Overview Day 4

- Introduction
- Generalized linear mixed models (GLMMs)
  - Combining GLM's with Mixed Models
  - Logistic and Poisson
  - Estimation procedure and software
- Extension to Non-linear models (*very* brief)

- Case studies and examples throughout

UMC Utrecht

Mixed Models Day 4: Beyond the Linear Mixed Model

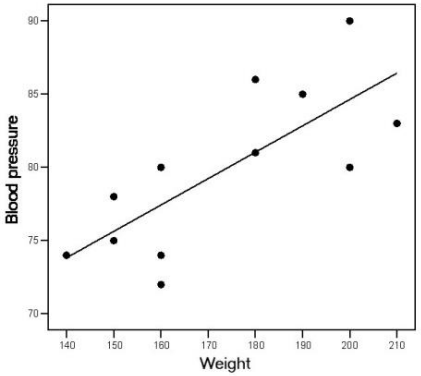
Rebecca Stellato
(Source: Cas Kruitwagen)

Generalized Linear Models

- Data
  - Outcome variable Y
  - Predictor variable(s) X
- Model
  - Left-hand side: Y (continuous, dichotomous, count, ordinal, categorical, etc., from the exponential family)
  - Right-hand side: linear equation $\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + + \beta_p X_{ip}$
  - Left- and right-hand side are linked together using an appropriate "link function"

Linear Regression

- Data
  - Continuous outcome variable Y:
    We assume the outcome for each individual i comes from $N(\mu_i, \sigma^2)$
  - Approach: we model μ_i given a (set of) predictor variable(s) X.
- Model
  - $Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$
  - $\varepsilon_i \sim N(0, \sigma^2)$
  - ε_i independent for i = 1, …, n
  - $\leftrightarrow \mu_i = \beta_0 + \beta_1 X_{i1}$



Generalized Linear Models

- Example: logistic regression
  - Dichotomous outcome variable Y (1/0)
  - Link function: logit

$$logit\big(P(Y=1)\big) = \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right)$$

  - Model:

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + + \beta_p X_{ip}$$

- For example:
  - Y = pregnant (1 = yes, 0 = no), X = age, weight, LHB/CGB genes, etc.
  - Y = heart disease (1 = yes, 0 = no), X = age, weight, exercise, blood pressure, cholesterol
- e^{β_p} is the odds ratio corresponding to the effect of X_p on Y

Generalized Linear Models

- Example: logistic regression
  - Dichotomous outcome variable Y (1/0), e.g.
    - pregnant (1 = yes, 0 = no)
    - heart disease (1 = yes, 0 = no)
  - Assumed distribution of the outcome: binomial
    - Each individual i that is drawn can be seen as the outcome of a "Bernoulli trial", with success probability $P(Y_i = 1)$
  - Principle: we model the success probability $P(Y_i = 1)$, given a set of predictor variables

Generalized Linear Models

- Example: Poisson regression
  - Count outcome variable Y
  - Link function: natural logarithm
  - Model:

$$\ln\big(E(Y_i)\big) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + + \beta_p X_{ip}$$

- For example:
  - Y = number of urinary tract infections per year, X = age, weight, antibiotics use, cranberry use, etc.
  - Y = number of telephone calls in NL on a given date, X = working day, season, temperature, economy, etc.

Generalized Linear Models

- Example: Poisson regression
  - Outcome variable Y: count within a given time or space, e.g.
    - Y = number of urinary tract infections per year
    - Y = number of telephone calls in NL on a given date
    - Y = number of insects on a plot of land
  - Assumed distribution of the outcome: Poisson
    - Parameter: rate λ (=mean, =variance)
    - Each individual i that is drawn can be seen as a draw from the Poisson distribution with rate λ_i
  - Principle: we model the rate λ_i, which is related to the expected count $E(Y_i)$, given a set of predictor variables

Linear Mixed Models

- Linear mixed model with levels i and j:

$$Y_{ij} = (\beta_0 + \upsilon_{0i}) + (\beta_1 + \upsilon_{1i}) \cdot X_{1ij} + \cdots + (\beta_p + \upsilon_{pi}) X_{pij} + \varepsilon_{ij}$$

  - Continuous outcome variable Y
  - p predictor variables X (X_{ij} on level 1, X_i on level 2)
  - Fixed effects $\beta_0 \ldots \beta_p$
  - Random effects $\upsilon_{0i} \ldots \upsilon_{pi}$ (multivariate normally distributed, with covariance matrix)
  - Residuals ε_{ij} (multivariate normally distributed, with covariance matrix)

