Module 3, part II: Generalized Linear Mixed Models

BIOS 526

Concepts

- Additional info on logistic regression
- Logistic and log-linear model for longitudinal data
- Conditional versus population effect estimates.

Reading

- You may find the following reference useful, specifically, the glmer() examples: Bolker, Ben. "GLMM Worked Examples." https://bbolker.github.io/mixedmodels-misc/ecostats_chap.html
- Sections 3.4-3.6 in Simon Wood, Generalized Additive Models, 2017, contains some information on glmms.

Example: 2×2 Crossover Trial

Data were obtained from a crossover trial on the disease cerebrovascular deficiency. The goal is to investigate the side effects of a treatment drug compared to a placebo.

Design:

- 34 patients: an active drug (A) and followed by a placebo (B)
- 33 patients: a placebo (B) and followed by an active drug (A).
- Outcome: normal (0) or abnormal (1) electrocardiogram.
- Each patient has a binary observation at period 1 and period 2
- Crossover design: can have "carryover" effects which confound treatment effect estimation. Test whether washout period was adequate.

Example: 2×2 Crossover Trial

Data:

```
> dat[1:5,]
  ID group period trt outcome
1  1     1      0   0       0
2  1     1      1   1       0
3  2     1      0   0       0
4  2     1      1   1       0
5  3     1      0   0       0
```

- ID i: subject id
- period j: 0 = period 1; 1 = period 2
- group : 0 = B then A; 1 = A then B
- outcome y_{ij}: 0 = normal ECG response; 1 = abnormal ECG response
- trt: 0 = placebo; 1 = active drug

More logit

Consider the following logistic model assuming responses within each subject are independent.

Model 1:

$$\text{logit } P(y_{ij} = 1) = \beta_0 + \beta_1 trt_{ij}$$

Model 2:

$$\text{logit } P(y_{ij} = 1) = \beta_0 + \beta_1 trt_{ij} + \beta_2 period_{ij}$$

Model 3:

$$\text{logit } P(y_{ij} = 1) = \beta_0 + \beta_1 trt_{ij} + \beta_2 period_{ij} + \beta_3 trt_{ij} * period_{ij}$$

- β_1: active drug versus placebo effects. (Note: in Model 3, active versus placebo for period 1.)
- β_2: second period versus first period effect
- β_3: carry-over effect. Does the effect of period differ between having the active drug during the second period versus having the active drug during the first period.

More logit review

Covariate	Model 1	Model 2	Model 3
Intercept β_0	-1.08 (0.28)	-1.22 (0.34)	-1.54 (0.45)
Treatment β_1	0.56 (0.38)	0.56 (0.38)	1.11 (0.57)
Period β_2		0.27 (0.38)	0.85 (0.58)
Treatment $\times$ Period β_3			-1.02 (0.77)

- Model 3: after controlling for period and carry-over effects, the estimated OR of abnormal ECG in period 1 was $3.03 = e^{1.11}$ and p-value $= 0.053$. The 95% confidence interval is

$$(e^{1.11-1.96*0.57}, e^{1.11+1.96*0.57}) = (0.99, 9.27).$$

- At $\alpha = 0.05$, we fail to reject the null hypothesis that the treatment in period has an impact on the probability of an abnormal ECG. However, future research is needed since the p-value is 0.053.
- β_3 is negative - the second period effect is smaller for those who received active drug during the second period; however, not significant.

2×2 Crossover Trial

Given the estimates in Model 3, calculate predicted probabilities:

$$\text{logit } P(y_{ij} = 1) = \beta_0 + \beta_1 trt_{ij} + \beta_2 period_{ij} + \beta_3 trt_{ij} * period_{ij}$$

For the treatment-placebo group:

$$\text{P}(outcome = 1 \mid period = 1,\ treat = 1) = \left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right) = 0.394$$

$$\text{P}(outcome = 1 \mid period = 2,\ treat = 0) = \left(\frac{e^{\beta_0+\beta_2}}{1+e^{\beta_0+\beta_2}}\right) = 0.333$$

For the placebo-treatment group:

$$\text{P}(outcome = 1 \mid period = 1,\ treatment = 0) = \left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right) = 0.176$$

$$\text{P}(outcome = 1 \mid period = 2,\ treatment = 1) = \left(\frac{e^{\beta_0+\beta_1+\beta_2+\beta_3}}{1+e^{\beta_0+\beta_1+\beta_2+\beta_3}}\right) = 0.353$$

Independence assumption

What's wrong with this model?

