment-center interaction

... non-specific effects

- suppose no treatment-center interaction
- example:

$$\text{logit}\{\text{pr}(Y_{ci} = 1)\} = \alpha_c + x_{ci}^T \beta$$

- should α_c be ?fixed? or ?random?
- effective use of a random-effects representation will require estimation of the variance component corresponding to the centre effects
- even under the most favourable conditions the precision achieved in that estimate will be at best that from estimating a single variance from a sample of a size equal to the number of centres
- very fragile unless there are at least, say, 10 centres and preferably considerably more

... non-specific effects

- if centres are chosen by an effectively random procedure from a large population of candidates, ... the random-effects representation has an attractive tangible interpretation. This would not apply, for example, to the countries of the EU in a social survey

- some general considerations in linear mixed models:
  - in balanced factorial designs, the analysis of treatment means is unchanged
  - in other cases, estimated effects will typically be ‘shrunk’, and precision improved
  - representation of the nonspecific effects as random effects involves independence assumptions which certainly need consideration and may need some empirical check

... non-specific effects

- if estimates of effect of important explanatory variables are essentially the same whether nonspecific effects are ignored, or are treated as fixed constants, then random effects model will be unlikely to give a different result
- it is important in applications to understand the circumstances under which different methods give similar or different conclusions
- in particular, if a more elaborate method gives an apparent improvement in precision, what are the assumptions on which that improvement is based, and are they reasonable?

... non-specific effects

- if there is an interaction between an explanatory variable [e.g. treatment] and a nonspecific variable
- i.e. the effects of the explanatory variable change with different levels of the nonspecific factor
- the first step should be to explain this interaction, for example by transforming the scale on which the response variable is measure or by introducing a new explanatory variable
  - example: two medical treatments compared at a number of centres show different treatment effects, as measured by an ratio of event rates
  - possible explanation: the difference of the event rates might be stable across centres
  - possible explanation: the ratio depends on some characteristic of the patient population, e.g. socio-economic status
- an important special application of random-effect models for interactions is in connection with overviews, that is, assembling of information from different studies of essentially the same effect

Generalized linear mixed models

ELM §10.1

- GLM:

$$f(y_i \mid \beta, \phi, \gamma) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{\phi a_i} + c(y_i; \phi a_i)\}$$

$b'(\theta_i) = \mu_i$

- random effects

$$g(\mu_i) = x_i^{\mathrm{T}}\beta + z_i^{\mathrm{T}}\gamma, \quad \gamma \sim N(0, D_\psi)$$

- likelihood

$$L(\beta, \phi, \psi; y) = \prod_{i=1}^{n} \int f(y_i \mid \beta, \gamma, \phi)\varphi(\gamma; 0, D_\psi)d\gamma$$

- ψ are parameters in the covariance matrix
  ϕ is the dispersion parameter for GLM
  $\varphi(x) \propto \exp(-x^2/2)$

... generalized linear mixed models

- likelihood

$$L(\beta, \phi, \psi; y) = \prod_{i=1}^{n} \int f(y_i \mid \beta, \gamma, \phi) \phi(\gamma; 0, D_\psi) d\gamma$$

- doesn't simplify unless $f(y_i \mid \gamma)$ is normal
- solutions proposed include
  - numerical integration, e.g. by quadrature
  - integration by MCMC
  - Laplace approximation to the integral – penalized quasi-likelihood

`MASS` library and book (§10.4):
`glmmNQ`, `GLMMGibbs`, `glmmPQL`, all in `library(MASS)`
`glmer` in `library(lme4)`

- several observations per subject: $g\{\mathrm{E}(y_{ij} \mid \gamma_i)\} = x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}}\gamma_i$, $L(\beta; y) = \prod_{i=1}^{n} \int \prod_{j=1}^{m_i} f(y_{ij}; \gamma_i) \phi(\gamma_i; 0, D_\psi) d\gamma$

Example: Balance experiment

Faraway, 10.1

- effects of surface and vision on balance; 2 levels of surface; 3 levels of vision
- surface: normal or foam
- vision: normal, eyes closed, domed
- 20 males and 20 females tested for balance, twice at each of 6 combinations of treatments
- auxiliary variables age, height, weight

