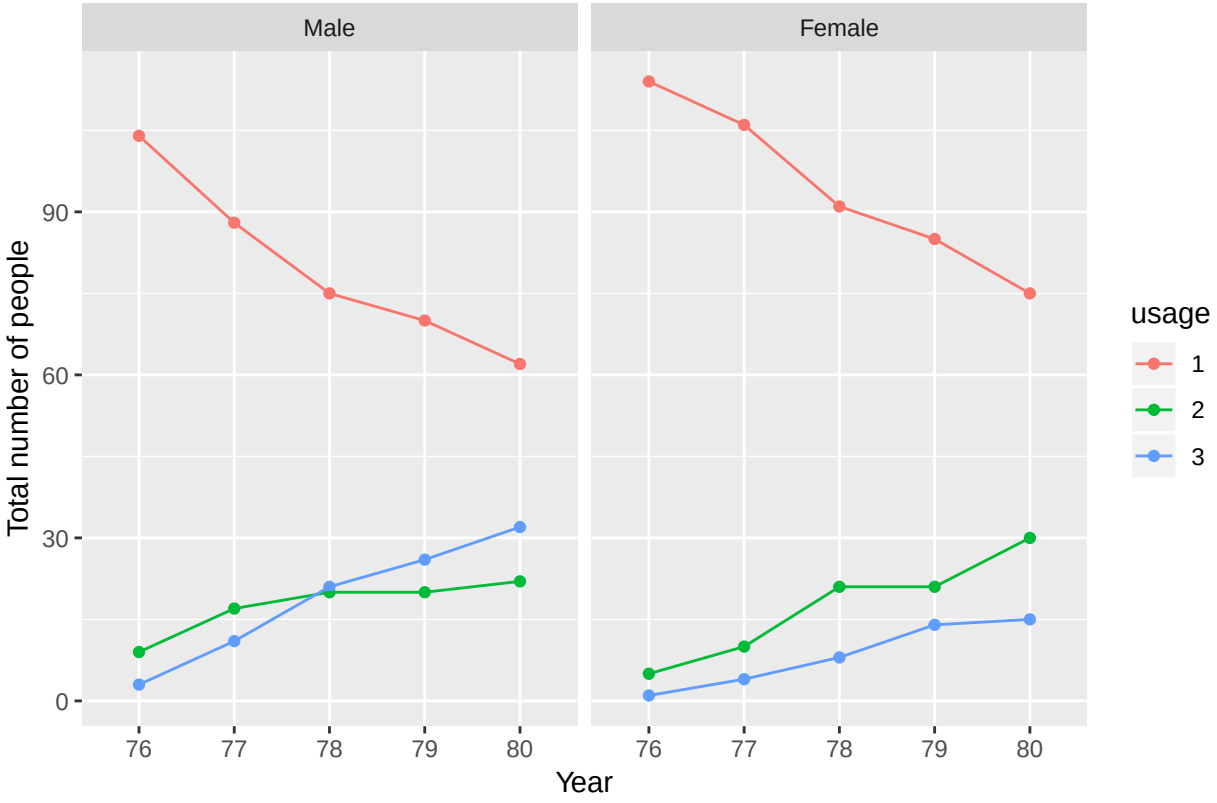
Homework 6

Sarah Adilijiang

Problem 3

(a)

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```



usage: 1=never used, 2=used no more than once a month, 3=used more than once a month.

(b)

```
# condense the levels of the response and change data type
potuse3 = sapply(potuse[,2:6], function(x) ifelse(x==1,0,1))
potuse3 = data.frame(potuse3, sex=potuse$sex, count=potuse$count)

n = sum(potuse$count)
potuse_binary = data.frame(matrix(NA,5*n,4))
colnames(potuse_binary) = c("person", "year", "sex", "usage")
potuse_binary$person = as.factor(rep(1:n, each=5))
potuse_binary$year = rep(76:80, times=n)
potuse_binary$sex = rep(c("Male","Female"), c(5*sum(potuse3$count[potuse3$sex=="Male"]),
                                              5*sum(potuse3$count[potuse3$sex=="Female"])))
potuse_binary$sex = relevel(as.factor(potuse_binary$sex), ref="Male")
```

```r
u = NULL
for (i in 1:nrow(potuse3)) {
    u = c(u, rep(potuse3[i,1:5], times=potuse3$count[i]))
}
potuse_binary$usage = unlist(u)
```

The GLMM model for Binary response here is:
likelihood: $P(usage_i = y_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i},\ y_i = 0, 1$, where p_i is the probability that a person did use marijuana.
link function (logit): $\eta_{ij} = \log \frac{p_i}{1-p_i}$
linear predictor: $\eta_{ij} = \mu + cyear_i + sex_j + cyear_i \times sex_j + \gamma_j^0 + \gamma_j^1 cyear_i$, where i indexes the year and j indexes the individual. $cyear_i$ and sex_j are fixed effects. Random effects $(\gamma_k^0\ \ \gamma_k^1)^T$ i.i.d. $\sim N(0, \sigma^2 D)$, error term ϵ_{ij} i.i.d. $\sim N(0, \sigma^2 I)$, and the random effects are independent with the error term.

```r
# GLMM model
library(lme4)
potuse_binary$cyear = potuse_binary$year - 78
glmm_mod = glmer(usage ~ sex*cyear + (cyear|person), family=binomial, potuse_binary)
summary(glmm_mod)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: usage ~ sex * cyear + (cyear | person)
##    Data: potuse_binary
##
##      AIC      BIC   logLik deviance df.resid
##   1004.4   1039.9   -495.2    990.4     1173
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8271 -0.2120 -0.1329  0.0752  3.2944
##
## Random effects:
##  Groups Name        Variance Std.Dev. Corr
##  person (Intercept) 11.4109  3.3780
##         cyear        0.9869  0.9934   0.74
## Number of obs: 1180, groups:  person, 236
##
## Fixed effects:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.9969     0.4743  -4.210 2.55e-05 ***
## sexFemale        -1.5073     0.6047  -2.493 0.012675 *
## cyear             0.6904     0.1981   3.486 0.000491 ***
## sexFemale:cyear  -0.1067     0.2495  -0.428 0.669012
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) sexFml cyear
## sexFemale   -0.287
## cyear        0.269 -0.202
## sexFeml:cyr -0.085  0.475 -0.355
```

Answer:

Here I centered the **year** at its median value 1978 so that the intercept will represent the predicted **usage** in 1978 thus aiding the interpretations. Also, this helps to reduce the magnitude of the values of **year**, since large numbers are not easy for the underlying fitting algorithm in **glmer** function.

In the model summary, the fixed effect **sex** is significant at 5% significance level, but the interaction term between **sex** and **cyear** is not. Therefore, the difference in general Marijuana usage between male and female is significant, but the difference in increasing rate of usage over year is not significant between male and female.

(c)

```
# test fixed effect: interaction term
 mod = glmer(usage ~ sex*cyear + (cyear|person), family=binomial, potuse_binary)
nmod = glmer(usage ~ sex+cyear + (cyear|person), family=binomial, potuse_binary)
LRT = as.numeric(-2*(logLik(nmod)-logLik(mod)))
1 - pchisq(LRT,1)  # chi-square p-value
```

```
## [1] 0.6621218
```

```
# test fixed effect: sex
 mod = glmer(usage ~ sex+cyear + (cyear|person), family=binomial, potuse_binary)
nmod = glmer(usage ~     cyear + (cyear|person), family=binomial, potuse_binary)
LRT = as.numeric(-2*(logLik(nmod)-logLik(mod)))
1 - pchisq(LRT,1)  # chi-square p-value
```

```
## [1] 0.005572652
```

Answer:

In the LRT test for dropping the interaction betweeen **sex** and **cyear**, the p-value > 0.05, so we do not reject the null model at 5% significance level. Thus the interaction term is not significant, we can drop it from the larger model.

In the LRT test for dropping **sex**, the p-value < 0.05, so we reject the null model at 5% significance level. Thus we **sex** is a significant fixed effect so we shoule keep it in the model. Therefore, our final model is:

$$\log \frac{p_i}{1-p_i} = \mu + cyear_i + sex_j + \gamma_j^0 + \gamma_j^1 cyear_i$$

where i indexes the year and j indexes the individual. $cyear_i$ and sex_j are fixed effects. Random effects $(\gamma_k^0 \ \ \gamma_k^1)^T$ i.i.d. $\sim N(0, \sigma^2 D)$, error term ϵ_{ij} i.i.d. $\sim N(0, \sigma^2 I)$, and the random effects are independent with the error term.

(d)

Here we use the final model from (c) to compare two models. But the computation of the algorithm for this final model with **cyear** as a factor is too time consuming and the algorithm do not converge. Therefore, we simplify this final model by removing the random slope term. So now the model is:

$$\log \frac{p_i}{1-p_i} = \mu + cyear_i + sex_j + \gamma_j$$

where i indexes the year and j indexes the individual. $cyear_i$ and sex_j are fixed effects. Random effects γ_j i.i.d. $\sim N(0, \sigma^2 D)$, error term ϵ_{ij} i.i.d. $\sim N(0, \sigma^2 I)$, and the random effect is independent with the error term.

```
# final model with "cyear" as a numerical linear term
glmm_num = glmer(usage ~ sex+cyear + (1|person), family=binomial, potuse_binary)
summary(glmm_num)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: usage ~ sex + cyear + (1 | person)
##    Data: potuse_binary
##
##      AIC      BIC   logLik deviance df.resid
##   1013.3   1033.6   -502.7   1005.3     1176
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.6905 -0.3186 -0.1278  0.1743  6.2645
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  person (Intercept) 7.508    2.74
## Number of obs: 1180, groups:  person, 236
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.84867    0.33285  -5.554 2.79e-08 ***
## sexFemale   -1.09035    0.44554  -2.447   0.0144 *
## cyear        0.91295    0.08828  10.341  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) sexFml
## sexFemale -0.556
## cyear     -0.252 -0.116
```

```r
# final model with "cyear" as a factor
potuse_binary$year_f = as.factor(potuse_binary$year)
glmm_fac = glmer(usage ~ sex+year_f + (1|person), family=binomial, potuse_binary)
summary(glmm_fac)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: usage ~ sex + year_f + (1 | person)
##    Data: potuse_binary
##
##      AIC      BIC   logLik deviance df.resid
##   1007.0   1042.5   -496.5    993.0     1173
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.4056 -0.2825 -0.1620  0.1498  8.9810
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  person (Intercept) 7.876    2.807
## Number of obs: 1180, groups:  person, 236
##
## Fixed effects:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -4.3878     0.5062  -8.668  < 2e-16 ***
## sexFemale    -1.1196     0.4554  -2.459   0.0139 *
## year_f77      1.7079     0.4072   4.194 2.74e-05 ***
## year_f78      3.0247     0.4175   7.245 4.31e-13 ***
## year_f79      3.4641     0.4254   8.144 3.82e-16 ***
## year_f80      4.1368     0.4404   9.392  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) sexFml yr_f77 yr_f78 yr_f79
## sexFemale -0.315
## year_f77  -0.602 -0.050
## year_f78  -0.682 -0.085  0.699
## year_f79  -0.700 -0.096  0.699  0.773
## year_f80  -0.718 -0.110  0.692  0.779  0.796
```

```
# standard likelihood-based method to construct a chi-squared test
anova(glmm_num, glmm_fac)
```

```
## Data: potuse_binary
## Models:
## glmm_num: usage ~ sex + cyear + (1 | person)
## glmm_fac: usage ~ sex + year_f + (1 | person)
##          Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## glmm_num  4 1013.3 1033.6 -502.67  1005.33
## glmm_fac  7 1007.0 1042.5 -496.48   992.96 12.372      3   0.006213 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

(1) Model Comparison:

From the summary of two models, we can see that the log-likelihood of the models are: -502.7 for linear **year** effect and -496.5 for factor **year** effect. So the factor model is slightly better due to a higher log-likelihood value. The same result is shown in the chi-squared test based on the standard likelihood, which shows a low p-value preferring the factor model.

Also, in the summary, the AIC values of the models are: 1013.3 for linear **year** effect and 1007.0 for factor **year** effect. So again the factor model is slightly preferred in terms of having a lower AIC value. However, the BIC values of the models are: 1033.6 for linear **year** effect and 1042.5 for factor **year** effect. Thus the linear model is slightly preferred in terms of having a lower BIC value. BIC does not agree with AIC because BIC method penalizes model complexity more heavily.

To sum up, the likelihood based chi-squared test and AIC method both prefer the factor model, while the BIC method prefers the linear model. Therefore, it is not very clear which one fits the data better and should be clearly preferred from this point of view.

On the other hand, we can think the linear **year** model as a submodel of the factor **year** model. In the factor **year** model, each year has its own parameter, so it allows different increasements between each two adjacent years. But in the linear **year** model, all the years has only one parameter, so it assumes an identical increasement between all the two adjacent years. From the plot in question (a), we can see that the increasements between each two adjacent years are actually slightly different for box male and female, but the difference is not that significant and can be modeled as a linear relationship.

Furthermore, thinking of the model complexity, when making **year** as a factor, it increases the degrees of the

freedom of the model, especially when adding the random slope term which results in the non-convergence of the model. So the linear model is better in fitting the model and would not make a too complex model.

Therefore, as a result, I would say the model with **year** as a linear term would be slightly preferred here.

(2) Factor Model Interpretation:

The estimate of the fixed effect intercept term is -4.3878, which means that the Marijuana usage odds of a male in year 1976 is exp(-4.3878)=0.0124.

The estimate of the fixed effect **sexFemale** is -1.1196, which means that the Marijuana usage odds of a female is exp(-1.1196)=0.3264 times as the odds of a male in the same year.

The estimate of the fixed effect **year_f77** is 1.7079, which means that the Marijuana usage odds of a person in year 1977 is exp(1.7079)=5.5174 times as the odds of a person in year 1976 when controlling for sex. The interpretation is similar for other levels of the year.

The estimate of the intercept random effect SD is 2.807, which represents the variation (SD) in overall Marijuana usage odds between individuals is exp(2.807)=16.5602.