Generalized Linear Models

- Poisson regression: offset
  - Varying exposure window, e.g.
    - Insects (not all plots of land which we observe have the same size -> insects/km^2).
    - Infections (not all patients were followed for the same length of time -> infections/year).
  - Formula:

$$\ln\left(\frac{E(Y_i)}{exposure}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} \leftrightarrow$$

$$\ln\big(E(Y_i)\big) - \ln(exposure) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} \leftrightarrow$$

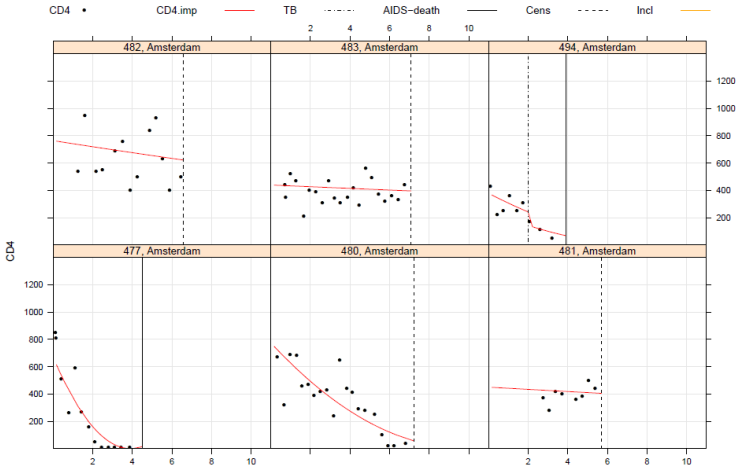
$$\ln\big(E(Y_i)\big) = \beta_0 + 1 * \ln(exposure) + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}$$

    - $\ln(exposure)$ is a so-called "offset" variable, with coefficient set to 1

Linear Mixed Models

- Example: CD4 count
  - Measured in HIV positive patients, over time (since seroconversion)
  - Level 1: repeated CD4 measurements (j)
  - Level 2: individual patients (i)
  - Level 1 covariate: having active tuberculosis (TB) (1=yes/0=no)
  - 6 example patients (next slide)

Linear Mixed Models



Linear Mixed Models

- Example: CD4 count
  - Model includes:
    - Square root of CD4 count as outcome
    - Fixed and random intercept
    - Fixed and random effect of time
    - Fixed effect of TB
  - Model:

$$\sqrt{CD4}_{ij} = (\beta_0 + \upsilon_{0i}) + (\beta_1 + \upsilon_{1i}) \cdot t_{1ij} + \beta_2 TB_{ij} + \varepsilon_{ij}$$

Generalized Linear Mixed Models (GLMMs)

- Example: logistic

$$\ln\left(\frac{P(Y_{ij}=1)}{1-P(Y_{ij}=1)}\right) = (\beta_0 + \upsilon_{0i}) + (\beta_1 + \upsilon_{1i}) \cdot X_{1ij} + \cdots + (\beta_p + \upsilon_{pi}) X_{pij}$$

- Example: Poisson

$$\ln\left(E(Y_{ij})\right) = (\beta_0 + \upsilon_{0i}) + (\beta_1 + \upsilon_{1i}) \cdot X_{1ij} + \cdots + (\beta_p + \upsilon_{pi}) X_{pij}$$

Generalized Linear Mixed Models (GLMMs)

- Similar to GLM:
  - Left-hand side: Y (continuous, dichotomous, count, ordinal, categorical, etc., from the exponential family)
  - Right-hand side: includes linear equation

$$(\beta_0 + \upsilon_{0i}) + (\beta_1 + \upsilon_{1i}) \cdot X_{1ij} + \cdots + (\beta_p + \upsilon_{pi}) X_{pij}$$

  - Left- and right-hand side are linked together using an appropriate link function.

Example cases

- These are analysed in R
  - Examples come from the mlmRev package:
    - install.packages("mlmRev")
  - Analysis using lme4 package:
    - install.packages("lme4")

Example case: contraception

- Data: Contraception

  A data frame with 1934 observations from married women <50 years old on the following 6 variables:

  - **woman** - Identifying code for each woman - a factor → *level 1*
  - **district** - Identifying code for each district - a factor → *level 2*
  - **use** - Contraceptive use at time of survey → outcome
  - **livch** - Number of living children at time of survey - ordered factor; levels are 0, 1, 2, 3+ → *level 1 covariate*
  - **age** - Age of woman at time of survey (in years), centered around mean → *level 1 covariate*
  - **urban** - Type of region of residence - a factor; levels are urban and rural → *level 1 covariate* (?)

