BIOST540HW3

Bryan Ng, 2427348

2025-05-28

Problme 1

a)

```
## Rows: 1101 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): sex_dv
## dbl (4): pidp, wave, age_dv, scghq1_dv
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: scghq1_dv ~ age_dv * sex + (1 | pidp)
##    Data: df
##
## REML criterion at convergence: 6549.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5497 -0.4924 -0.1191  0.3433  4.8971
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  pidp     (Intercept) 19.05    4.364
##  Residual             16.07    4.009
## Number of obs: 1101, groups:  pidp, 188
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)      12.31082    1.75825   7.002
## age_dv           -0.03155    0.03008  -1.049
## sexFemale         0.84137    2.25175   0.374
## age_dv:sexFemale  0.01690    0.03894   0.434
##
## Correlation of Fixed Effects:
##             (Intr) age_dv sexFml
## age_dv      -0.947
## sexFemale   -0.781  0.739
## ag_dv:sxFml  0.732 -0.772 -0.947
```

```
## refitting model(s) with ML (instead of REML)

## Data: df
## Models:
## m.noInt: scghq1_dv ~ age_dv + sex + (1 | pidp)
## m.lmm: scghq1_dv ~ age_dv * sex + (1 | pidp)
##         npar    AIC    BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
## m.noInt    5 6549.8 6574.8 -3269.9    6539.8
## m.lmm      6 6551.6 6581.6 -3269.8    6539.6 0.1864  1     0.6659
```

The age×sex interaction estimate is

$$\hat{\beta}_{\text{age}\times\text{sex}} = 0.017 \quad (\text{SE} = 0.039,\ t \approx 0.43),$$

and the likelihood-ratio test comparing models with vs. without this term yields

$$\chi^2(1) = 0.186,\ p = 0.67.$$

Therefore, we fail to reject the null hypothesis of no interaction: the decline in well-being with age is essentially the same for Males and Females in this sample.

b)

We fitted three GEE models with independence, exchangeable, and AR(1) working correlations and compared them using QIC.

```
##
## Call:
## geeglm(formula = scghq1_dv ~ age_dv * sex, family = gaussian,
##     data = df, id = pidp, corstr = "independence")
##
##  Coefficients:
##                   Estimate   Std.err   Wald Pr(>|W|)
## (Intercept)      12.673395  1.433761 78.133   <2e-16 ***
## age_dv           -0.037437  0.026034  2.068    0.150
## sexFemale         1.311518  2.207334  0.353    0.552
## age_dv:sexFemale  0.002707  0.037675  0.005    0.943
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = independence
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)       34   4.401
## Number of clusters:   188  Maximum cluster size: 9

##
## Call:
## geeglm(formula = scghq1_dv ~ age_dv * sex, family = gaussian,
##     data = df, id = pidp, corstr = "exchangeable")
##
##  Coefficients:
```

```
##                  Estimate Std.err  Wald Pr(>|W|)
## (Intercept)       12.3099  1.4466 72.41   <2e-16 ***
## age_dv            -0.0315  0.0264  1.42     0.23
## sexFemale          0.8845  2.0736  0.18     0.67
## age_dv:sexFemale   0.0160  0.0366  0.19     0.66
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     34.1    4.42
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.518  0.0644
## Number of clusters:   188  Maximum cluster size: 9

##
## Call:
## geeglm(formula = scghq1_dv ~ age_dv * sex, family = gaussian,
##     data = df, id = pidp, corstr = "ar1")
##
##  Coefficients:
##                  Estimate Std.err  Wald Pr(>|W|)
## (Intercept)       12.1943  1.5226 64.14  1.1e-15 ***
## age_dv            -0.0272  0.0272  1.00     0.32
## sexFemale          0.7383  2.0761  0.13     0.72
## age_dv:sexFemale   0.0144  0.0364  0.16     0.69
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = ar1
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     34.1    4.42
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.799  0.0388
## Number of clusters:   188  Maximum cluster size: 9
```

We observed that all three specifications lead to the same conclusion: a modest, non-significant decline in well-being with age and no evidence of a sex difference in that slope.

c)

Comparing the LMM and GEE results shows that all estimated effects are virtually identical: a modest negative slope for age, a small positive main effect of sex, and a near-zero age×sex interaction; and in none of the models is the age×sex interaction statistically significant.

