UMC Utrecht

Mixed Models Day 4: Beyond the Linear Mixed Model

Cas Kruitwagen

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)

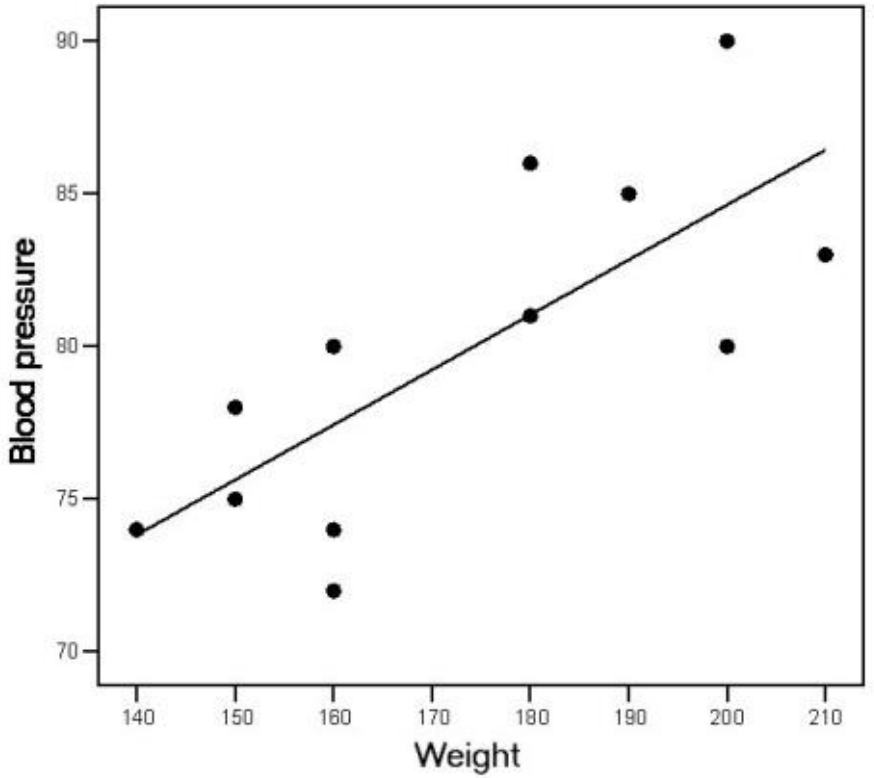
- Case studies and examples throughout

1

Generalized linear mixed models (GLMMs)

Linear Regression

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



Generalized Linear Models

broader

- Data
  - Outcome variable Y left hand side
  - Predictor variable(s) X
- Model
  - Left-hand side: Y (continuous, dichotomous, count, ordinal, categorical, etc., from the exponential family)
    linear reg logistic reg
  - 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"

Generalized Linear Models

- Example: logistic regression
  - Dichotomous outcome variable Y (1/0), e.g.
    - pregnant (1 = yes, 0 = no), left hand side of equation
    - 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 $P(Y_i=1)$. probability of a outcome being Yes/No
  - Principle: we model the success probability $P(Y_i=1)$, given a set of predictor variables.

Generalized Linear Models

- Example: logistic regression

not directly probability of it being success (yes) but logit of the probability

left

  - Dichotomous outcome variable Y (1/0).

link

  - 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:

ß estimated log odds ratio corresponding to the effect of that variable to outcome

  - 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

probability of becoming pregnant related to age, weight, BMI, etc.

Generalized Linear Models

counts will follow a poisson distribution (count of anything in given space/time) e.g. how many seagulls per square metre on a soccer field

- 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. exponential
    - Parameter: rate λ (=mean, =variance) mean = variance; less mean = less variation etc.
    - Each individual i that is drawn can be seen as a draw from the Poisson distribution with rate λ_i depends on covariance of the individual
  - Principle: we model the rate λ_i, which is related to the expected count $E(Y_i)$, given a set of predictor variables

    mean of Y given a set of predictors

Generalized Linear Models

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

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

log of expected values of Y

linear function of predictor variables

- For example:
  - Y = number of urinary tract infections per year, X = age, weight, antibiotics use, cranberry use, etc.

time-frame!

  - Y = number of telephone calls in NL on a given date, X = working day, season, temperature, economy, etc.

