Mixed Effect Logistic Regression
STAT 506 Spring 2013

Example from Doug Bates talk at UseR 2009

Data on individual women in Bangladesh. Response is use of contraception.

These observational data are unbalanced (some districts have only 2 observations, some have nearly 120). They are not longitudinal (no time variable).

Binary responses have low per-observation information content (exactly one bit per observation). Districts with few observations will not contribute strongly to estimates of random effects.

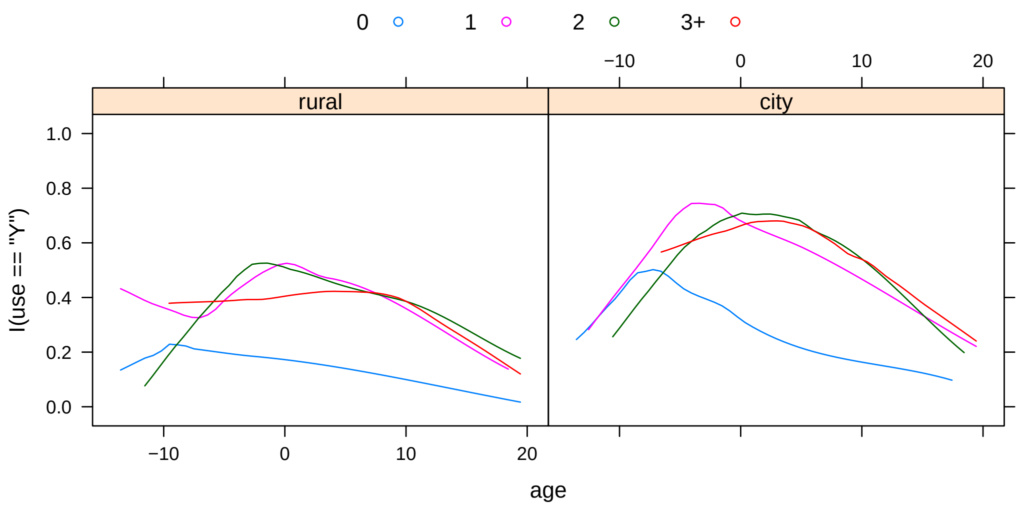
Within-district plots will be too imprecise so we only examine the global effects in plots.

The form of the curves suggests at least a quadratic in age.

The urban versus rural differences may be additive.

It appears that the livch factor could be dichotomized into 0 versus 1 or more.

```
data(Contraception, package = "mlmRev")
require(lattice)
levels(Contraception$urban) <- c("rural", "city")
xyplot(I(use == "Y") ~ age | urban, group = livch, data = Contraception, type = "smooth", auto.key = list(columns = 4))
```



Show smoothers, not the 0-1 data.

```
require(lme4)
conFit1 <- lmer(use ~ age + I(age^2) + urban + livch + (1 | district), data = Contraception, family = binomial)

conFit1

## Generalized linear mixed model fit by the Laplace approximation
## Formula: use ~ age + I(age^2) + urban + livch + (1 | district)
##    Data: Contraception
##   AIC  BIC logLik deviance
##  2389 2433  -1186     2373
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  district (Intercept) 0.226    0.475
## Number of obs: 1934, groups: district, 60
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.035073   0.174361   -5.94  2.9e-09
## age          0.003533   0.009231    0.38      0.7
```

```
## I(age^2)    -0.004562   0.000725   -6.29  3.2e-10
## urbancity    0.697269   0.119879    5.82  6.0e-09
## livch1       0.815045   0.162190    5.03  5.0e-07
## livch2       0.916511   0.185100    4.95  7.4e-07
## livch3+      0.915021   0.185769    4.93  8.4e-07
##
## Correlation of Fixed Effects:
##           (Intr) age    I(g^2) urbnct livch1 livch2
## age        0.572
## I(age^2)  -0.547 -0.478
## urbancity -0.270 -0.054  0.022
## livch1    -0.610 -0.274  0.253  0.068
## livch2    -0.681 -0.456  0.357  0.100  0.525
## livch3+   -0.763 -0.698  0.325  0.108  0.565  0.652
```

Note similarity of curves for more than 0 children. Let's combine those.