Problem 2

```
## New names:
## Rows: 2148 Columns: 6
## -- Column specification
## -------------------------------------------------------- Delimiter: "," dbl
## (6): ...1, id, resp, age, smok, aXs
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

GLMM assume that conditional independence given the random intercept; random effects $\sim N(0, \sigma^2)$.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: resp ~ age + smok + aXs + (1 | id)
##    Data: df
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##      1599      1628      -795      1589      2143
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -1.399 -0.178 -0.159 -0.128  2.602
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 5.5      2.35
## Number of obs: 2148, groups:  id, 537
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.4017     0.2789  -12.20   <2e-16 ***
## age          -0.2170     0.0868   -2.50    0.012 *
## smok          0.4782     0.2993    1.60    0.110
## aXs           0.1046     0.1391    0.75    0.452
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr) age    smok
## age   0.272
## smok -0.442 -0.193
## aXs  -0.146 -0.621  0.280
```

GEE assume that correct specification of the marginal mean.

```
##
## Call:
## geeglm(formula = resp ~ age + smok + aXs, family = binomial,
##     data = df, id = id, corstr = "exchangeable")
##
```

```
##  Coefficients:
##             Estimate Std.err   Wald Pr(>|W|)
## (Intercept)  -1.9005  0.1191 254.69   <2e-16 ***
## age          -0.1412  0.0582   5.89    0.015 *
## smok          0.3138  0.1878   2.79    0.095 .
## aXs           0.0708  0.0883   0.64    0.422
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    0.999   0.114
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.355   0.063
## Number of clusters:   537  Maximum cluster size: 4
```

Transition model assume that each outcome depends on the immediate past outcome.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: resp ~ prev_resp + age + smok + aXs + (1 | id)
##    Data: df2
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##      1596      1630      -792      1584      2142
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -1.413 -0.217 -0.191 -0.146  2.868
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 3.46     1.86
## Number of obs: 2148, groups:  id, 537
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.1337     0.2448  -12.80   <2e-16 ***
## prev_resp     0.6707     0.2950    2.27   0.0230 *
## age          -0.2636     0.0872   -3.02   0.0025 **
## smok          0.4127     0.2592    1.59   0.1113
## aXs           0.0840     0.1355    0.62   0.5351
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) prv_rs age    smok
## prev_resp  0.386
```

```
## age        0.193 -0.248
## smok      -0.474 -0.104 -0.193
## aXs       -0.185 -0.062 -0.585  0.337
```

In summary, the three approaches give subtly different smoking coefficients because they rest on different conditioning and correlation assumptions: the GLMM estimates a subject-specific effect of $\beta_{\text{smok}} = 0.478$ by modeling a normal random intercept; the GEE yields a marginal effect of $\beta_{\text{smok}} = 0.314$ using an exchangeable working correlation and robust "sandwich" variances; and the first-order Markov transition model—by also conditioning on the immediately prior outcome plus a random intercept—produces $\beta_{\text{smok}} = 0.413$.

The GLMM's subject-specific odds ratio of

$$\exp(\beta_{\text{smok}}) = \exp(0.478) \approx 1.61$$

tells you that, for a given individual holding their baseline risk constant, picking up smoking raises their odds of the outcome by 61%—a clear, personalized message to motivate cessation.

At the population level, the GEE's marginal OR of

$$\exp(\beta_{\text{smok}}) = \exp(0.314) \approx 1.37$$

means that, on average across everyone, smokers have 37% higher odds, so cutting smoking prevalence by, say, 10% could yield roughly a 13% drop in overall disease odds—insight critical for public-health policy.

And if you're building a longitudinal risk tool that updates at each visit, the first-order Markov transition model's OR of

$$\exp(\beta_{\text{smok}}) = \exp(0.413) \approx 1.51$$

indicates that, conditional on someone's last outcome, smoking boosts their next visit's odds by 51%, which helps you dynamically recalibrate risk as history unfolds.

Problem 3

a)

Treatment	Visit	Mean rate (seizures/week)	SD	N
placebo	Baseline (8 wk)	3.85	3.26	28
placebo	Visit 1 (2 wk)	4.68	5.07	28
placebo	Visit 2 (2 wk)	4.39	7.34	28
placebo	Visit 4 (2 wk)	3.98	3.81	28
progabide	Baseline (8 wk)	3.95	3.50	31
progabide	Visit 1 (2 wk)	4.29	9.12	31
progabide	Visit 2 (2 wk)	4.21	5.93	31
progabide	Visit 3 (2 wk)	4.06	6.95	31
progabide	Visit 4 (2 wk)	3.35	5.63	31

Table 1: Seizure rates by treatment arm and visit

b)

