2019 Q1

codename

(Note: For clarity of the report, part of the output results are hidden. And the utilized packages are as shown in the Appendix.)

(a)

```
d <- read.csv("cholesterol.csv", header = T)
d <- within(d , {
  treat <- as.factor(treat)
  drug <- as.factor(drug)
  statin <- as.factor(statin)
  gender <- as.factor(gender)
  race <- as.factor(race)
  ldl.chg <- ldl - ldl_0
  ldl.pchg <- ldl.chg / ldl_0 * 100
})
```

The experiment is randomized, so it would fine to use either unadjusted model or adjusted model. Here, I use adjusted model for variance reduction. This is a longitudinal data, so we should fit a linear mixed model, with fixed effects for the factorial treatment structure and with a random intercept per participant to model within-subject correlation. After model selection, the treatment effects are estimated by using contrasts.

As for the response, I try both "change in LDL from baseline" and "percent change in LDL", and use residual plots to tell which one is better. Figure 1a-1c show the mean ldl, mean change in ldl, mean percent change in ldl by treatment.

Since we are supposed to estimate the effects of the treatment, the other covariates should be additive. Start with a full model, and use p-value 0.1 as the cutoff to select variables, and then use `anova` to validate model choices. Since `gender` and `female` have the same effect, I would only include `gender` in the model. Also, I treat `visit` as a factor. (Using `correlation = corAR1()` would lead to convergence error, so I just use identity correlation structure.)

Starting with a general model, try using both percent change and change from baseline as the response, and as shown in Figure 2a - 2c, the residuals show some problem when using percent change. And when using change from baseline, the residuals look fine, as shown in Figure 3a - 3c. So for the rest of the analysis, I use `ldl.chg` as the response.

```
options(contrasts = c("contr.treatment", "contr.poly"))
da <- d %>% filter(visit > 0)
# try using percent change as the response
ma.lme0 <- lme(ldl.pchg ~ drug * statin + factor(visit) + ldl_0 + age +
                gender + race + smoking + disease_yrs, random = ~1|id,
              data = da)
Anova(ma.lme0)
# try using change from baseline as the response
ma.lme1 <- lme(ldl.chg ~ drug * statin + factor(visit) + ldl_0 + age +
                gender + race + smoking + disease_yrs, random = ~1|id,
              data = da)

Anova(ma.lme1)
```