Y covariates

Generalized Linear Models

- Poisson regression: offset (adding an offset variable)
  - Varying exposure window, e.g. (different sizes or time frame)
    - 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: (corrected for exposure window)

expected mean

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

variable (longer shorter times)

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

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

log expected count = + 1 * log of exposure

coefficient for the log of exposure set to be 1

Linear Mixed Models

- Linear mixed model with levels i and j:

outcome dependent on jth level 1 within ith unit level 2

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

random effect added for variables (depends on model but one could add it for any variable)

  - 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)

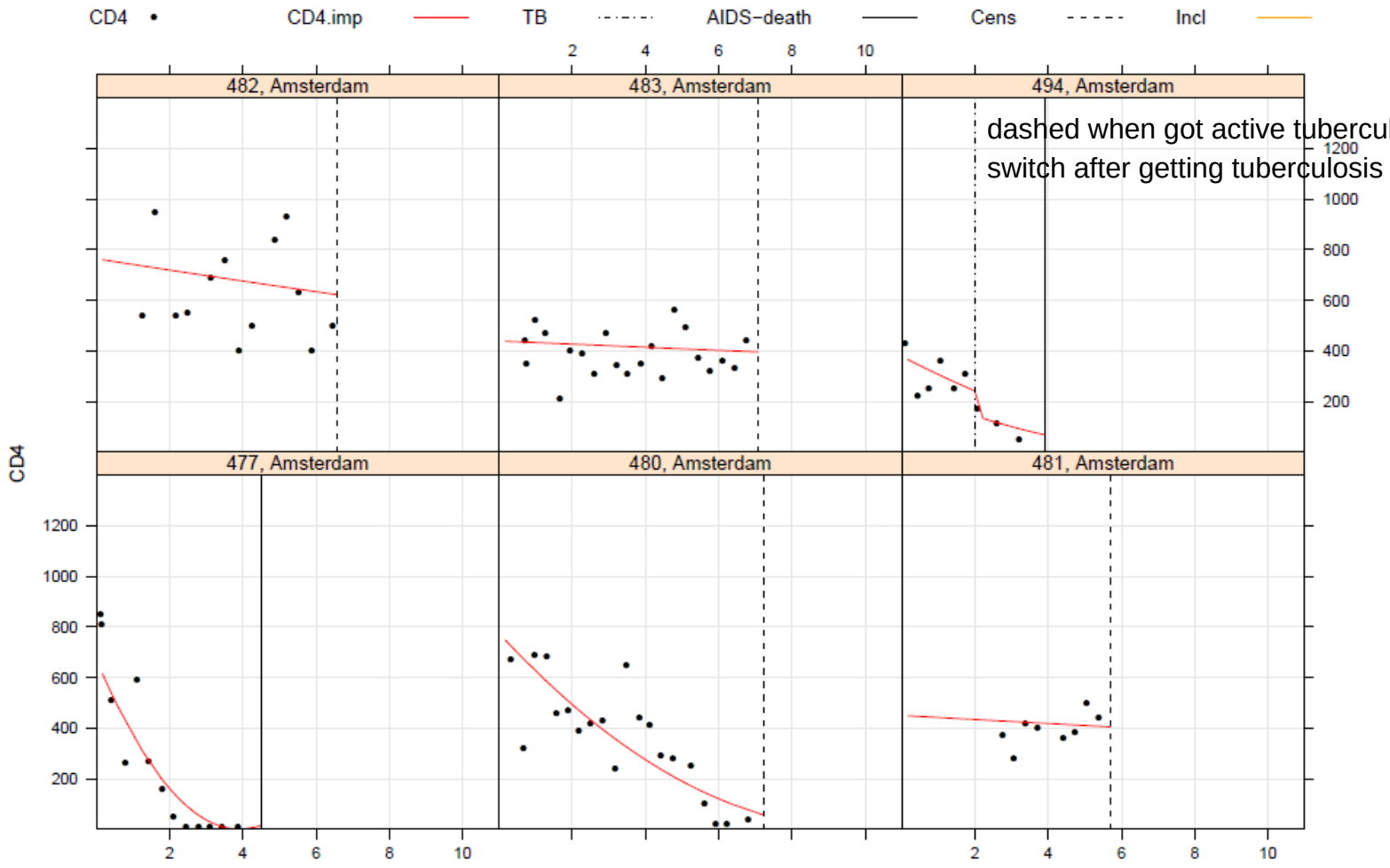
often when linear mixed model it's just sigma squared times identity matrix

Linear Mixed Models

- Example: CD4 count
  - Measured in HIV positive patients, over time (since seroconversion).
  - Level 1: repeated CD4 measurements (j). within individuals
  - Level 2: individual patients (i).
  - Level 1 covariate: having active tuberculosis (TB) (1=yes/0=no). at this point in time yes or no? time dependent variable
  - 6 example patients (next slide).

