Poisson Regression: Mixed-Effects Model

2019-04-22

In this set of notes, you will learn about generalized mixed-effects models and how we can extend the Poisson regression model to account for non-independence.

Dataset and Research Question

In this set of notes, we will use data from the file *rapi.csv*. These data are from Atkins, Baldwin, Zheng, Gallop, & Neighbors (2013) and are:

> drawn from an intervention study aimed at reducing problematic drinking in college students (Neighbors et al., 2010). The current paper focuses on gender differences across two years in alcohol-related problems, as measured by the Rutgers Alcohol Problem Index (RAPI). The dataset includes 3,616 repeated measures across five time points from 818 individuals (p. 167).

```
# Load libraries
library(AICcmodavg)
library(broom)
library(corrr)
library(dplyr)
library(ggplot2)
library(glmmTMB)
library(gridExtra)
library(lme4)
library(readr)
library(sm)
library(tidyr)

# Read in data
rapi = read_csv(file = "~/Documents/github/epsy-8252/data/rapi.csv")

# View data
head(rapi)
```

```
# A tibble: 6 x 4
     id problems month  male
  <dbl>    <dbl> <dbl> <dbl>
1     1        0     0     1
2     1        0     6     1
3     1        0    18     1
4     2        3     0     0
5     2        6     6     0
6     2        5    12     0
```

The variables include:

- `id`: The ID for the college student

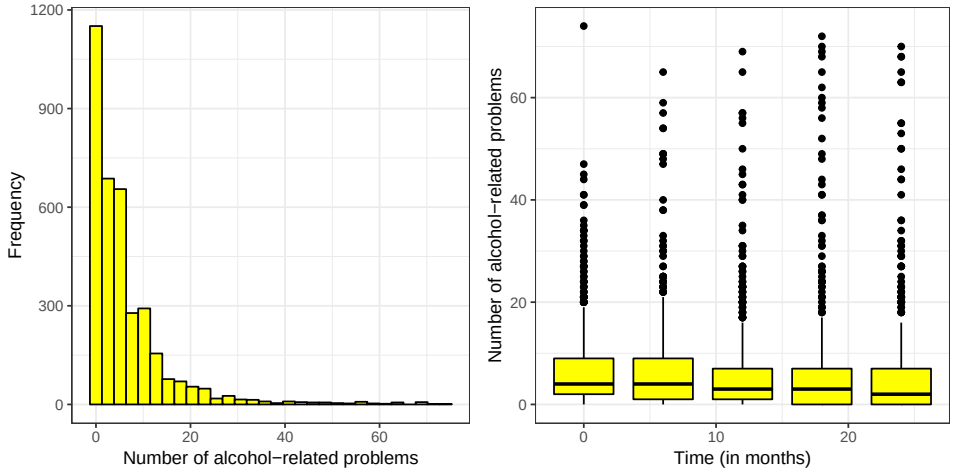


Figure 1: Histogram showing the number of alcohol-related problems over the entire study (left). Box-and-whiskers plots showing the distribution of alcohol-related problems at different points in time during the study (right).

- `problems`: The RAPI score indicating the number of self-reported alcohol-related problems for the student at a given time point.
- `month`: Time in months since the introduction of intervention (0, 6, 12, 18, 24)
- `male`: Dummy coded gender variable (0 = Female; 1 = Male)

We will use these data to explorethe longitudinal variation in alcohol-related problems.

Exploration

We begin by exploring both the marginal distribution of RAPI scores and the conditional distributions of RAPI scores over time.

```
# Plot of marginal distribution
p1 = ggplot(data = rapi, aes(x = problems)) +
  geom_histogram(fill = "yellow", color = "black") +
  theme_bw() +
  xlab("Number of alcohol-related problems") +
  ylab("Frequency")

# Plot of conditional distributions
p2 = ggplot(data = rapi, aes(x = month, y = problems, group = month)) +
  geom_boxplot(fill = "yellow", color = "black") +
  theme_bw() +
  xlab("Time (in months)") +
  ylab("Number of alcohol-related problems")

# Create side-by-side plots; need gridExtra package
grid.arrange(p1, p2, ncol = 2)
```