Steele 1998, OzDASL

- response measured on a 4 point scale; converted by Faraway to binary (stable/not stable)
- analysed using linear models at OzDASL

www.statsci.org/data/oz/ctsibrm.txt

Subject	Sex	Age	Height	Weight	NO1	NO2	NC1	NC2	ND1	ND2	FO1	FO2	FC1	FC2	FD1	FD2
1	male	22	176	68.2	1	1	2	2	1	2	2	2	2	2	2	2
2	male	22	181	67.6	1	1	2	2	2	2	2	2	3	3	3	3
3	male	22	175.5	72	2	2	2	2	2	2	2	2	3	3	2	3
4	male	21	180	73.2	1	2	2	2	2	2	2	2	3	3	3	3
5	female	20	166	63.8	1	2	2	2	3	2	2	2	3	3	3	3
6	male	18	177	78.8	1	1	1	1	1	2	2	2	2	2	2	2
7	male	29	183	86.4	1	1	2	2	2	2	2	2	2	2	2	2
8	female	22	150	44.6	1	1	2	2	2	2	2	2	3	3	2	2
9	female	29	154	57.8	1	1	2	2	2	2	2	2	3	3	3	3
10	male	31	176.5	80.8	1	1	2	2	1	1	2	2	2	2	2	2
11	male	24	176	91	1	1	2	2	2	2	2	2	3	3	2	2
12	male	33	184	89.8	1	1	2	2	2	2	2	2	2	2	3	2
13	male	18	187	85	1	1	2	2	2	2	2	2	3	3	3	2
14	female	34	168	54.4	2	2	2	2	2	2	2	2	3	3	3	3
15	female	27	173	60.8	1	1	2	2	2	2	2	2	3	3	3	3
16	female	20	142	44.2	1	1	1	2	1	1	2	2	2	2	2	2
17	male	36	183	88.6	1	1	1	1	1	1	2	2	2	3	3	2
18	female	34	170	67.8	1	1	2	2	2	2	2	2	2	2	2	2
19	male	18	190	78.6	1	1	1	2	2	2	2	2	2	2	2	2
20	male	30	168.5	64.2	1	1	1	1	2	2	2	2	3	3	3	3
21	female	19	167.5	69.4	1	1	2	2	1	1	2	2	2	2	2	2
22	female	20	167	50	1	2	2	2	2	2	2	2	3	3	3	3
23	male	36	184	102.4	1	1	2	2	2	2	2	2	2	2	2	2
24	male	31	182.5	83	1	1	2	2	2	2	2	2	3	2	3	2
25	male	24	187	86.8	1	1	1	1	1	1	1	1	2	2	2	2
26	male	29	176	101.4	1	1	2	2	1	1	1	1	2	2	2	2
27	female	23	163	65.8	1	1	1	1	1	1	1	1	2	2	2	2
28	female	36	159	51	1	1	2	2	2	2	2	2	3	3	3	3
29	female	36	161.5	45.6	1	1	2	2	1	1	1	1	3	3	2	2
30	female	33	174.5	64.8	2	2	2	2	2	2	2	2	3	3	3	3
31	male	38	185	96	1	1	1	1	2	2	1	2	2	2	2	2
32	female	28	165.5	60.2	1	1	2	2	2	2	2	1	3	3	2	2
33	female	35	159	53	1	1	2	2	1	1	2	2	3	3	2	3
34	male	28	172	74	1	2	2	1	2	2	2	2	2	2	2	2
35	male	24	177	89.8	1	1	1	1	1	1	2	2	2	2	2	2
36	female	19	167	60.2	1	1	2	2	2	2	2	2	3	3	3	3
37	female	37	167	62.2	2	2	2	2	2	2	2	2	3	3	2	2
38	female	24	173	81.4	2	2	2	2	2	2	2	2	3	4	3	3
39	female	20	170	66.2	1	1	2	2	2	2	2	2	2	2	2	3
40	female	22	164	65	1	2	2	2	2	2	2	2	2	2	2	2

... balance

```
> balance <- glmer(stable ~ Sex + Age + Height + Weight + Surface + Vision +
+   (1|Subject), family = binomial, data = ctsib)

# Subject effect is random

> summary(balance)
Generalized linear mixed model fit by maximum likelihood ['glmerMod']

 ...

Random effects:
 Groups  Name        Variance Std.Dev.
 Subject (Intercept) 8.197    2.863
Number of obs: 480, groups: Subject, 40

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  9.920750  13.358013   0.743    0.458
Sexmale      2.825305   1.762383   1.603    0.109
Age         -0.003644   0.080928  -0.045    0.964
Height      -0.151012   0.092174  -1.638    0.101
Weight       0.058927   0.061958   0.951    0.342
Surfacenorm  7.524423   0.888827   8.466  < 2e-16 ***
Visiondome   0.683931   0.530654   1.289    0.197
Visionopen   6.321098   0.839469   7.530 5.08e-14 ***
---
```

$$\text{logit}(p_{ij}) = \mu + gender_i + age_i + height_i + weight_i + surface_{ij} + vision_{ij} + \gamma_i$$