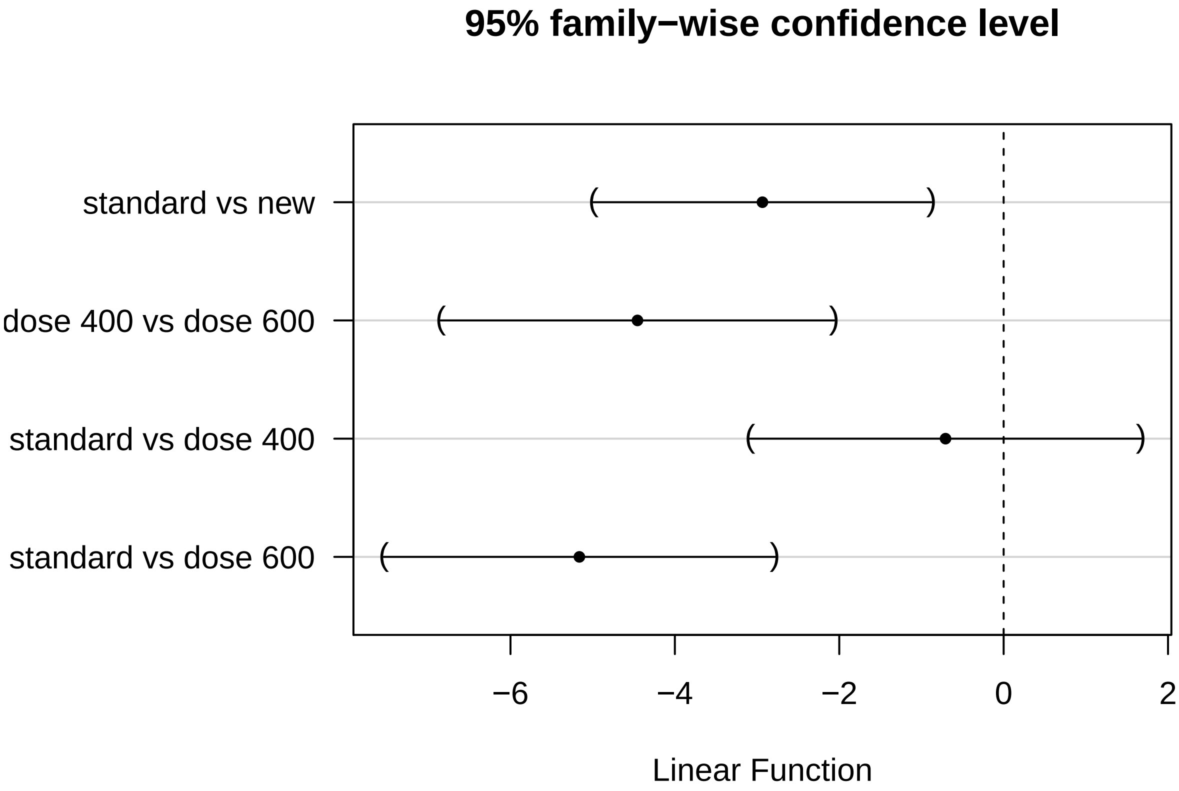
```
ma.lme2 <- update(ma.lme1, method = "ML")
ma.lme3 <- update(ma.lme2, fixed = . ~ drug * statin + factor(visit) +
                    ldl_0 + gender)
ma.lme4 <- update(ma.lme2, fixed = . ~ drug + statin + factor(visit) +
                    ldl_0 + gender)
anova(ma.lme2, ma.lme3, ma.lme4)
```

```
        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
ma.lme2     1 19 18976.12 19084.28 -9469.058                        
ma.lme3     2 13 18966.85 19040.85 -9470.423 1 vs 2 2.728863  0.8420
ma.lme4     3 11 18968.92 19031.54 -9473.461 2 vs 3 6.076951  0.0479
```

The interaction effect is borderline significant, so there can be two strategies to estimate the treatment effect. First, remove the interaction effect, and analyze the drug effect and the statin effect separately. Second, use `treat` to refit the model and consider the effect of drug at different levels of statin, and the effect of statin at different levels of drug.

If we use the first strategy, then we should use `ma.lme4`. Also, in order to adjust for multiple comparison, `glht{multcomp}` is used to estimate the contrasts for `drug`. And for `statin`, use `fit.contrast` to estimate its effect. We can also use `intervals()`, since there are only two levels, but do note that when using `intervals()`, we need to make sure the options for contrasts is `contr.treatment`.

```
m4a <- update(ma.lme4, method = "REML")
contr <- rbind("standard vs new" = c(1, -0.5, -0.5),
               "dose 400 vs dose 600" = c(0, 1, -1),
               "standard vs dose 400" = c(1, -1, 0),
               "standard vs dose 600" = c(1, 0, -1))
g <- glht(m4a, linfct = mcp(drug = contr), alternative = "two.sided")
par(mar=c(5.1, 9.1, 4.1, 2.1))
plot(g)
```



```
par(mar=c(5.1, 4.1, 4.1, 2.1))
summary(g)
```

```
...
                          Estimate Std. Error z value Pr(>|z|)    
standard vs new == 0       -2.9345     0.8679  -3.381  0.00251 ** 
dose 400 vs dose 600 == 0  -4.4552     1.0102  -4.410  < 0.001 ***
standard vs dose 400 == 0  -0.7070     1.0040  -0.704  0.76609    
standard vs dose 600 == 0  -5.1621     1.0043  -5.140  < 0.001 ***
...
```

```r
confint(g)
```

```
...
                          Estimate lwr     upr
standard vs new == 0      -2.9345  -4.9927 -0.8764
dose 400 vs dose 600 == 0 -4.4552  -6.8509 -2.0594
standard vs dose 400 == 0 -0.7070  -3.0879  1.6740
standard vs dose 600 == 0 -5.1621  -7.5439 -2.7803
...
```

```r
# estimate the effect of statin
fit.contrast(m4a, "statin", coeff = c(1, -1), conf.int = 0.95)
```

```
                  Estimate Std. Error  t-value Pr(>|t|) lower CI upper CI
statin c=( 1 -1 ) 13.07006  0.8218642 15.90294        0 11.45594 14.68417
attr(,"class")
[1] "fit_contrast"
```