We can also compute summary measures of the average number of alcohol-related problems over time and the variation at each time point.

```r
 # Conditional Mean, SD, and Variance by time point
rapi %>%
  group_by(month) %>%
  summarize(
    M = mean(problems),
    SD = sd(problems),
    Var = var(problems)
    )
```

```
# A tibble: 5 x 4
  month     M    SD   Var
  <dbl> <dbl> <dbl> <dbl>
1     0  6.91  7.75  60.0
2     6  6.51  8.55  73.1
3    12  5.98  8.92  79.5
4    18  6.26 10.3  107.
5    24  5.73 10.1  101.
```

Not surprisingly, the marginal and conditional distributions are right-skewed with several individuals reporting zero or low numbers of alcohol-related problems. Summary statistics reveal that while the average number of alcohol-related problems seems to be decreasing throughout the duration of the study, there is a great deal of variability in the number of alcohol-related problems being reported at each wave ofthe study.

We can plot the mean profile to get a sense of the functional form of the change curve.

```r
ggplot(data = rapi, aes(x = month, y = problems)) +
  stat_summary(aes(group = month), geom = "line", fun.y = mean, color = "black", group = 1) +
  stat_summary(aes(group = month), geom = "point", fun.y = mean, color = "black", group = 1) +
  theme_bw() +
  xlab("Time (in months)") +
  ylab("Average Number of alcohol-related problems")
```

This plot suggests that there is potentially some non-linearity in the change curve.

Individual Profiles

Below we plot the individual change profiles for a random sample of 20 students. The individual linear fitted regressions are also displayed.

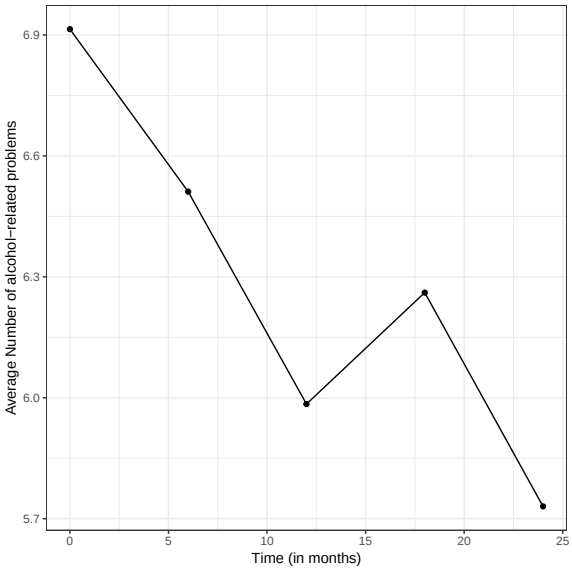
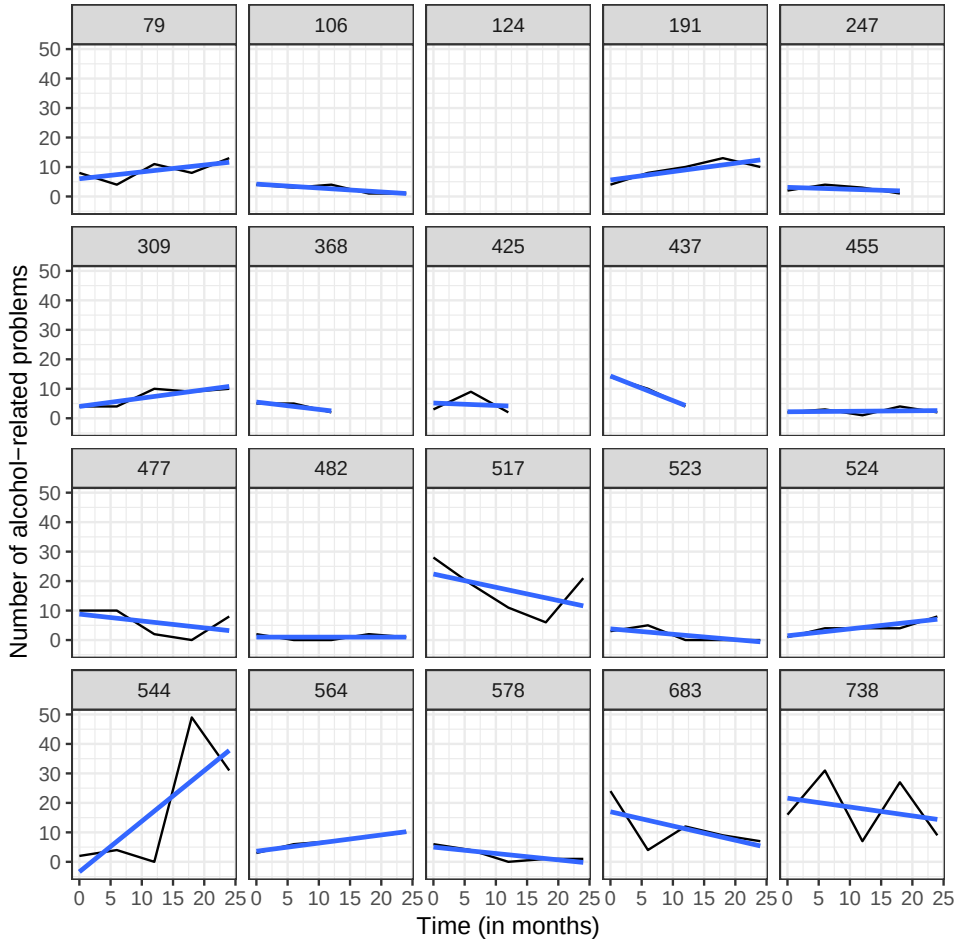