It assumes observations are independent:

$$L(\boldsymbol{\beta}; \mathbf{y}, \boldsymbol{x}) = \prod_{i=1}^{n} \prod_{j=1}^{r_i} p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}$$

Generalized Linear Mixed Model

We will now extend the generalized linear model framework to analyze clustered binary data.

Let index $i = 1, \ldots, n$ denote group ID, $j = 1, \ldots, r_i$ denote observation within group i, $N = \sum_{i=1}^{n} r_i$.

Consider the random-intercept logistic regression model:

$$y_{ij} \sim \text{Binomial } (p_{ij})$$

$$\text{logit } (p_{ij}) = \beta_0 + \theta_i + \boldsymbol{x}_{ij}'\boldsymbol{\beta}$$

$$\theta_i \overset{iid}{\sim} N(0, \tau^2)$$

- β_0 is the overall baseline log odds.
- θ_i is the difference between group-specific baseline log odds and β_0.
- $\boldsymbol{x}_{ij}$ is the $p \times 1$ vector of covariates and $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients.
- τ^2 is the variation of baseline log odds between groups (e.g., each group is an individual).

Likelihood in GLMM

Consider a logistic regression with random intercept:

$$y_{ij} \sim \text{Binomial } (p_{ij})$$

$$\text{logit } (p_{ij}) = \beta_0 + \theta_i + \boldsymbol{x}'_{ij}\boldsymbol{\beta} \qquad \theta_i \overset{iid}{\sim} N(0, \tau^2).$$

Let $\mathbf{y}$ be all the data and $\boldsymbol{\theta}$ the vector of all random effects. Let $[\mathbf{y}, \boldsymbol{\theta}]$ denote their joint density.

$$\begin{aligned}
[\mathbf{y}, \boldsymbol{\theta}] &= \prod_{i=1}^{n} [\mathbf{y}_i, \theta_i] \quad (\mathbf{y}_i \text{ independent}) \\
&= \prod_{i=1}^{n} [\mathbf{y}_i | \theta_i][\theta_i] \\
&= \prod_{i=1}^{n} \left(\prod_{j=1}^{r_i} [y_{ij} | \theta_i][\theta_i] \right) \quad (y_{ij} \text{ conditionally independent}) \\
&= \prod_{i=1}^{n} \left(\prod_{j=1}^{r_i} [y_{ij} | \theta_i] \right) (2\pi\tau^2)^{-1/2} \exp\left(-\frac{1}{2\tau^2} \theta_i^2 \right)
\end{aligned}$$

The y_{ij} are conditionally independent given the random effects.

Likelihood in GLMM Estimation

We define the likelihood for the fixed parameters. Integrating each θ_i, the likelihood is

$$L(\boldsymbol{\beta}, \beta_0, \tau^2 \,|\, \mathbf{y}) = \prod_{i=1}^{n} \int \prod_{j=1}^{r_i} [y_{ij}|\theta_i] \times (2\pi\tau^2)^{-1/2} e^{\left(-\frac{1}{2\tau^2}\theta_i^2\right)} d\theta_i$$

Likelihood in GLMM Estimation

For Bernoulli outcome, the data likelihood for group i is

$$\prod_{j=1}^{r_i}[y_{ij}|\theta_i] = \prod_{j=1}^{r_i} p_{ij}^{y_{ij}} \times (1-p_{ij})^{1-y_{ij}}$$

$$= \prod_{j=1}^{r_i} \left(\frac{e^{\beta_0+\theta_i+\boldsymbol{x}'_{ij}\boldsymbol{\beta}}}{1+e^{\beta_0+\theta_i+\boldsymbol{x}'_{ij}\boldsymbol{\beta}}} \right)^{y_{ij}} \times \left(\frac{1}{1+e^{\beta_0+\theta_i+\boldsymbol{x}'_{ij}\boldsymbol{\beta}}} \right)^{1-y_{ij}}$$