... balance

- if we allow γ_i to be a fixed effect for each subject, then model fit fails

why?

```
> gfs <- glm(stable ~ Sex + Age + Height + Weight + Surface + Vision
+            + factor(Subject),
+            family = binomial,
+            data = ctsib)
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(gfs)

Call:
glm(formula = stable ~ Sex + Age + Height + Weight + Surface +
    Vision + factor(Subject), family = binomial, data = ctsib)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.62183  -0.08595  -0.00170   0.00000   3.11251

Coefficients: (2 not defined because of singularities)
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.408e+14  7.907e+14   0.178    0.859
Sexmale            1.130e+13  5.662e+13   0.200    0.842
Age               -3.723e+12  1.772e+13  -0.210    0.834
Height            -2.139e+12  9.517e+12  -0.225    0.822
Weight             4.491e+12  1.829e+13   0.246    0.806
Surfacenorm        9.550e+00  1.435e+00   6.654 2.86e-11 ***
Visiondome         8.211e-01  5.819e-01   1.411    0.158
Visionopen         8.241e+00  1.370e+00   6.016 1.79e-09 ***
```

... balance: random effects models

```
> library(MASS)

> balance2 <- glmmPQL(stable ~ Sex + Age + Height + Weight + Surface + Vision,
+ random = ~1 | Subject, family = binomial, data = ctsib)

> summary(balance2)

Random effects:
 Formula: ~1 | Subject
        (Intercept)  Residual
StdDev:    3.060712 0.5906232

Variance function:
 Structure: fixed weights
 Formula: ~invwt

Fixed effects: stable ~ Sex + Age + Height + Weight + Surface + Vision
                Value Std.Error  DF   t-value p-value
(Intercept) 15.571494 13.498304 437  1.153589  0.2493
Sexmale      3.355340  1.752614  35  1.914478  0.0638
Age         -0.006638  0.081959  35 -0.080992  0.9359
Height      -0.190819  0.092023  35 -2.073601  0.0455
Weight       0.069467  0.062857  35  1.105155  0.2766
Surfacenorm  7.724078  0.573578 437 13.466492  0.0000
Visiondome   0.726464  0.325933 437  2.228873  0.0263
Visionopen   6.485257  0.543980 437 11.921876  0.0000
```

... balance

```
> balance4 <- glmer(stable ~ Sex + Age + Height + Weight + Surface + Vision +
+ (1|Subject), family = binomial, data = ctsib, nAGQ = 9)

> summary(balance4)

Random effects:
 Groups  Name        Variance Std.Dev.
 Subject (Intercept) 7.8      2.793
Number of obs: 480, groups: Subject, 40

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 13.551847  13.067369   1.037   0.2997
Sexmale      3.109307   1.724797   1.803   0.0714 .
Age         -0.001804   0.079161  -0.023   0.9818
Height      -0.175061   0.090239  -1.940   0.0524 .
Weight       0.065742   0.060606   1.085   0.2780
Surfacenorm  7.428046   0.872416   8.514  < 2e-16 ***
Visiondome   0.682509   0.527836   1.293   0.1960
Visionopen   6.210825   0.822012   7.556 4.17e-14 ***
```

See `Mar11.R` for more details and to fit other versions

... balance

	`glmer`(Laplace)	`glmer`(Quad. 5)	`glmer`(Quad. 9)	`glmmPQL`
$\tilde{\sigma}_\gamma$	2.86	2.72	2.79	3.07
`Surface`	7.5	7.3	7.4	7.7
`norm`	(1.16)	(1.05)	(1.09)	(0.57)
`Height`	-0.15	-0.19	-0.17	-0.19
	(0.09)	(0.09)	(0.09)	(0.09)

– note: no analogue of `REML` for generalized linear mixed models

References: MASS Book §10.4; online resource for R and mixed models

Generalized Estimating Equations

ELM §10.2

- GLM's have $E(y_i) = \mu_i$; $\quad \text{var}(y_i) = \phi V(\mu_i)$
- ML equation of the form $\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_i}{g'(\mu_i)V(\mu_i)} = 0$