Figure 2: Mean profile showing the average Number of alcohol-related problems over time.



The plot shows:

- A general pattern of decline in problems over the course of the study
- Variation in individual participant's intercepts
- Variation in individual participant's change over time (slopes)

This suggests that we may want to adopt a model that not only includes random-effects of intercept, but also random-effects of slope.

Mixed-Effects Poisson Model

To begin, we will fit the unconditional random-intercepts model. To do this, we will use the `glmer()` function from the **lme4** package. We will also include a Poisson-distributed error struture with the log-link function.

```
glmer.0 = glmer(problems ~ 1 + (1|id), data = rapi, family = poisson(link = "log"))
summary(glmer.0)
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: poisson  ( log )
Formula: problems ~ 1 + (1 | id)
   Data: rapi

     AIC      BIC   logLik deviance df.resid
   24206    24219   -12101    24202     3614

Scaled residuals:
   Min     1Q Median     3Q    Max
-6.323 -1.103 -0.323  0.761 13.886

Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 1.07     1.04
Number of obs: 3616, groups:  id, 818

Fixed effects:
            Estimate Std. Error z value            Pr(>|z|)
(Intercept)   1.3639     0.0379      36 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The fitted equation is:

$$\ln(\mu_{ij}) = 1.36$$

Based on this output,

- The model-predicted average number of self-reported alcohol-related problems is 3.91 ($e^{1.36}$), conditional on the random-effects.

In GLMER models, the *conditional on the random-effects* part of the interpretation is rather important. In LMER models, we do not say this because the interpretation of a fixed-effect is considered to be "averaged over" the random-effects. (It is a marginal effect.) Consider the fitted LMER model:

$$\hat{Y}_{ij} = \hat{\beta}_0 + \hat{b}_{0j}$$

The random-effects are assumed to have a mean of 0. Thus, when we average across them, they are not contributing anything to the population-average model; the average random-effect of intercept is 0, so adding it to the fixed-effects part just gives the fixed-effects). Thus the interpretations of the fixed-effects are the population-average model.

This is not true for GLMER models. The link function changes everything,

$$\ln(\mu_{ij}) = \hat{\beta}_0 + \hat{b}_{0j}$$

The random-effects are still assumed to have a mean of 0, but only on the linear predictor scale; not on the original scale of Y. For that, we need to exponentiate the right-hand side of the equation. Now the random-effects are adding to the fixed-effects ($e^0 \neq 0$). Because of this, when we interpretat the fixed-effects, we need to do so *conditioned on the random-effects* included in the population-average model.

Variance Components and Random-Effects

Since the GLMER models use a link function which relates the mean to the set of linear predictors, there is no level-1 error term in the fitted equation for the link. This also means that there is not a variance component estimated for the level-1 residuals either. The only variance component estimated is those for any included random-effects.

```
Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 1.07     1.04
Number of obs: 3616, groups:  id, 818
```

You can access the estimated random-effects using the `ranef()` function, the same way we did for `lmer()` models.

```
head(ranef(glmer.0)$id)
```

```
  (Intercept)
1      -1.895
2       0.156
3      -0.146
4      -0.379
5       0.812
6      -1.494
```

We can use these to write the individual participants' fitted equations. For example, the fitted equation for Student 01 is:

$$\begin{aligned}\ln(\mu_{ij}) &= \big[1.36 - 1.90\big] \\ &= -0.54\end{aligned}$$

For Student 01, the estimated average number of alcohol-related problems reported is 0.58 ($e^{-0.54}$). For the individual interpretations we no longer need the *conditioned on the random-effects* part of the interpretation as we are not interpreting the population-average effects anymore.

Change Over Time?

To examine change over time, we also fit a model that further includes the fixed-effect of time (`month`).

```
glmer.1 = glmer(problems ~ 1 + month + (1|id), data = rapi, family = poisson(link = "log"))
summary(glmer.1)
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: poisson  ( log )
Formula: problems ~ 1 + month + (1 | id)
   Data: rapi

     AIC      BIC   logLik deviance df.resid
   24149    24168   -12072    24143     3613

Scaled residuals:
   Min     1Q Median     3Q    Max
-6.568 -1.102 -0.309  0.744 14.692

Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 1.07     1.04
Number of obs: 3616, groups:  id, 818

Fixed effects:
             Estimate Std. Error z value             Pr(>|z|)
(Intercept)  1.429575   0.038706   36.93 < 0.0000000000000002 ***
month       -0.006164   0.000795   -7.75   0.0000000000000092 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
      (Intr)
month -0.214
convergence code: 0
Model is nearly unidentifiable: very large eigenvalue
 - Rescale variables?
```

The fitted equation is:

$$\ln(\mu_{ij}) = 1.43 - 0.006(\text{Month}_{\text{ij}})$$

This model suggests that:

- The estimated average number of alcohol-related problems reported at the onset of the study is 4.78, conditional on the random-effects ($e^1.56$) ($p < .001$).
- Each month of the intervention is associated with a 0.6% decrease, on average, in the number of alcohol-related problems reported, conditional on the random-effects ($p < .001$).

The warning message that `Model is nearly unidentifiable: very large eigenvalue` indicates that the mthematics in the estimation procedure almost didn't work. This might lead estimates that are unstable (e.g., large SEs).

Random-Effects

Obtaining the estimated random-effects to obtain the fitted equation for Student 01:

```
head(ranef(glmer.1)$id)
```

```
  (Intercept)
1      -1.905
2       0.162
3      -0.140
4      -0.420
5       0.801
6      -1.488
```

$$\begin{aligned}\ln(\mu_{ij}) &= \big[1.43 - 1.91\big] - 0.006(\text{Month}_{\text{ij}}) \\ &= -0.48 - 0.006(\text{Month}_{\text{ij}})\end{aligned}$$

For Student 01:

- The estimated average number of alcohol-related problems reported at the onset of the study is 0.61 ($e^{-.048}$).
- Each month of the intervention is associated with a 0.6% decrease, on average, for Student 01 in the number of alcohol-related problems reported.

Note Student 01's rate-of-change over time is the same as the population average rate-of-change. This was expected since we did not fit a model that allowed for individual differences in the slope coefficient.

Random-Effect of Slope

We can also fit a model that includes a random-effect of the `month` predictor to allow individual participants to have rates-of-change that vary from the average.

```
glmer.2 = glmer(problems ~ 1 + month + (1 + month|id), data = rapi, family = poisson(link = "log"))
summary(glmer.2)
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: poisson  ( log )
Formula: problems ~ 1 + month + (1 + month | id)
   Data: rapi

     AIC      BIC   logLik deviance df.resid
   21497    21528   -10744    21487     3611

Scaled residuals:
   Min     1Q Median     3Q    Max
-6.207 -0.850 -0.242  0.572  9.338

Random effects:
 Groups Name        Variance Std.Dev. Corr
```

```
 id     (Intercept) 0.93692  0.9679
        month       0.00395  0.0628   -0.14
Number of obs: 3616, groups:  id, 818

Fixed effects:
            Estimate Std. Error z value            Pr(>|z|)
(Intercept)  1.56355    0.03736    41.9 <0.0000000000000002 ***
month       -0.03130    0.00268   -11.7 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
      (Intr)
month -0.261
```

This model suggests that:

- The estimated average number of alcohol-related problems reported at the onset of the study is 4.78, conditional on the random-effects ($e^1.56$) ($p < .001$).
- Each month of the intervention is associated with a 3% decrease, on average, in the number of alcohol-related problems reported, conditional on the random-effects ($p < .001$).

Moreover, the warning message has now disappeared.

Random-Effects

Obtaining the estimated random-effects to obtain the fitted equation for Student 01:

```
head(ranef(glmer.2)$id)
```

```
  (Intercept)    month
1     -1.6741 -0.02306
2     -0.0643  0.03349
3      0.4419 -0.08513
4     -0.4458 -0.00617
5      1.7140 -0.20874
6     -0.8490 -0.07710
```

$$\begin{aligned}\ln(\mu_{ij}) &= \big[1.56 - 1.67\big] - \big[0.031 - 0.02\big](\text{Month}_{\text{ij}}) \\ &= -0.11 - 0.011(\text{Month}_{\text{ij}})\end{aligned}$$

For Student 01:

- The estimated average number of alcohol-related problems reported at the onset of the study is 0.90 ($e^{-0.11}$).
- Each month of the intervention is associated with a 1% decrease, on average, in the number of alcohol-related problems reported.

Student 01's rate-of-change over time is is now lower than the population average rate-of-change.

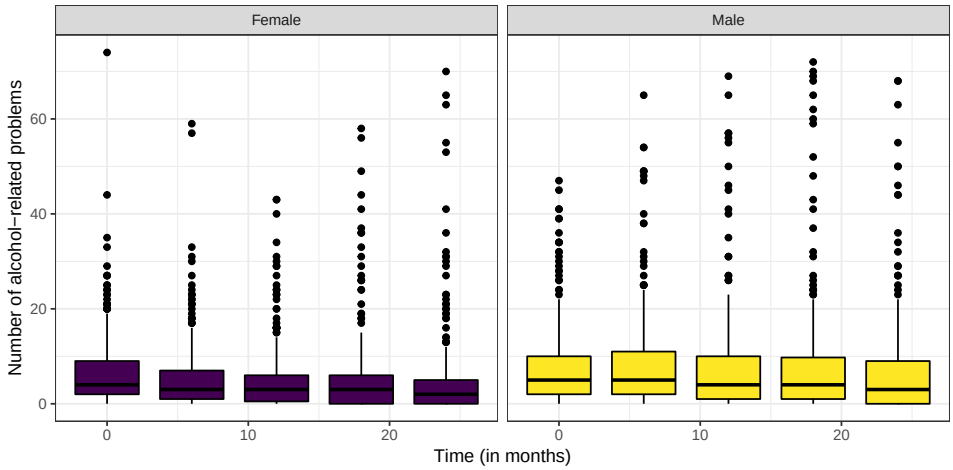


Figure 3: Side-by-side box-and-whiskers plots showing the distribution of alcohol-related problems at different points in time during the study for females and males.

Effect of Sex

To examine whether there is a main-effect of gender, we include male as a fixed-effect in the model. Before we fit the model, we will graphically explore this by examining side-by-side box-and-whisker plots.

```
rapi = rapi %>%
  mutate(
    sex = if_else(male == 1, "Male", "Female")
  )

# Plot of conditional distributions by sex
ggplot(data = rapi, aes(x = month, y = problems, group = month)) +
  geom_boxplot(aes(fill = sex), color = "black") +
  theme_bw() +
  xlab("Time (in months)") +
  ylab("Number of alcohol-related problems") +
  facet_wrap(~sex) +
  guides(fill = FALSE) +
  scale_fill_viridis_d()
```

We can also compute summary measures of conditioned on time (`month`) and sex, as well as plot the mean profiles.

```
 # Conditional Mean, SD, and Variance by time point
rapi %>%
  group_by(month, sex) %>%
  summarize(
    M = mean(problems),
    SD = sd(problems),
    Var = var(problems)
    ) %>%
  arrange(sex, month)
```

```
# A tibble: 10 x 5
```

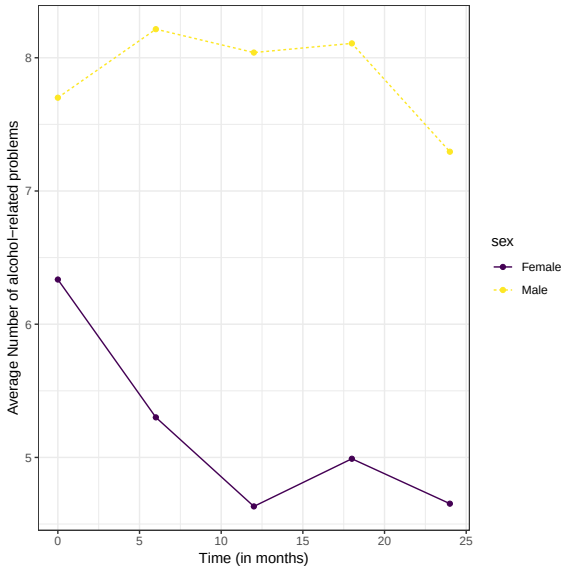


Figure 4: Mean profile by sex showing the average number of alcohol-related problems over time.

```
# Groups:   month [5]
   month sex        M    SD   Var
   <dbl> <chr>  <dbl> <dbl> <dbl>
 1     0 Female  6.34  7.01  49.2
 2     6 Female  5.30  6.78  46.0
 3    12 Female  4.63  6.43  41.4
 4    18 Female  4.99  7.90  62.3
 5    24 Female  4.65  8.90  79.3
 6     0 Male    7.70  8.59  73.8
 7     6 Male    8.21 10.3  106.
 8    12 Male    8.04 11.4  131.
 9    18 Male    8.11 12.9  166.
10    24 Male    7.29 11.4  130.
```

```
# Profile plot
ggplot(data = rapi, aes(x = month, y = problems, color = sex, linetype = sex)) +
  stat_summary(geom = "line", fun.y = mean) +
  stat_summary(geom = "point", fun.y = mean) +
  theme_bw() +
  xlab("Time (in months)") +
  ylab("Average Number of alcohol-related problems") +
  scale_color_viridis_d()
```

All of this evidence suggests that males and females report different numbers of alcohol-related problems at the onset of the study and also have differeing average rates-of-change. First, we explore the sex differences at the onset of the study by including the main-effect of `male` in the GLMER model. Then we fit a model that also allows for the rate-of-change to differ by sex by including an interaction term between `male` and `month`.

```
# Fit main-effect of sex model
glmer.3 = glmer(problems ~ 1 + month + male + (1 + month|id),
                data = rapi, family = poisson(link = "log"))
```

```
# Fit sex x month interaction model
glmer.4 = glmer(problems ~ 1 + month + male + male:month + (1 + month|id),
                data = rapi, family = poisson(link = "log"))

# Table of model evidence
aictab(
  cand.set = c(glmer.2, glmer.3, glmer.4),
  modnames = c("Time", "Time + Sex", "Time x Sex")
)
```

```
Model selection based on AICc:

           K  AICc Delta_AICc AICcWt Cum.Wt     LL
Time x Sex 7 21479       0.00   0.98   0.98 -10732
Time + Sex 6 21487       8.11   0.02   1.00 -10737
Time       5 21497      18.55   0.00   1.00 -10744
```

The empirical evidence supports the time by sex interaction model. Examiningthis model's output:

```
summary(glmer.4)
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: poisson  ( log )
Formula: problems ~ 1 + month + male + male:month + (1 + month | id)
   Data: rapi

     AIC      BIC   logLik deviance df.resid
   21478    21522   -10732    21464     3609

Scaled residuals:
   Min     1Q Median     3Q    Max
-6.205 -0.854 -0.245  0.574  9.329

Random effects:
 Groups Name        Variance Std.Dev. Corr
 id     (Intercept) 0.92689  0.9628
        month       0.00388  0.0623   -0.16
Number of obs: 3616, groups:  id, 818

Fixed effects:
            Estimate Std. Error z value            Pr(>|z|)
(Intercept)  1.48036    0.04886   30.30 <0.0000000000000002 ***
month       -0.03805    0.00348  -10.94 <0.0000000000000002 ***
male         0.19651    0.07378    2.66              0.0077 **
month:male   0.01650    0.00515    3.21              0.0013 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
           (Intr) month  male
```

```
month      -0.273
male       -0.648  0.166
month:male  0.169 -0.644 -0.251
```

This model suggests that:

- The estimated average number of alcohol-related problems reported at the onset of the study for females is 4.39, conditional on the random-effects ($e^{1.48}$) ($p < .001$).
- For female, each month of the intervention is associated with a 3.8% decrease, on average, in the number of alcohol-related problems reported, conditional on the random-effects ($p < .001$).
- Males report 19.6% more alcohol-related problems reported at the onset of the study than females, conditional on the random-effects ($p < .001$) .
- For males, each month of the intervention is associated with a 2.1% decrease, on average, in the number of alcohol-related problems reported, conditional on the random-effects. This is a lower rate-of-change than for females by 1.7% ($p < .001$).

Because there is a significant interaction, it is useful to plot the fitted growth profiles. However, we will hold off until we fit one last model.

Overdispersion

Recall that the Poisson distribution posits that the mean and variance are identical. From our summary measures (presented earlier) it seems that the variances are greater than the means. This suggests overdispersion. To account for overdispersion, we can fit a negative binomial model. To do this we will use the `glmmTMB()` function from the **glmmTMB** package. This function is used in a similar manner as `glmer()`. To fit the negative binomial model we set the `family=` argument to `nbinom2`.

```
# Fit negative binomial model
glmer.5 = glmmTMB(problems ~ 1 + month + male + month:male + (1 + month | id), data = rapi, family = nbinom2)

# Examine output
summary(glmer.5)
```

```
 Family: nbinom2  ( log )
Formula:
problems ~ 1 + month + male + month:male + (1 + month | id)
Data: rapi

     AIC      BIC   logLik deviance df.resid
   19390    19439    -9687    19374     3608

Random effects:

Conditional model:
 Groups Name        Variance Std.Dev. Corr
 id     (Intercept) 0.6285   0.7928
        month       0.0021   0.0458   0.20
Number of obs: 3616, groups:  id, 818

Overdispersion parameter for nbinom2 family (): 2.52
```

```
Conditional model:
             Estimate Std. Error z value            Pr(>|z|)
(Intercept)  1.57806    0.04886    32.3 <0.0000000000000002 ***
month       -0.03807    0.00342   -11.1 <0.0000000000000002 ***
male         0.19920    0.07318     2.7               0.0065 **
month:male   0.01649    0.00501     3.3               0.0010 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These estimates are interpreted similarly to the Poisson model estimates. This model suggests that:

- The estimated average number of alcohol-related problems reported at the onset of the study for females is 4.85, conditional on the random-effects ($e^{1.57}$) ($p < .001$).
- For female, each month of the intervention is associated with a 3.8% decrease, on average, in the number of alcohol-related problems reported, conditional on the random-effects ($p < .001$).
- Males report 20% more alcohol-related problems reported at the onset of the study than females, conditional on the random-effects ($p < .001$) .
- For males, each month of the intervention is associated with a 2.1% decrease, on average, in the number of alcohol-related problems reported, conditional on the random-effects. This is a lower rate-of-change than for females by 1.7% ($p < .001$).

Because there is a significant interaction, it is useful to plot the fitted growth profiles. The `predict()` function for glmmTMB models only allows us to obtain predictions of the fixed- *and* random-effects. Since we want ONLY the estimates of the fixed-effects, we need to compute them from the fitted model manually.

```r
# Create plotting data
plot_data = expand.grid(
  month = seq(from = 0, to = 24, by = 0.1),
  male = c(0, 1)
) %>%
  mutate(
    log_mu = 1.57806 - 0.03807*month + 0.19920*male + 0.01649*month*male,
    mu = exp(log_mu),
    sex = factor(male, levels = c(0, 1), labels = c("Female", "Male"))
  )

# View plotting data
head(plot_data)
```

```
  month male log_mu   mu    sex
1   0.0    0   1.58 4.85 Female
2   0.1    0   1.57 4.83 Female
3   0.2    0   1.57 4.81 Female
4   0.3    0   1.57 4.79 Female
5   0.4    0   1.56 4.77 Female
6   0.5    0   1.56 4.75 Female
```

```r
# Plot
ggplot(data = plot_data, aes(x = month, y = mu, color = sex, linetype = sex)) +
  geom_line() +
  theme_bw() +
  xlab("Time (in months)") +
```

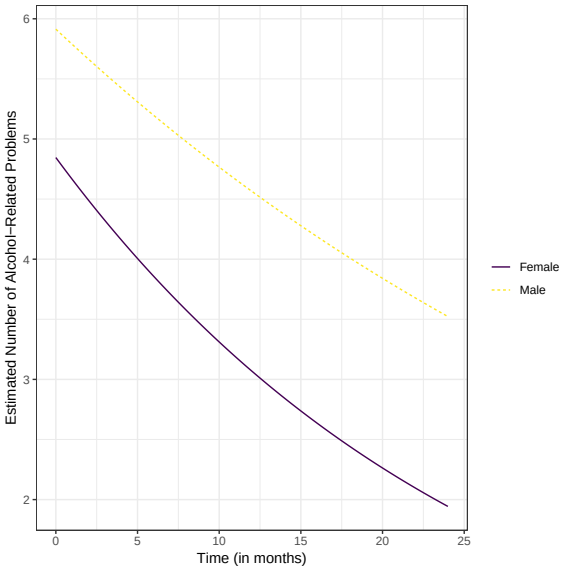


Figure 5: Model predicted mean profile by sex showing the average number of alcohol-related problems over time. A negative binomial model was fitted to the data to account for overdispersion.

```
  ylab("Estimated Number of Alcohol-Related Problems") +
  scale_color_viridis_d(name = "") +
  scale_linetype(name = "")
```

The predicted mean profiles show that in general, females report fewer alcohol-related problems than males at each time point in the study, and that the number of alcohol-related problems for both sexes decrease over the course of the study. Females show a greater-rate of decrease than males.

Other Resources

In addition to the notes and what we cover in class, there many other resources for learning about using binomial logistic regression models for analyzing binary data. Here are some resources that may be helpful in that endeavor:

- Section 3.2.1: *The Binomial and Bernoulli Distributions* in Fox (2009) [Required Textbook]

Atkins, D. C., Baldwin, S. A., Zheng, C., Gallop, R. J., & Neighbors, C. (2013). A tutorial on count regression and zero-altered count models for longitudinal substance use data. *Psychology of Addictive Behaviors*, *27*(1), 166–177. https://doi.org/10.1037/a00296508