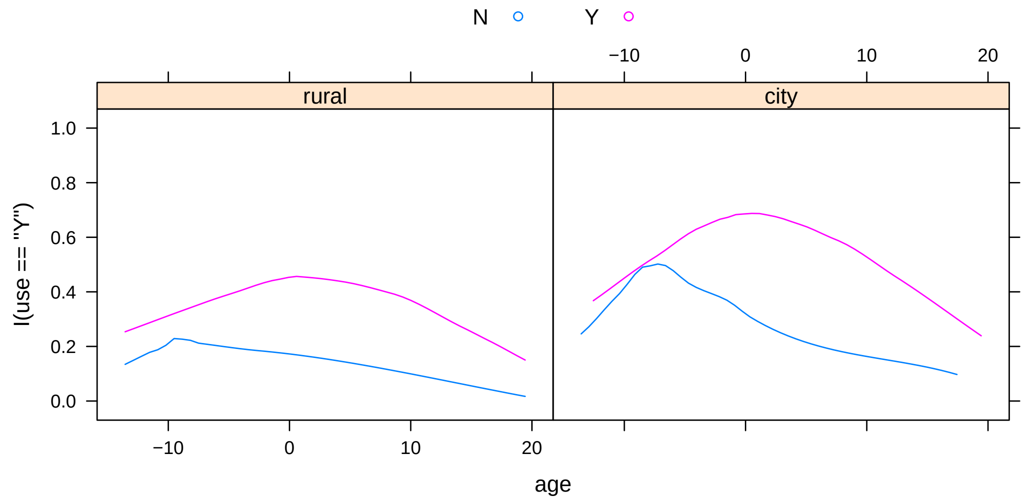
```
Contraception$ch <- factor(ifelse(Contraception$livch == "0", "N", "Y"))

conFit2 <- lmer(use ~ age + I(age^2) + urban + ch + (1 | district), data = Contraception, family = binomial)

print(conFit2, corr = F)

## Generalized linear mixed model fit by the Laplace approximation
## Formula: use ~ age + I(age^2) + urban + ch + (1 | district)
##    Data: Contraception
##   AIC  BIC logLik deviance
##  2385 2419  -1187     2373
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  district (Intercept) 0.225    0.474
## Number of obs: 1934, groups: district, 60
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.006426   0.167895   -5.99  2.0e-09
## age          0.006256   0.007840    0.80     0.42
## I(age^2)    -0.004635   0.000716   -6.47  9.7e-11
## urbancity    0.692950   0.119669    5.79  7.0e-09
## chY          0.860376   0.147354    5.84  5.3e-09
```

and compare

```
xtable(anova(conFit2, conFit1))
```

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
conFit2	6	2385.19	2418.59	-1186.59			
conFit1	8	2388.73	2433.27	-1186.36	0.46	2	0.7957

```
xyplot(I(use == "Y") ~ age | urban, group = ch, data = Contraception, type = "smooth", auto.key =
```



Allow age pattern to change with ch

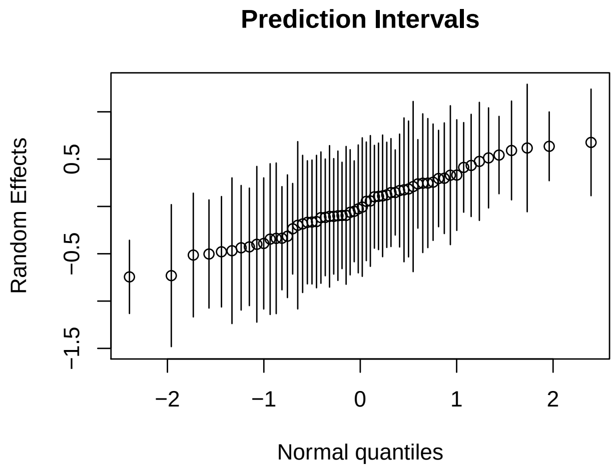
```
conFit3 <- lmer(use ~ age * ch + I(age^2) + urban + (1 | district), data = Contraception, family = binomial)

print(conFit3, corr = F)

## Generalized linear mixed model fit by the Laplace approximation
## Formula: use ~ age * ch + I(age^2) + urban + (1 | district)
##    Data: Contraception
##   AIC  BIC logLik deviance
##  2379 2418  -1183     2365
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  district (Intercept) 0.223    0.472
## Number of obs: 1934, groups: district, 60
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.323318   0.214447   -6.17  6.8e-10
## age         -0.047296   0.021839   -2.17   0.0303
## chY          1.210786   0.206994    5.85  4.9e-09
## I(age^2)    -0.005757   0.000836   -6.89  5.6e-12
## urbancity    0.714033   0.120258    5.94  2.9e-09
## age:chY      0.068352   0.025435    2.69   0.0072
```

```
xtable(anova(conFit2, conFit3))
```

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
conFit2	6	2385.19	2418.59	-1186.59			
conFit3	7	2379.18	2418.15	-1182.59	8.00	1	0.0047

Random effects are assumed normal. Are they?

```
quant <- qnorm((0.5 + 0:59)/60)
rFX <- ranef(conFit3)[[1]][[1]]
se <- se.ranef(conFit3)[[1]][order(rFX), 1]
rFX <- sort(rFX)
plot(quant, rFX, xlab = "Normal quantiles", ylab = "Random Effects", main = "Prediction Intervals", ylim = c(-1.5,
    1.3))
segments(quant, rFX - 2 * se, quant, rFX + 2 * se)
```



Some districts have no city folks, but we can still posit a model where the urban/rural effect varies by district.

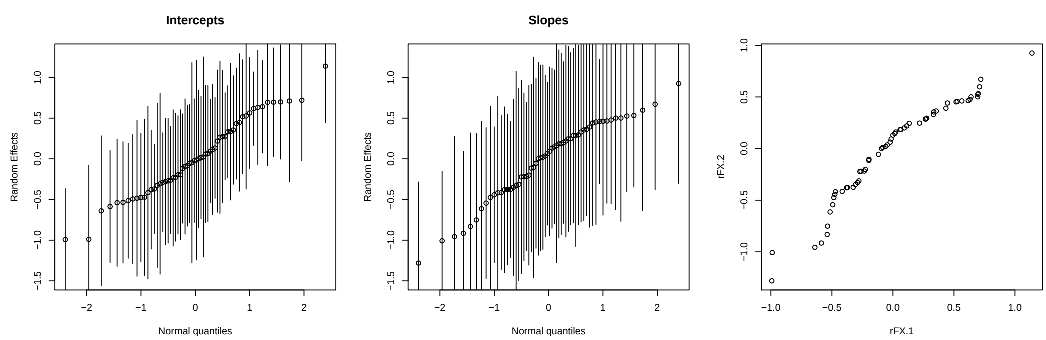
```
with(Contraception, table(urban, district))

##        district
## urban     1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23
##   rural  54  20   0  19  37  58  18  35  20  13  21  23  16  17  14  18  24  33  22  15  10  20  15
##   city   63   0   2  11   2   7   0   2   3   0   0   6   8 101   8   2   0  14   4   0   8   0   0
##        district
## urban    24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46
##   rural  14  49  13  39  45  25  45  27  24   7  26  28  14  13   7  24  12  23   6  28  27  34  74
##   city    0  18   0   5   4   7  16   6   0   7   9  20   3   0   7   2  29   3   5  17   0   5  12
##        district
## urban    47  48  49  50  51  52  53  55  56  57  58  59  60  61
##   rural   9  26   4  15  20  42   0   0  24  23  20  10  22  31
##   city    6  16   0   4  17  19  19   6  21   4  13   0  10  11


conFit4 <- lmer(use ~ age * ch + I(age^2) + urban + (1 + urban | district), data = Contraception, family = binomial)

print(conFit4)

## Generalized linear mixed model fit by the Laplace approximation
## Formula: use ~ age * ch + I(age^2) + urban + (1 + urban | district)
##    Data: Contraception
##   AIC  BIC logLik deviance
##  2372 2422  -1177     2354
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  district (Intercept) 0.378    0.615
##           urbancity   0.526    0.725    -0.793
## Number of obs: 1934, groups: district, 60
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.344263   0.222767   -6.03  1.6e-09
## age         -0.046184   0.021945   -2.10   0.0353
## chY          1.211653   0.208237    5.82  5.9e-09
## I(age^2)    -0.005651   0.000843   -6.70  2.0e-11
## urbancity    0.790210   0.160048    4.94  7.9e-07
## age:chY      0.066468   0.025567    2.60   0.0093
##
## Correlation of Fixed Effects:
##           (Intr) age    chY    I(g^2) urbnct
## age        0.696
## chY       -0.855 -0.792
## I(age^2)  -0.091  0.301 -0.097
## urbancity -0.370 -0.061  0.088 -0.018
## age:chY   -0.575 -0.930  0.676 -0.496  0.055
```

```r
xtable(anova(conFit3, conFit4))
```

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
conFit3	7	2379.18	2418.15	-1182.59			
conFit4	9	2371.53	2421.64	-1176.76	11.65	2	0.0030

Check random effects again

```r
par(mfrow = c(1, 3))
rFX.1 <- ranef(conFit4)[[1]][[1]]
rFX.2 <- ranef(conFit4)[[1]][[2]]
se.1 <- se.ranef(conFit4)[[1]][order(rFX.1), 1]
se.2 <- se.ranef(conFit4)[[1]][order(rFX.2), 2]
rFX.1 <- sort(rFX.1)
rFX.2 <- sort(rFX.2)
plot(quant, rFX.1, xlab = "Normal quantiles", ylab = "Random Effects", main = "Intercepts", ylim = c(-1.5,
    1.3))
segments(quant, rFX.1 - 2 * se.1, quant, rFX.1 + 2 * se.1)
plot(quant, rFX.2, xlab = "Normal quantiles", ylab = "Random Effects", main = "Slopes", ylim = c(-1.5,
    1.3))
segments(quant, rFX.2 - 2 * se.2, quant, rFX.2 + 2 * se.2)
plot(rFX.1, rFX.2)
```



Flat spots for districts with no urban women.
Note strong correlation. Can we get it down to 1 random effect?

```r
conFit5 <- lmer(use ~ age * ch + I(age^2) + urban + (1 | urban:district), data = Contraception, family = binomial)

print(conFit5)

## Generalized linear mixed model fit by the Laplace approximation
## Formula: use ~ age * ch + I(age^2) + urban + (1 | urban:district)
##    Data: Contraception
##   AIC  BIC logLik deviance
##  2368 2407  -1177     2354
## Random effects:
##  Groups         Name        Variance Std.Dev.
##  urban:district (Intercept) 0.323    0.568
## Number of obs: 1934, groups: urban:district, 102
##
## Fixed effects:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.340841   0.221139   -6.06  1.3e-09
## age         -0.046160   0.022022   -2.10   0.0361
```

```
## chY          1.212970   0.208994    5.80  6.5e-09
## I(age^2)    -0.005626   0.000844   -6.67  2.6e-11
## urbancity    0.786756   0.170590    4.61  4.0e-06
## age:chY      0.066466   0.025639    2.59   0.0095
##
## Correlation of Fixed Effects:
##           (Intr) age    chY    I(g^2) urbnct
## age        0.705
## chY       -0.865 -0.793
## I(age^2)  -0.093  0.299 -0.096
## urbancity -0.333 -0.057  0.082 -0.016
## age:chY   -0.583 -0.930  0.677 -0.495  0.049
```

Lessons (from Bates):

- Again, carefully plotting the data is enormously helpful in formulating the model.
- Observational data tend to be unbalanced and have many more covariates than data from a designed experiment.
- Formulating a model is typically more difficult than in a designed experiment.
- A generalized linear model is fit by adding a value, typically binomial or poisson, for the optional argument family in the call to lmer.
- MCMC sampling is not available for GLMMs at present but will be added.
- We use likelihood-ratio tests and z-tests in the model building.