If we use the second strategy, we need to refit the model with `treat` and do multiple comparison to estimate effect of drug at different levels of statin, and the same for statin. For convenience, I would go for the first strategy.

```r
m5a <- update(m4a, fixed = .~ treat + factor(visit) + ldl_0 + gender)
contr <- rbind("standard vs new(no statin)" = c(1, -0.5, -0.5, 0, 0, 0),
               "dose 400 vs dose 600(no statin)" = c(0, 1, -1, 0, 0, 0),
               "standard vs dose 400(no statin)" = c(1, -1, 0, 0, 0, 0),
               "standard vs dose 600(no statin)" = c(1, 0, -1, 0, 0, 0),
               "standard vs new(with statin)" = c(0, 0, 0, 1, -0.5, -0.5),
               "dose 400 vs dose 600(with statin)" = c(0, 0, 0, 0, 1, -1),
               "standard vs dose 400(with statin)" = c(0, 0, 0, 1, -1, 0),
               "standard vs dose 600(with statin)" = c(0, 0, 0, 1, 0, -1))
g <- glht(m5a, linfct = mcp(treat = contr), alternative = "two.sided")
summary(g)
confint(g)

contr <- rbind("no vs statin(standard)" = c(-1, 0, 0, 1, 0, 0),
               "no vs statin(dose 400)" = c(0, -1, 0, 0, 1, 0),
               "no vs statin(dose 600)" = c(0, 0, -1, 0, 0, 1))
g <- glht(m5a, linfct = mcp(treat = contr), alternative = "two.sided")
summary(g)
confint(g)
```

(b)

For this part, we need to refit the model with interaction terms included. As suggested in part (a), the interaction between `drug` and `statin` is not significant, so I will not include it in the model. Starting with a large plausible model, I then do manual model selection, using 0.05 as a cutoff to select significant terms.

```r
mb.lme1 <- lme(ldl.chg ~ (drug + statin) * (visit + ldl_0 + age + gender +
                                              race + smoking + disease_yrs),
               random = ~1|id, data = da, method = "ML")
Anova(mb.lme1)
```

```
...
Analysis of Deviance Table (Type II tests)
```

```
drug:smoking         2.8038  2   0.246123
drug:disease_yrs     2.7427  2   0.253766
statin:visit         5.9357  1   0.014837 *
statin:ldl_0        16.7102  1  4.355e-05 ***
statin:age           3.4509  1   0.063216 .
statin:gender        0.3172  1   0.573322
statin:race          0.4441  3   0.930995
...
```

```
mb.lme2 <- update(mb.lme1, fixed = .~ drug * visit + statin * visit +
                    statin * ldl_0  + gender)
anova(mb.lme1, mb.lme2)
```

```
        Model df      AIC      BIC    logLik   Test L.Ratio p-value
mb.lme1     1 42 18983.53 19222.62 -9449.767
mb.lme2     2 13 18947.75 19021.76 -9460.876 1 vs 2 22.2176  0.8111
```

```
Anova(mb.lme2)
```

```
Analysis of Deviance Table (Type II tests)

Response: ldl.chg
                Chisq Df Pr(>Chisq)
drug          34.2667  2  3.623e-08 ***
visit          3.6696  1    0.05542 .
statin       263.8305  1  < 2.2e-16 ***
ldl_0         68.6306  1  < 2.2e-16 ***
gender        19.1600  1  1.202e-05 ***
drug:visit     8.2683  2    0.01602 *
visit:statin   5.8820  1    0.01530 *
statin:ldl_0  15.9476  1  6.512e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fixef(mb.lme2)
```

```
  (Intercept)         drug2         drug3         visit       statin1
  20.54425747    4.84704717   11.16528704   -0.01848083   -6.18423585
        ldl_0       gender2   drug2:visit   drug3:visit visit:statin1
  -0.05053141   -3.91674930   -0.14079164   -0.20427423    0.14515801
statin1:ldl_0
  -0.09920914
```