CD4 counts are number of cells that show health

Linear Mixed Models



dashed when got active tuberculosis so it would switch after getting tuberculosis

Linear Mixed Models

also example of generalized mixed effects model

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

$$\underbrace{\sqrt{CD4}_{ij}}_{\text{link}} = (\beta_0 + \upsilon_{0i}) + (\beta_1 + \upsilon_{1i}) \cdot t_{1ij} + \underbrace{\beta_2 TB_{ij}}_{\substack{\text{fixed effect} \\ \text{for tuberculosis}}} + \varepsilon_{ij}$$

doesn't make sense to add random effect for tuberculosis because then we would have to assume that the effect of TB differs for each person

Generalized Linear Mixed Models (GLMMs)

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

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

(fixed and random effects again)

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

Generalized Linear Mixed Models (GLMMs)

adding mixed effects to generalized mixed models

- Example: logistic

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

log of odds

- Example: Poisson

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

log of counts

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

```
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

level 1 = per person

- Data: Contraception

  A data frame with 1934 observations 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 are they using it yes or no (dich)
  - **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* (?)

0 = average age which was around 30,5

because centered around the mean

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
       . . .
```

quick check to see if it is a level 1 covariate
it varies across districts

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

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
```

age can be negative because mean age at 30,5 so age 17 women are included in the survey

Example case: contraception

- Let's think about the analysis
  - Dichotomous outcome (yes/no) → 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? for the districts (some districts might have higher/lower proportion)
  - Random slope for predictors, at level 2?
    depends on how much we know about the topic -

Example case: contraception

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

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

only varies per women not district

but it's 3 dummies

so 3 ß

adding

  - Random intercept per district:

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

all level 1 varying across women within districts

random intercept

random effect for urban/rural may vary for each district

  - Random intercept + random slope urban per district:

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

random effect for urban saying that differences between urban and rural might vary over the districts

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:

random intercept 1 per district

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

automatically logit link when binomial

odds ratios dangerous because always odds of contr. use are this much higher than...
but difficult to use it as it's interpreted wrong, calling it chance instead of odds so relative risk is better
but logistic regression gives us odds
options: binomial distribution with logik instead of log ->relative risk
but can only add dich. variables or it'll be a mess

Example case: contraception

assuming linear connection between age and log contraception - which is questionable as younger women are likely to take more contraception as older women - so not negative over all ages

```
> 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 ***
```

modeling odds ratios with log odds not probability

estimates on log odds

Intercept is the log odds of contr. use for a woman that has 0 on all covariates (= no living children, rural area and average age) -1.68 is log odds for that woman

always compared to the reference group

odds of using contr. lich1 is e to the power of 1.109 -> 3 times higher than Intercept

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: 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)

compared to men exp. to die

  - **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

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
```

offset variable

varies in regions in countries = more or less sun

```
> 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

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
```