Example case: contraception

- Data: Contraception

```
library(mlmRev)
data(Contraception)
?Contraception
```

- These data on the use of contraception by women in urban and rural areas (within districts) come from the 1988 Bangladesh Fertility Survey.

Example case: contraception

Examine the dataset:

```
> Contraception[1:4,]
  woman district use livch      age urban
1     1        1   N     3+ 18.4400     Y
2     2        1   N      0 -5.5599     Y
3     3        1   N      2  1.4400     Y
4     4        1   N     3+  8.4400     Y

> Contraception[501:504,]
    woman district use livch     age urban
501   501       14   Y      2 -4.5599     Y
502   502       14   Y      1 -5.5599     Y
503   503       14   N      1 -8.5599     Y
504   504       14   Y      2  0.4400     Y
```

Example case: contraception

Is urban constant within district?

```
> with(Contraception, table(district, urban))
        urban
district   N   Y
       1   54  63
       2   20   0
       3    0   2
       4   19  11
       5   37   2
       6   58   7
       7   18   0
       8   35   2
       9   20   3
       . . .
```

-> No, urban varies within district, so is indeed a *level 1* covariate.

Example case: contraception

- Let's think about the analysis
  - Dichotomous outcome → logistic regression
  - Predictors: number of living children (factor), age, urban
  - Women (=level 1) live within districts (sample of all districts in Bangladesh, = level 2)
  - Random intercept at level 2?
  - Random slope for predictors, at level 2?

Example case: contraception

Some descriptives

```
> table(Contraception$use)
   N    Y
1175  759
> table(Contraception$livch)
  0   1   2  3+
530 356 305 743
> summary(Contraception$age)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-13.560000  -7.560000  -1.560000   0.002198   6.440000  19.440000
> table(Contraception$urban)
   N    Y
1372  562
```


Example case: contraception

Logistic model for contraception use, regressed on main effects of livch, age and urban, and with a random intercept for each district:

```
> mod1 <- glmer(use ~ livch + age + urban + (1 | district), family =
binomial, data = Contraception)
```


Example case: contraception

- Some possible models (livch as factor variable, 3 dummies*)
  - Fixed effects only, don't take district into account*:

$$\ln\left(\frac{P(use_i = 1)}{1 - P(use_i = 1)}\right) = \beta_0 + \beta_1 livch_i + \beta_2 age_i + \beta_3 urban_i$$

  - Random intercept per district:

$$\ln\left(\frac{P(use_{ij} = 1)}{1 - P(use_{ij} = 1)}\right) = (\beta_0 + \upsilon_{0i}) + \beta_1 livch_{ij} + \beta_2 age_{ij} + \beta_3 urban_{ij}$$

  - Random intercept + random effect urban per district:

$$\ln\left(\frac{P(use_{ij} = 1)}{1 - P(use_{ij} = 1)}\right) = (\beta_0 + \upsilon_{0i}) + \beta_1 livch_{ij} + \beta_2 age_{ij} + (\beta_3 + \upsilon_{3i}) urban_{ij}$$

  - *Right-hand side should actually read:

$$\beta_0 + \beta_1(livch_i = 1) + \beta_2(livch_i = 2) + \beta_3(livch_i = 3) + \beta_4 age_i + \beta_5 urban_i$$

Example case: Melanoma Mortality

- Data: Mmmec

```
library(mlmRev)
data(Mmmec)
?Mmmec
```

- Malignant Melanoma Mortality in the European Community associated with the impact of UV radiation exposure.

Example case: contraception

```
> mod1
AIC  BIC logLik deviance
 2428 2467  -1207     2414
Random effects:
 Groups   Name        Variance Std.Dev.
 district (Intercept) 0.21239  0.46086
Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.689710   0.145496 -11.613  < 2e-16 ***
livch1       1.109184   0.156825   7.073 1.52e-12 ***
livch2       1.376396   0.173309   7.942 1.99e-15 ***
livch3+      1.345234   0.177772   7.567 3.81e-14 ***
age         -0.026595   0.007828  -3.398  0.00068 ***
urbanY       0.732918   0.118419   6.189 6.05e-10 ***
```