In `mb.lme1` results, the p-value for `stain:gender` is 0.573322 and the one for `statin:ldl_0` is 4.355e-05. Comparing `mb.lme1` and `mb.lme2`, there is no significant difference, so the effect of statins doesn't differ between men and women, but differ by baseline of LDL. The estimated fixed effect is -0.09920914, and it means that if using statin, with every unit increase in `ldl_0`, the `ldl.chg` would decrease -0.09920914 unit more than the case without statin.

(c)

A less conservative alternative is the FDR method of Benjamin & Hochberg. The idea is: start with the largest p-value, work down, and reject $H_{0(j)}$ if $p_{(j)} \leq \frac{\epsilon j}{K}$, and $K = 6$ in this case. This method controls false discovery rate and is less conservative than Bonferroni. From the results, we can see that $p_1 = 0.002, p_4 = 0.01$ remain significant after the adjustment.

```
p <- c(0.002, 0.25, 0.04, 0.01, 0.08, 0.6)
p.ordered <- sort(p)
threshold <- 0.05 * c(1:6) / 6
```

```
p.ordered[which(p.ordered <= threshold)]
```

```
[1] 0.002 0.010
```

(d)

In this part, the response is binary, so we need to fit a generalized linear mixed model. Also, since the new response is based on `ldl`, not `ldl.chg`, we need to create a new variable `fu` and use `statin:fu` to estimate the effect of statin use. There are multiple candidate models to use: `glmer`, `glmmPQL`, and `geeglm`. Since the results from GEE is usually more stable, I use `geeglm` here, and include other covariates to reduce variance.

```
dd <- d %>% mutate(ldl.high = ifelse(ldl > 150, 1, 0),
                   fu = ifelse(visit > 0, 1, 0))
md.gee1 <- geeglm(ldl.high ~ drug * statin * fu + age +
                    gender + race + smoking + disease_yrs, id = id,
                  family = binomial, data = dd, corstr = "exchangeable")
summary(md.gee1)
md.gee2 <- geeglm(ldl.high ~ drug * statin + drug * fu + statin * fu +
                    age + race, id = id,
                  family = binomial, data = dd, corstr = "exchangeable")
md.gee3 <- geeglm(ldl.high ~ drug * statin + statin * fu +
                    age + race, id = id,
                  family = binomial, data = dd, corstr = "exchangeable")
```

```
anova(md.gee1, md.gee2)
```

```
Analysis of 'Wald statistic' Table

Model 1 ldl.high ~ drug * statin * fu + age + gender + race + smoking + disease_yrs
Model 2 ldl.high ~ drug * statin + drug * fu + statin * fu + age + race
  Df   X2 P(>|Chi|)
1  5 3.29      0.66
```

```
anova(md.gee2, md.gee3)
```

```
Analysis of 'Wald statistic' Table

Model 1 ldl.high ~ drug * statin + drug * fu + statin * fu + age + race
Model 2 ldl.high ~ drug * statin + statin * fu + age + race
  Df    X2 P(>|Chi|)
1  2 0.149      0.93
```

```
summary(md.gee3)
```

```
...
              Estimate  Std.err   Wald Pr(>|W|)
(Intercept)   -4.14760  0.29895 192.49  < 2e-16 ***
drug2          0.09084  0.16455   0.30   0.5809
drug3         -0.16606  0.16533   1.01   0.3152
statin1       -0.42497  0.30296   1.97   0.1607
fu             0.96805  0.19510  24.62  7.0e-07 ***
age            0.05007  0.00506  97.79  < 2e-16 ***
race2          0.18511  0.18609   0.99   0.3199
race3          1.42659  0.18969  56.56  5.5e-14 ***
race4          0.33210  0.12356   7.22   0.0072 **
drug2:statin1 -0.13607  0.26785   0.26   0.6114
```

```
drug3:statin1  0.64135  0.25300   6.43   0.0112 *
statin1:fu    -0.67473  0.28548   5.59   0.0181 *
...
```