```
##
## Call:
## geeglm(formula = seizures ~ tx + age, family = poisson(link = "log"),
##     data = df_long, offset = log(weeks), id = id, corstr = "exchangeable")
##
##  Coefficients:
##             Estimate Std.err  Wald Pr(>|W|)
## (Intercept)   2.2394  0.4461 25.20  5.2e-07 ***
## txprogabide   0.0292  0.2049  0.02    0.887
## age          -0.0332  0.0147  5.10    0.024 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     19.8    8.25
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.764  0.0867
## Number of clusters:   59  Maximum cluster size: 5
```

Estimated treatment effect:

$$\hat{\beta}_{\text{txprogabide}} = 0.0292,\ \text{SE} = 0.2049,\ p = 0.887, \quad \text{IRR} = e^{0.0292} \approx 1.03 \text{ (with 95\% CI 0.69–1.54).}$$

Estimated age effect:

$$\hat{\beta}_{\text{age}} = -0.0332,\ \text{SE} = 0.0147,\ p = 0.024, \quad \text{IRR} = e^{-0.0332} \approx 0.97 \text{ (with 95\% CI 0.94–0.996).}$$

Estimated working correlation:

$$\hat{\alpha} = 0.764.$$

Under this GEE, we observed that there is no significant treatment effect of progabide on seizure rates, but patient age does modestly reduce seizure incidence.

Appendix: R Code

```
library(lme4)
library(geepack)
library(readr)
library(dplyr)
library(tidyverse)

#Problem 1 a)
df <- read_csv("C:/Users/ncwbr/Desktop/uk_sub.csv")
df$sex <- relevel(factor(df$sex_dv), ref="Male")

m.lmm <- lmer(scghq1_dv ~ age_dv * sex + (1 | pidp), data = df)
summary(m.lmm)

m.noInt <- update(m.lmm, . ~ . - age_dv:sex)
anova(m.noInt, m.lmm)

#Problem 1 b)

gee.ind <- geeglm(scghq1_dv ~ age_dv * sex,
                  id = pidp, corstr = "independence",
                  data = df, family = gaussian)

gee.exc <- update(gee.ind, corstr = "exchangeable")

gee.ar1 <- update(gee.ind, corstr = "ar1")

summary(gee.ind)
summary(gee.exc)
summary(gee.ar1)

#Problem 2

df <- read_csv("C:/Users/ncwbr/Desktop/sixcity.csv")
m.glmm <- glmer(resp ~ age + smok + aXs + (1 | id),
                data = df, family = binomial)

summary(m.glmm)
m.gee <- geeglm(resp ~ age + smok + aXs,
                id       = id,
                data     = df,
                family   = binomial,
                corstr   = "exchangeable")

summary(m.gee)
df2 <- df %>%
  arrange(id, age) %>%
  group_by(id) %>%
  mutate(prev_resp = lag(resp, default=0)) %>%
  ungroup()

m.trans <- glmer(resp ~ prev_resp + age + smok + aXs + (1 | id),
```

```
                data = df2, family = binomial)

summary(m.trans)

#Problem 3 a)

df <- read_csv("C:/Users/ncwbr/Desktop/Seizure.csv")
df_long <- df %>%
  pivot_longer(y0:y4, names_to="visit", values_to="seizures") %>%
  mutate(
    weeks     = if_else(visit=="y0", 8, 2),
    rate      = seizures / weeks,
    visit_lab = factor(visit,
                       levels=c("y0","y1","y2","y3","y4"),
                       labels=c("Baseline\n(8 wk)", "Visit 1\n(2 wk)",
                                "Visit 2\n(2 wk)", "Visit 3\n(2 wk)", "Visit 4\n(2 wk)"))
  )

summary_tbl <- df_long %>%
  group_by(tx, visit_lab) %>%
  summarize(
    mean_rate  = mean(rate),
    sd_rate    = sd(rate),
    n          = n(),
    .groups     = "drop"
  )
knitr::kable(summary_tbl,
             col.names=c("Treatment","Visit","Mean rate\n(seizures/week)","SD","N"),
             digits=2)
             gee_mod <- geeglm(
  seizures ~ tx + age,
  offset = log(weeks),
  id     = id,
  data   = df_long,
  family = poisson(link="log"),
  corstr = "exchangeable"
)

summary(gee_mod)
```