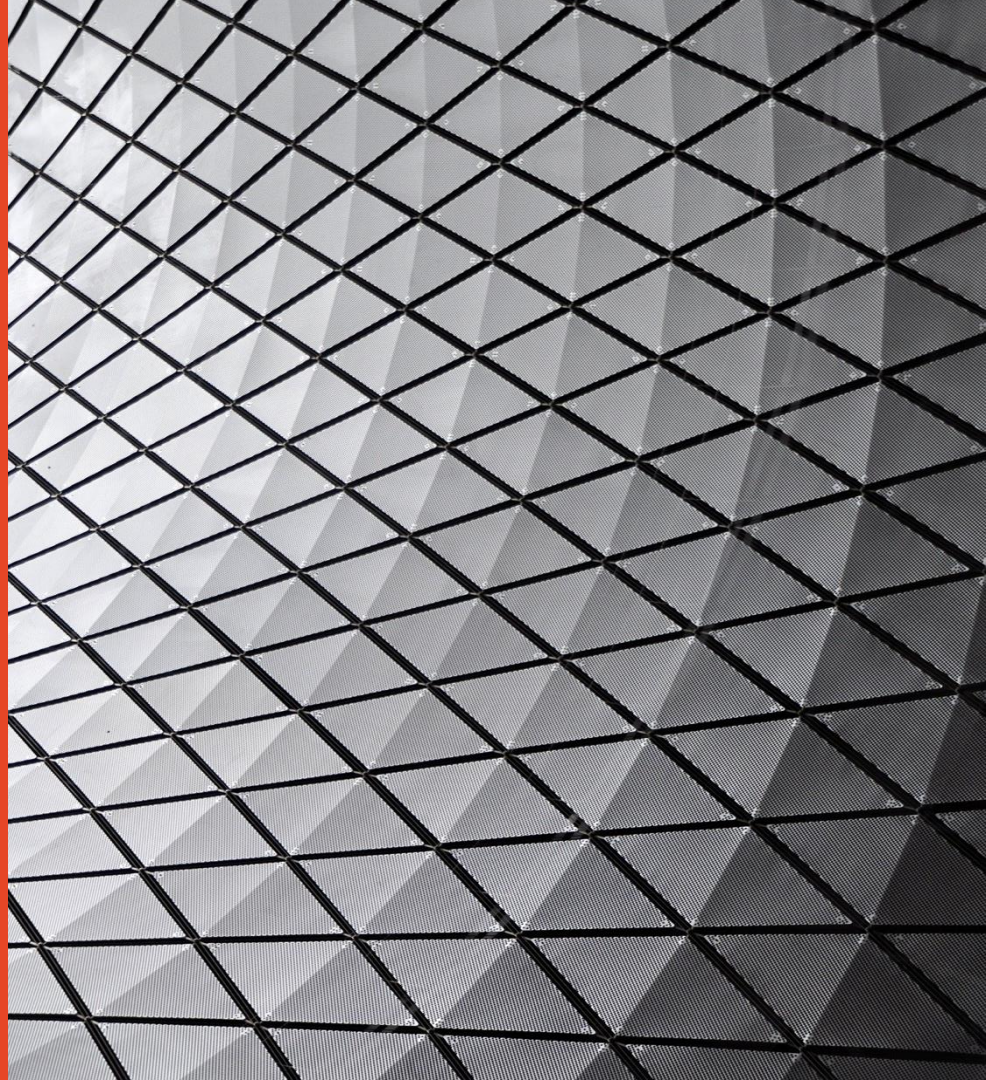
Generalised linear mixed Models and GEEs

Workshop: Analysis of Longitudinal Data

12th Nov 2024

Jaroslaw Harezlak
Armando Teixeira-Pinto

THE UNIVERSITY OF SYDNEY



Generalised Linear Models

- We have been talking about the linear model

$$y_i = \beta_0 + \beta X_i + \varepsilon_i$$

$$E(y_i) = \mu_i = \beta_0 + \beta X_i$$

- And how to extend this model to accommodate correlated observations
- We can use the same ideas for the generalised linear models

$$g(\mu_i) = \beta_0 + \beta X_i$$

Generalised Linear Models

$$g(\mu_i) = \beta_0 + \beta X_i$$

- Here, μ_i can be the usual mean, a proportion, or a rate
- Notice that the **proportion is the mean** of a dichotomous variable Y coded as 0,1
- If we take the mean of $0, 0, 1, 0, 1, 0, 1$ … it will correspond to the proportions of 1's

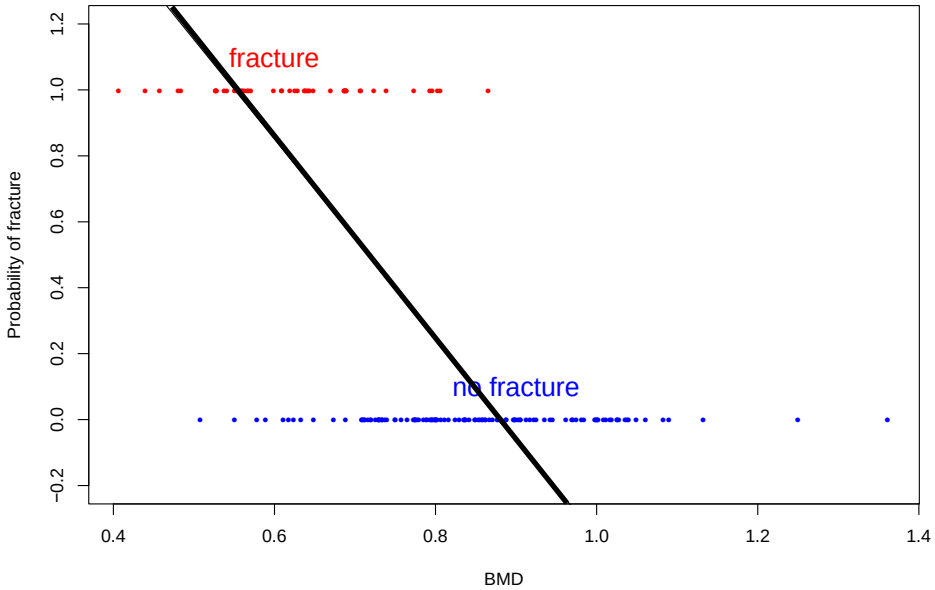
$$\mu_i = \Pr(Y_i = 1)$$

- The same idea for a count outcome Y

Generalised Linear Models

$$\mu_i = \Pr(Y_i = fracture) = \beta_0 + \beta_1 BMD_i$$

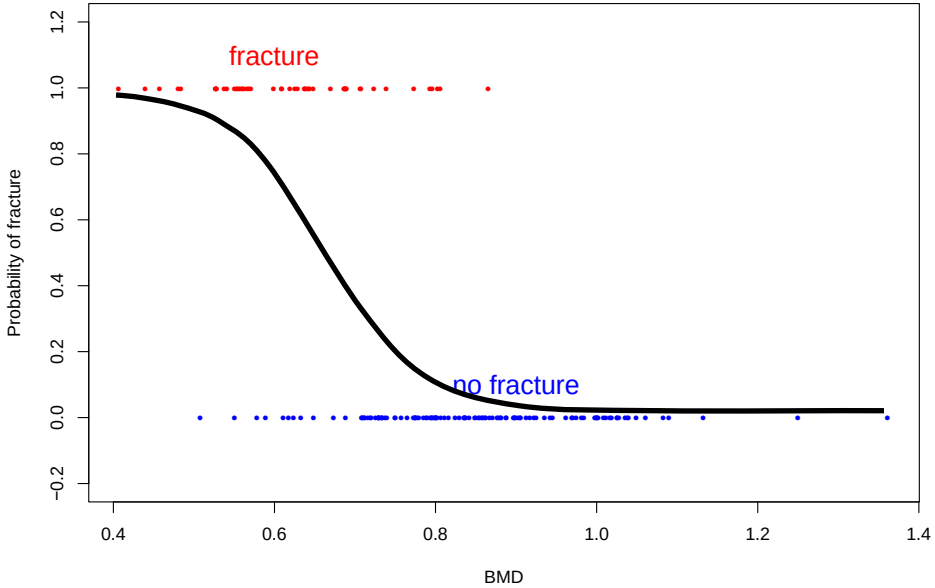
- Consider that we want to **model the probability of hip fracture** based on the bone mineral density
- A linear model for the probability of fracture will produce probabilities above 1 and below 0



Generalised Linear Models

$$\mu_i = \Pr(Y_i = fracture) = \frac{\exp(\beta_0 + \beta_1 BMD_i)}{1 + \exp(\beta_0 + \beta_1 BMD_i)}$$

- A better alterative is to consider a line that is bounded by 0 and 1
- The **logistic (or logit)** line is one of many options



Generalised Linear Models

$$\mu_i = \Pr(Y_i = fracture) = \frac{\exp(\beta_0 + \beta_1 BMD_i)}{1 + \exp(\beta_0 + \beta_1 BMD_i)}$$

- The equation above can be re-written as

$$log\left(\frac{\Pr(Y_i = fracture)}{1 - \Pr(Y_i = fracture)}\right) = \beta_0 + \beta_1 BMD_i$$

Generalised Linear Models

$$\mu_i = \Pr(Y_i = fracture) = \frac{\exp(\beta_0 + \beta_1 BMD_i)}{1 + \exp(\beta_0 + \beta_1 BMD_i)}$$

- The equation above can be re-written as

$$\underbrace{log\left(\frac{\Pr(Y_i = fracture)}{1 - \Pr(Y_i = fracture)}\right)}_{\text{logit}(\Pr(Y_i = fracture))} = \beta_0 + \beta_1 BMD_i$$

Generalised Linear Models

$$\mu_i = \Pr(Y_i = fracture) = \frac{\exp(\beta_0 + \beta_1 BMD_i)}{1 + \exp(\beta_0 + \beta_1 BMD_i)}$$

- The equation above can be re-written as

$$\text{logit}(\Pr(Y_i = fracture)) = \boldsymbol{g(\mu_i) = \beta_0 + \beta_1 BMD_i}$$

- A linear model for a transformation of the outcome's mean, i.e., a **generalised linear model**

Generalised Linear Models

$$g(\mu_i) = \beta_0 + \beta X_i$$

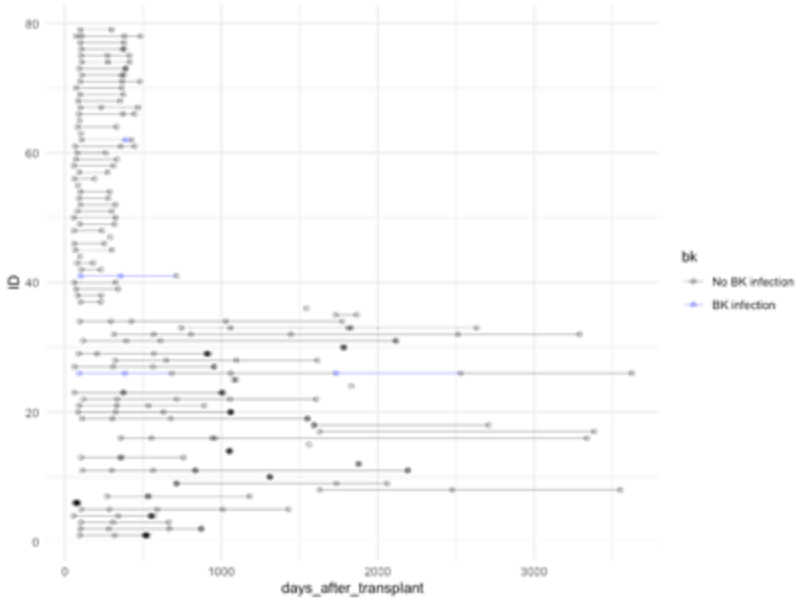
- Now, we would use maximum likelihood to estimate the regression parameters
- This would be done under the assumption of independence of observations, so that the likelihood simplifies to:

$$\mathcal{L}(\beta_0, \beta; y_1, y_2, y_3, \dots) = f(y_1, y_2, y_3, \dots | \beta_0, \beta) = \prod_i f(y_i | \beta_0, \beta)$$

independent

Generalised Estimation Equations

- **Example:**
  - After kidney transplant, patients are at risk of infection due to the immunosuppression therapy
  - BK virus poses an important risk in this population
  - The dataset bk.csv includes the variable **bk** that identifies the infection status of patients over time – infected vs not infected
  - The observations within patients are likely correlated



Generalised Estimation Equations

ID	bk	gendercode	donoragecat	DonorSource	ischaemia	days_after_transplant	egfr	time
116973	0	M	30-39	Deceased	15.00	340	69	0.931
116976	0	M	30-39	Deceased	4.00	61	67	0.167
116976	0	M	30-39	Deceased	4.00	324	66	0.887
132133	1	M	50-59	Living	3.00	99	52	0.271
132133	1	M	50-59	Living	3.00	355	55	0.972
132133	0	M	50-59	Living	3.00	713	44	1.952
143035	0	M	40-49	Living	4.00	104	59	0.285
143035	0	M	40-49	Living	4.00	229	66	0.627
143064	0	M	40-49	Deceased	5.00	78	42	0.214

Generalised Estimation Equations

- If we run a simple logistic regression

```
logit.bk <- glm(bk ~ time,
                family=binomial,
                data=bk.Data)
summary(logit.bk)
```

Generalised Estimation Equations

- If we run a simple logistic regression

```
logit.bk <- glm(bk ~ time,
                family=binomial,
                data=bk.Data)
summary(logit.bk)
```

- The OR estimate should be correct, but the SE is wrong

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.1968     0.1108  -28.84  < 2e-16 ***
time         -0.4667     0.0754   -6.19  5.9e-10 ***
---

> exp(logit.bk$coefficients)
(Intercept)        time
     0.0409      0.6271
```

Generalised Estimation Equations

- We will instead use **GEE to fit the logistic regression**

```
gee.bk.ind <- geeglm(bk ~ time,
                     family=binomial, id=ID,
                     corst="independence", data=bk.Data)
summary(gee.bk.ind)
```

- The OR will be similar but now the **SE is robust to correlations in the data**

```
Coefficients:
            Estimate Std.err  Wald Pr(>|W|)
(Intercept)  -3.1968  0.1239 665.5  < 2e-16 ***
time         -0.4667  0.0818  32.6  1.2e-08 ***
---

> exp(gee.bk.ind$coefficients)
(Intercept)        time
     0.0409      0.6271
```

Generalised Estimation Equations

- As in the continuous outcome case, we can specify other **working correlation structures**

```
gee.bk.exch <- geeglm(bk ~ time,
                      family=binomial, id=ID,
                      corst="exchangeable", data=bk.Data)
summary(gee.bk.exch)
```

- *Exchangeable* assumes that the correlation of having a BK infection in two time points is always the same

$$\begin{array}{c} \\ 1^{st} \\ 2^{nd} \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} 1^{st} & 2^{nd} & \cdots & \end{array} \\ \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} \end{array}$$

Generalised Estimation Equations

- As in the continuous outcome case, we can specify other **working correlation structures**

```
gee.bk.exch <- geeglm(bk ~ time,
                      family=binomial, id=ID,
                      corst="exchangeable", data=bk.Data)
summary(gee.bk.exch)
```

- The OR will change slightly depending on the structure

```
Coefficients:
            Estimate Std.err  Wald Pr(>|W|)
(Intercept)  -3.0406  0.1256 585.9  < 2e-16 ***
time         -0.5692  0.0972  34.3  4.7e-09 ***
---

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.144   0.273
Number of clusters:   2009  Maximum cluster size: 13
```

Comparing models fitted with GEE

```
> QIC(gee.bk.ind, gee.bk.exch, gee.bk.unst)
              QIC QICu Quasi Lik  CIC params QICC
gee.bk.ind   1642 1640      -818 2.81      2 1642
gee.bk.exch  1645 1642      -819 3.33      2 1645
gee.bk.unst  1644 1641      -819 3.23      2 1650
```

Generalised Estimation Equations

- As before, the OR obtained from GEE is a **population average odds ratio**
- Meaning that the is the average effect across all the individual
- The Pearson correlation is not a "common" measure of association between binary measurements
- In particular, the correlations between binary measurements have smaller upper and lower limits (away from 1 and -1)
- Another option is to **parametrise the correlation matrix in terms of odds ratios** (a more natural way of establishing the association between binary measurements)
- Unfortunately, R does not have this implemented (SAS does!)

Generalised Linear Mixed Models

- Let's consider now the use of random effects to model the longitudinal measurements of BK
- These models are called **generalised linear mixed models** (GLMM)

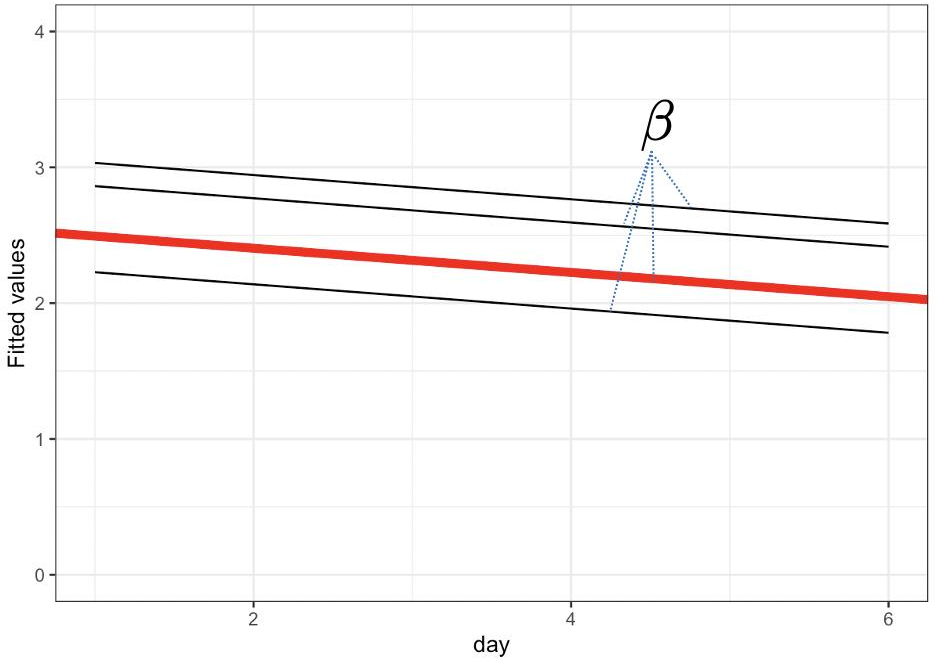
$$g(\mu_i|b_{0i}, b_i) = \underbrace{\beta_0 + \beta X_i}_{\text{Fixed effects}} + \underbrace{b_{i0} + b_i Z_i}_{\text{Random effects}}$$

- With the **random effects normally distributed**

Generalised Linear Mixed Models

- The interpretation of the parameters is not as straightforward as in the linear case.
- Recall that for the **random intercept linear model,** the effect of time, β_1, has both a subject-specific and population-averaged interpretation

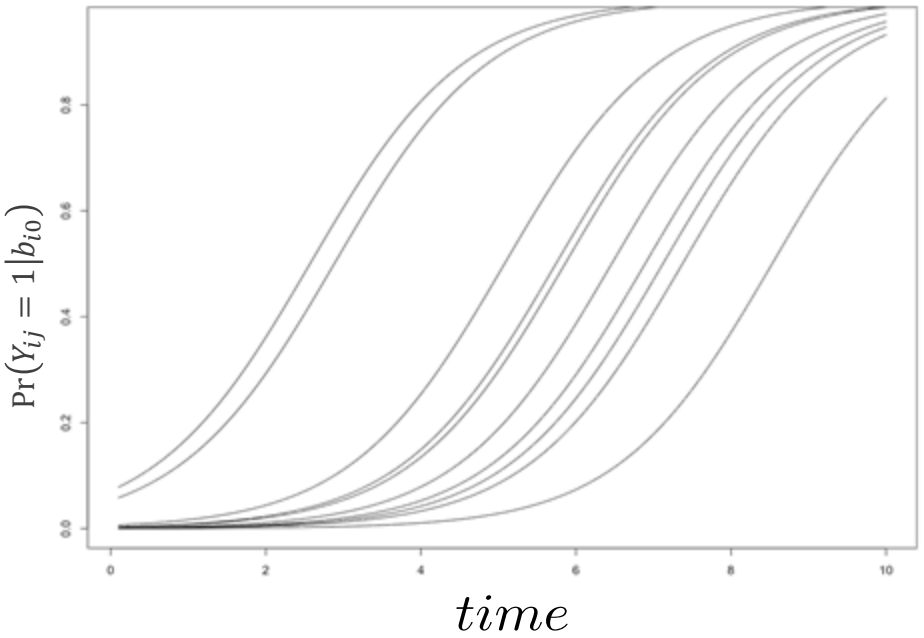
$$Y_{ij} = \beta_0 + \beta_1 \text{time}_{ij} + b_{i0} + \varepsilon_{ij}$$



Generalised Linear Mixed Models

- For the logistic random effect model, this is not the case.
- Consider the random intercept logistic regression
- The $\exp(\beta_1)$ is the change in odds (odds ratio) per unit of time, for **each individual**

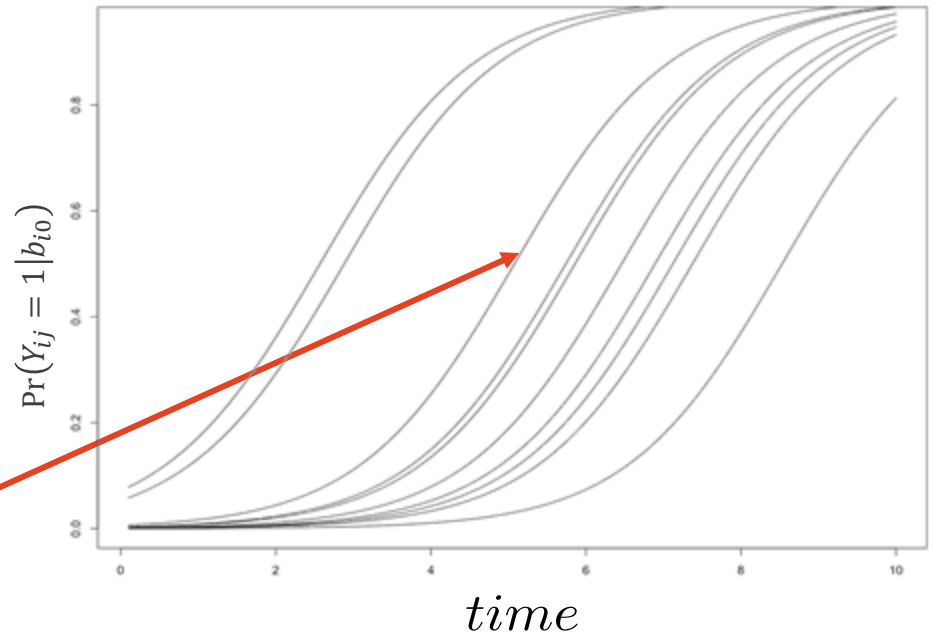
$$\Pr(Y_{ij} = 1 | b_{i0}) = \frac{\exp(\beta_0 + \beta_1 time_{ij} + b_{i0})}{1 + \exp(\beta_0 + \beta_1 time_{ij} + b_{i0})}$$



Generalised Linear Mixed Models

- For the logistic random effect model, this is not the case.
- Consider the random intercept logistic regression
- The $\exp(\beta_1)$ is the change in odds (odds ratio) per unit of time, for **each individual**

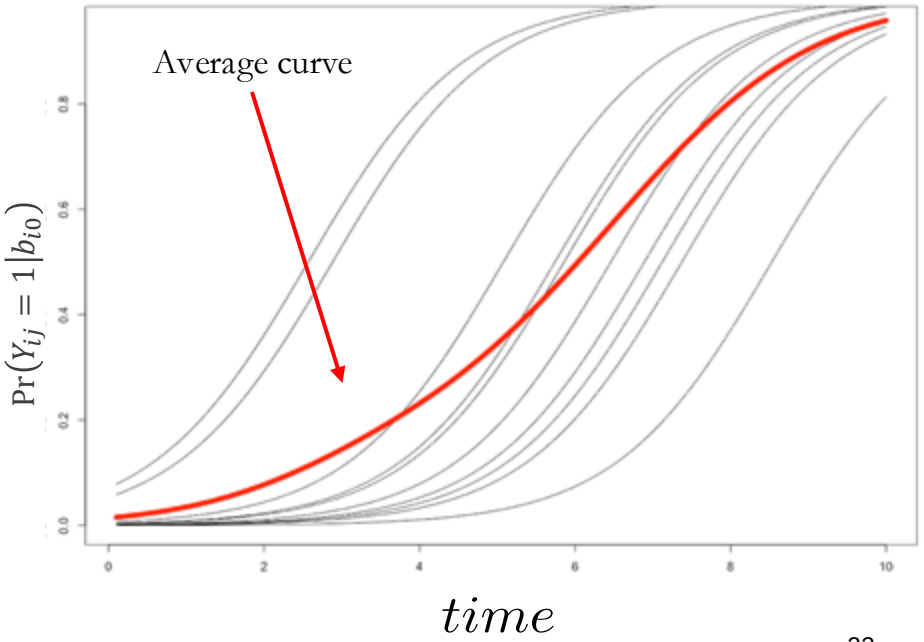
$$\Pr(Y_{ij} = 1 | b_{i0}) = \frac{\exp(\beta_0 + \beta_1 time_{ij} + b_{i0})}{1 + \exp(\beta_0 + \beta_1 time_{ij} + b_{i0})}$$



Generalised Linear Mixed Models

- If we **average over all the individual logit curves, the** average is not similar to other curves
- In fact, it is not even a logit curve
- This means that the subject specific OR is not the same as the population-average OR

$$\Pr(Y_{ij} = 1|b_{i0}) = \frac{\exp(\beta_0 + \beta_1 time_{ij} + b_{i0})}{1 + \exp(\beta_0 + \beta_1 time_{ij} + b_{i0})}$$



Generalised Linear Mixed Models

- We will use a logistic model with a random intercept for the BK example

```
bk.glmm <- glmer(bk ~ time + (1|ID),
                 family=binomial,
                 data=bk.Data))
```

- The patient specific odds of infection decreases almost 65% per year (0.36 times per year)

```
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -8.226      0.390  -21.11   <2e-16 ***
time          -1.019      0.123   -8.25   <2e-16 ***
---
 > exp(bk.glmm@beta)
[1] 0.000268 0.360996
```

Generalised Linear Mixed Models

- Notice that this is quite different from the result using the GEE

GEE

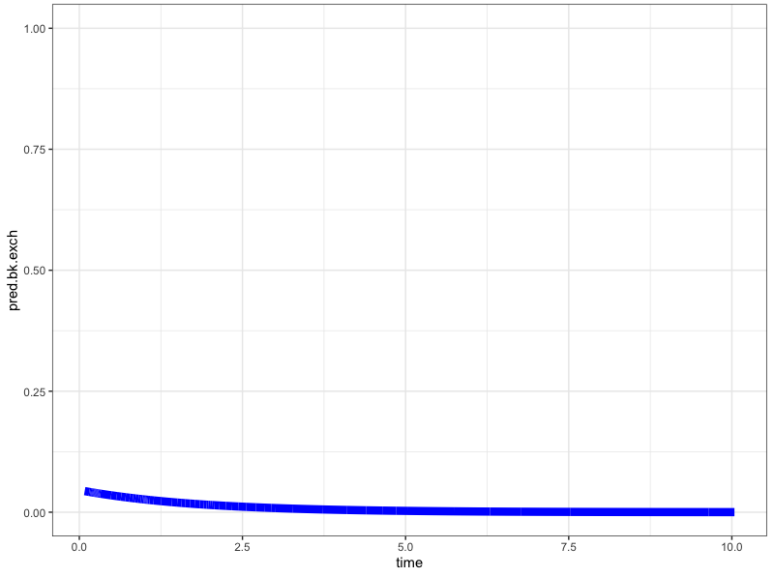
```
Coefficients:
             Estimate Std.err  Wald Pr(>|W|)
(Intercept)  -3.0406  0.1256 585.9  < 2e-16 ***
time         -0.5692  0.0972  34.3  4.7e-09 ***
---

> exp(gee.bk.exch$coefficients)
(Intercept)        time
     0.0478      0.5660
```

Random intercept model

```
Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -8.226      0.390  -21.11   <2e-16 ***
time          -1.019      0.123   -8.25   <2e-16 ***
---
 > exp(bk.glmm@beta)
[1] 0.000268 0.360996
```

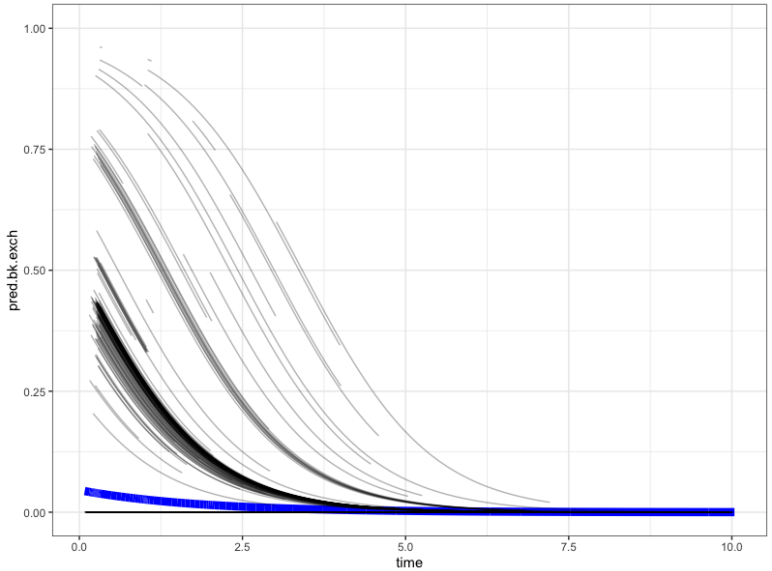
Generalised Linear Mixed Models

- We can plot the predicted probabilities
- This will correspond to the logit curve
- First for the **marginal model** from GEE
- *(here we just have the lower end of the logit curve given the low incidence of the outcome)*



Generalised Linear Mixed Models

- We can plot the predicted probabilities
- This will correspond to the logit curve
- First for the marginal model from GEE
- *(here we just have the lower end of the logit curve given the low incidence of the outcome)*
- And then, the subject specific predictions given by the **random intercept logistic model**



Generalised Estimation Equations

- Let's now compare the risk of BK for men and women (sex assigned at birth) using a marginal model

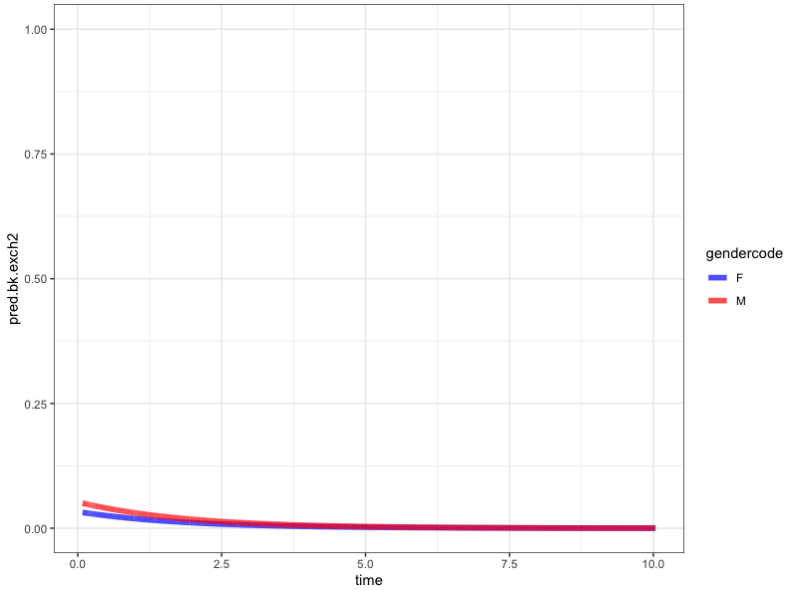
```
gee.bk.exch2 <- geeglm(bk ~ time + gendercode,
                       family=binomial,
                       id=ID,
                       corst="exchangeable",
                       data=bk.Data)
summary(gee.bk.exch2)
```

```
Coefficients:
             Estimate Std.err   Wald Pr(>|W|)
(Intercept)  -3.3759  0.1857 330.53  < 2e-16 ***
time         -0.5638  0.0962  34.36  4.6e-09 ***
gendercodeM   0.4828  0.2122   5.18    0.023 * ---

> exp(gee.bk.exch2$coefficients)
(Intercept)        time gendercodeM
     0.0342      0.5690      1.6206
```

Generalised Estimation Equations

- Let's now compare the risk of BK for men and women (sex assigned at birth) using a marginal model

```
Coefficients:
             Estimate Std.err   Wald Pr(>|W|)
(Intercept)  -3.3759  0.1857 330.53  < 2e-16 ***
time         -0.5638  0.0962  34.36  4.6e-09 ***
gendercodeM   0.4828  0.2122   5.18    0.023 * ---

> exp(gee.bk.exch2$coefficients)
(Intercept)        time gendercodeM
     0.0342      0.5690      1.6206
```



Generalised Linear Mixed Models

- And the same analysis comparing the risk of BK for men and women but using a random intercept model

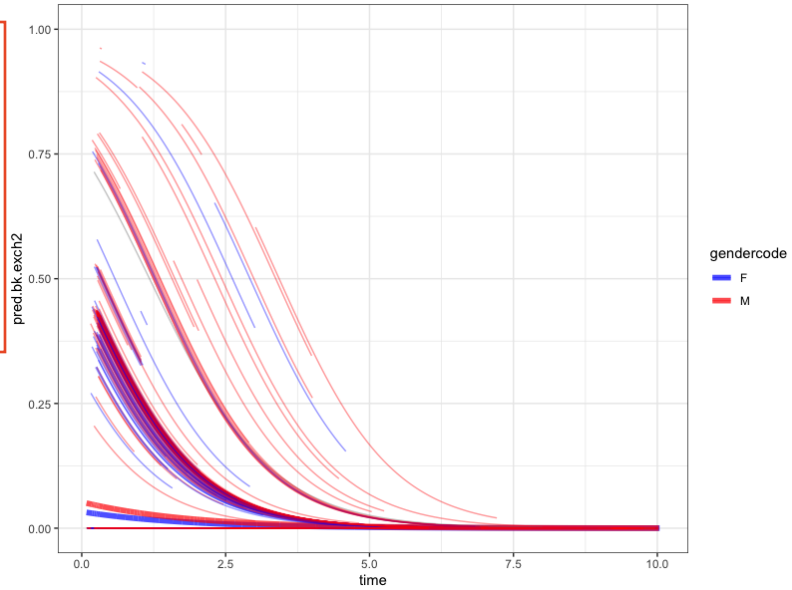
```
rInt.bk2 <- glmer(bk ~ time + gendercode + (1|ID),
                  family=binomial,
                  data=bk.Data)
summary(rInt.bk2)
```

```
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -8.774      0.570  -15.39   <2e-16 ***
time          -1.022      0.124   -8.26   <2e-16 ***
gendercodeM    0.774      0.498    1.55     0.12

> exp(rInt.bk2@beta)
[1] 0.000155 0.359772 2.168564
```

Generalised Linear Mixed Models

- And the same analysis comparing the risk of BK for men and women but using a random intercept model

```
Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -8.774      0.570  -15.39   <2e-16 ***
time           -1.022      0.124   -8.26   <2e-16 ***
gendercodeM     0.774      0.498    1.55     0.12

> exp(rInt.bk2@beta)
[1] 0.000155 0.359772 2.168564
```



Generalised Estimation Equations

- Finally, let's consider a marginal model with the interaction between sex and time

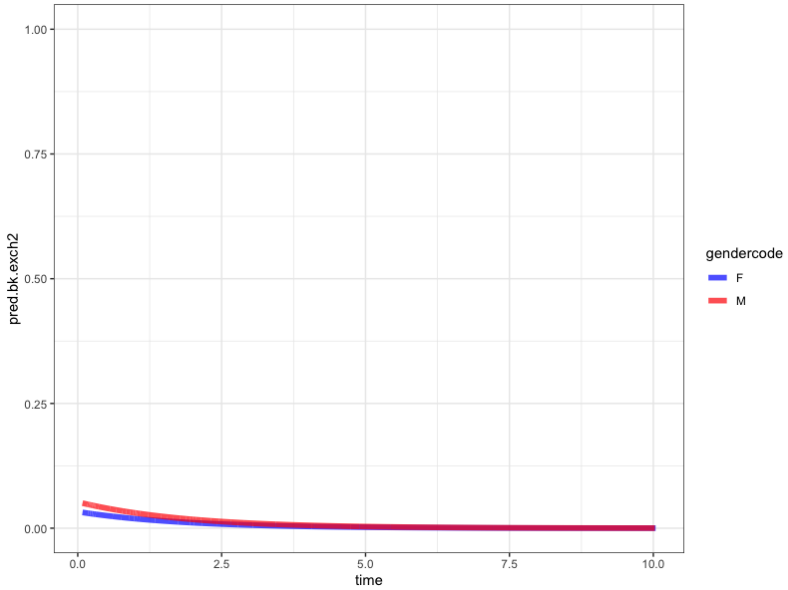
```
gee.bk.exch3 <- geeglm(bk ~ time * gendercode,
                       family=binomial,
                       id=ID,
                       corst="exchangeable",
                       data=bk.Data)
   summary(gee.bk.exch3)
```

```
 Coefficients:
                Estimate Std.err   Wald Pr(>|W|)
(Intercept)       -3.483   0.225 239.33   <2e-16 ***
time              -0.446   0.166   7.24   0.0071 **
gendercodeM        0.641   0.271   5.61   0.0179 *
time:gendercodeM  -0.175   0.207   0.71   0.3982
---
> exp(-.446) #Female
[1] 0.64
> exp(-.446 -.175) #Male
[1] 0.537
```

Generalised Estimation Equations

- Finally, let's consider a marginal model with the interaction between sex and time

```
 Coefficients:
                 Estimate Std.err   Wald Pr(>|W|)
(Intercept)        -3.483   0.225 239.33   <2e-16 ***
time               -0.446   0.166   7.24   0.0071 **
gendercodeM         0.641   0.271   5.61   0.0179 *
time:gendercodeM   -0.175   0.207   0.71   0.3982

> exp(-.446) #Female
[1] 0.64
> exp(-.446 -.175) #Male
[1] 0.537
```



Generalised Linear Mixed Models

- And the same analysis comparing the risk of BK for men and women but using a random intercept model

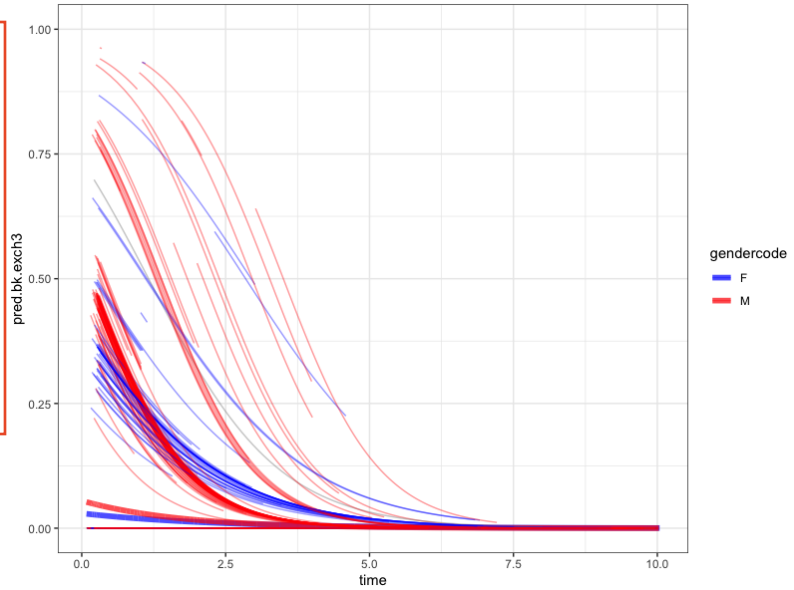
```
rInt.bk3 <- glmer(bk ~ time * gendercode + (1|ID),
                  family=binomial,
                  data=bk.Data)
```