Example case: Melanoma Mortality

Examine the dataset

```
> Mmmec[1:4,]
   nation region county deaths expected     uvb
1 Belgium      1      1     79  51.2220 -2.9057
2 Belgium      2      2     80  79.9560 -3.2075
3 Belgium      2      3     51  46.5169 -2.8038
4 Belgium      2      4     43  55.0530 -3.0069

> Mmmec[301:304,]
    nation region county deaths expected    uvb
301  Italy     66    302      5   8.2140 6.0751
302  Italy     66    303     11   7.1600 6.6938
303  Italy     67    304     13  13.6230 1.2744
304  Italy     67    305     15  13.9220 1.6140
```

Example case: Melanoma Mortality

- Data: Mmmec

  data frame with 354 observations on the following 6 variables:

  - **nation** - a factor with levels Belgium, W. Germany, Denmark, France, UK, Italy, Ireland, Luxembourg, and Netherlands → *level 3*
  - **region** - region ID - a factor. → *level 2*
  - **county** - county ID - a factor. → *level 1*
  - **deaths** - number of male deaths due to MM during 1971–1980 → outcome (number of deaths within county)
  - **expected** - number of expected deaths due to MM → measure for exposure (based on total number of deaths and person years at risk, used as *offset variable*).
  - **uvb** - *centered* measure of the UVB dose reaching the earth's surface in each county → *level 1 covariate*

Example case: Melanoma Mortality

Some descriptives

```
> as.data.frame(table(Mmmec$nation)) #table in nice format
         Var1 Freq
1     Belgium   11
2   W.Germany   30
3     Denmark   14
4      France   94
5          UK   70
6       Italy   95
7     Ireland   26
8  Luxembourg    3
9 Netherlands   11

> length(unique(Mmmec$region)) #number of regions
[1] 78
```


Example case: Melanoma Mortality

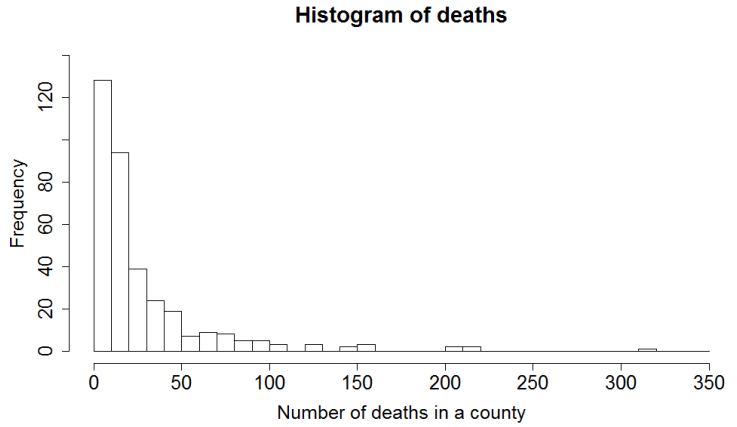
Some more descriptives

```
> summary(Mmmec$deaths)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    8.00   14.50   27.83   31.00  313.00

> summary(Mmmec$expected)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.69   11.02   18.76   27.80   34.39  258.90

> summary(Mmmec$uvb)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-8.900000 -4.158000 -0.886400  0.000204  3.276000 13.360000
```


Example case: Melanoma Mortality

- Histogram of the outcome variable

```
> hist(Mmmec$deaths, xlim = c(0, 320), breaks = 320)
```



Example case: Melanoma Mortality

- Let's think about the analysis
  - Deaths in county (count) → Poisson regression
  - Counties (=level 1) within regions (sample of regions in EU, = level 2)
  - Predictor: UVB dose
  - Random intercept per region?
  - Random slope for UVB per region?

Example case: Melanoma Mortality

- Some possible models
  - Fixed effect only:
    $$\ln\left(E(deaths_i)\right) = \ln(expected_i) + \beta_0 + \beta_1 uvb_{ij}$$
  - Random intercept per region:
    $$\ln\left(E\left(deaths_{ij}\right)\right) = \ln\left(expected_{ij}\right) + (\beta_0 + \upsilon_{0i}) + \beta_1 uvb_{ij}$$
  - Random intercept + random slope of UVB per region:
    $$\ln\left(E\left(deaths_{ij}\right)\right) = \ln\left(expected_{ij}\right) + (\beta_0 + \upsilon_{0i}) + (\beta_1 + \upsilon_{1i}) uvb_{ij}$$

Example case: Melanoma Mortality

- Expected deaths -> Use as offset in Poisson model

$$\ln\left(\frac{E(deaths_i)}{expected_i}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} \leftrightarrow$$

$$\ln\left(E(deaths_i)\right) - \ln(expected_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} \leftrightarrow$$

$$\ln\left(E(deaths_i)\right) = \beta_0 + 1 * \ln(expected_i) + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$$