Therefore the likelihood is

$$L(\boldsymbol{\beta}, \beta_0, \tau^2 \,|\, \mathbf{y}) = \prod_{i=1}^{n} \int \prod_{j=1}^{r_i} \left(\frac{e^{\beta_0+\theta_i+\boldsymbol{x}'_{ij}\boldsymbol{\beta}}}{1+e^{\beta_0+\theta_i+\boldsymbol{x}'_{ij}\boldsymbol{\beta}}} \right)^{y_{ij}} \times \left(\frac{1}{1+e^{\beta_0+\theta_i+\boldsymbol{x}'_{ij}\boldsymbol{\beta}}} \right)^{1-y_{ij}}$$
$$\times (2\pi\tau^2)^{-1/2} e^{\left(-\frac{1}{2\tau^2}\theta_i^2\right)} d\theta_i$$

Finding the MLE

Note that in the Gaussian case, we can similarly specify a model using conditional independence.

There, we can easily evaluate the integral and obtain a nice form for the multivariate normal distribution.

The covariance matrix nicely captures dependence via the block diagonal structure.

In GLMMs, we are stuck with an integral. Trickier optimization.

Finding the MLE

$$L(\boldsymbol{\beta}, \beta_0, \tau^2 \,|\, \mathbf{y}) = \prod_{i=1}^{n} \int \prod_{j=1}^{r_i} \left(\frac{e^{\beta_0 + \theta_i + \boldsymbol{x}'_{ij}\boldsymbol{\beta}}}{1 + e^{\beta_0 + \theta_i + \boldsymbol{x}'_{ij}\boldsymbol{\beta}}} \right)^{y_{ij}} \times \left(\frac{1}{1 + e^{\beta_0 + \theta_i + \boldsymbol{x}'_{ij}\boldsymbol{\beta}}} \right)^{1-y_{ij}} \times (2\pi\tau^2)^{-1/2} e^{\left(-\frac{1}{2\tau^2}\theta_i^2\right)} d\theta_i$$

Because of our non-linear link function, maximizing the above function that involves an integral is quite challenging.

Statistical software performs numerical integration that involves some approximation. Convergence issues are common in glmms.

Example: 2×2 Crossover Trial

Using the crossover trial, we now model subject specific random baseline odds:

Model 4:

$$\text{logit } P(y_{ij} = 1) = \beta_0 + \theta_i + \beta_1 trt_{ij} + \beta_2 period_{ij} + \beta_3 trt_{ij} * period_{ij}$$

$$\theta_i \overset{iid}{\sim} N(0, \tau^2).$$

Fitting GLMMs

The random intercept logistic model can be fit using the `glmer ( )` function with the binomial family:

```
> fit4 = glmer (outcome~trt*period+(1|ID), family=binomial(link='logit'),

data = cbv)
> summary(fit4)

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod

Random effects:
 Groups Name        Variance Std.Dev.
 ID     (Intercept) 551      23.47
Number of obs: 134, groups:  ID, 67

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -18.925      3.500  -5.407 6.39e-08 ***
trt            9.988      3.102   3.220  0.00128 **
period         8.214      3.219   2.551  0.01073 *
trt:period    -8.234      4.577  -1.799  0.07205 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note on GLMM Estimation

The *glmer ( )* function has an *nAGQ* option:
nAGQ integer scalar - the number of points per axis for evaluating the adaptive Gauss-Hermite approximation to the log-likelihood. Defaults to 1, corresponding to the Laplace approximation. Values greater than 1 produce greater accuracy in the evaluation of the log-likelihood at the expense of speed. A value of zero uses a faster but less exact form of parameter estimation for GLMMs by optimizing the random effects and the fixed-effects coefficients in the penalized iteratively reweighted least squares step. (See Details.)

Even if your model converges, it's often a good idea to increase the numerical integration accuracy and see whether the estimates are robust.

Refit with nAGQ=2

```
> fit5 = glmer (outcome~trt*period+(1|ID), family=binomial(link='logit'),

data = cbv, nAGQ = 2)
> summary(fit5)
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature
  nAGQ = 2) [glmerMod]