```
Fixed effects:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       -9.082      0.603  -15.07  < 2e-16 ***
time              -0.710      0.181   -3.93  8.5e-05 ***
gendercodeM        1.194      0.550    2.17    0.030 *
time:gendercodeM  -0.489      0.240   -2.03    0.042 *

> exp(-.710)            #FEMALES
[1] 0.492
> exp(-.710-.489)       #MALES
[1] 0.301
```

Generalised Linear Mixed Models

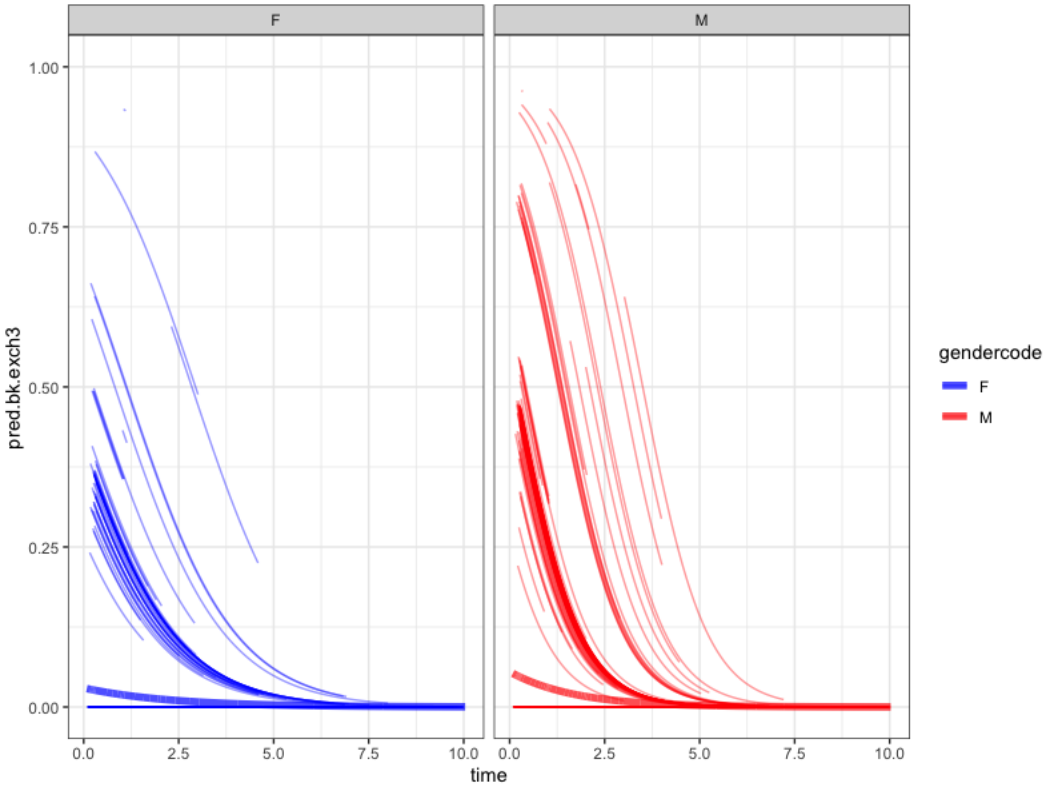
- And the same analysis comparing the risk of BK for men and women but using a random intercept model

```
Fixed effects:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)        -9.082      0.603  -15.07  < 2e-16 ***
time               -0.710      0.181   -3.93  8.5e-05 ***
gendercodeM         1.194      0.550    2.17    0.030 *
time:gendercodeM   -0.489      0.240   -2.03    0.042 *

> exp(-.710)            #FEMALES
[1] 0.492
> exp(-.710-.489)       #MALES
[1] 0.301
```



Generalised Linear Mixed Models

- We can see that the odds of BK decreases with time
- Men start at a higher risk (odds)
- But the risk drops faster than in women



Generalised Linear Mixed Models

- It is not uncommon for the GLMM (and even the LMM) not to converge
- There are multiple reasons for this and it can really be a difficult problem
- Many times, changing the numerical method (”optimiser” methods) solves the problem
- https://rstudio-pubs-static.s3.amazonaws.com/33653_57fc7b8e5d484c909b615d8633c01d51.html

Final thoughts

- GLS (repeated measures) – "old fashin" although useful for well structured data
- Mixed models put emphasis in the longitudinal nature of the data and treat it as an important feature of the data
- GEE treats the correlation in the data as nuisance
- Loss to follow-up might be associated with the trend in previous observations. This would mean that the missingness is at random and not completely at random. GEE would not be appropriate for those cases

Final thoughts

- GEE are designed for many clusters
- Due to robust standard errors, the choice of correlation structure is not very important
- Choosing the correct correlation structure is more efficient, i.e. leads to smaller SEs and tighter CIs
- In the linear case (continuous outcomes), the fixed part of mixed effects models and marginal models tend to be similar.
- Not the case for binary outcomes

Final thoughts

- The GEE and mixed models estimate different effects.
- In the linear model, this is not an issue but in the logistic or Poisson case, the two approaches are not comparable
- When we the interest is in estimating average risk, the GEE is a better choice
- On the other hand, if the interest is in the individual risk, mixed models would be the common approach