- extend to vector $y_i = (y_{i1}, \ldots, y_{in_i})$ Liang & Zeger, 1986
- $\text{var}(y_i) = V_i(\beta, \alpha)$ is now $n_i \times n_i$ matrix
- estimating equation for β:

$$\sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \beta} \right)^{\mathrm{T}} V_i(\beta; \alpha)^{-1}(y_i - \mu_i) = 0$$

- LZ suggest using a working covariance matrix e.g. AR(1)
- estimates of β are consistent, even if covariance is mis-specified

- correlation between measurements on the same subject are modelled/assumed
- not generated from random effects

... GEE

	`glmer`(Quad. 5)	`glmmPQL`	GEE
$\tilde{\sigma}_\gamma$	2.72	3.07	
`Surface`	7.5	7.7	3.92
`norm`	(1.05)	(0.57)	(0.57)
`Height`	-0.19	-0.19	-0.10
	(0.09)	(0.09)	(0.04)

β has a different interpretation under GEE: it is the marginal effect on the population average

by assumption: $\mathrm{E}(y_i) = \mu_i(\beta)$, $\mathrm{Var}(y_i) = V(\beta, \alpha)$, y is a vector

in the GLMM model β is the conditional effect on an individual subject's response y_{ij}

Diggle, Liang & Zeger, Ch. 7

Marginal and conditional models

Diggle et al. Ch. 7

- Marginal model for binary data
  - $E(y_{ij}) = \mu_{ij}, \quad \text{logit}(\mu_{ij}) = x_{ij}^T\beta = \beta_0 + \beta_1 x_{ij},$ for example
  - $\text{var}(y_{ij}) = \phi V(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$
  - $\text{corr}(y_{ij}, y_{ik}) = \rho(\mu_{ij}, \mu_{ik}, \alpha) = \alpha$
  - $\exp(\beta_0)$ is the ratio of $Pr(1)$ to $Pr(0)$ when $x_{ij} = 0$
  - $\exp(\beta_1)$ is the increase in odds associated with an increase in x
- Random effects model for binary data
  - $\text{logit}\{\Pr(y_{ij}\} = 1 \mid \gamma_i) = (\beta_0^* + \gamma_i) + \beta_1^* x_{ij}$
  - baseline ($x = 0$) ratio: $\exp(\beta_0^* + U_i)$, for subject i
  - increase with x: $\exp(\beta_1^*)$