Random effects:
 Groups Name        Variance Std.Dev.
 ID     (Intercept) 7.538    2.746
Number of obs: 134, groups:  ID, 67

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.313      1.076  -3.078  0.00208 **
trt            2.384      1.233   1.933  0.05326 .
period         1.780      1.194   1.490  0.13615
trt:period    -2.173      1.937  -1.122  0.26199
```

Convergence issues

Note the differences in the description of the optimizers.

p.148 in Wood GAMs book says Laplace approximation should not be used if ≤ 3 observations per subject.

Note estimate of τ^2 in `fit4` exploded.
Also note the intercept estimate with Laplace approximation is very negative, $\frac{e^{-18.925}}{1+e^{-18.925}} = 6.0e-09$, an extremely small probability that leads to numerical instability.

Some statistical programs will provide warnings, but also give results. DO NOT use them. Different programs can give different results.

Different versions of `glmer()` may give different results.

Number of quadrature points

```
> fit6 = glmer (outcome~trt*period+(1|ID), family=binomial(link='logit'),

data = cbv, nAGQ = 25)
> summary(fit6)
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature
  nAGQ = 25) [glmerMod]
 Family: binomial  ( logit )
Formula: outcome ~ trt * period + (1 | ID)
   Data: cbv

     AIC      BIC   logLik deviance df.resid
   145.1    159.6    -67.5    135.1      129

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.1399 -0.2140 -0.1462  0.2435  1.3149

Random effects:
 Groups Name        Variance Std.Dev.
 ID     (Intercept) 24.4     4.94
Number of obs: 134, groups:  ID, 67

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.004      2.176  -2.299   0.0215 *
trt            3.595      2.140   1.680   0.0929 .
period         2.786      2.042   1.364   0.1726
trt:period    -3.338      3.303  -1.011   0.3122
```

Example: 2×2 Crossover Trial

```
> fit7 = glmer (outcome~trt*period+(1|ID), family=binomial(link='logit'),

data = cbv, nAGQ = 100)
> summary(fit7)
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature
  nAGQ = 100) [glmerMod]

Random effects:
 Groups Name        Variance Std.Dev.
 ID     (Intercept) 24.15    4.915
Number of obs: 134, groups:  ID, 67

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.981      2.116  -2.354   0.0186 *
trt            3.578      2.107   1.698   0.0895 .
period         2.772      2.015   1.376   0.1690
trt:period    -3.319      3.270  -1.015   0.3101
```

Interpretations

Covariate	GLMM	GLM
Intercept β_0	-4.98 (2.12)	-1.54 (0.45)
Treatment β_1	3.58 (2.11)	1.11 (0.57)
Period β_2	2.78 (2.02)	0.85 (0.58)
Treatment $\times$ Period β_3	-3.32 (3.27)	-1.02 (0.77)
τ^2	4.92^2	

- The point estimates from the random intercept model are larger. However, the standard errors also increased such that inference on direction and significance remain the same.
- The baseline (period 1, placebo) log odds across subjects has a population mean of -4.98 and a standard deviation of 2.12. The middle 95% of subjects have baseline log odds between

$$-4.98 \pm 1.96 \times 4.92 = (-11.6, 2.9)$$

or a baseline probabilities of $(9 \times 10^{-6}, 0.95)$. Very large between-subject heterogeneity!

Population versus Conditional Interpretations

The GLM is estimating the marginal model (integrating out the RE):

$$g(E\left[y_{ij}\right]) = \boldsymbol{\beta}'\boldsymbol{x}_{ij}$$

This is known as the population-averaged effect or marginal effect.

The GLMM is estimating the slopes conditioned on the random effects:

$$g(E\left[y_{ij}|\theta_i\right]) = \boldsymbol{\beta}'\boldsymbol{x}_{ij} + \theta_i$$

These slopes are estimated controlling for subject effects, which are called conditional effects.

The two approaches are estimating different slopes.

Note: the GLM likelihood assumes independence, resulting in incorrect SE. Later in the course, we will see how to make marginal inference accounting for within-group correlation using generalized estimating equations (GEE).

Population versus Conditional Interpretations

We are modeling transformations of the expectations:

$$E[y_{ij}|\theta_i] = g^{-1}(\beta_0 + \theta_i + \sum_{k=1}^{p} \beta_k x_{ijk}).$$

For Gaussian, $g()$ is the identity function, so the slopes in the marginal model (integrating out RE) have the same interpretation as the conditional model:

$$E[y_{ij}] = E(E[y_{ij}|\theta_i]) = E(\beta_0 + \theta_i + \sum_{k=1}^{p} \beta_k x_{ijk}) = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ijk}.$$

But for GLMMs, we have

$$\begin{aligned} E[y_{ij}] &= E(E[y_{ij}|\theta_i]) \\ &= E\left\{ g^{-1}(\beta_0 + \theta_i + \sum_{k=1}^{p} \beta_k x_{ijk}) \right\} \neq g^{-1}(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ki}). \end{aligned}$$

Population versus Conditional Interpretations

Covariate	GLMM	GLM
Treatment (Period 1) $\exp \beta_1$	exp(3.58)=35.9	exp(1.11)=3.03

Here the OR from conditional inference is about 12 times larger than that from marginal inference. The CI are (0.57,2243) and (0.99, 9.27),

respectively. (Note also the GLM CI is incorrect due to violations of independence.)

Clearly we have a lot of uncertainty in the models.

To gain some insight into the marginal versus conditional models, see the simulated mixed model in the R code.

Poisson Regression: Modeling Cancer Incidence

Let s index one of the 88 counties in Ohio, t index year, and k index a population sex-race stratum.

Variables:

- $death_{stk}$: stratified lung cancer death counts for population k in county s during year t.
- sex_k: 1 = female; 0 = male.
- $race_k$: 1 = white; 0 = nonwhite.
- $year_t$: 1, 2, ..., 9 for year 1980 till 1988.
- pop_{stk}: at risk population size.

Questions:

- What were the associations between lung cancer death counts and sex/race.
- Estimate the between-county variation in lung cancer risks.

Ohio Cancer Surveillance Data

```
> dat[1:20,]
   county sex race year death   pop
1       1   1    1    1    11 12006
2       1   1    1    2     7 12142
3       1   1    1    3    12 12085
4       1   1    1    4     7 11944
5       1   1    1    5     9 11875
6       1   1    1    6    15 11915
7       1   1    1    7     9 12074
8       1   1    1    8    12 12325
9       1   1    1    9    13 12443
10      1   1    0    1     0    51
11      1   1    0    2     0    52
12      1   1    0    3     0    70
13      1   1    0    4     0    84
14      1   1    0    5     0    89
15      1   1    0    6     0   100
16      1   1    0    7     0   104
17      1   1    0    8     0   111
18      1   1    0    9     0   120
19      1   2    1    1     3 12196
20      1   2    1    2     4 12409
```

Poisson Regression: Modeling Cancer Incidence

Consider the following random-intercept Poisson model where we treat all stratified death counts within the same county as a group.

$$death_{stk} \sim \text{ Poisson}(\lambda_{stk})$$

$$\log \lambda_{stk} = \beta_0 + \theta_s + \beta_1 sex_k + \beta_2 race_k + \beta_3 sex_k \times race_k + \beta_4 year_t$$

$$\theta_s \overset{iid}{\sim} N(0, \tau^2)$$

- β_0 = log expected lung cancer death counts at baseline (non-white males in *1979*) for the average county.
- θ_s = county-specific deviation in baseline log expected lung cancer counties.
- e^{β_1} = relative rate of lung cancer deaths for female compared to male.
- e^{β_2} = relative rate of lung cancer deaths for white compared to non-white.
- e^{β_3} = relative rate change in lung cancer deaths for white females
- e^{β_4} = relative rate of lung cancer deaths increase per year.

Poisson Regression: Modeling Cancer Incidence

```
> fit = glmer (death~sex*race+year + (1|county), family = poisson, data = cancer)
1> summary (fit)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod
 Family: poisson  ( log )
Formula: death ~ sex * race + year + (1 | county)
   Data: cancer

     AIC      BIC   logLik deviance df.resid
 16418.7  16455.0  -8203.3  16406.7     3162

Scaled residuals:
    Min      1Q  Median      3Q     Max
-7.0824 -1.0647 -0.6006  0.4144 11.6463