counties in the countries

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
```

average number of deaths is scewed

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

right scewed because median smaller than mean

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

because it's centered

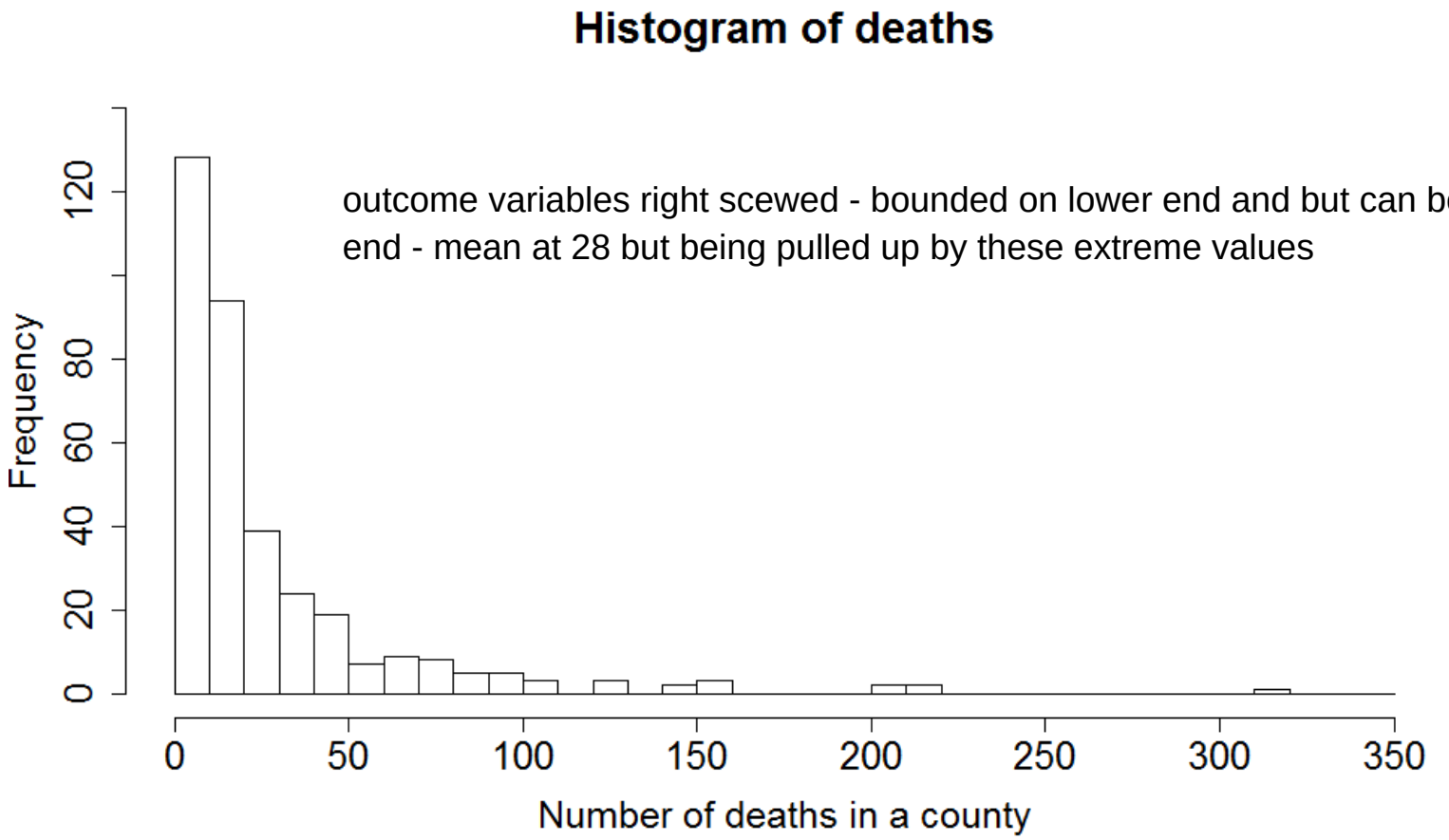
Example case: Melanoma Mortality

- Let's think about the analysis
  - Deaths in county (count) → Poisson regression count = poisson regression
  - Counties (=level 1) within regions (sample of regions in EU, = level 2)
  - Predictor: UVB dose measured at level 1 so level 1 covariate
  - Random intercept per region? because multilevel it makes sense to add
  - Random slope for UVB per region? maybe association between UV dose and log odds varies within the counties

Example case: Melanoma Mortality

- Histogram of the outcome variable

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



outcome variables right scewed - bounded on lower end and but can be much higher on other end - mean at 28 but being pulled up by these extreme values

Example case: Melanoma Mortality

- Expected deaths -> Use offset in Poisson model

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

number of deaths per county

only one explanatory variable

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

log of expected moves over to this side again

offset in poisson model is to account for followup time as a "normalizing" variable to add number of events per person year

Example case: Melanoma Mortality

- Some possible models
  - Fixed effect only:

linear effect for uvb exposure

$$\ln\left(E(deaths_i)\right) = \ln(expected_i) + \beta_0 + \beta_1 uvb_i$$

log of mean of deaths — intercept

  - Random intercept per region:

add random intercept

$$\ln\left(E\left(deaths_{ij}\right)\right) = \ln\left(expected_{ij}\right) + \beta_0 + \upsilon_{0i} + \beta_1 uvb_{ij}$$

  - Random intercept + random slope of UVB per region:

random slope for uvb

$$\ln\left(E\left(deaths_{ij}\right)\right) = \ln\left(expected_{ij}\right) + \beta_0 + \upsilon_{0i} + (\beta_1 + \upsilon_{1i})uvb_{ij}$$

allow linear association between uvb and deaths to vary within region

Example case: Melanoma Mortality

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

1| nation to add another intercept per nation

random intercept

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

most used and implemented in packages (lme4)

coefficient is 1

which is why we don't get a coefficient for the offset

if one thinks it's a "true 3 level" design one should try for two intercepts but one can check after adding a random intercept per nation and then see if the model fits well or not

Example case: Melanoma Mortality

when 3 level design at least add a random intercept for that level

```
> 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 ***
```

log of the mean number of deaths per expected when uvb is zero (average) because centered

for every one unit increase in uvb the log of the mean number of deaths per expected is decreasing by -0.0344 (exponentiate = incidence rate ratio of .97)

GLMM: parameter estimation

tricky to figure out the likelihood

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

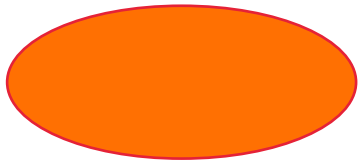
GLMM: commonly used software

- R
  - MASS package: glmPQL (possible bias, no likelihood/AIC/BIC)
  - lme4 package: glmer (Laplace approximation)
  - MCMglmm package (MCMC)
- SAS
  - PROC GLIMMIX (Laplace)
- WinBUGS
  - Bayesian inference (MCMC)
- MLwiN

Comparing GLMMs with Laplace approximation

interpretation Laplace approximation

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



3

Non-Linear Mixed Models (NLMMs)

Non-exponential non-linear models

- We covered some often-used GLMM's
- 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: