GLMER: Count Data

2017-11-26

Preparation

We will use the data in the *RAPI.csv* file. These data are from Atkins, Baldwin, Zheng, Gallop, & Neighbors (2013) and

> …is 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).

```r
# Load libraries
library(tidyverse)
library(corrr)
library(gridExtra)
library(lme4)
library(AICcmodavg)



# This is to disable scientific notation
options(scipen = 99)

# Read in long data
rapi = read_csv("~/Dropbox/epsy-8282/data/RAPI.csv")
head(rapi)
```

```
## # A tibble: 6 x 4
##      id  rapi  male  time
##   <int> <int> <int> <int>
## 1     1     0     1     0
## 2     1     0     1     6
## 3     1     0     1    18
## 4     2     3     0     0
## 5     2     6     0     6
## 6     2     5     0    12
```

The variables include:

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

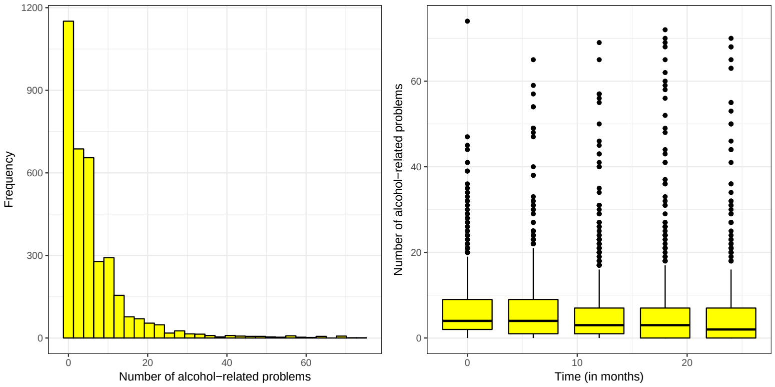
Exploration

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

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

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

grid.arrange(p1, p2, ncol = 2)
```



```
# Marginal Mean, SD, and Variance
rapi %>% summarize(M = mean(rapi), SD = sd(rapi), Var = var(rapi))
```

```
## # A tibble: 1 x 3
##          M       SD      Var
##      <dbl>    <dbl>    <dbl>
## 1 6.310288 9.106954 82.93661
```

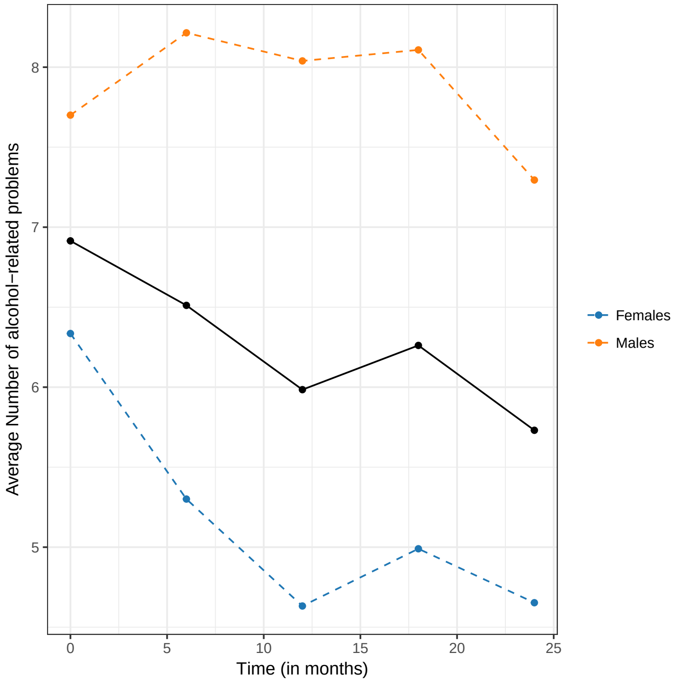
```
# Conditional Mean, SD, and Variance by gender and time point
rapi %>% group_by(male, time) %>% summarize(M = mean(rapi), SD = sd(rapi), Var = var(rapi))
```

```
## # A tibble: 10 x 5
## # Groups:   male [?]
##     male  time        M        SD       Var
##    <int> <int>    <dbl>     <dbl>     <dbl>
##  1     0     0 6.335456  7.011389  49.15957
##  2     0     6 5.300683  6.782045  45.99614
##  3     0    12 4.632319  6.432983  41.38327
##  4     0    18 4.990099  7.895402  62.33737
##  5     0    24 4.652956  8.904729  79.29420
##  6     1     0 7.700288  8.591034  73.80587
##  7     1     6 8.214744 10.316629 106.43284
##  8     1    12 8.039146 11.438121 130.83060
##  9     1    18 8.107914 12.870749 165.65618
## 10     1     24 7.294776 11.394807 129.84162
```

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 there is a great deal of variability in the number of alcohol-related problems being reported, and in general, men report more alcohol-related problems than women (and also display more variation). They also suggest that the number of problems is diminishing over time, albeit slightly.

Plotting the mean profile for each gender (and for the entire sample) suggests some non-linearity in the change curve.

```
ggplot(data = rapi, aes(x = time, y = rapi)) +
  stat_summary(aes(color = factor(male)), geom = "line",  fun.y = mean, linetype = "dashed") +
  stat_summary(aes(color = factor(male)), geom = "point", fun.y = mean) +
  stat_summary(aes(group = time), geom = "line", fun.y = mean, color = "black", group = 1) +
  stat_summary(aes(group = time), geom = "point", fun.y = mean, color = "black") +
  theme_bw() +
  xlab("Time (in months)") +
  ylab("Average Number of alcohol-related problems") +
  ggsci::scale_color_d3(name = "", labels = c("Females", "Males"))
```



Examining Change Profiles

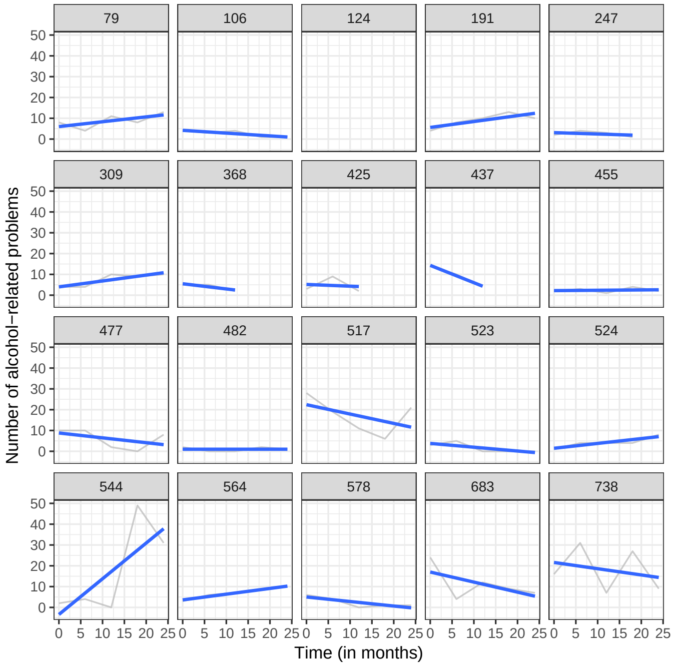
Below we plot the change profiles for a random sample of 20 students. The individual linear fitted regressions are also displayed.

```
# Obtain ID numbers
IDs = unique(rapi$id)

# Choose random sample of ID numbers (Sample w/o replacement)
set.seed(200)
my_samp = sample(IDs, size = 20, replace = FALSE)
```

```
# Get data for students in the sample
rapi_srs = rapi %>% filter(id %in% my_samp)

# Plot
ggplot(data = rapi_srs, aes(x = time, y = rapi, group = id)) +
  geom_line(color = "grey", alpha = 0.8) +
  theme_bw() +
  xlab("Time (in months)") +
  ylab("Number of alcohol-related problems") +
  facet_wrap(~id) +
  geom_smooth(method = "lm", se = FALSE)
```



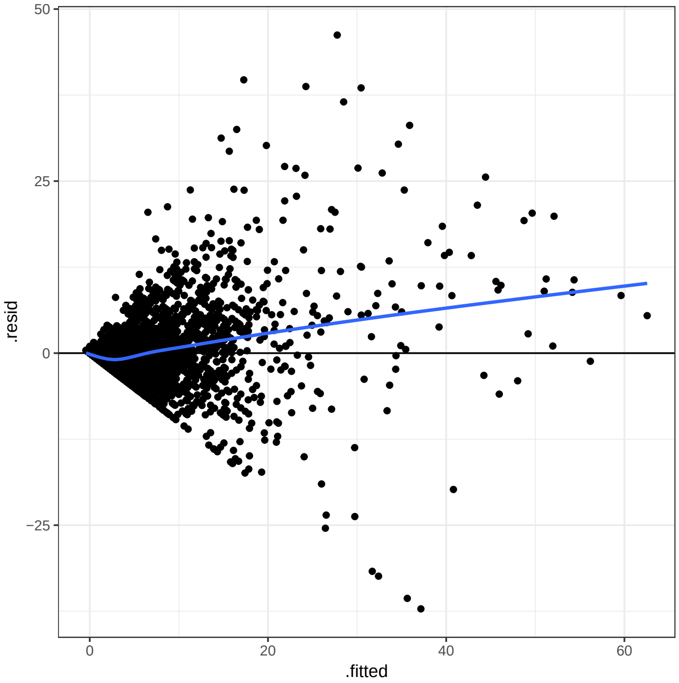
The plot shows:

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

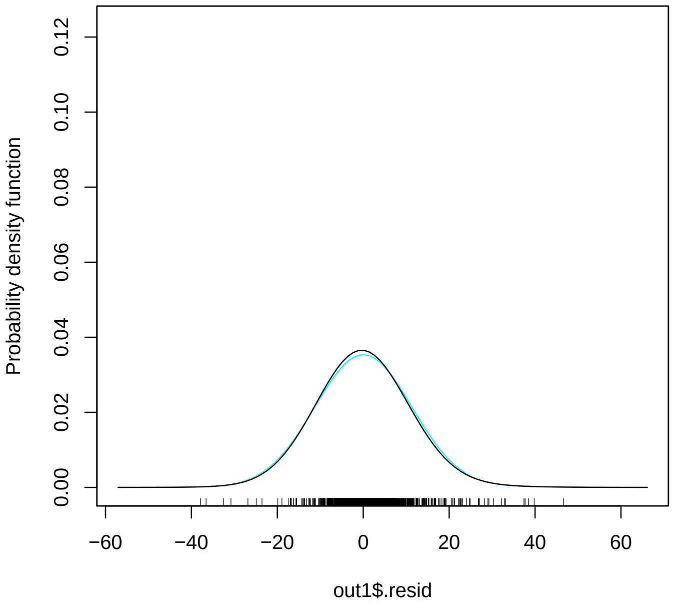
This leads us to adopt random-effects of both intercept and time. We will begin by fitting a linear mixed-effects model with random intercepts and slopes to the data.

```
lmer.1 = lmer(rapi ~ 1 + time + (1 + time | id), data = rapi, REML = FALSE)
out1 = broom::augment(lmer.1)

ggplot(data = out1, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  geom_smooth(se = FALSE) +
  theme_bw()
```



```
sm::sm.density(out1$.resid, model = "normal")
```



The residuals versus the fitted values show major violations of the models' assumptions. The plot displays growing variance, and non-linearity.

Poisson Regression: Modeling Count Data

We could try to transform the data to alleviate these problems. We could also fit a model that allows a more flexible error structure (e.g., non-normal errors). A better model when dealing with count data is to use a *Generalized Linear Mixed-Effects Regression Model* (GLMER). These models are the mixed-effects counterpart to the Generalized Linear Models.

Recall that Generalized Linear Models have three major components: (1) a linear predictor specifying a linear function of predictors; (2) a link function which transforms the expectation of the outcome to the linear predictor; and (3) a random component specifiying the conditional distributions of the outcome.

For count data, there are several possibilities, but we will start by fitting a model that assumes Poisson distributed outcomes with a log-link function. The multilevel unconditional growth model is:

$$\ln\Big(E(\mathrm{RAPI}_{ij})\Big) = \beta_0 + \beta_1(\mathrm{Time_{ij}}) + b_{0i} + b_{1i}(\mathrm{Time_{ij}}) + \epsilon_{ij}$$

where ϵ_{ij} are distributed as Poisson with λ. (The parameter λ is the rate parameter). To fit this model as a fixed-effects only model, we use the `glm()` function.

```
glm.1 = glm(rapi ~ 1 + time, data = rapi, family = poisson(link = "log"))
summary(glm.1)
```

```
##
## Call:
## glm(formula = rapi ~ 1 + time, family = poisson(link = ”log”),
##     data = rapi)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6939  -2.7065  -1.2683   0.5512  15.0194
##
## Coefficients:
##               Estimate Std. Error z value            Pr(>|z|)
## (Intercept)  1.9202426  0.0107420 178.761 <0.0000000000000002 ***
## time        -0.0070348  0.0007802  -9.016 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 32797  on 3615  degrees of freedom
## Residual deviance: 32715  on 3614  degrees of freedom
## AIC: 42645
##
## Number of Fisher Scoring iterations: 6
```

The fitted equation is

$$\ln\Big(E(\mathrm{R\hat{A}PI}_{ij})\Big) = 1.92 - 0.007(\mathrm{Time_{ij}})$$

To interpret these, we back-transform using the inverse link function. Since here we used a log-link, to back-transform, we exponentiate. Back-transforming the fitted model we get

$$\begin{aligned} e^{\ln\Big(E(\mathrm{R\hat{A}PI}_{ij})\Big)} &= e^{1.92 - 0.007(\mathrm{Time_{ij}})} \\ E(\mathrm{R\hat{A}PI}_{ij}) &= e^{1.92} \times e^{-0.007(\mathrm{Time_{ij}})} \end{aligned}$$

At time = 0, the estimated average RAPI score is

$$
\begin{aligned}
&= e^{1.92} \times e^{-0.007(0)} \\
&= 6.82 \times 1 \\
&= 6.82
\end{aligned}
$$

The estimated average number of alcohol-related problems at time 0 is 6.82. This is the same as directly back-transforming the intercept:

$$
\begin{aligned}
e^{1.92} \times e^{-0.007(1)} &= 6.82 \\
&= 6.82 \times 0.9930244 \\
&= 6.77
\end{aligned}
$$

The estimated RAPI decreases by a factor of 0.99. This is the value we get if we back-transform the slope directly,

$$
e^{-0.007} = 0.99
$$

Another way to interpret this is to use the coefficient of -0.007 and interpret that as the percent change. Each one month difference is associated with a 0.7% decrease in the number of alcohol-related problems, on average.

Mixed-Effects Poisson Model

To include random-effects in the generalized model we use the `glmer()` function (from the **lme4** package). For example, to include both fixed- and random-effects of time we use:

```
# Fit random-intercept model
glmer.1 = glmer(rapi ~ 1 + time + (1 + time | 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: rapi ~ 1 + time + (1 + time | id)
##    Data: rapi
##
##      AIC      BIC   logLik deviance df.resid
##  21497.1  21528.0 -10743.5  21487.1     3611
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.2070 -0.8503 -0.2421  0.5722  9.3384
##
## Random effects:
##  Groups Name        Variance Std.Dev. Corr
##  id     (Intercept) 0.936916 0.96794
##         time        0.003949 0.06284  -0.14
## Number of obs: 3616, groups:  id, 818
##
## Fixed effects:
##             Estimate Std. Error z value            Pr(>|z|)
## (Intercept)  1.56355    0.03736   41.85 <0.0000000000000002 ***
## time        -0.03130    0.00268  -11.68 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr)
## time -0.261
```

Based on the output,

- The estimated average number of alcohol-related problems reported at the onset of the study is 4.77 ($e^{1.56}$), conditional on the random-effects.
- Each month of the intervention is associated with a 3% decrease in the number of alcohol-related problems reported, 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 don't 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

$$\hat{Y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1(\text{Time}_{ij}) + \hat{b}_{0i} + \hat{b}_{1i}(\text{Time}_{ij})$$

Since the random-effects have mean=0, when we average across them, they are not contributing anything to the population-average model (the two random-effects are both 0, so addign them 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.

$$\ln\left(\hat{Y}_{ij}\right) = \hat{\beta}_0 + \hat{\beta}_1(\text{Time}_{ij}) + \hat{b}_{0i} + \hat{b}_{1i}(\text{Time}_{ij})$$

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 ($e^0 \neq 0$).

Random-Effects

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

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

```
##   (Intercept)         time
## 1 -1.67405323 -0.023058974
## 2 -0.06428572  0.033488273
## 3  0.44190844 -0.085133921
## 4 -0.44579678 -0.006170376
## 5  1.71397388 -0.208740963
## 6 -0.84899403 -0.077095744
```

The fitted equation for Student 01 is:

$$\begin{aligned}\ln\left(E(\text{RAPI}_{ij})\right) &= 1.563 - 0.031(\text{Time}_{\text{ij}}) - 1.674 - 0.023(\text{Time}_{\text{ij}}) \\ &= -0.111 - 0.054(\text{Time}_{\text{ij}})\end{aligned}$$

- The estimated average number of alcohol-related problems reported at the onset of the study for Student 01 is 0.89 ($e^{-0.111}$)
- Each month of the intervention is associated with a 5% decrease in the number of alcohol-related problems reported for Student 01.

Gender Effect

To examine whether there is a main-effect of gender, we include `male` as a fixed-effect in the model.

```r
# Fit random-intercept model
glmer.2 = glmer(rapi ~ 1 + time + male + (1 + time | 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: rapi ~ 1 + time + male + (1 + time | id)
##    Data: rapi
##
##      AIC      BIC   logLik deviance df.resid
##  21486.6  21523.8 -10737.3  21474.6     3610
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.2039 -0.8509 -0.2468  0.5691  9.3226
##
## Random effects:
##  Groups Name        Variance Std.Dev. Corr
##  id     (Intercept) 0.927706 0.96317
##         time        0.003935 0.06273  -0.16
## Number of obs: 3616, groups:  id, 818
##
## Fixed effects:
##              Estimate Std. Error z value             Pr(>|z|)
## (Intercept)  1.453842   0.048708  29.848 < 0.0000000000000002 ***
## time        -0.031019   0.002677 -11.587 < 0.0000000000000002 ***
## male         0.256010   0.071999   3.556             0.000377 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr) time
## time -0.234
## male -0.645  0.036
```

Based on the output,

- The estimated average number of alcohol-related problems reported at the onset of the study for females is 4.28 ($e^{1.45}$), conditional on the random-effects.
- Each month of the intervention is associated with a 3% decrease in the number of alcohol-related problems reported, controlling for differences in gender, conditional on the random-effects.
- Males report 29% more problems than females, controlling for differences in time and conditional on the random-effects.

Interaction between Time and Gender

```
# Fit random-intercept model
glmer.3 = glmer(rapi ~ 1 + time + male + time:male + (1 + time | id), data = rapi, family = poisson(link = "log"))
summary(glmer.3)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: rapi ~ 1 + time + male + time:male + (1 + time | id)
##    Data: rapi
##
##      AIC      BIC   logLik deviance df.resid
##  21478.5  21521.8 -10732.2  21464.5     3609
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.2046 -0.8545 -0.2446  0.5745  9.3291
##
## Random effects:
##  Groups Name        Variance Std.Dev. Corr
##  id     (Intercept) 0.926891 0.96275
##         time        0.003878 0.06227  -0.16
## Number of obs: 3616, groups:  id, 818
##
## Fixed effects:
##               Estimate Std. Error z value             Pr(>|z|)
## (Intercept)  1.480357   0.048860  30.298 < 0.0000000000000002 ***
## time        -0.038050   0.003478 -10.942 < 0.0000000000000002 ***
## male         0.196512   0.073777   2.664              0.00773 **
## time:male    0.016502   0.005145   3.207              0.00134 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) time   male
## time      -0.273
## male      -0.648  0.166
## time:male  0.169 -0.644 -0.251
```

Based on the output,

- There is an interaction-effect between time and gender. The longitudinal change is differernt for males and females.

This suggests that males and females have different longitudinal profiles. We can compute the fitted equations for males and females by substituting in the aprropriate 0/1 value for the `male` predictor.

$$\textbf{Females}: \quad \ln\left(E(\text{RAPI}_{ij})\right) = 1.480 - 0.038(\text{Time}_{\text{ij}})$$

$$\textbf{Males}: \quad \ln\left(E(\text{RAPI}_{ij})\right) = \big[1.480 + 0.197\big] - \big[0.038 + 0.017\big](\text{Time}_{\text{ij}})$$

Back-transforming these results leads us to the direct coefficient interpretations:

- The estimated average number of alcohol-related problems reported at the onset of the study for females is 4.39 ($e^{1.480}$), conditional on the random-effects.
- For females, each month of the intervention is associated with a 3.8% decrease in the number of alcohol-related problems reported, conditional on the random-effects.
- Males report 21% more problems than females at the onset of the study, conditional on the random-effects.
- For males, each month of the intervention is associated with a decrease that is 1.7% greater than that of females, conditional on the random-effects.

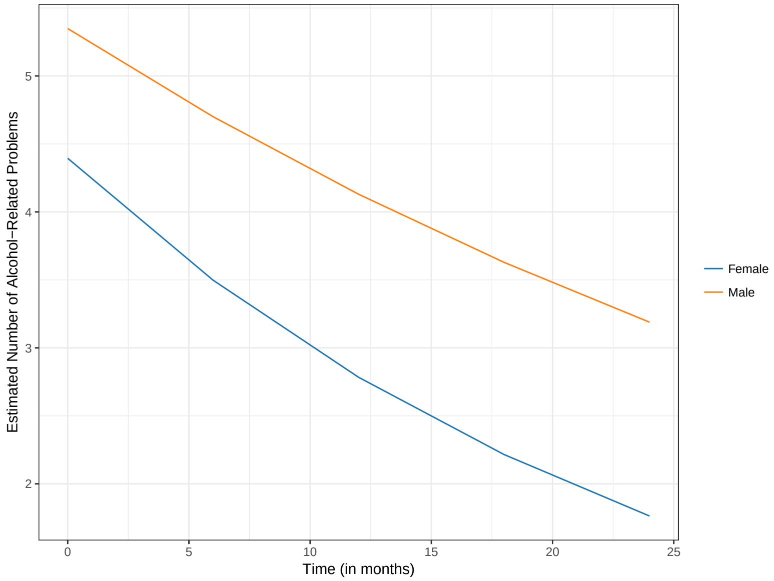
To understand this further, we will plot the fitted growth patterns for males and females. We do this the the same way we did for LMERs. The only differences is that when we use the `predict()` function for a GLMER, we specify `type="response"` to back-transform the estimates back to the scale of the response variable.

```
my_data = expand.grid(
  time = c(0, 6, 12, 18, 24),
  male = c(0, 1)
)

# Get y-hat values (use type="response")
my_data$yhat = predict(glmer.3, newdata = my_data, re.form = NA, type = "response")

# Coerce male into a categorical factor for better plotting
my_data$gender = factor(my_data$male, levels = c(0, 1), labels = c("Female", "Male"))

# Plot the fitted growth patterns
ggplot(data = my_data, aes(x = time, y = yhat, color = gender)) +
  geom_line() +
  theme_bw() +
  xlab("Time (in months)") +
  ylab("Estimated Number of Alcohol-Related Problems") +
  ggsci::scale_color_d3(name = "")
```



The estimated change profile show that in general, females report fewer alcohol-related problems than males, and that the number of alcohol-related problems for both sexes decrease over the course of the study. Females also show a greater-rate of decrease than males.

Overdispersion

Recall that the Poisson distribution posits that the mean and varaince are identical. From our summary measures (presented earlier) it seems that the variances are greater than the means. This suggests *overdispersion*. To fit an overdispersed Poisson model, we (1) create a per-observation term in the data and (2) include that term as a second random-effect in the model.

```
# Create the per-observation error term
rapi$over = 1:nrow(rapi)

# Include the term as a second random-effect in the model
glmer.4 = glmer(rapi ~ 1 + time + male + time:male + (1 + time | id) + (1 | over),
                data = rapi, family = poisson(link = "log"))
```

Then, since the original Poisson interaction model is nested in the overdispersed model, we can examine the necessity of the overdispersion by carrying out a likelihood ratio test.

```
anova(glmer.3, glmer.4)
```

```
## Data: rapi
## Models:
## glmer.3: rapi ~ 1 + time + male + time:male + (1 + time | id)
## glmer.4: rapi ~ 1 + time + male + time:male + (1 + time | id) + (1 | over)
##         Df   AIC   BIC logLik deviance  Chisq Chi Df            Pr(>Chisq)
## glmer.3  7 21478 21522 -10732    21464
## glmer.4  8 19448 19498  -9716    19432 2032.5      1 < 0.00000000000000022
##
## glmer.3
## glmer.4 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This test suggests that there is a statistically significant reduction in the deviance when including the overdispersion parameter. The fixed-effects are interpreted the same way from this model, only now they account for the overdispersion in the data.

```
# Examine output
summary(glmer.4)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: rapi ~ 1 + time + male + time:male + (1 + time | id) + (1 | over)
##    Data: rapi
##
##      AIC      BIC   logLik deviance df.resid
##  19448.0  19497.5  -9716.0  19432.0     3608
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.03682 -0.53729 -0.03497  0.24217  1.29823
##
## Random effects:
##  Groups Name        Variance Std.Dev. Corr
##  over   (Intercept) 0.376001 0.61319
##  id     (Intercept) 0.676816 0.82269
##         time        0.002328 0.04825  0.11
## Number of obs: 3616, groups:  over, 3616; id, 818
##
## Fixed effects:
##              Estimate Std. Error z value             Pr(>|z|)
## (Intercept)  1.387128   0.048956  28.334 < 0.0000000000000002 ***
## time        -0.037891   0.003383 -11.201 < 0.0000000000000002 ***
## male         0.200662   0.073876   2.716              0.00660 **
## time:male    0.015946   0.005021   3.176              0.00149 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) time   male
## time      -0.283
## male      -0.647  0.173
## time:male  0.176 -0.643 -0.263
```

Substituting the appropriate 0/1 value for `male` and back-transforming these results:

- The estimated average number of alcohol-related problems reported at the onset of the study for females is 4.00 ($e^{1.387}$), conditional on the random-effects.
- For females, each month of the intervention is associated with a 3.8% decrease in the number of alcohol-related problems reported, conditional on the random-effects.
- Males report 20% more problems than females at the onset of the study, conditional on the random-effects.
- For males, each month of the intervention is associated with a decrease that is 1.6% greater than that of females, conditional on the random-effects.

Again, to help understand these interpretations, we can plot these results.

```
# Plot the fitted growth patterns
# Get y-hat values (use type="response")
my_data$yhat2 = predict(glmer.4, newdata = my_data, re.form = NA, type = "response")

ggplot(data = my_data, aes(x = time, y = yhat2, color = gender)) +
  geom_line(linetype = "dashed") +
  geom_line(aes(y = yhat)) +
  theme_bw() +
  xlab("Time (in months)") +
  ylab("Estimated Number of Alcohol-Related Problems") +
  ggsci::scale_color_d3(name = "")
```

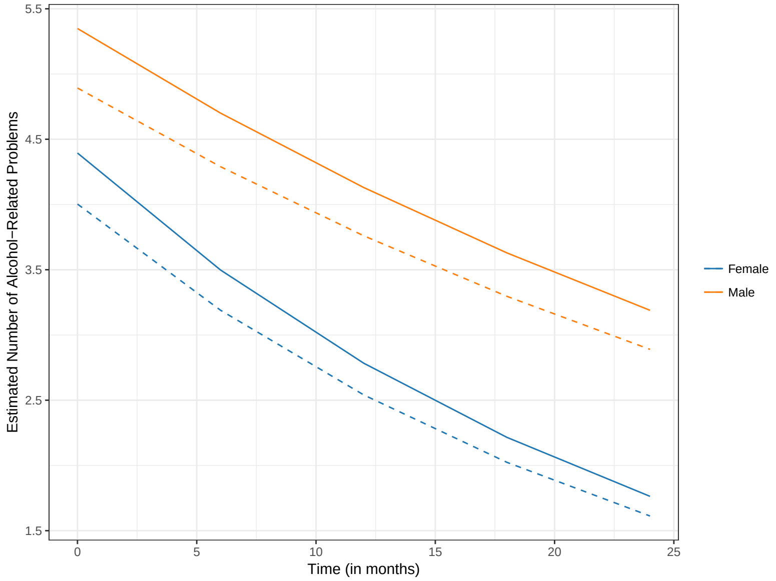


Figure X. Plot of the fitted longitudinal profiles for the Poisson model (solid lines) and the Poisson model with an overdispersion parameter (dashed lines).

Negative Binomial Random-Effects Model

Another model that directly allows for overdispersion is the negative binomial model. Because it directly fits overdispersion, it does not require the overdispersion random-effect in the model.

```
library(glmmTMB)
```

```
## Warning in checkMatrixPackageVersion(): Package version inconsistency detected.
## TMB was built with Matrix version 1.2.10
## Current Matrix version is 1.2.12
## Please re-install 'TMB' from source or restore original 'Matrix' package
```

```
glmer.5 = glmmTMB(rapi ~ 1 + time + male + time:male + (1 + time | id), data = rapi, family = nbinom2)
summary(glmer.5)
```

```
##  Family: nbinom2  ( log )
## Formula:          rapi ~ 1 + time + male + time:male + (1 + time | id)
## Data: rapi
## 
##      AIC      BIC   logLik deviance df.resid
##  19389.5  19439.1  -9686.8  19373.5     3608
## 
## Random effects:
## 
## Conditional model:
##  Groups Name        Variance Std.Dev. Corr
##  id     (Intercept) 0.628487 0.79277      
##         time        0.002099 0.04581  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)  1.578063   0.048858   32.30 < 0.0000000000000002 ***
## time        -0.038071   0.003419  -11.14 < 0.0000000000000002 ***
## male         0.199203   0.073177    2.72              0.00648 ** 
## time:male    0.016486   0.005013    3.29              0.00101 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These estimates are interpreted similarly to the Poisson model estimates. Because there is a significant interaction, we will again plot the fitted growth profiles.

Substituting the appropriate 0/1 value for `male` and back-transforming these results:

- The estimated average number of alcohol-related problems reported at the onset of the study for females is 4.85 ($e^{1.578}$), conditional on the random-effects.
- For females, each month of the intervention is associated with a 3.8% decrease in the number of alcohol-related problems reported, conditional on the random-effects.
- Males report 20% more problems than females at the onset of the study, conditional on the random-effects.
- For males, each month of the intervention is associated with a decrease that is 1.6% greater than that of females, conditional on the random-effects.

Again, to help understand these interpretations, we can plot these results. 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.

```r
# Get y-hat values (use type="response")
my_data$yhat3 = exp(1.578063 + -0.038071*my_data$time + 0.199203*my_data$male + 0.016486*my_data$time*my_data$male)

# Plot the fitted growth patterns
ggplot(data = my_data, aes(x = time, y = yhat3, color = gender)) +
  geom_line(linetype = "dotted") +
  geom_line(aes(y = yhat)) +
  geom_line(aes(y = yhat2), linetype = "dashed") +
  theme_bw() +
  xlab("Time (in months)") +
  ylab("Estimated Number of Alcohol-Related Problems") +
  ggsci::scale_color_d3(name = "")
```

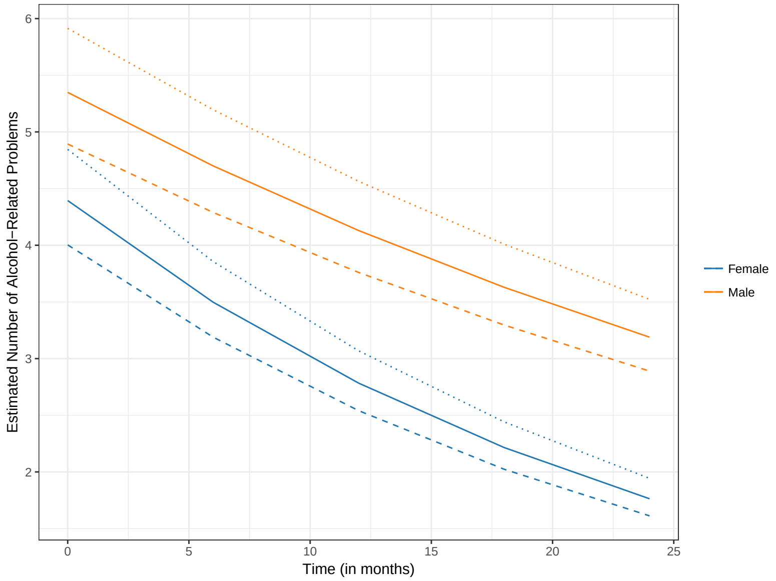


Figure X. Plot of the fitted longitudinal profiles for the (1) Poisson model (solid lines); (2) Poisson model with an overdispersion parameter (dashed lines); and (3) negative binomial model.

To determine fit between the Poisson model with overdispersion and the negative binomial model we could examine the AIC indices.

```
# Poisson
AIC(glmer.3)
```

```
## [1] 21478.49
```

```
# Poisson with overdispersion
AIC(glmer.4)
```

```
## [1] 19447.98
```

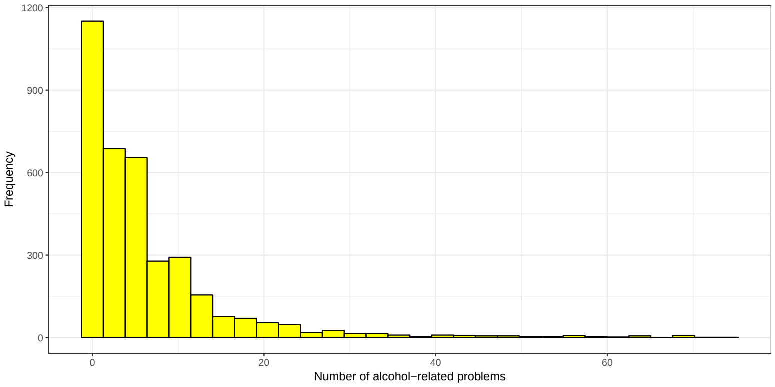
```
# Negative binomial
AIC(glmer.5)
```

```
## [1] 19389.52
```

Zero Inflated Model

One issue with count data that can affect the fit of the Poisson model is a large number of zeros. The histogram shows several zeros in the observed data.

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



From this plot we see zero is the most commmon self-reported score. Are there an excess number of zeros? To determine this, we can fit a zero inflated model and compare it to the non-zero inflated negative binomial model. The UCLA Institute for Digital Research and Education says this about zero inflated models:

> The zero-inflated negative binomial regression generates two separate models and then combines them. First, a logit model is generated for the "certain zero" cases described above, predicting whether or not a student would be in this group. Then, a negative binomial model is generated predicting the counts for those students who are not certain zeros. Finally, the two models are combined.

To fit a zero inflated model, we need to specify both models: the count model and the model predicting the certain zeros. The second model is specified via the argument `ziformula=` in the `glmmTMB()` function. This takes a model formula that specifies the model to estimate for the excess zeros. Below we are predicting count using time, gender, and the interaction between time and gender. We predict the certain zeros using the same set of predictors. Note that the model that included random-effects of intercept and time did not converge, so we dropped the RE of time.

```
# Model does not converge
# glmer.6 = glmmTMB(rapi ~ 1 + time + male + time:male + (1 + time | id), data = rapi,
#                   family = nbinom2, zi = ~ 1 + time + male + time:male)


glmer.7 = glmmTMB(rapi ~ 1 + time + male + time:male + (1 | id), data = rapi,
                  family = nbinom2, zi = ~ 1 + time + male + time:male)

summary(glmer.7)
```

```
##  Family: nbinom2  ( log )
## Formula:          rapi ~ 1 + time + male + time:male + (1 | id)
## Zero inflation:        ~1 + time + male + time:male
## Data: rapi
##
##      AIC      BIC   logLik deviance df.resid
##  19479.5  19541.4  -9729.8  19459.5     3606
##
## Random effects:
##
## Conditional model:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 0.8929   0.945
## Number of obs: 3616, groups:  id, 818
##
## Overdispersion parameter for nbinom2 family (): 2.76
##
## Conditional model:
##              Estimate Std. Error z value             Pr(>|z|)
## (Intercept)  1.543823   0.054276  28.444 < 0.0000000000000002 ***
## time        -0.017680   0.002545  -6.946     0.00000000000377 ***
## male         0.182329   0.082589   2.208             0.027267 *
## time:male    0.013077   0.003710   3.525             0.000423 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##             Estimate Std. Error z value             Pr(>|z|)
## (Intercept) -3.48294    0.28949 -12.031 < 0.0000000000000002 ***
## time         0.07983    0.01550   5.152          0.000000258 ***
## male        -0.54382    0.55137  -0.986                0.324
## time:male    0.01251    0.02819   0.444                0.657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There are two sets of estimated coefficents in the output. The estimates under the "Conditional model" part of the output are the regression coefficients for the count part of the model. We interpret these coefficients as we would interpret coefficients from a standard negative binomial model: the expected number of alcohol-related problems changes by exp(Coef.) for each unit increase in the corresponding predictor. Here again, we could plot the interaction results to help make that interpretation.

The results from the "Zero-inflation model" refer to the logistic model predicting whether or not a student is a certain zero. These are interpreted similar to logistic regression results: the expected log-odds that a student is a "certain zero" changes by Coef. for each unit increase in the corresponding predictor.

In our fitted model, the results from the zero inflated part of the output suggests that gender and the gender by time interaction are not statistically significant predictors of whether or not a student is a certain zero, and could be dropped from the zero inflation part of the model.

```
glmer.8 = glmmTMB(rapi ~ 1 + time + male + time:male + (1 | id), data = rapi,
                  family = nbinom2, zi = ~ 1 + time)

summary(glmer.8)
```

```
##  Family: nbinom2  ( log )
## Formula:          rapi ~ 1 + time + male + time:male + (1 | id)
## Zero inflation:           ~1 + time
## Data: rapi
##
##      AIC      BIC   logLik deviance df.resid
##  19477.9  19527.5  -9731.0  19461.9     3608
##
## Random effects:
##
## Conditional model:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 0.8918   0.9444
## Number of obs: 3616, groups:  id, 818
##
## Overdispersion parameter for nbinom2 family (): 2.76
##
## Conditional model:
##              Estimate Std. Error z value             Pr(>|z|)
## (Intercept)  1.540120   0.054019  28.511 < 0.0000000000000002 ***
## time        -0.018113   0.002485  -7.290     0.00000000000031 ***
## male         0.193247   0.081819   2.362             0.018182 *
## time:male    0.013822   0.003566   3.876             0.000106 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##             Estimate Std. Error z value             Pr(>|z|)
## (Intercept) -3.66421    0.25465 -14.389 < 0.0000000000000002 ***
## time         0.08258    0.01291   6.396        0.00000000016 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Substituting the appropriate 0/1 value for `male` and back-transforming the results of the count model:

- The estimated average number of alcohol-related problems reported at the onset of the study for females is 4.66 ($e^{1.540}$), conditional on the random-effects.
- For females, each month of the intervention is associated with a 1.8% decrease in the number of alcohol-related problems reported, conditional on the random-effects.
- Males report 19.3% more problems than females at the onset of the study, conditional on the random-effects.
- For males, each month of the intervention is associated with a decrease that is 1.3% greater than that of females, conditional on the random-effects.

Again, to help understand these interpretations, we can plot these results.

```
# Get y-hat values (use type="response")
my_data$yhat4 = exp(1.540120 + -0.018113*my_data$time + 0.193247*my_data$male + 0.013822*my_data$time*my_data$male

# Plot the fitted growth patterns
ggplot(data = my_data, aes(x = time, y = yhat4, color = gender)) +
  geom_line() +
  theme_bw() +
  xlab("Time (in months)") +
  ylab("Estimated Number of Alcohol-Related Problems") +
  ggsci::scale_color_d3(name = "")
```

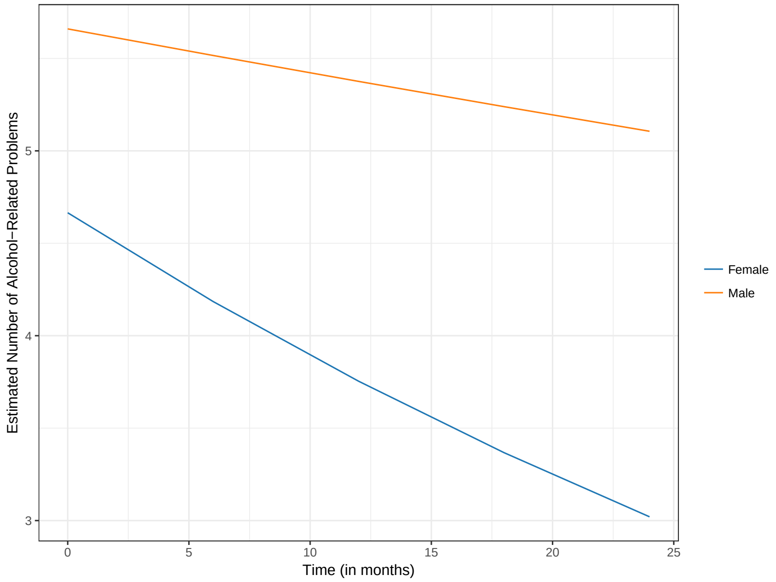


Figure X. Plot of the fitted longitudinal profiles for the count part of the zero-inflated negative binomial model.

To interpret the zero inflated portion of the model, we will write out the fitted logistic model

$$\ln\left[\frac{Pr(\text{Certain Zero})}{Pr(\text{Non Certain Zero})}\right] = -3.664 + 0.083(\text{Time})$$

We can interpret these in the log-odds scale:

- At the study's onset, the log-odds of a student (regardless of gender) being a certain zero is -3.664, conditioned on the random-effects.
- Each month, the predicted log-odds of a student (regardless of gender) being a certain zero increases by 0.083, on average, conditioned on the random-effects.

We can also convert log-odds to odds by exponentiating the coefficients:

- At the study's onset, the odds of a student (regardless of gender) being a certain zero is 0.026, conditioned on the random-effects.
- Each month, the predicted odds of a student (regardless of gender) being a certain zero increases by 8.6%, on average, conditioned on the random-effects.

References

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. doi:10.1037/a00296508