The use of statins indeed decreases the prevalence of LDL>150mg/dL, and the p-value is 3.3e-06. The point estimate of the effect is -1.47731, with 95% CI being [-2.100, -0.855]. Note that this is a binomial model, so the effects represents the ratio of odds, and it means that using statins will reduce the odds of high LDL to 22.8% of the original level, namely, reduce by 77.2%.

```r
c(-0.67473 - 1.96 * 0.28548, -0.67473 + 1.96 * 0.28548)
```

```
[1] -1.234 -0.115
```

```r
exp(-0.67473)
```

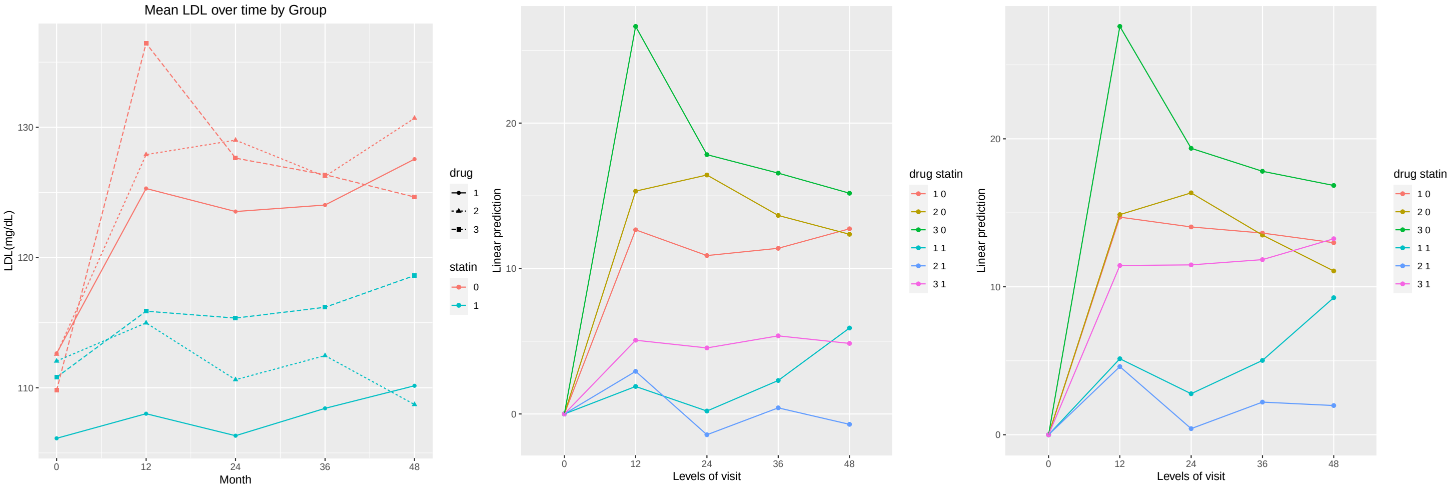
```
[1] 0.509
```

Appendix

Packages

```r
library(gmodels)
library(MASS)
library(car)
library(ggplot2)
library(dplyr)
library(emmeans)
library(nlme)
library(geepack)
library(emmeans)
library(multcomp)
```

Figures

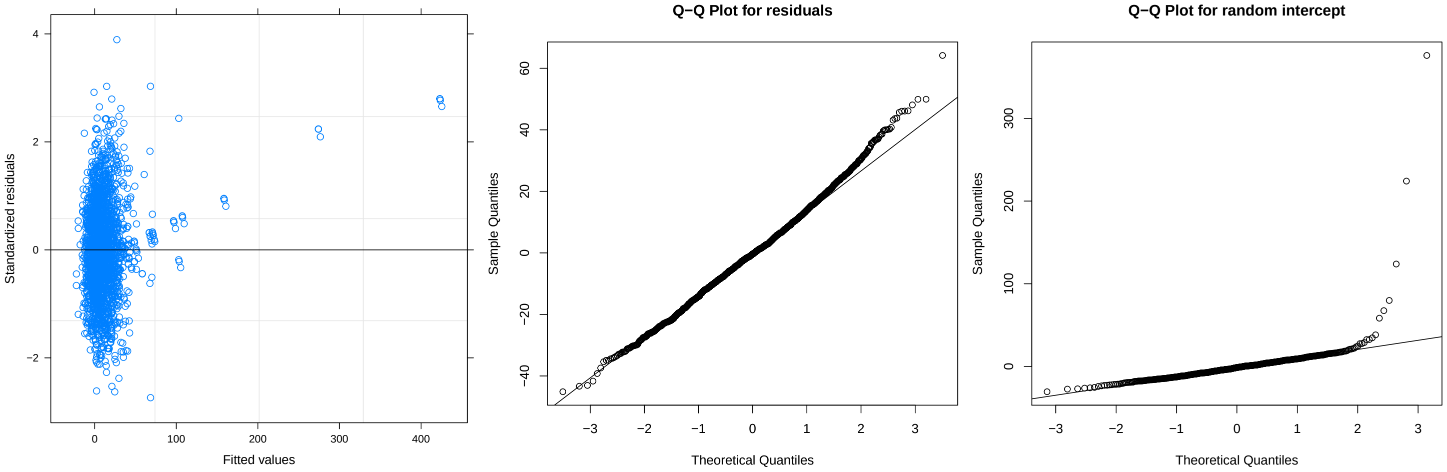


(a) residual plot of ma.lme0

(b) qqnorm of residuals

(c) qqnorm of random effects of ma.lme0

Figure 1: Model Diagnostics

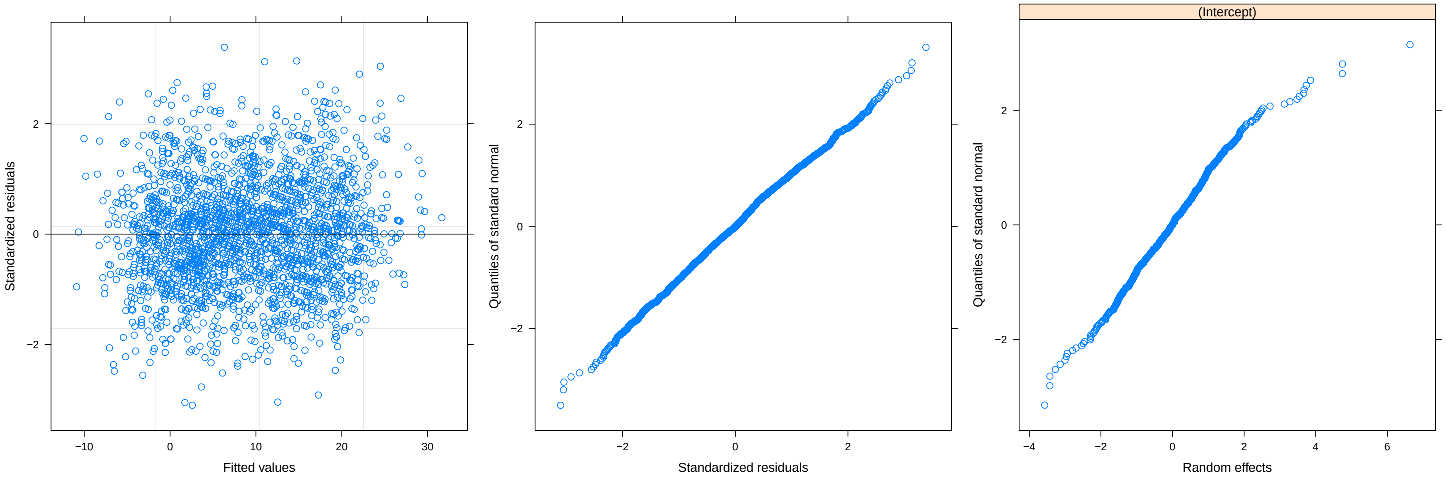


(a) residual plot of ma.lme0

(b) qqnorm of residuals

(c) qqnorm of random effects of ma.lme0

Figure 2: Model Diagnostics



(a) residual plot of ma.lme1

(b) qqnorm of residuals

(c) qqnorm of random effects of ma.lme1

Figure 3: Model Diagnostics