Random effects:
 Groups Name        Variance Std.Dev.
 county (Intercept) 1.067    1.033
Number of obs: 3168, groups:  county, 88

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.860181   0.111647   7.704 1.31e-14 ***
sex         -0.979603   0.029002 -33.778  < 2e-16 ***
race         2.036423   0.016111 126.402  < 2e-16 ***
year         0.022249   0.001667  13.349  < 2e-16 ***
sex:race     0.161888   0.030642   5.283 1.27e-07 ***
```

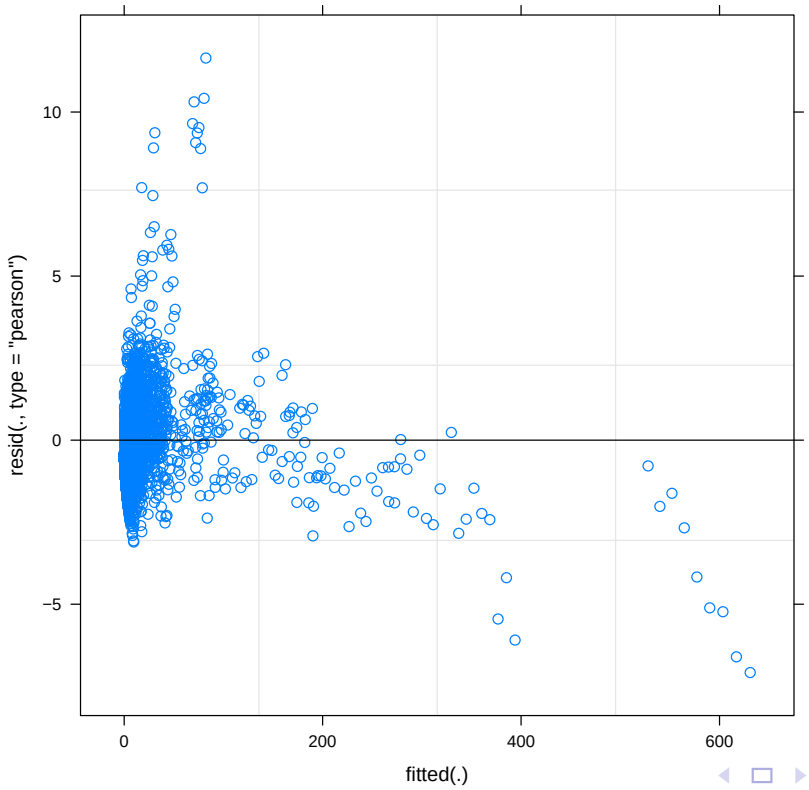
Checking convergence

```
 > fit.check = glmer (death~sex*race+year + (1|county), family = poisson, data = cancer,nA
> summary (fit.check)
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature
  nAGQ = 25) [glmerMod]
 Family: poisson  ( log )
Formula: death ~ sex * race + year + (1 | county)
   Data: cancer

     AIC      BIC   logLik deviance df.resid
  7781.2   7817.6  -3884.6   7769.2     3162

Scaled residuals:
    Min      1Q  Median      3Q     Max
-7.0824 -1.0647 -0.6006  0.4144 11.6462

Random effects:
 Groups Name        Variance Std.Dev.
 county (Intercept) 1.067    1.033
Number of obs: 3168, groups:  county, 88

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.860183   0.111652   7.704 1.32e-14 ***
sex         -0.979603   0.029013 -33.764  < 2e-16 ***
race         2.036422   0.016117 126.351  < 2e-16 ***
year         0.022250   0.001667  13.343  < 2e-16 ***
sex:race     0.161889   0.030654   5.281 1.28e-07 ***
```

Checking Goodness of Fit

Approximate test for overdispersion
https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#overdispersion

```
> overdisp_fun(fit)
        chisq         ratio           rdf             p
 6.117391e+03  1.934659e+00  3.162000e+03 2.492817e-191
```



Regression Coefficient Interpretations

- Note: we will fit a better model next, but the following provides information about interpretation.
- The baseline expected count was $e^{0.86} = 2.36$ cases for non-white males in 1979 in an typical county.
- There exists considerable heterogeneity in baseline counts with a between-county standard deviation of 1.03. So 95% of the counties have baseline counts between $e^{0.86 \pm 1.96 \times 1.03} = (0.3, 17.8)$
- There is evidence that lung cancer rate was increasing by $e^{0.022} = 1.022$ (approximately 2.2%) per year.
- We found that when conditioning on county effects and controlling for year, cancer rates were higher in males compared to females, and higher in the white population compared to non-white.
- The expected lung cancer death count for white females in a typical county in 1980 is $e^{0.860 - 0.979 + 0.022 * 1} = 0.907$.

Compare to the GLM

- The marginal versus conditional interpretation impacts the intercept in Poisson.
- The marginal model estimates $\beta_{0*} = \tau^2/2 + \beta_0$, where β_0 is the intercept in the conditional model. See R Code.
- Slopes are comparable (the SEs in the GLM are wrong).

```
> fit.poisson.glm = glm(death~sex*race+year,family=poisson,data=cancer)
> summary(fit.poisson.glm)

Call:
glm(formula = death ~ sex * race + year, family = poisson, data = cancer)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-9.481  -3.463  -2.260  -1.401  41.151

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.590932   0.017420  91.326  < 2e-16 ***
sex         -0.979605   0.029013 -33.764  < 2e-16 ***
race         2.036422   0.016117 126.352  < 2e-16 ***
year         0.022249   0.001667  13.343  < 2e-16 ***
sex:race     0.161890   0.030654   5.281 1.28e-07 ***
---
```

Poisson Regression: Modeling Cancer Incidence

Consider an alternative random-intercept Poisson model where we incorporate the population size.

$$y_{stk} \sim \text{Poisson}(\lambda_{stk})$$

$$\log \lambda_{stk} = \log pop_{stk} + \beta_0 + \theta_s + \beta_1 sex_k + \beta_2 race_k + \beta_3 sex_k \times race_k + \beta_4 year_t$$

$$\theta_s \overset{iid}{\sim} N(0, \tau^2)$$

- We assume the coefficient on $\log pop_{stk}$ is 1. This is known as an offset variable.

$$\begin{aligned}\lambda_{stk} &= e^{\log pop_{stk} + \beta_0 + \theta_s + \beta_1 sex_k + \beta_2 race_k + \beta_3 sex_k \times race_k + \beta_4 year_t} \\ &= pop_{stk} \times e^{\beta_0 + \theta_s + \beta_1 sex_k + \beta_2 race_k + \beta_3 sex_k \times race_k + \beta_4 year_t}\end{aligned}$$

$$\lambda_{stk}/pop_{stk} = e^{\beta_0 + \theta_s + \beta_1 sex_k + \beta_2 race_k + \beta_3 sex_k \times race_k + \beta_4 year_t}$$

Here e^{β_0} is interpreted is the baseline per capita deaths, instead of the expected counts (for a non-white male in year 1979 conditioning on county).

Note on offset

Consider the simple model:

$$\begin{aligned} \log \lambda_i &= \beta_0 + \log pop_i \\ \log \lambda_i - \log pop_i &= \beta_0 \\ \log(\lambda_i / pop_i) &= \beta_0 \\ \lambda_i / pop_i &= e^{\beta_0} \end{aligned}$$

e^{β_0} is the fraction of deaths per person, i.e., per capita death rate.

Poisson Regression: Modeling Cancer Incidence

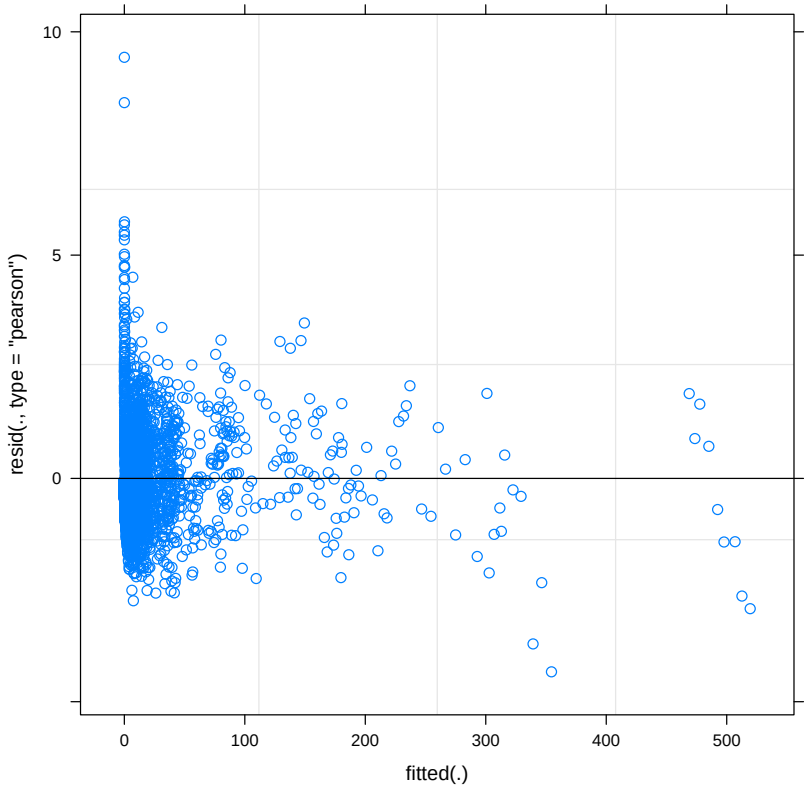
```
> cancer$logpop = log (cancer$pop)
> fit = glmer (death~offset(logpop) + sex*race+year + (1|county), family = poisson, data =
> summary (fit)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod
 Family: poisson  ( log )
Formula: death ~ offset(logpop) + sex * race + year + (1 | county)
   Data: cancer

     AIC      BIC   logLik deviance df.resid
 11932.5  11968.9  -5960.3  11920.5     3162

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.3306 -0.5816 -0.2218  0.4209  9.4296

Random effects:
 Groups Name        Variance Std.Dev.
 county (Intercept) 0.03905  0.1976
Number of obs: 3168, groups:  county, 88

Fixed effects:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -7.361892   0.028682 -256.670  < 2e-16 ***
sex         -1.103919   0.029011  -38.051  < 2e-16 ***
race         0.029238   0.016512    1.771   0.0766 .
year         0.022775   0.001666   13.672  < 2e-16 ***
sex:race     0.219027   0.030651    7.146 8.95e-13 ***
```

Check convergence

```
   >  fit.check = glmer (death~offset(logpop) + sex*race+year + (1|county), family = poisso
>   summary(fit.check)
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature
  nAGQ = 25) [glmerMod]
 Family: poisson  ( log )
Formula: death ~ offset(logpop) + sex * race + year + (1 | county)
   Data: cancer

     AIC      BIC   logLik deviance df.resid
  3295.1   3331.5  -1641.6   3283.1     3162

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.3306 -0.5816 -0.2218  0.4209  9.4297

Random effects:
 Groups Name        Variance Std.Dev.
 county (Intercept) 0.03906  0.1976
Number of obs: 3168, groups:  county, 88

Fixed effects:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -7.361893   0.028689 -256.607  < 2e-16 ***
sex         -1.103915   0.029015  -38.047  < 2e-16 ***
race         0.029239   0.016514    1.771   0.0766 .
year         0.022774   0.001666   13.667  < 2e-16 ***
sex:race     0.219024   0.030655    7.145 9.01e-13 ***
```

Goodness of fit

```
> overdisp_fun(fit)
       chisq        ratio          rdf            p
3.383166e+03 1.069945e+00 3.162000e+03 3.194105e-03
```



Poisson Regression: Modeling Cancer Incidence

	With Population Offset	
Coef Estimates	No	Yes
Intercept β_0	0.86	-7.36
sex β_1	-0.98	-1.10
race β_2	2.04	0.03
sex $\times$ race β_3	0.162	0.219
year β_4	0.022	0.023
τ^2	1.03^2	0.198^2

- With population offset, β_0 becomes extremely small. It reflects the baseline (male, non-white, year 1979) rates ($e^{-7.36} = 0.0006$).
- The coefficient for race dropped considerably! This is because the high number of deaths seen in the white population is accounted for by the larger white population counts (89% of the total pop).

Random effects in GLMMs