Epilepsy data

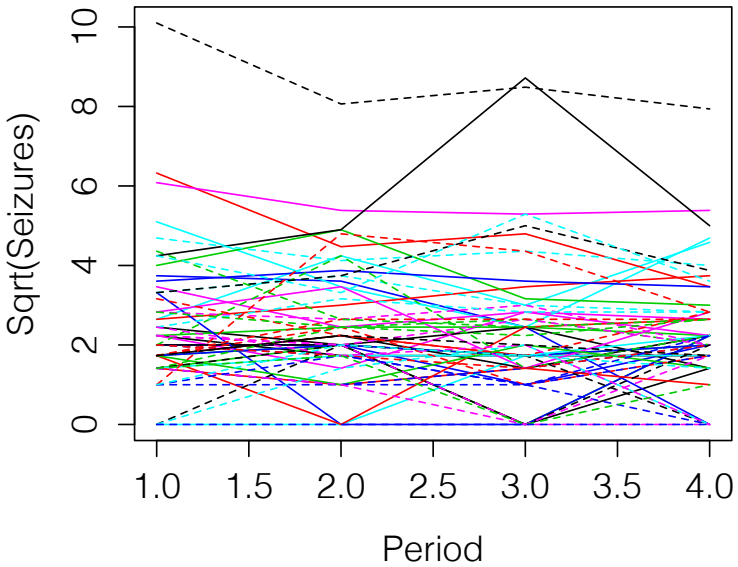
ELM §10.2

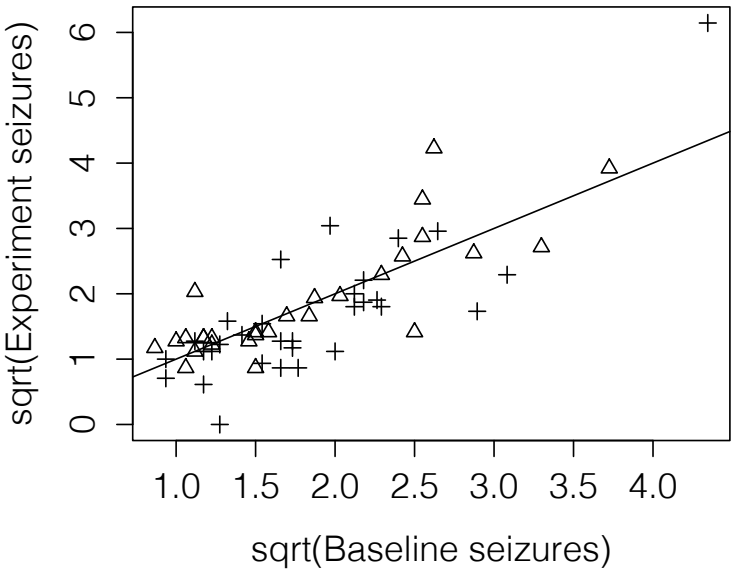
- 59 patients, 5 measurements per patient, over time
- first measurement: number of seizures in an eight-week period
- next four measurements: number of seizures in consecutive two-week periods
- 31 patients randomized to drub Progabide; 28 to placebo
- see `Mar11.R` for R code as in ELM
-

	baseline	experiment
placebo	3.85	4.30
treatment	3.96	3.98

- is the drug beneficial?

... epilepsy data





	estimate	robust s.e.	robust z
(Intercept)	1.320	0.161	
period(exposure)	0.143	0.108	1.33
treatment(drug)	−0.079	0.197	−0.403
interaction	−0.377	0.168	−2.242

... epilepsy

	estimate	robust s.e.	robust z
(Intercept)	1.320	0.161	
period(exposure)	0.143	0.108	1.33
treatment(drug)	−0.079	0.197	−0.403
interaction	−0.377	0.168	−2.242

Interaction between exposure period and treatment is the effect of the drug
why?

marginally significant
with patient 49 included, becomes insignificant Diggle et al. (2002)

`Estimated Scale Parameter:  10.687`: automatically incorporates over-dispersion

```
Working Correlation
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 1.0000000 0.8102249 0.6564644 0.5318838 0.4309455
[2,] 0.8102249 1.0000000 0.8102249 0.6564644 0.5318838
[3,] 0.6564644 0.8102249 1.0000000 0.8102249 0.6564644
[4,] 0.5318838 0.6564644 0.8102249 1.0000000 0.8102249
[5,] 0.4309455 0.5318838 0.6564644 0.8102249 1.0000000
```

Variance-stabilizing transformations

- suppose $\mathrm{E}(y) = \mu, \quad \mathrm{var}(y) \propto V(\mu)$
- is there a transformation of y for which variance is constant?
- $g(y) \doteq g(\mu) + (y - \mu)g'(\mu)$
- $\mathrm{E}\{g(y)\} \doteq g(\mu), \quad \mathrm{var}\{g(y)\} \doteq cV(\mu)\{g'(\mu)\}^2$
- choose $g(\mu) \propto \int \frac{1}{V^{1/2}(\mu)} d\mu$ variance-stabliziing transf.
- example: Poisson $V(\mu) = \mu, \quad g(\mu) = \int \mu^{-1/2} d\mu \propto \mu^{1/2}$
- example: exponential
  $V(\mu) = \mu^2, \quad g(\mu) = \int \mu^{-1} d\mu \propto \log(\mu)$

Box-Cox transformation

- an older approach to regression uses variance-stabilizing transformations
- followed by linear model fitting
- instead of GLM

- Box & Cox (1964) formalized this approach with the model

$$y^{(\lambda)} = x^{\mathrm{T}}\beta + \epsilon,$$

λ is a parameter to be estimated

-

$$y^{(\lambda)} = \begin{cases} (y^{\lambda} - 1)/\lambda, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

- λ to be estimated by maximum likelihood; then fixed for linear regression
- usually GLM approach preferred in most settings

In the News

Significance website: review of methods of UK polling firms
Election Forecast UK: aggregates results of all polls
FiveThirtyEight: planning to forecast the UK election

SRS: sample $y_1, \ldots, y_n$. Estimate population mean by $\bar{y} = \Sigma y_i / n$ and population total by $N\bar{y}$

stratified RS: $y_{hj}, h = 1, \ldots, H; j = 1, \ldots, n_h$. Estimation population total by $\Sigma_h \Sigma_{j \in S_h} (N_h / n_h) y_{jj}$ – each unit in stratum h represents N_h / n_h of the proportion in the population in stratum h

both estimates can be expressed as $\sum w_i y_i$, where $w_i = 1/\pi_i$ and π_i is the probability of selection

complex sample surveys weight each sampled unit to ensure that the sample has the same age/sex/SES/... as the full population