- Offsets used in Poisson regression when outcome is a *rate*: count (numerator) is outcome, denominator (expected deaths, person-years, etc.) is added as offset

Example case: Melanoma Mortality

```
> pmod1
Generalized linear mixed model fit by the Laplace approximation
Formula: deaths ~ uvb + (1 | region)
   Data: Mmmec
   AIC BIC logLik deviance
 661.4 673 -327.7    655.4
Random effects:
 Groups Name        Variance Std.Dev.
 region (Intercept) 0.16968  0.41192
Number of obs: 354, groups: region, 78

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.138601   0.049330  -2.810 0.004959 **
uvb         -0.034434   0.009734  -3.538 0.000404 ***
```


Example case: Melanoma Mortality

Poisson regression model for deaths, regressed on a main effect of uvb, and including a random intercept for region

```
> pmod1 <- glmer(deaths ~ uvb + (1|region), family = poisson,
data = Mmmec, offset = log(expected))
```

GLMM: parameter estimation

- Marginal quasi-likelihood (MQL) -> biased.
- Penalized/predictive quasi-likelihood (PQL) -> biased.
- Laplace approximation -> accurate, fast, likelihood/AIC/BIC obtainable.
- Gauss-Hermite quadrature -> accurate, likelihood/AIC/BIC obtainable, but computationally intensive.
- Markov chain Monte Carlo (MCMC) -> very flexible, but computationally intensive.

Example case: Melanoma Mortality

Interpretation parameter estimates

- Intercept = ln(mean number of deaths/expected for a county with a mean UV-B exposure) (uvb = 0)
  - exp(-0.1386) = 0.87 is mean "rate" or #deaths/expected
- Coefficient for uvb is a ln(RR) for a 1-unit increase in UV-B
  - So exp(-0.0344) = 0.97: RR for melanoma mortality/expected mortality for 1-unit increase in UV-B

GLMM: commonly used software

- R
  - MASS package: glmPQL (possible bias, no likelihood/AIC/BIC)
  - lme4 package: glmer (Laplace approximation)
  - MCMCglmm package (MCMC)
- SAS
  - PROC GLIMMIX (Laplace)
  - PROC NLMIXED (adaptive Gaussian quadrature, first-order Taylor series approximation)
- WinBUGS
  - Bayesian inference (Markov chain Monte Carlo)
- MLwiN
  - MQL, PQL, MCMC

GLMM: commonly used software

Published in final edited form as:
Stat Med. 2011 September 10; 30(20): 2562–2572. doi:10.1002/sim.4265.

On Fitting Generalized Linear Mixed-effects Models for Binary Responses using Different Statistical Packages

Hui Zhang[1], **Naiji Lu**[2,3], **Changyong Feng**[2], **Sally W. Thurston**[2], **Yinglin Xia**[2,3], and **Xin M. Tu**[2,3,4]

- In most procedures, estimates are biased (exception SAS NLMIXED)
- "We are a bit surprised by the performance of the R lme4 and glmmML packages, as neither seems to yield comparable inference as its SAS NLMIXED counterpart given that it implements the same integral approximation approach, albeit using different algorithms."

Comparing GLMMs with Laplace approximation

- Comparing the models
  - AIC: lower is better
  - Model with -2LL significantly lower is better
  - Model with -2LL not significantly different, but with fewer parameters is better

GLMM vs GEE

- Beyond the scope of this course, *but*
- GLMM's give *conditional* parameter estimates (given your random effects)
- GEE gives population-averaged parameter estimates (generally preferred)

- Also: mixed models okay when outcomes are MCAR, MAR; GEE only give unbiased estimates when when outcomes are MCAR
  - Some authors recommend first imputing, then using GEE

Non-exponential non-linear models

- We've covered two frequently used GLMM's
  - Logistic (dichotomous outcomes)
  - Poisson (count outcomes)

- Other random effect-models can be defined, e.g. non-linear models not from the exponential family, with random effects
- Example: children with development of motor function
  - Motor function distribution defined by asymptote (maximum level), and rate of change (increase with age in motor function)
  - Asymptote and rate can differ between children
  - Non-linear asymptotic regression with random effects
- Software: nlme package (R) -> nlme function with *SSasymp* term

Non-exponential non-linear models

Fitted curve (fixed effect), with individual data points: