2008 Q2

November

(a)

This is a 2^k fractional factorial design, so we need to specify the contrast to be `contr.sum`.

```
# read and preprocess data: factorization
d <- read.table('circuit.txt', header = T)
cols <- names(d)[1:8]
d[cols] <- lapply(d[cols], factor)
# set contrast option
options(contrasts = c('contr.sum', 'contr.poly'))
```

```
# check the alias structure
m <- lm(y ~ blocks + a*b*c*d*e*f*g, data = d)
dr <- alias(m)$Complete[, "blocks1"]
dr[dr %in% c("1", "-1")]
```

```
##          b1:c1:d1          a1:d1:f1       b1:c1:e1:f1       d1:e1:f1:g1
##                 1                 1                -1                -1
##    a1:b1:c1:f1:g1 a1:b1:c1:d1:e1:g1                g1             a1:e1
##                 1                -1                 1                -1
```

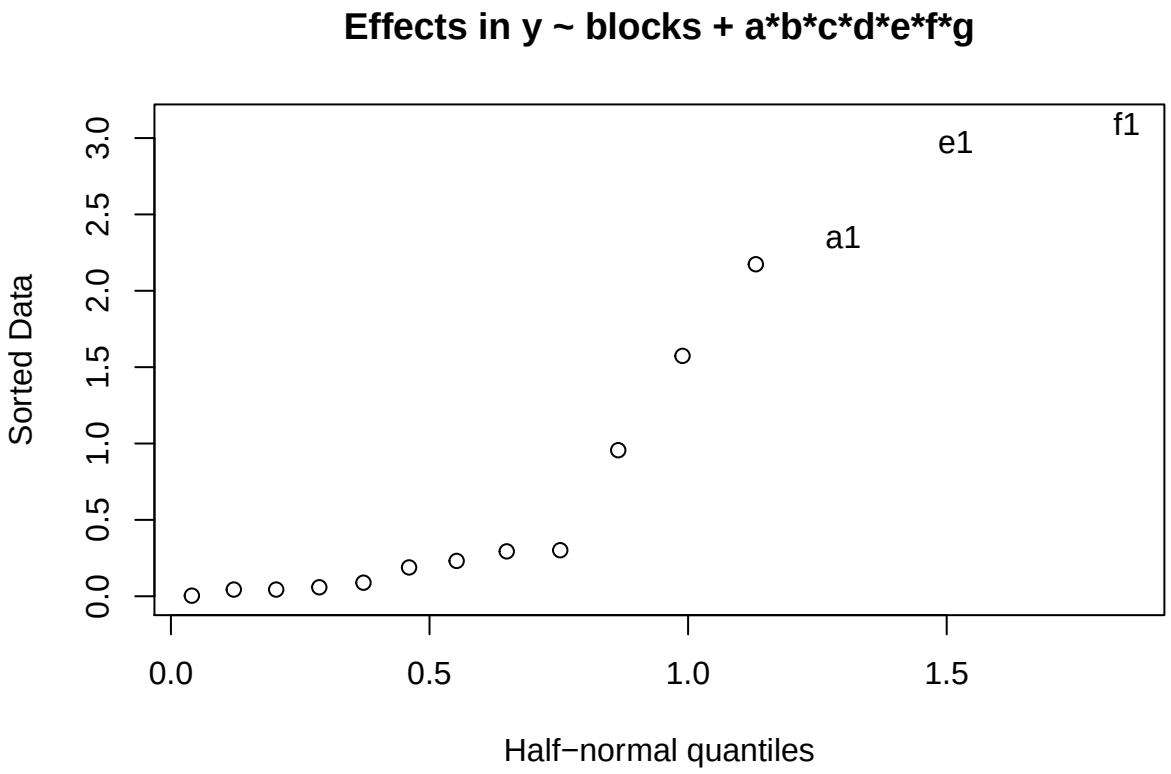
```
# try to find a better block generator
m0 <- lm(y ~ a*b*c*d*e*f*g, data = d)
aliases(m0)
```

```
##
##  a = e:g = b:c:f = d:f:g = b:c:d:e = a:d:e:f = a:b:c:d:g = a:b:c:e:f:g
##  b = a:c:f = c:d:g = a:c:d:e = b:d:e:f = a:b:e:g = c:e:f:g = a:b:d:f:g
##  c = a:b:f = b:d:g = a:b:d:e = c:d:e:f = a:c:e:g = b:e:f:g = a:c:d:f:g
##  d = e:f = b:c:g = a:f:g = a:b:c:e = a:d:e:g = a:b:c:d:f = b:c:d:e:f:g
##  e = d:f = a:g = a:b:c:d = b:c:f:g = a:b:c:e:f = b:c:d:e:g = a:d:e:f:g
##  f = d:e = a:b:c = a:d:g = b:c:e:g = a:e:f:g = b:c:d:f:g = a:b:c:d:e:f
##  g = b:c:d = a:d:f = b:c:e:f = d:e:f:g = a:b:c:f:g = a:b:c:d:e:g = a:e
##  a:b = c:f = c:d:e = b:e:g = a:c:d:g = b:d:f:g = a:b:d:e:f = a:c:e:f:g
##  a:c = b:f = b:d:e = c:e:g = a:b:d:g = c:d:f:g = a:c:d:e:f = a:b:e:f:g
##  b:c = a:f = d:g = a:d:e = e:f:g = b:c:d:e:f = a:b:c:e:g = a:b:c:d:f:g
##  a:d = f:g = b:c:e = a:e:f = d:e:g = b:c:d:f = a:b:c:g = a:b:c:d:e:f:g
##  b:d = c:g = a:c:e = b:e:f = a:c:d:f = a:b:f:g = a:b:d:e:g = c:d:e:f:g
##  c:d = b:g = a:b:e = c:e:f = a:b:d:f = a:c:f:g = a:c:d:e:g = b:d:e:f:g
##  b:e = a:c:d = b:d:f = a:b:g = c:f:g = a:c:e:f = c:d:e:g = a:b:d:e:f:g
##  c:e = a:b:d = c:d:f = a:c:g = b:f:g = a:b:e:f = b:d:e:g = a:c:d:e:f:g
```

The block generator used in the experiment is `BCD`, and the issue is that it will confound with the main effect `G`, which is undesirable. To find a block generator that will not confound with main effects, and even not with two-way interaction terms, we can check the alias structure. From the results, we can see that `ACD` or `ABD` would be good choices, since they would confound with only one two-way interaction term, and the others are three-way or higher order.

(b)

```
# check the halfnorm plot
halfnorm(-2 * coef(m)[-1], labs = names(coef(m)[-1]), nlab = 3)
title(main = "Effects in y ~ blocks + a*b*c*d*e*f*g")
```




```
# aliases of A, E, F
dr <- alias(m)$Complete[, "a1"]
dr[dr %in% c("1", "-1")]
```

```
##             e1:g1          b1:c1:f1          d1:f1:g1       b1:c1:d1:e1
##                -1                 1                 1                -1
##       a1:d1:e1:f1    a1:b1:c1:d1:g1 a1:b1:c1:e1:f1:g1
##                -1                 1                -1
```

```
dr <- alias(m)$Complete[, "e1"]
dr[dr %in% c("1", "-1")]
```

```
##          d1:f1          a1:g1    a1:b1:c1:d1    b1:c1:f1:g1 a1:b1:c1:e1:f1
##            -1             -1             -1             -1              1
## b1:c1:d1:e1:g1 a1:d1:e1:f1:g1
##             1              1
```

```
dr <- alias(m)$Complete[, "f1"]
dr[dr %in% c("1", "-1")]
```

```
##             d1:e1          a1:b1:c1          a1:d1:g1       b1:c1:e1:g1
##                -1                 1                 1                -1
##       a1:e1:f1:g1    b1:c1:d1:f1:g1 a1:b1:c1:d1:e1:f1
##                -1                 1                -1
```

The top3 promising main effects are: A, E, F. For this problem, I assume all three-way interactions terms are not important. Factor A is aliased with EG, factor E is aliased with DF and AG, and factor F is aliased with DE. We can not separate the effect of these three main effects with the two-way interaction effects, so it is likely that these two-way interaction terms are also important. Further analysis should concentrate on A, E, F.

(c)

```
m1 <- lm(y ~ blocks + a * e * f, data = d)
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## blocks     1 18.901  18.901 10.5877 0.011634 *
## a          1 22.114  22.114 12.3873 0.007851 **
## e          1 35.373  35.373 19.8148 0.002135 **
## f          1 38.347  38.347 21.4809 0.001678 **
## a:f        1  0.363   0.363  0.2033 0.664006
## e:f        1  0.014   0.014  0.0077 0.932084
## a:e:f      1  0.032   0.032  0.0176 0.897594
## Residuals  8 14.281   1.785
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
alias(m1)
```

```
## Model :
## y ~ blocks + a * e * f
##
## Complete :
##       (Intercept) blocks1 a1 e1 f1 a1:f1 e1:f1 a1:e1:f1
## a1:e1  0          -1       0  0  0  0     0     0
```

```
m2 <- lm(y ~ blocks + a + e + f, data = d)
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## blocks     1 18.901  18.901  14.153 0.0031432 **
## a          1 22.114  22.114  16.559 0.0018536 **
## e          1 35.373  35.373  26.488 0.0003199 ***
## f          1 38.347  38.347  28.715 0.0002307 ***
## Residuals 11 14.690   1.335
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m1, m2)
```

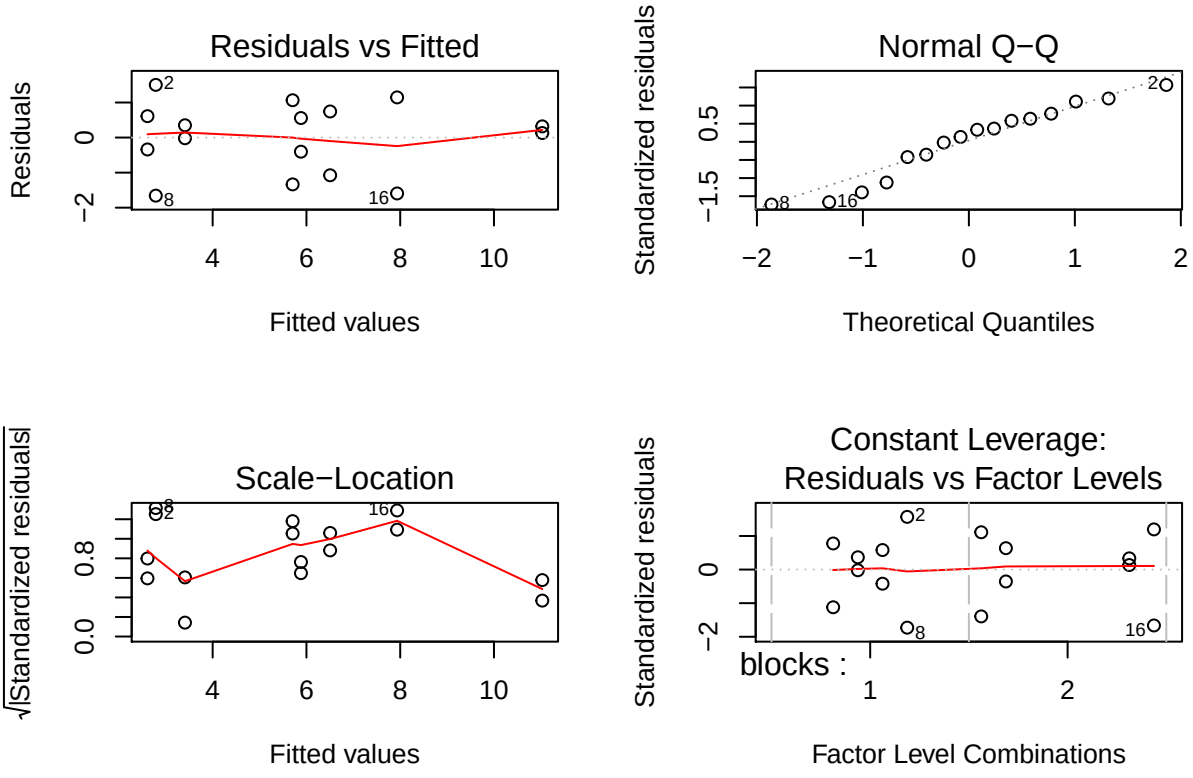
```
## Analysis of Variance Table
##
## Model 1: y ~ blocks + a * e * f
## Model 2: y ~ blocks + a + e + f
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      8 14.281
## 2     11 14.690 -3  -0.40832 0.0762 0.9711
```

```
summary(m2)
```

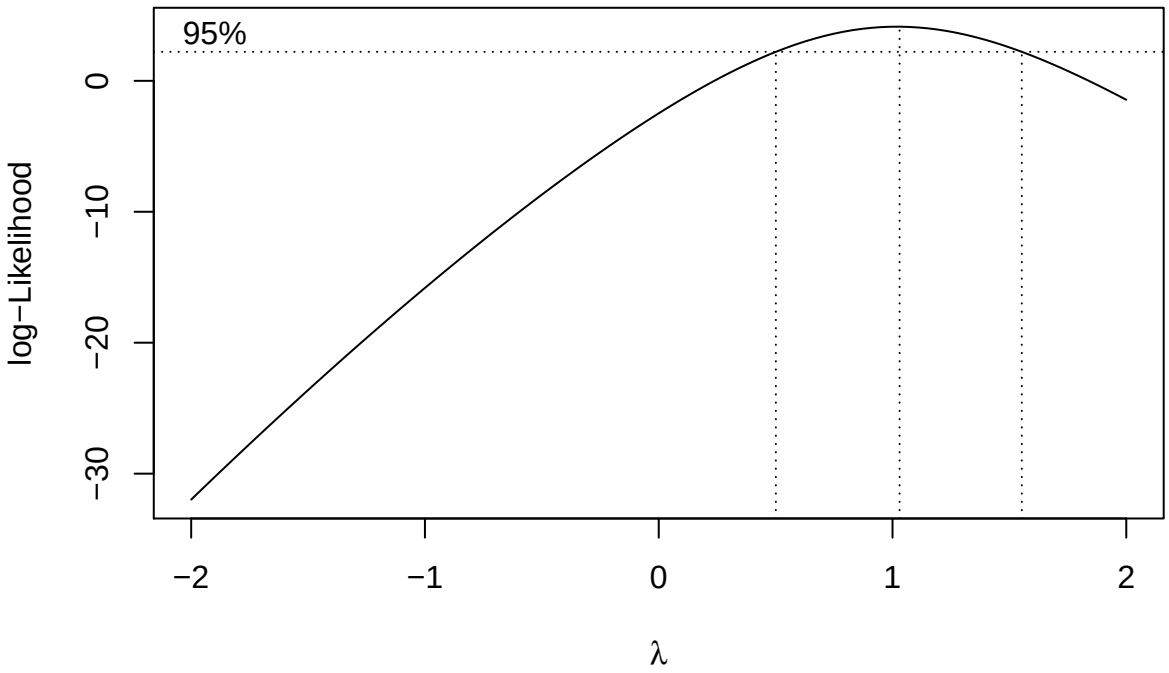
```
##
```

```
## Call:
## lm.default(formula = y ~ blocks + a + e + f, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6569 -0.5713  0.2244  0.6441  1.5031
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.7331     0.2889  19.845 5.81e-10 ***
## blocks1      -1.0869     0.2889  -3.762 0.003143 **
## a1           -1.1756     0.2889  -4.069 0.001854 **
## e1           -1.4869     0.2889  -5.147 0.000320 ***
## f1            1.5481     0.2889   5.359 0.000231 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.156 on 11 degrees of freedom
## Multiple R-squared:  0.8865, Adjusted R-squared:  0.8452
## F-statistic: 21.48 on 4 and 11 DF,  p-value: 3.729e-05
```

```
# model diagnostics
par(mfrow = c(2, 2))
plot(m2)
```




```
par(mfrow = c(1, 1))
boxcox(m2)
```




```
linear.contrast(m2, term = a, contr.coefs = c(1, -1))
```

```
##   estimates        se   t-value     p-value  lower-ci  upper-ci
## 1  -2.35125 0.5778028 -4.069295 0.001853647 -3.622985 -1.079515
```

```
linear.contrast(m2, term = e, contr.coefs = c(1, -1))
```

```
##   estimates        se   t-value      p-value  lower-ci  upper-ci
## 1  -2.97375 0.5778028 -5.146652 0.0003199055 -4.245485 -1.702015
```

```
linear.contrast(m2, term = f, contr.coefs = c(1, -1))
```

```
##   estimates        se  t-value    p-value lower-ci upper-ci
## 1   3.09625 0.5778028 5.358662 0.00023074 1.824515 4.367985
```

Note that effect AE confounds with the block, so we cannot separate these two. The ANOVA table shows that the interaction terms are not significant, so I delete them and refit the model. Comparing `m1` and `m2`, I would use `m2` as the final model.

All three main effects are significant, with p-value being ....

The model diagnostic results look good.

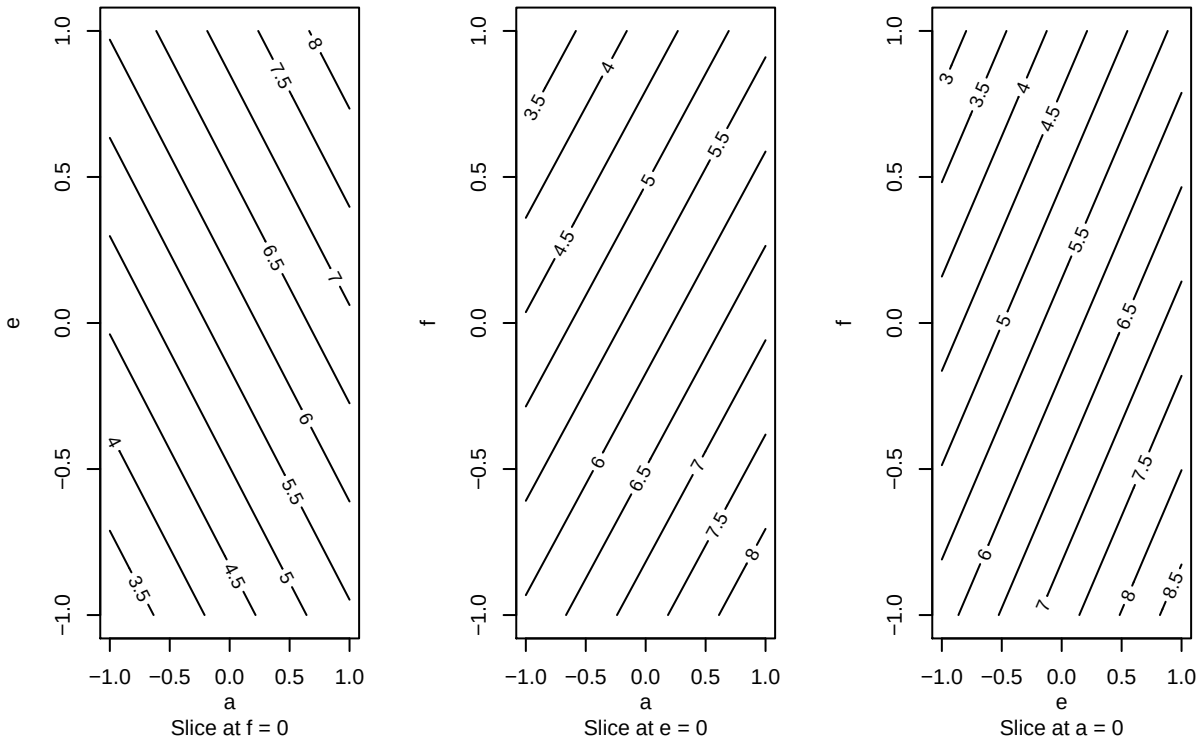
(d)

```
# response surface
dd <- within(d, {
  a <- ifelse(a=='-1', -1, 1)
  e <- ifelse(e=='-1', -1, 1)
  f <- ifelse(f=='-1', -1, 1)
})
m.rsm <- rsm(y ~ blocks + FO(a, e, f), data = dd)
summary(m.rsm)
```

```
##
## Call:
## rsm(formula = y ~ blocks + FO(a, e, f), data = dd)
```

```
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   5.7331     0.2889 19.8446 5.811e-10 ***
## blocks1      -1.0869     0.2889 -3.7621 0.0031432 **
## a             1.1756     0.2889  4.0693 0.0018536 **
## e             1.4869     0.2889  5.1467 0.0003199 ***
## f            -1.5481     0.2889 -5.3587 0.0002307 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Multiple R-squared:  0.8865, Adjusted R-squared:  0.8452
## F-statistic: 21.48 on 4 and 11 DF,  p-value: 3.729e-05
##
## Analysis of Variance Table
##
## Response: y
##             Df Sum Sq Mean Sq F value   Pr(>F)
## blocks       1 18.901  18.901 14.1534 0.003143
## FO(a, e, f)  3 95.833  31.944 23.9208 4.02e-05
## Residuals   11 14.690   1.335
## Lack of fit  3  0.408   0.136  0.0762 0.971079
## Pure error   8 14.281   1.785
##
## Direction of steepest ascent (at radius 1):
##          a          e          f
##  0.4803641  0.6075418 -0.6325688
##
## Corresponding increment in original units:
##          a          e          f
##  0.4803641  0.6075418 -0.6325688
```

```
par(mfrow = c(1, 3))
contour(m.rsm, ~ a + e + f)
```




```
par(mfrow = c(1, 1))
```

From the contour plot, I don't think ≤ 1.5 is achievable with current settings.

(e)

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

Reminder

- Read through the problem, understand the problem, write down the solution sketch and highlight note, check understanding, then start coding.
- If there are covariates, start with plotting scatter plots.

Checklist

(1) Randomized or observational? Fit an adjusted or unadjusted model?
(2) Balanced or unbalanced?
(3) Data preprocessing. Factorization.
(4) For change from baseline problems, do remember to remove the baseline observation.
(5) `contr.sum` or `contr.treatment`? If 2^k factorial, we have to use `contr.sum`.
(6) Fit a large model, check model assumptions: typical 4 plots, residual plots, qqnorm for error, qqnorm for random effects, boxcox for transformation. If specified, check outliers. Check P17.1 in the assignments.R for fancy plots. But note that we need to refit the model and check whether the inference will change in order to determine outliers.
(7) Correlation structure? Which response to use? Interaction and polynomial terms? Random slope or random intercept? Try and use model diagnostics to help choose.
(8) If there are covariates, plot the scatter plot of the response vs. covariate. Check lecture code Chapter 17. Center the covariate for better interpretation.
(9) `anova()` or `Anova(, type = 2)`?
(10) Multiple comparison. Use `glht()`, use `linear.contrast()` and `fit.contrast()` to check. If using `intervals()`, then the contrast option needs to be `contr.treatment`.
(11) Copy the library code block to the appendix.

(12) Follow-up or use diff?
(13) If possible, plot fitted to check the goodness of the model.
(14) Always use `REML` to make inference.
(15) Do not factorized `fu`.
(16) Check previous problems to guide writing.
(17) Model selection: interaction terms, quadratic terms, random slope/intercept, correlation structure. AIC, or manual.
(18) Show all your findings and considerations, let the graders know your understandings.

Notes (1) `Anova(type=2)` means independent effect. (2) We can treat `month` as factor or numerics for variance reduction. Factor is more general. And if we want to try random slope, then we need to use numerics.

As shown in Figure 1a in Appendix, there is a non-linear relationship between `age` and `np.chg`, so I include the quadratic term. Also, for easier interpretation, I center `age` around 40.

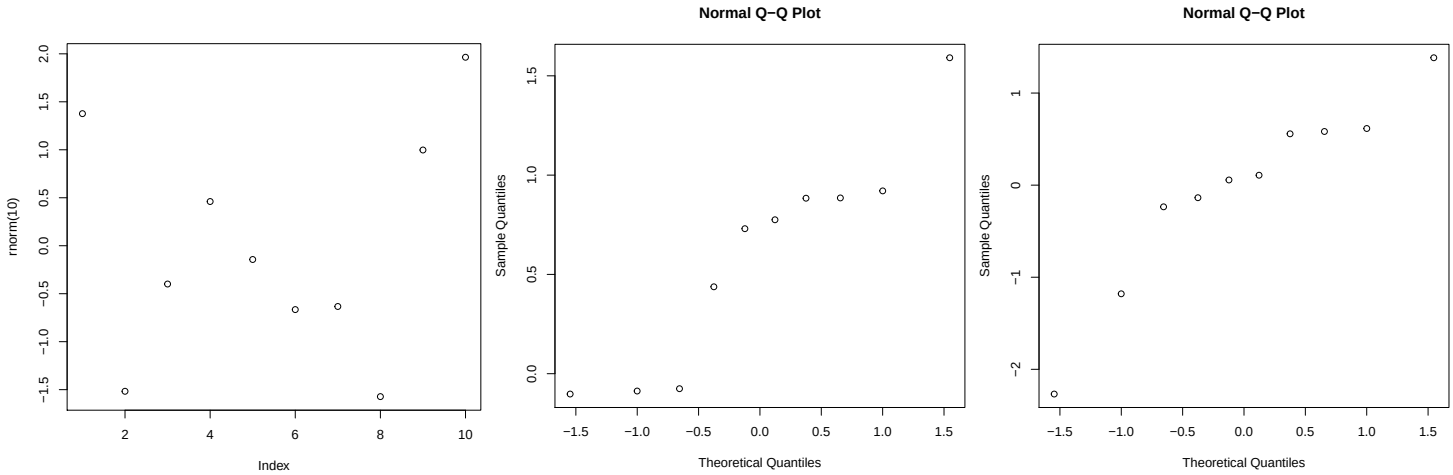
Packages

All R packages used in this problem are listed below.

```
library(gmodels)
library(MASS)
library(car)
library(dplyr)
```

Appendix

Figures



(a) residual plot of m.lme4 (b) qqnorm of residuals (c) qqnorm of random effects

Figure 1: Model Diagnostics

2009 Q2

November

```
d <- read.table("http://users.stat.umn.edu/~wangx346/bmd.txt", header = T)
# check types of the dataframe
sapply(d, class)
```

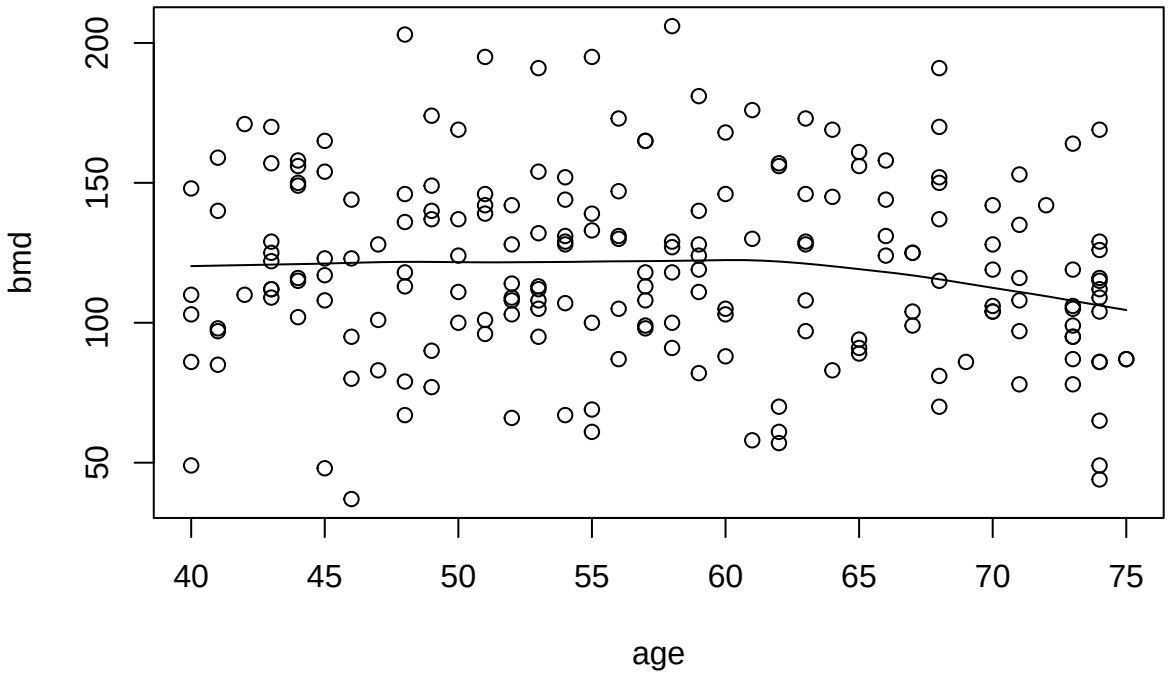
```
##        id    gender       age  diabetes   smoking      drug     visit       bmd
## "integer" "integer" "integer" "integer" "integer" "integer" "integer" "integer"
```

```
# factorize columns
d <- within(d, {
  gender <- as.factor(gender)
  drug <- as.factor(drug)
})
d0 <- d[d$visit == 0, ]
```

```
median(d0$age)
```

```
## [1] 57
```

```
with(d0, scatter.smooth(x=age, y=bmd))
```




```
m <- lm(bmd ~ (gender + smoking + diabetes + I(age-57) + I((age-57)^2))^2,
        data = d0)
Anova(m, type=2)
```

```
## Anova Table (Type II tests)
##
## Response: bmd
##                            Sum Sq  Df F value    Pr(>F)
## gender                      59137   1 74.7831 2.547e-15 ***
## smoking                      1309   1  1.6557   0.19980
## diabetes                       80   1  0.1018   0.75006
## I(age - 57)                  2306   1  2.9158   0.08940 .
```

```
## I((age - 57)^2)               2345   1  2.9659   0.08672 .
## gender:smoking                 598   1  0.7564   0.38560
## gender:diabetes               1481   1  1.8733   0.17277
## gender:I(age - 57)             582   1  0.7364   0.39192
## gender:I((age - 57)^2)         285   1  0.3607   0.54888
## smoking:diabetes               423   1  0.5352   0.46535
## smoking:I(age - 57)            270   1  0.3415   0.55969
## smoking:I((age - 57)^2)        634   1  0.8022   0.37160
## diabetes:I(age - 57)           441   1  0.5577   0.45615
## diabetes:I((age - 57)^2)         1   1  0.0007   0.97860
## I(age - 57):I((age - 57)^2)    614   1  0.7767   0.37929
## Residuals                   145505 184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m2 <- step(m, trace=0)
Anova(m2, type=2)
```

```
## Anova Table (Type II tests)
##
## Response: bmd
##                 Sum Sq  Df F value    Pr(>F)
## gender           58943   1 76.4031 1.064e-15 ***
## smoking           1633   1  2.1171   0.14727
## I(age - 57)       2334   1  3.0253   0.08356 .
## I((age - 57)^2)   2183   1  2.8303   0.09410 .
## Residuals       150437 195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

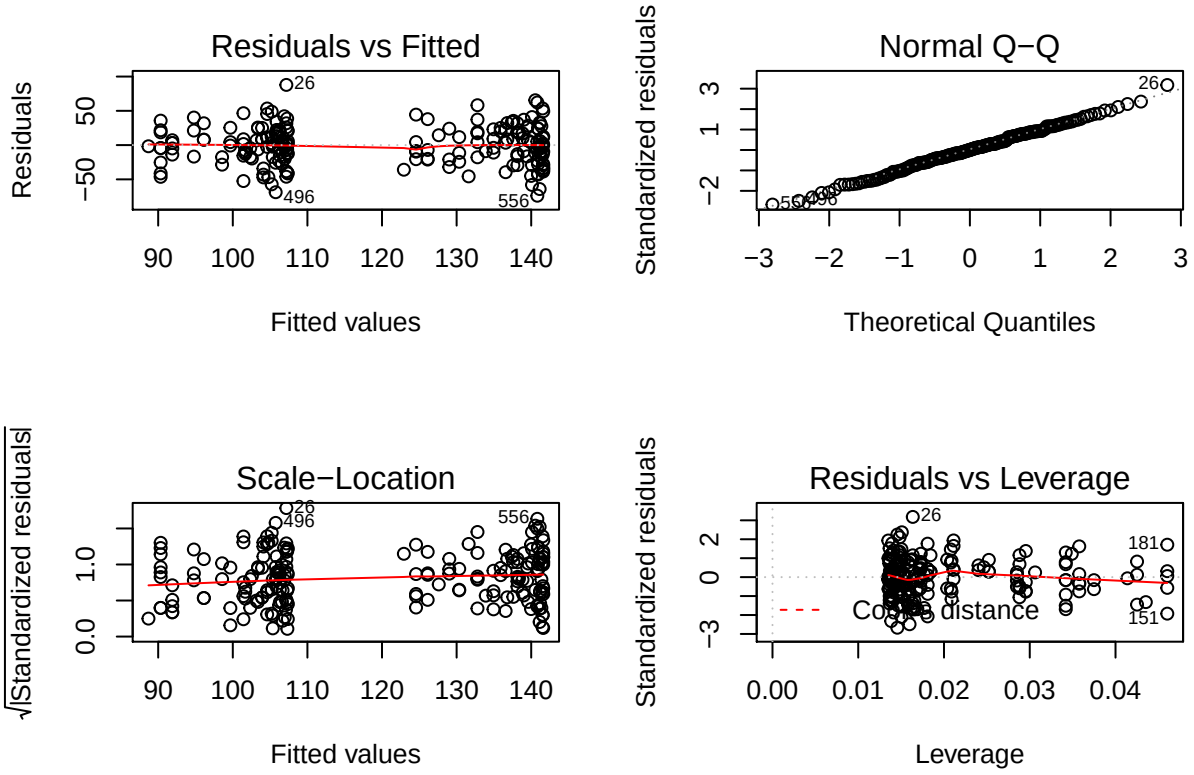
```
anova(m, m2)
```

```
## Analysis of Variance Table
##
## Model 1: bmd ~ (gender + smoking + diabetes + I(age - 57) + I((age - 57)^2))^2
## Model 2: bmd ~ gender + smoking + I(age - 57) + I((age - 57)^2)
##   Res.Df    RSS  Df Sum of Sq     F Pr(>F)
## 1    184 145505
## 2    195 150437 -11   -4932.2 0.567 0.8539
```

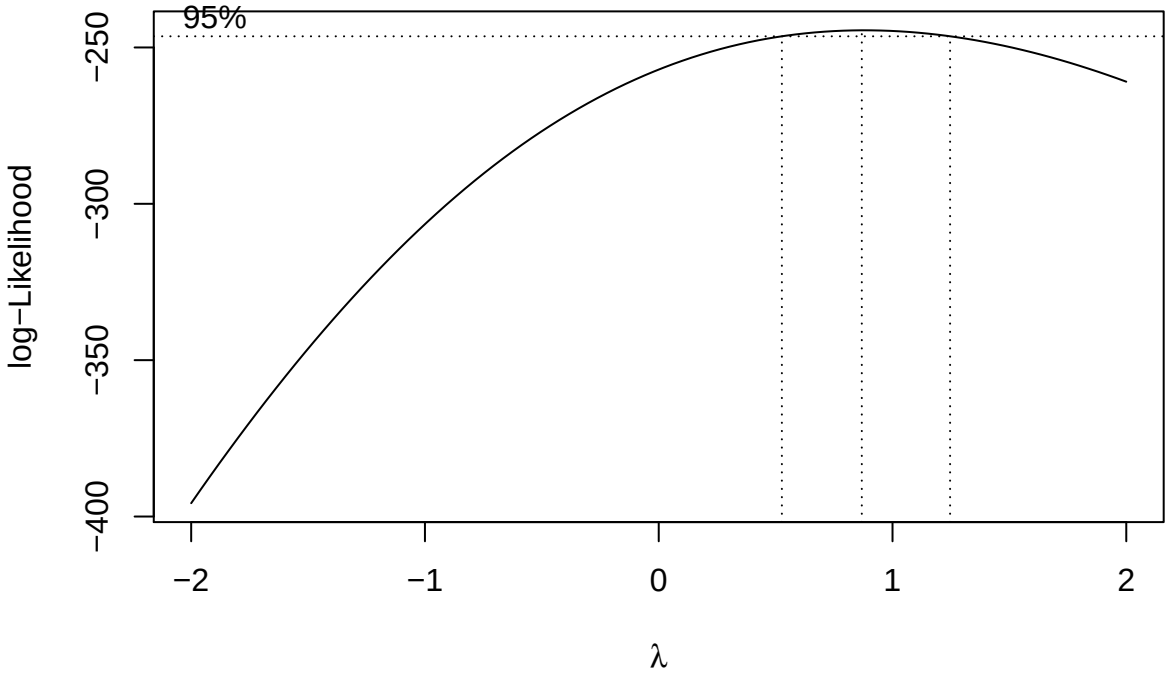
```
m3 <- lm(bmd ~ gender + I(age-57) + I((age-57)^2), data = d0)
m4 <- lm(bmd ~ gender + I(age-57), data = d0)
anova(m4, m3, m2)
```

```
## Analysis of Variance Table
##
## Model 1: bmd ~ gender + I(age - 57)
## Model 2: bmd ~ gender + I(age - 57) + I((age - 57)^2)
## Model 3: bmd ~ gender + smoking + I(age - 57) + I((age - 57)^2)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    197 154691
## 2    196 152070  1    2621.2 3.3977 0.06681 .
## 3    195 150437  1    1633.3 2.1171 0.14727
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow = c(2, 2))
plot(m3)
```




```
par(mfrow = c(1, 1))
boxcox(m3)
```



(b)

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

Reminder

- Read through the problem, understand the problem, write down the solution sketch and highlight note, check understanding, then start coding.
- If there are covariates, start with plotting scatter plots.

Checklist

(1) Randomized or observational? Fit an adjusted or unadjusted model?
(2) Balanced or unbalanced?
(3) Data preprocessing. Factorization.
(4) For change from baseline problems, do remember to remove the baseline observation.
(5) `contr.sum` or `contr.treatment`? If 2^k factorial, we have to use `contr.sum`.
(6) Fit a large model, check model assumptions: typical 4 plots, residual plots, qqnorm for error, qqnorm for random effects, boxcox for transformation. If specified, check outliers. Check P17.1 in the assignments.R for fancy plots. But note that we need to refit the model and check whether the inference will change in order to determine outliers.
(7) Correlation structure? Which response to use? Interaction and polynomial terms? Random slope or random intercept? Try and use model diagnostics to help choose.
(8) If there are covariates, plot the scatter plot of the response vs. covariate. Check lecture code Chapter 17. Center the covariate for better interpretation.
(9) `anova()` or `Anova(, type = 2)`?
(10) Multiple comparison. Use `glht()`, use `linear.contrast()` and `fit.contrast()` to check. If using `intervals()`, then the contrast option needs to be `contr.treatment`.
(11) Copy the library code block to the appendix.
(12) Follow-up or use diff?
(13) If possible, plot fitted to check the goodness of the model.
(14) Always use `REML` to make inference.
(15) Do not factorized `fu`.
(16) Check previous problems to guide writing.
(17) Model selection: interaction terms, quadratic terms, all two-way interactions, 3rd-order terms, center covariates, random slope/intercept, correlation structure. AIC, or manual.
(18) Show all your findings and considerations, let the graders know your understandings.
(19) Interaction plot can be obtained by `interaction.plot` or `emmip`.
(20) Try both adjusted and unadjusted models.
(21) For continuous covariates, try 3rd-order terms and start with a large model with all possible interactions if allowed.

Notes (1) `Anova(type=2)` means independent effect. (2) We can treat `month` as factor or numerics for variance reduction. Factor is more general. And if we want to try random slope, then we need to use numerics.

As shown in Figure 1a in Appendix, there is a non-linear relationship between `age` and `np.chg`, so I include the quadratic term. Also, for easier interpretation, I center `age` around 40.

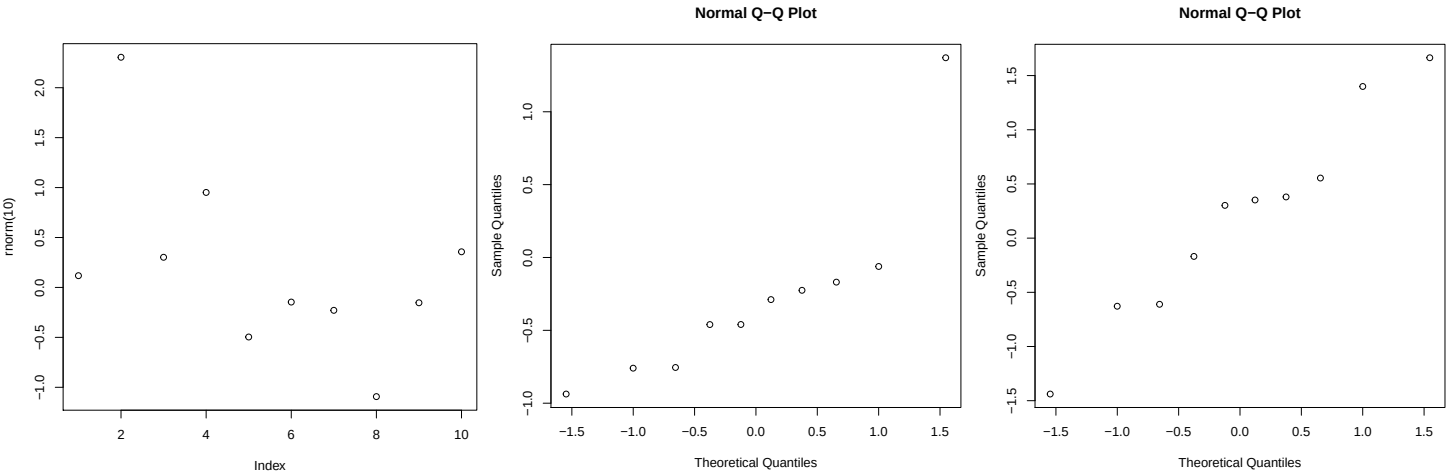
Packages

All R packages used in this problem are listed below.

```
library(gmodels)
library(MASS)
library(car)
library(dplyr)
```

Appendix

Figures



(a) residual plot of m.lme4 (b) qqnorm of residuals (c) qqnorm of random effects

Figure 1: Model Diagnostics

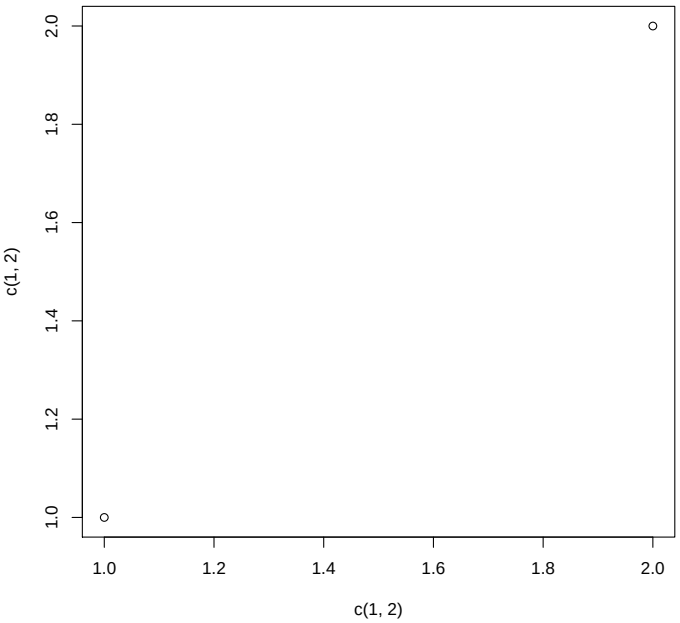


Figure 2: Scatter Plot Age vs. np.chg

2013 Q2

codename

```
library(gmodels)
library(MASS)
# library(car)
```

(a)

Since the patients are randomized with treatment, it would fine to use a model without adjustment to compare the treatment effects.

```
da <- read.table("http://users.stat.umn.edu/~wangx346/artstudy-widedata.txt",
                 header = T)

da <- within(da, {
  treat <- as.factor(treat)
  country <- as.factor(country)
  gender <- as.factor(gender)
  np.chg <- np.8 - 0.5 * (np.0 + np.2)
})

ma <- lm(np.chg ~ treat, data = da)
fit.contrast(ma, "treat", coeff = c(1, -1), conf.int = 0.95)
```

```
##                   Estimate Std. Error  t value    Pr(>|t|)  lower CI  upper CI
## treat c=( 1 -1 ) -2.073679  0.6986437 -2.96815 0.003177036 -3.447172 -0.700186
## attr(,"class")
## [1] "fit_contrast"
```

In the analysis, participants on continuous ART have lower NP score than those on intermittent ART, by 2.07 points(p-value = 0.003). The 95% confidence of the difference is [-3.447172, -0.700186].

(b)

```
pdf("images/scatter_age_np.pdf")
with(da, scatter.smooth(x = age, y = np.chg))
dev.off()
```

As shown in Figure 1 in Appendix, there is a non-linear relationship between `age` and `np.chg`, so I include the quadratic term. Also, for easier interpretation, I center `age` around 40.

```
mb <- lm(np.chg ~ (treat + I(age-40) + I((age - 40)^2) + gender + country +
                     art.pre0 + artdur2)^2, data = da)
stepAIC(mb, scope = list(lower = ~ treat), direction = "backward", trace = F)$call
```

```
## lm(formula = np.chg ~ treat + I(age - 40) + I((age - 40)^2) +
##     gender + country + art.pre0 + treat:I((age - 40)^2) + I(age -
##     40):I((age - 40)^2) + I((age - 40)^2):art.pre0 + gender:country,
##     data = da)
```

Since the interaction of the treatment effect with the quadratic effect of age is included, to get a hierarchical model, I repeat the model selection, forcing the interaction between the treatment effect and the linear effect of age included.

```
stepAIC(mb, scope = list(lower = ~ treat * I(age - 40)),
        direction = "backward", trace = F)$call
```

```
## lm(formula = np.chg ~ treat + I(age - 40) + I((age - 40)^2) +
##     gender + country + art.pre0 + treat:I(age - 40) + treat:I((age -
##     40)^2) + I(age - 40):I((age - 40)^2) + I((age - 40)^2):art.pre0 +
##     gender:country, data = da)
```

Fit the selected model and check the summary results.

```
mb.2 <- lm(np.chg ~ (treat + art.pre0) * (I(age - 40) + I((age - 40)^2)) +
             I((age - 40)^3) + gender * country, data = da)
summary(mb.2)
```

```
...
## Call:
## lm(formula = np.chg ~ (treat + art.pre0) * (I(age - 40) + I((age -
##     40)^2)) + I((age - 40)^3) + gender * country, data = da)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.9221  -3.1022  -0.0879   3.0640  12.7128
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.0433726  0.8284734   1.259  0.20865
## treat2                     1.3147355  0.6827635   1.926  0.05489 .
## art.pre0                   2.3829338  0.8346416   2.855  0.00454 **
## I(age - 40)               -0.7062102  0.0759048  -9.304  < 2e-16 ***
## I((age - 40)^2)           -0.0098514  0.0050513  -1.950  0.05187 .
## I((age - 40)^3)            0.0012366  0.0002283   5.416 1.08e-07 ***
## gender2                    0.9828502  0.9598437   1.024  0.30649
## country2                   1.2801042  0.7851565   1.630  0.10384
## country3                  -0.5521595  0.8225370  -0.671  0.50244
## treat2:I(age - 40)        -0.0155535  0.0551045  -0.282  0.77790
## treat2:I((age - 40)^2)     0.0167229  0.0051727   3.233  0.00133 **
## art.pre0:I(age - 40)      -0.0386780  0.0646655  -0.598  0.55011
## art.pre0:I((age - 40)^2) -0.0107209  0.0056883  -1.885  0.06022 .
## gender2:country2          -3.4958794  1.3707794  -2.550  0.01115 *
## gender2:country3          -1.2171009  1.4334416  -0.849  0.39637
...
```

From the results, we can see that there is no evidence that the treatment effect depends on gender, country, ART use at study entry, and duration of ART use. But the treatment effect does depend on age.

(c)

Due to the existence of the interaction between the treatment effect and age, we cannot simply use `fit.contrast` to compare the two treatment effects. But from the estimated coefficients above, we can see that at every level of `age`, treatment 2, namely, intermittent ART has higher NP score than continuous ART. Also, from the results in (a), we can also conclude that intermittent ART is better(p-value = 0.003).

(d)

i. The treatment effect is confounded with duration of ART use during the study, so the model with all covariates is not suitable for estimating the treatment effect.

ii.

iii. This model doesn't recognize that the treatment effect would change with age.

Appendix

Figures

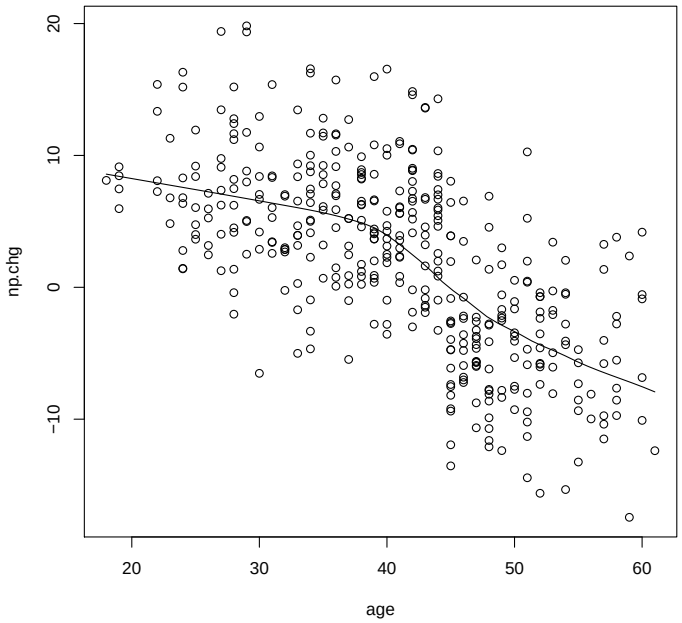


Figure 1: Scatter Plot Age vs. np.chg

2015 Q1

codename

```
library(nlme)
library(car)
library(gmodels)
library(cfcdae)
library(dplyr)
library(mgcv)
```

(a)

```
### Data Preprocessing
d <- read.table('fev1.txt', header = T)
d <- within(d, {
  grp <- as.factor(grp)
  gender <- as.factor(gender)
  fev1.chg <- fev1 - fev1.0
})
d <- d %>% filter(year != 0)

### Simple estimate of treatment effect without adjustment to other covariates
ma.lme <- lme(fev1.chg ~ grp, random = ~ 1|ID, data = d)
Anova(ma.lme)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: fev1.chg
##      Chisq Df Pr(>Chisq)
## grp 17.764  3   0.000492 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeff.matrix <- rbind("1 vs 2" = c(1, -1, 0, 0),
                      "1 vs 3" = c(1, 0, -1, 0),
                      "1 vs 4" = c(1, 0, 0, -1))
fit.contrast(ma.lme, "grp", coeff = coeff.matrix, conf.int = 1 - (0.05/3))
```

```
##             Estimate Std. Error   t-value     Pr(>|t|)     lower CI  upper CI
## grp1 vs 2 0.51448507  0.1389639 3.7022917 0.0002421447  0.180470170 0.8485000
## grp1 vs 3 0.32508532  0.1385313 2.3466566 0.0194065343 -0.007889595 0.6580602
## grp1 vs 4 0.05572438  0.1403292 0.3970976 0.6914972345 -0.281572038 0.3930208
## attr(,"class")
## [1] "fit_contrast"
```

(b)

```
getVarCov(ma.lme, type = "random.effects")
```

```
## Random effects variance covariance matrix
##             (Intercept)
## (Intercept)      0.7912
##   Standard Deviations: 0.88949
```

```
getVarCov(ma.lme, individuals = "1", type = "conditional")
```

```
## ID 1
## Conditional variance covariance matrix
##         1       2       3
## 1 0.83885 0.00000 0.00000
## 2 0.00000 0.83885 0.00000
## 3 0.00000 0.00000 0.83885
##   Standard Deviations: 0.91589 0.91589 0.91589
```

(c)

```
mc.lme1 <- lme(fev1.chg ~ (grp + I(age - 60) + I((age - 60)^2) +
                           gender + smoking + bpmeds + fev1.0)^2 + year,
               random = ~ year|ID, data = d)
Anova(mc.lme1)
```

```
...
##                                Chisq Df Pr(>Chisq)
## grp                          57.6134  3  1.901e-12 ***
## I(age - 60)                   1.5030  1    0.22022
## I((age - 60)^2)               0.0368  1    0.84797
## gender                        0.0774  1    0.78089
## smoking                       0.0053  1    0.94213
## bpmeds                        0.1743  1    0.67636
## fev1.0                     1202.5442  1  < 2.2e-16 ***
## year                          6.4525  1    0.01108 *
## grp:I(age - 60)              10.1911  3    0.01701 *
## gender:fev1.0                 0.4030  1    0.52552
## smoking:bpmeds                4.4114  1    0.03570 *
## smoking:fev1.0                1.7426  1    0.18681
...
```

```
mc.lme2 <- update(mc.lme1, method = "ML")
mc.lme3 <- update(mc.lme2, fixed = fev1.chg ~ grp * I(age - 60) + fev1.0 +
                    year + smoking * bpmeds)
anova(mc.lme2, mc.lme3)
```

```
##         Model df      AIC      BIC    logLik   Test L.Ratio p-value
## mc.lme2     1 48 3992.080 4246.632 -1948.040
## mc.lme3     2 17 3959.192 4049.346 -1962.596 1 vs 2 29.1121  0.5634
```

```
mc.lme4 <- update(mc.lme3, method = "REML")
mc.lme5 <- update(mc.lme4, random = ~ 1|ID)
anova(mc.lme4, mc.lme5)
```

```
##         Model df      AIC      BIC    logLik   Test      L.Ratio p-value
## mc.lme4     1 17 4036.132 4126.137 -2001.066
## mc.lme5     2 15 4032.132 4111.548 -2001.066 1 vs 2 1.897006e-07       1
```

```
Anova(mc.lme5)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: fev1.chg
##                     Chisq Df Pr(>Chisq)
```

```
## grp               65.3512  3  4.219e-14 ***
## I(age - 60)        2.4778  1   0.115463
## fev1.0          1322.7572  1  < 2.2e-16 ***
## year               6.5178  1   0.010680 *
## smoking            0.1656  1   0.684057
## bpmeds             0.5744  1   0.448530
## grp:I(age - 60)    9.3300  3   0.025210 *
## smoking:bpmeds    10.4865  1   0.001202 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
d <- within(d, {
  dose <- as.ordered(dose)
})
mc.lme6 <- lme(fev1.chg ~ dose * I(age - 60) + fev1.0 + year +
                smoking * bpmeds, random = ~ 1|ID, data = d)
```

```
summary(mc.lme6)
```

```
...
## Fixed effects: fev1.chg ~ dose * I(age - 60) + fev1.0 + year + smoking * bpmeds
##                          Value  Std.Error   DF   t-value p-value
## (Intercept)          2.7685307 0.10771155 1060  25.70319  0.0000
## dose.L               0.1183948 0.04850057  412   2.44110  0.0151
## dose.Q               0.1900586 0.04932388  412   3.85328  0.0001
## dose.C              -0.3125340 0.04938725  412  -6.32823  0.0000
## I(age - 60)         -0.0071003 0.00420450  412  -1.68875  0.0920
## fev1.0              -1.0062925 0.02766841  412 -36.36973  0.0000
## year                -0.0586302 0.02296525 1060  -2.55300  0.0108
## smoking              0.1420019 0.06898850  412   2.05834  0.0402
## bpmeds               0.2007069 0.07271454  412   2.76020  0.0060
## dose.L:I(age - 60) -0.0182902 0.00737097  412  -2.48138  0.0135
## dose.Q:I(age - 60)  0.0108024 0.00725328  412   1.48931  0.1372
## dose.C:I(age - 60)  0.0045118 0.00708608  412   0.63671  0.5247
## smoking:bpmeds      -0.3138557 0.09692011  412  -3.23829  0.0013
...
```

```
d <- within(d, {
  dose <- as.numeric(dose)
})
mc.lme7 <- lme(fev1.chg ~ dose + I(dose^2) + I(dose^3) + smoking * bpmeds +
                 fev1.0 + year, random = ~ 1|ID, data = d)
Anova(mc.lme7)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: fev1.chg
##                     Chisq Df Pr(>Chisq)
## dose              51.9470  1  5.702e-13 ***
## I(dose^2)         47.5565  1  5.344e-12 ***
## I(dose^3)         42.4957  1  7.084e-11 ***
## smoking            0.3056  1   0.580406
## bpmeds             0.0636  1   0.800886
## fev1.0          1362.3180  1  < 2.2e-16 ***
## year               6.7056  1   0.009611 **
## smoking:bpmeds     8.4462  1   0.003658 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mc.lme7)
```

```
...
## Linear mixed-effects model fit by REML
##  Data: d
##        AIC      BIC    logLik
##   4003.555 4061.823 -1990.778
##
## Random effects:
##  Formula: ~1 | ID
##          (Intercept)  Residual
## StdDev: 0.0002173172 0.9136655
##
## Fixed effects: fev1.chg ~ dose + I(dose^2) + I(dose^3) + smoking * bpmeds +      fev1.0 + year
##                    Value Std.Error   DF   t-value p-value
## (Intercept)     5.613466 0.4111610 1060  13.65272  0.0000
## dose           -4.357695 0.6046119  416  -7.20742  0.0000
## I(dose^2)       1.844808 0.2675138  416   6.89612  0.0000
## I(dose^3)      -0.231858 0.0355672  416  -6.51887  0.0000
## smoking         0.116237 0.0684375  416   1.69845  0.0902
## bpmeds          0.143028 0.0657693  416   2.17469  0.0302
## fev1.0         -0.992214 0.0268823  416 -36.90959  0.0000
## year           -0.059612 0.0230206 1060  -2.58953  0.0097
## smoking:bpmeds -0.277222 0.0953889  416  -2.90623  0.0039
...
```

The effect of drug doesn't differ by age, gender, smoking status, use of blood pressure medication, or baseline level of FEV1. The effect of the drug does is cubic, not linear. ...

(d)

```
d2 <- d %>% filter(year == 1, grp %in% c(2, 4))
m.lm <- lm(fev1.chg ~ grp, data = d2)
summary(m.lm)$sigma
```

```
## [1] 1.31876
```

```
power.t.test(n = NULL, delta = 0.2, sd = 1.32, sig.level = 0.05, power = 0.8,
             type = "two.sample", alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 684.7561
##           delta = 0.2
##              sd = 1.32
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Appendix

Figures

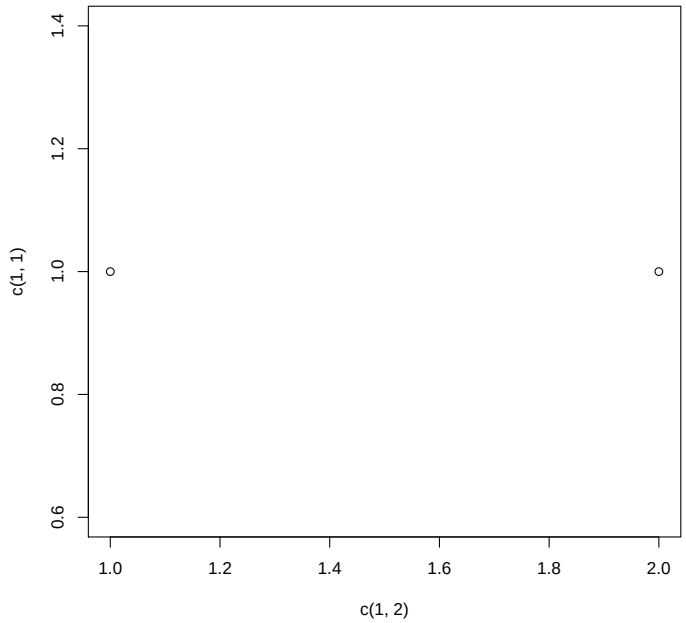


Figure 1: Scatter Plot Age vs. np.chg

2016 Q3

codename

```
library(gmodels)
library(MASS)
# library(car)
```

```
pdf("images/scatter_age_np.pdf")
plot(c(1, 2), c(1, 2))
dev.off()
```

As shown in Figure 1 in Appendix, there is a non-linear relationship between `age` and `np.chg`, so I include the quadratic term. Also, for easier interpretation, I center `age` around 40.

Appendix

Figures

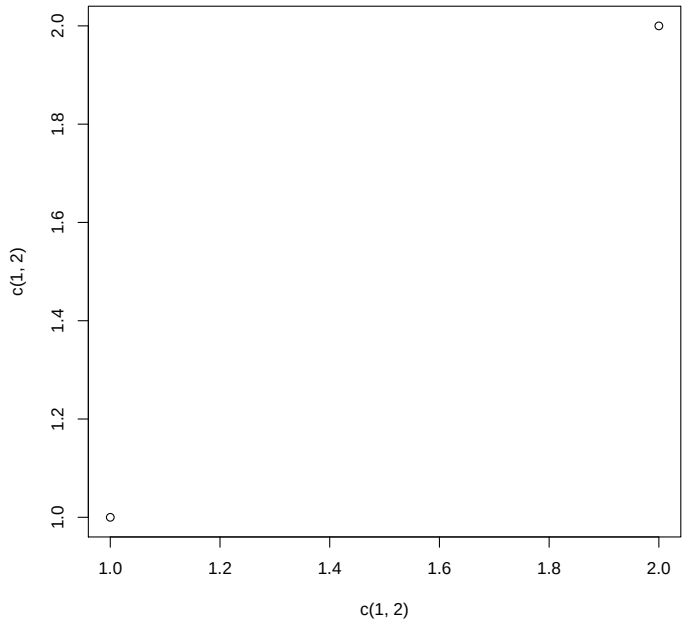


Figure 1: Scatter Plot Age vs. np.chg

2017 Applied Exam Q2

November

(i)

```
d <- read.csv("2017QualDataQ2.csv", header = T)
code <- function(x) {
  y <- x
  y[x == min(x)] <- -1
  y[x == max(x)] <- 1
  y[x > min(x) & x < max(x)] <- 0
  y
}
d <- within(d, {
  Design <- as.factor(Design)
  Machine <- as.factor(Machine)
  block <- as.factor(block)
  cTime <- code(Time)
  cP <- code(P)
  cTemp <- code(Temp)
  y.penal <- ifelse(Effort <= 4, Strength, 0)
})
```

Test whether `Design` interacts with the three settings. It shows that we can include `Design` as additive in the model.

```
m1.1 <- lm(Strength ~ block + Machine + Design * (cTime + cP + cTemp)^2 +
             Design * (I(cTime^2) + I(cP^2) + I(cTemp^2)), data = d)
m1.2 <- lm(Strength ~ block + Machine + Design + (cTime + cP + cTemp)^2 +
             (I(cTime^2) + I(cP^2) + I(cTemp^2)), data = d)
anova(m1.1, m1.2)
```

```
...
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     42 0.43117
## 2     51 0.50125 -9 -0.070076 0.7584 0.6543
...
```

Now, use `{rsm}` to fit the response surface model.

```
m1 <- rsm(Strength ~ block + Machine + Design + FO(cTime, cP, cTemp),
          data = d)
summary(m1)
```

```
...
## Analysis of Variance Table
##
## Response: Strength
##                       Df  Sum Sq Mean Sq  F value    Pr(>F)
## block                  1  0.5175  0.5175  16.2675 0.0001654
## Machine                1  0.0113  0.0113   0.3548 0.5537451
## Design                 1  0.4016  0.4016  12.6243 0.0007727
## FO(cTime, cP, cTemp)   3 16.3636  5.4545 171.4472 < 2.2e-16
## Residuals             57  1.8134  0.0318
## Lack of fit           10  1.3313  0.1331  12.9793 1.762e-10
## Pure error            47  0.4821  0.0103
...
```

```
m2 <- rsm(Strength ~ block + Machine + Design + SO(cTime, cP, cTemp),
          data = d)
summary(m2)
```

```
...
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  1.3709375  0.0391878 34.9838 < 2.2e-16 ***
## block2      -0.0219792  0.0314589 -0.6987   0.48794
## Machine2     0.0265625  0.0247845  1.0717   0.28888
## Design2     -0.1584375  0.0247845 -6.3926 4.953e-08 ***
## cTime        0.5935000  0.0156751 37.8626 < 2.2e-16 ***
## cP           0.2050000  0.0156751 13.0781 < 2.2e-16 ***
## cTemp       -0.1217500  0.0156751 -7.7671 3.366e-10 ***
## cTime:cP    -0.0303125  0.0175253 -1.7296   0.08974 .
## cTime:cTemp -0.0096875  0.0175253 -0.5528   0.58283
## cP:cTemp     0.0153125  0.0175253  0.8737   0.38636
## cTime^2      0.3082292  0.0314589  9.7978 2.576e-13 ***
## cP^2        -0.0317708  0.0314589 -1.0099   0.31730
## cTemp^2      0.0744792  0.0314589  2.3675   0.02174 *
## Analysis of Variance Table
##
## Response: Strength
##                        Df  Sum Sq Mean Sq  F value    Pr(>F)
## block                   1  0.5175  0.5175  52.6584 2.141e-09
## Machine                 1  0.0113  0.0113   1.1486    0.2889
## Design                  1  0.4016  0.4016  40.8653 4.953e-08
## FO(cTime, cP, cTemp)    3 16.3636  5.4545 554.9788 < 2.2e-16
## TWI(cTime, cP, cTemp)   3  0.0399  0.0133   1.3535    0.2675
## PQ(cTime, cP, cTemp)    3  1.2723  0.4241  43.1499 5.001e-14
## Residuals              51  0.5012  0.0098
## Lack of fit             4  0.0191  0.0048   0.4667    0.7598
## Pure error             47  0.4821  0.0103
...
```

```
m3 <- rsm(Strength ~ block + Machine + Design + FO(cTime, cP, cTemp) +
            PQ(cTime, cTemp), data = d)
summary(m3)
```

```
##
## Call:
## rsm(formula = Strength ~ block + Machine + Design + FO(cTime,
##     cP, cTemp) + PQ(cTime, cTemp), data = d)
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  1.357791   0.037324 36.3786 < 2.2e-16 ***
## block2      -0.014310   0.030827 -0.4642   0.64433
## Machine2     0.026563   0.025027  1.0614   0.29316
## Design2     -0.158437   0.025027 -6.3307 4.680e-08 ***
## cTime        0.593500   0.015828 37.4960 < 2.2e-16 ***
## cP           0.205000   0.015828 12.9514 < 2.2e-16 ***
## cTemp       -0.121750   0.015828 -7.6919 2.776e-10 ***
## cTime^2      0.300560   0.030827  9.7499 1.372e-13 ***
## cTemp^2      0.066810   0.030827  2.1673   0.03456 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Multiple R-squared:  0.9712, Adjusted R-squared:  0.967
## F-statistic: 231.5 on 8 and 55 DF,  p-value: < 2.2e-16
##
## Analysis of Variance Table
##
## Response: Strength
##                       Df  Sum Sq Mean Sq  F value    Pr(>F)
## block                  1  0.5175  0.5175  51.6437 1.867e-09
## Machine                1  0.0113  0.0113   1.1265    0.2932
## Design                 1  0.4016  0.4016  40.0779 4.680e-08
## FO(cTime, cP, cTemp)   3 16.3636  5.4545 544.2855 < 2.2e-16
## PQ(cTime, cTemp)       2  1.2623  0.6311  62.9776 5.981e-15
## Residuals             55  0.5512  0.0100
## Lack of fit            8  0.0691  0.0086   0.8419    0.5710
## Pure error            47  0.4821  0.0103
##
## Stationary point of response surface:
##      cTime          cP      cTemp
## -0.9873225  0.0000000  0.9111613
##
## Eigenanalysis:
## eigen() decomposition
## $values
## [1] 0.30056034 0.06681034 0.00000000
##
## $vectors
##       [,1] [,2] [,3]
## cTime    1    0    0
## cP       0    0    1
## cTemp    0    1    0
```

The estimated coefficient for `Design2` is negative, suggesting that at the same level of the covariates, Design2 has smaller strength. The contour plots suggest border would give larger strength. Use grid search to find the optimal setting, and it gives the same result. `cTemp=-1, cTime=1, cP=1, Design=1` would provide the strongest seal. From the contour plot, we can see no big difference between the machines.

`cTime` and `cTemp` affects the strength quadratically, while `cP` has linear influence.

```
# grid search
newdata1 <- expand.grid(cTemp = seq(-1, 1, by = 0.1),
                        cTime = seq(-1, 1, by = 0.1),
                        cP = seq(-1, 1, by = 0.1))
n <- dim(newdata1)[1]
newdata2 <- data.frame(Design = rep("1", n),
                       block = rep("1", n),
                       Machine = rep("1", n))
newdata <- cbind(newdata1, newdata2)
pred <- predict(m3, newdata)
ind <- which.max(pred)
newdata[ind, ]
```

```
##      cTemp cTime cP Design block Machine
## 9241    -1     1  1      1     1       1
```

```
pred[ind]
```

```
##     9241
## 2.645412
```

(ii)

As shown in (i), the first-order model doesn't fit well, and in order to fit a second-order model, we need more points, both axial and center points.

(iii)

As shown in Figure 5 in Appendix, for Design 2, Effort is acceptable for whatever `Strength`. So I would recommend the same setting as in (i) to achieve strong seals with effort ≤ 4.

(iv)

```
m4 <- rsm(y.penal ~ block + Machine + Design + SO(cTime, cP, cTemp),
          data = d)
```

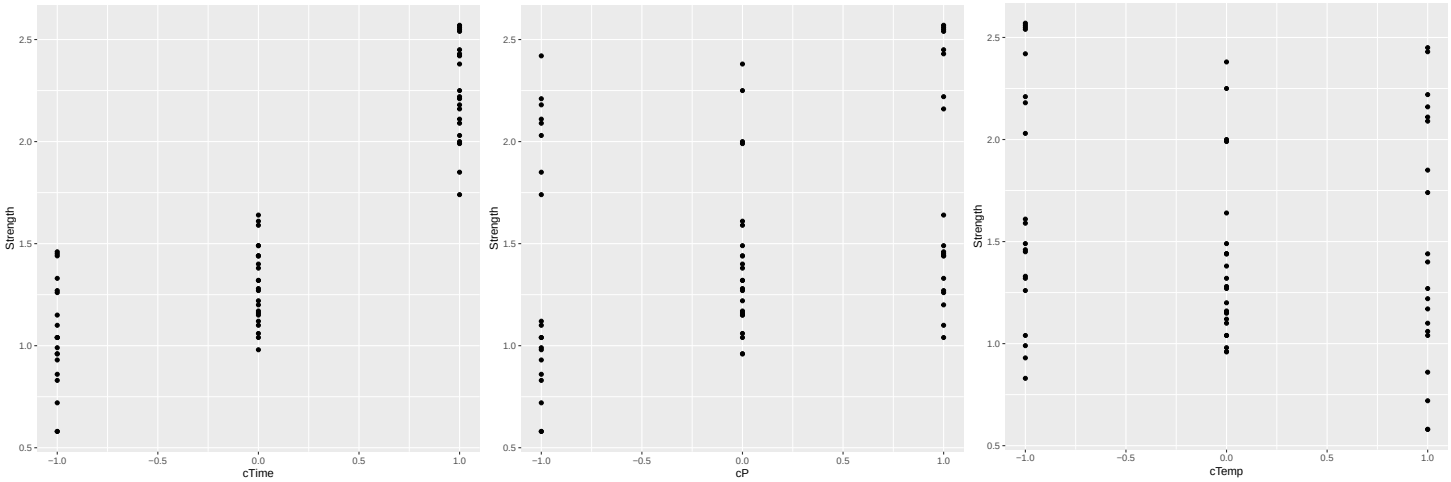
Packages

All R packages used in this problem are listed below.

```
library(gmodels)
library(MASS)
library(car)
library(dplyr)
```

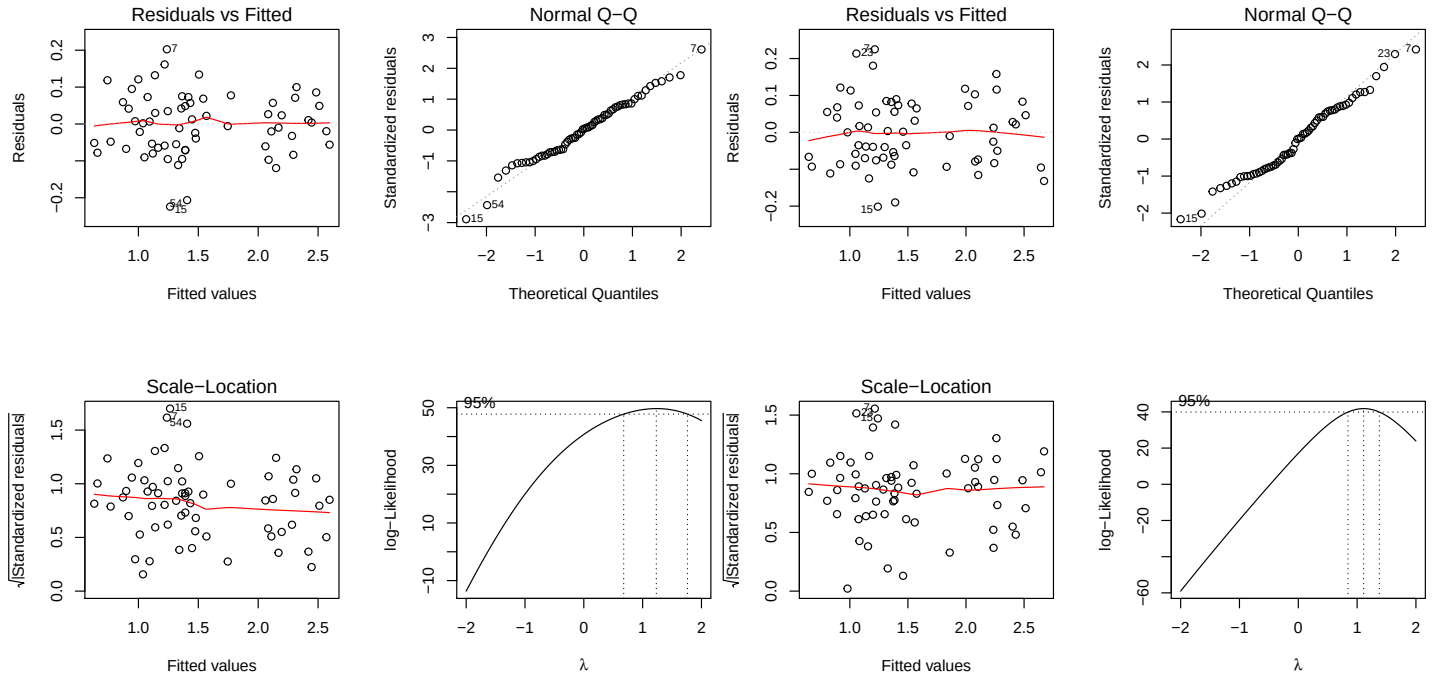
Appendix

Figures



(a) cTime vs Strength (b) cP vs Strength (c) cTemp vs Strength

Figure 1: Scatter plots



(a) Diagnostics for m1.1

(b) Diagnostics for m3

Figure 2: Model Diagnostics

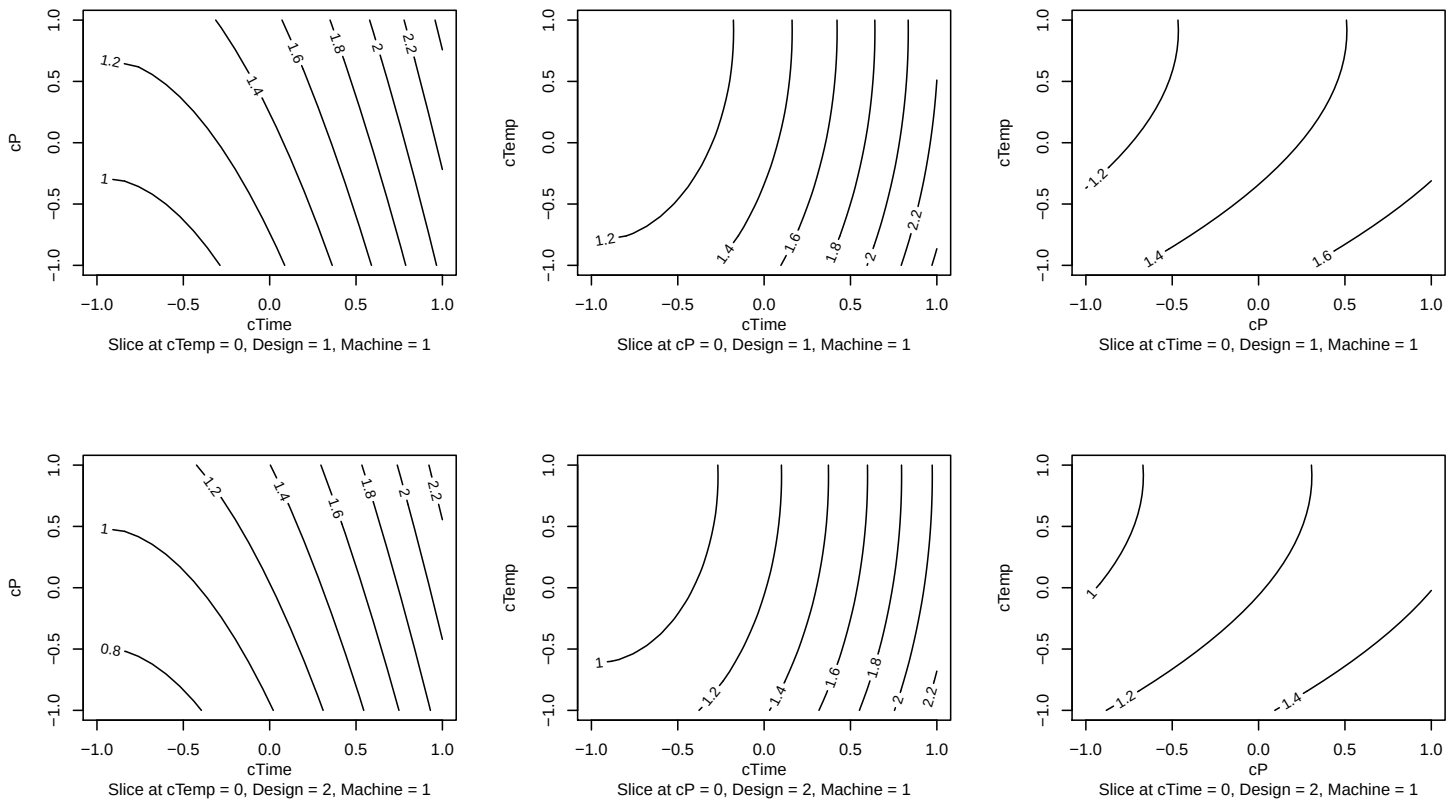


Figure 3: Contour Plots: Different Designs

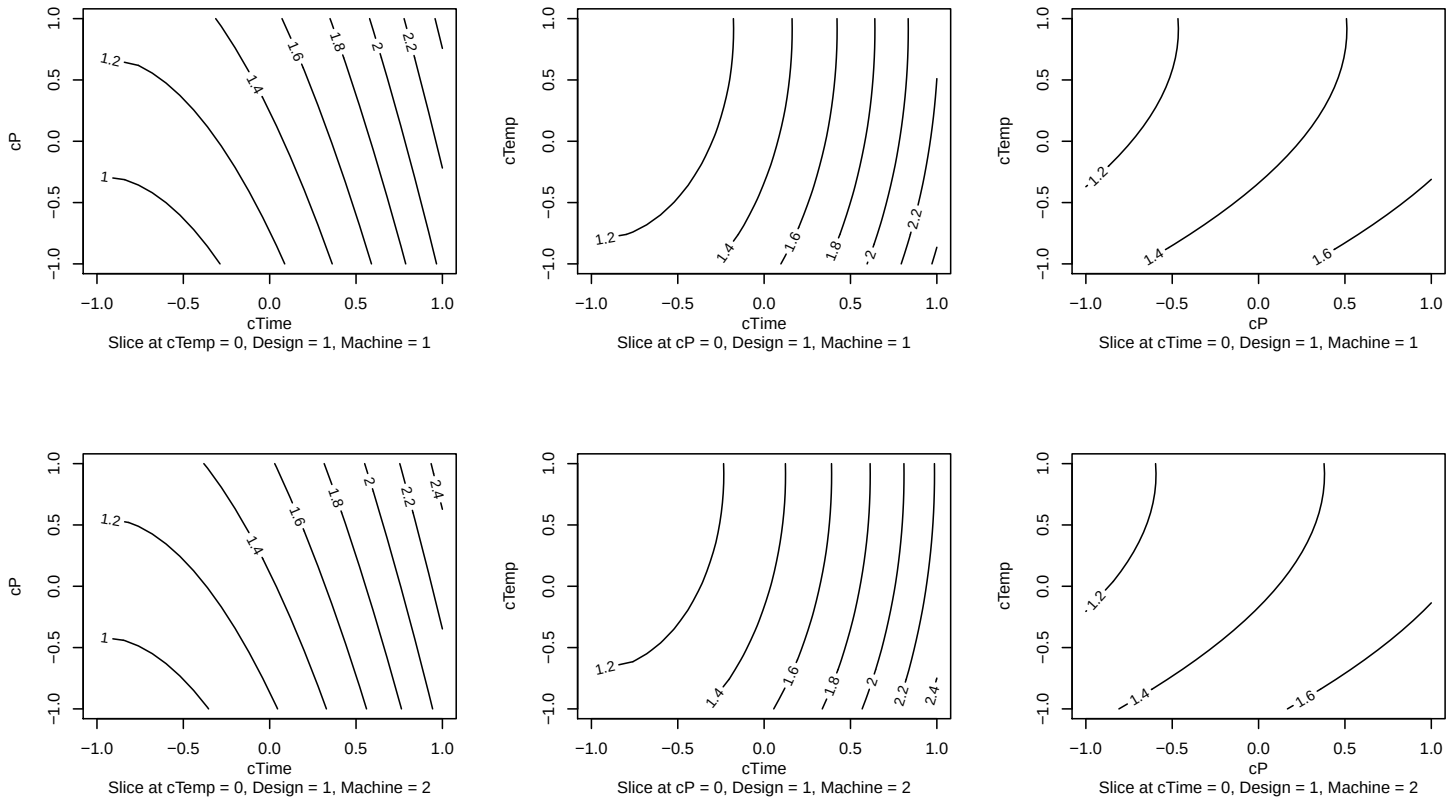


Figure 4: Contour Plots: Different Machines

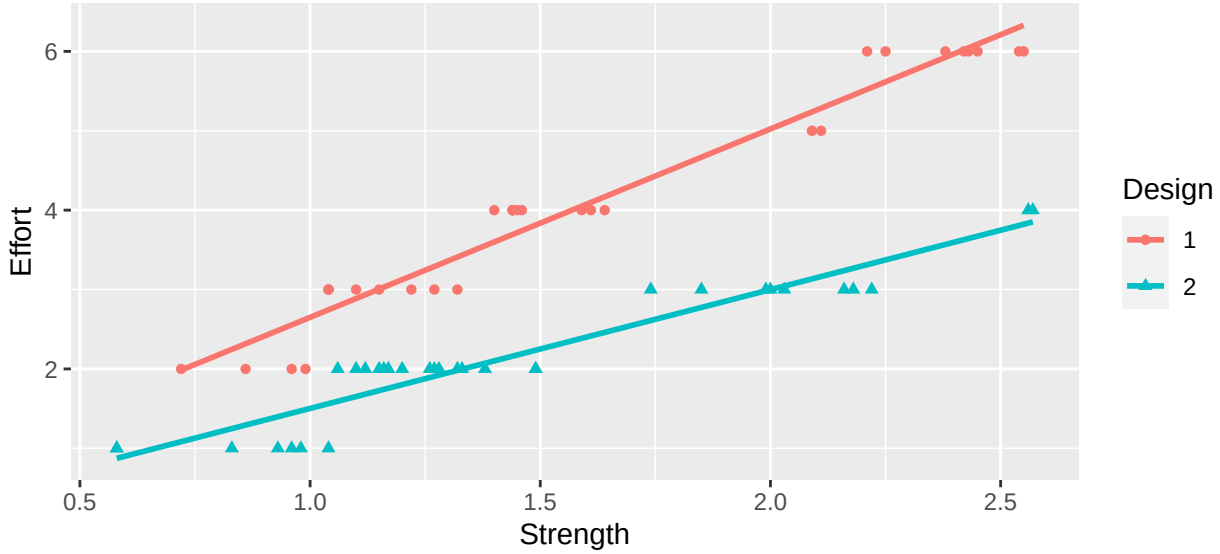


Figure 5: Strength vs Effort

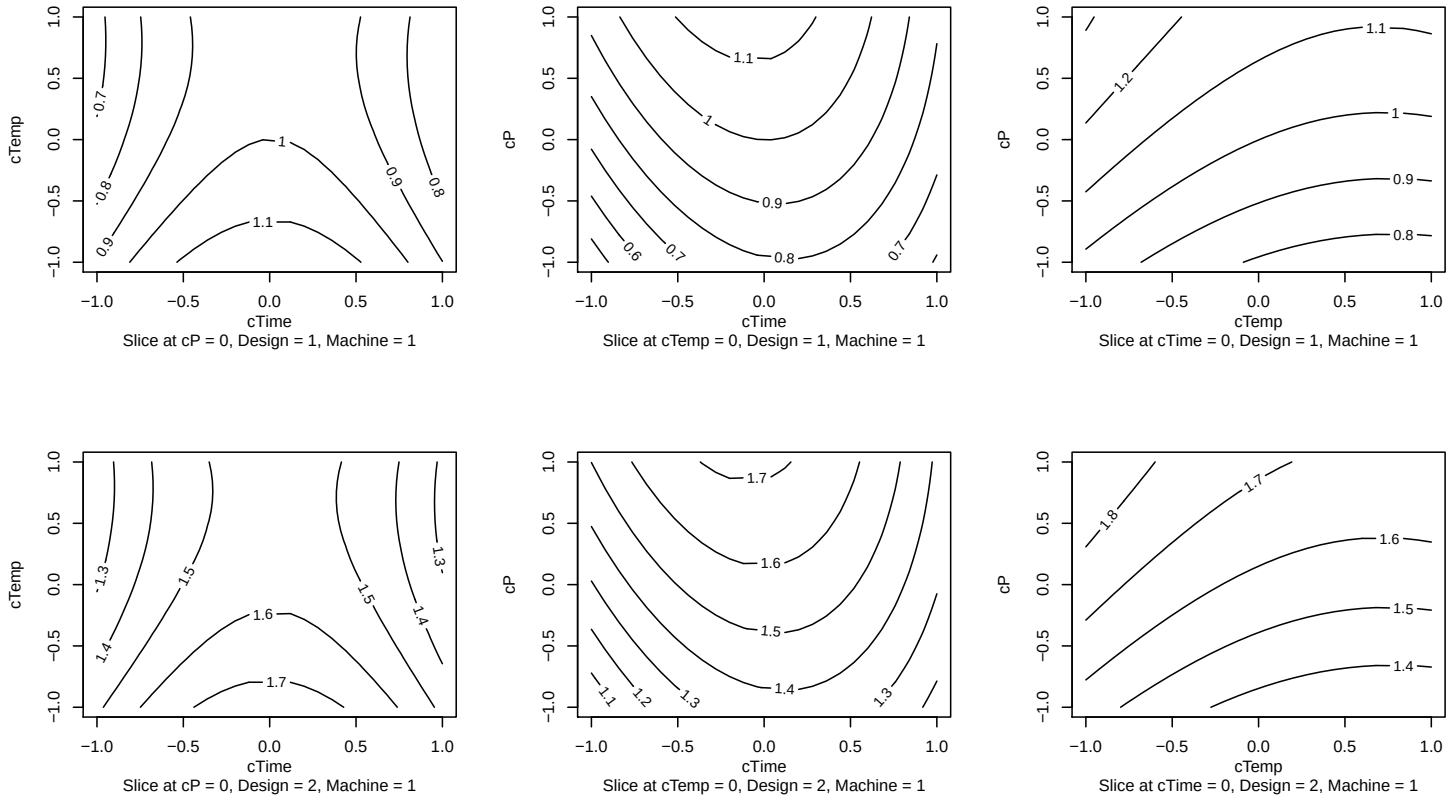


Figure 6: Contour Plots for m4

2018 Q1

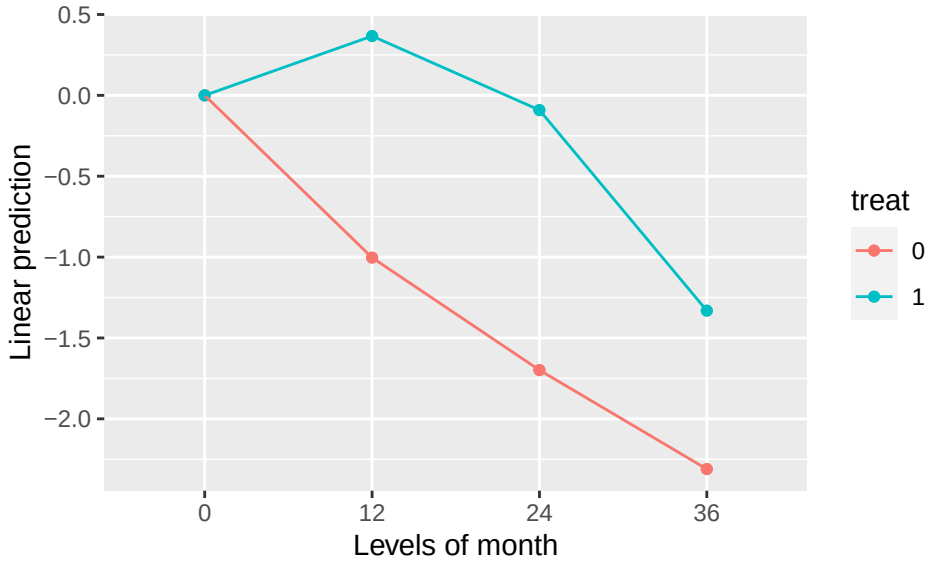
codename

```
library(gmodels)
library(MASS)
library(car)
library(nlme)
library(emmeans)
library(dplyr)
```

(1)

```
d <- read.csv("q1.csv", header = T)
d <- within(d, {
  treat <- as.factor(treat)
  gender <- as.factor(gender)
  race <- as.factor(race)
  scanner <- as.factor(scanner)
  clinic <- as.factor(clinic)
})
```

```
m <- lm(bmd.chg ~ treat * factor(month), data = d)
emmip(m, treat ~ factor(month), style = "factor")
```



For treat0, namely, placebo group, the mean percent change in BMD keeps going down, while for the active drug group, it first goes up, then goes down after month 12. Also, the differences between the two treatment are not the same in the follow-up months.

(2)

This is a longitudinal data, so we need to fit a linear mixed model. The model including all two-way interactions suffers singularity issue. So I then starting with the largest plausible model, and do manual selection. Keep the items with p-value $<$ 0.1. Also, for the same reason as that in 2015 Q1, it is justifiable to use identity as the correlation matrix.

```
d2 <- d %>% filter(month > 0)
m.lme1 <- lme(bmd.chg ~ treat + factor(month) + bmd0 + age + gender +
                          race + scanner + clinic + smoking + bmibl +
                          il6bl + crpbl + cd4bl, random = ~1|id,
              data = d2, method = "ML")
anova(m.lme1)
m.lme2 <- update(m.lme1, fixed = . ~ treat + factor(month) +
                   bmd0 + scanner)
Anova(m.lme2)
m.lme3 <- update(m.lme2, fixed = . ~ treat + factor(month) + scanner)
Anova(m.lme3)
```

```
anova(m.lme1, m.lme2, m.lme3)
```

```
##        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m.lme1     1 21 2292.178 2378.566 -1125.089
## m.lme2     2  8 2275.707 2308.616 -1129.853 1 vs 2 9.528595  0.7320
## m.lme3     3  7 2278.206 2307.002 -1132.103 2 vs 3 4.499149  0.0339
```

m.lme4 is the selected model. Checking the residuals, as shown in Figure 1a-1c in the Appendix, no transformation is needed.

```
m.lme4 <- update(m.lme2, method = "REML")
fit.contrast(m.lme4, "treat", coeff = c(1, -1), conf.int = 0.95)
```

```
##                    Estimate Std. Error   t-value    Pr(>|t|)  lower CI
## treat c=( 1 -1 ) -1.378845  0.4326775 -3.186772 0.001673193 -2.232119
##                     upper CI
## treat c=( 1 -1 ) -0.5255704
## attr(,"class")
## [1] "fit_contrast"
```

```
intervals(m.lme4)
```

```
## Approximate 95% confidence intervals
##
##  Fixed effects:
##                        lower       est.       upper
## (Intercept)     -1.6693674  1.5299496  4.72926666
## treat1           0.5255704  1.3788446  2.23211887
## factor(month)24 -0.9694324 -0.4998415 -0.03025054
## factor(month)36 -1.9549666 -1.3871094 -0.81925218
## bmd0            -6.2463261 -3.2297186 -0.21311109
## scannerType2     0.5472170  1.4066934  2.26616990
## attr(,"label")
## [1] "Fixed effects:"
##
##  Random Effects:
##   Level: id
##                     lower     est.    upper
## sd((Intercept)) 2.330979 2.650146 3.013014
##
##  Within-group standard error:
##    lower     est.    upper
## 1.976879 2.153265 2.345389
```

The placebo group has BMD measure 1.39 lower than the active drug group, with the p-value being 0.0015. The 95% CI is [-2.25, -0.54].

(3)

It is the treatment difference after adjusting for `scanner` and `month`, namely, the effect of treatment after excluding the effect of `scanner` and `month`. And in the plot in (1), the mean percentage change in BMD is raw mean, without forcing the difference among `month` to be the same.

```
xtabs(~treat+month, d2)
```

```
##      month
## treat  12  24  36
##     0  92  73  45
##     1 107  84  51
```

(4)

Firstly, there might be model bias in our proposed model. Secondly, augmenting the missing data with fitted values would reduce the variance, and reduce the width of the confidence interval. In summary, the new CI would have shorter coverage, as well as possible biased center point.

(5)

Now the data is grouped under `clinic`, instead of `id`, so we need to fit another linear mixed model. Again, start with the largest plausible model and do manual model selection.

```
d5 <- d %>% filter(month == 0)
m5.lme1 <- lme(bmd ~ age + gender + race + scanner + smoking +
                                  bmibl + il6bl + crpbl + cd4bl,
               random = ~1|clinic, data = d5, method = "ML")
Anova(m5.lme1)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: bmd
##           Chisq Df Pr(>Chisq)
## age      1.3372  1  0.2475258
## gender   6.3565  1  0.0116950 *
## race    14.6242  3  0.0021676 **
## scanner 13.2566  1  0.0002716 ***
## smoking  0.2475  1  0.6188063
## bmibl   38.8908  1  4.482e-10 ***
## il6bl    5.1939  1  0.0226660 *
## crpbl    0.9944  1  0.3186631
## cd4bl    0.9088  1  0.3404293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m5.lme2 <- update(m5.lme1, . ~ gender + race + scanner + bmibl + il6bl)
Anova(m5.lme2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: bmd
##           Chisq Df Pr(>Chisq)
## gender   7.4305  1  0.0064128 **
## race    15.3174  3  0.0015645 **
## scanner 14.5440  1  0.0001369 ***
## bmibl   39.0579  1  4.114e-10 ***
```

```
## il6bl   11.1122  1  0.0008576 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m5.lme1, m5.lme2)
```

```
##         Model df       AIC       BIC   logLik   Test  L.Ratio p-value
## m5.lme1     1 14 -338.6556 -288.4867 183.3278
## m5.lme2     2 10 -342.8162 -306.9813 181.4081 1 vs 2 3.839411  0.4282
```

```
m5.lme2 <- update(m5.lme2, method = "REML")
fixef(m5.lme2)
```

```
##  (Intercept)      gender2        race2        race3        race4 scannerType2
##  0.804434743 -0.055813565 -0.057781692 -0.023362277 -0.081163322  0.065781308
##        bmibl        il6bl
##  0.009691822 -0.006214962
```

Based on the estimated effects, (1) female would have lower BMD than male; (2) the ranking of race effects is race4 < race2 < race3 < race1; (3) scanner Type 2 would have higher BMD than Type 1; (4) higher `bmibl` is associated with higher BMD; (5) higher `il6bl` is associated with lower BMD. There is no interaction among the covariates in the selected model, so all of these effects are independent.

```
summary(m5.lme2)
```

```
...
## Random effects:
##  Formula: ~1 | clinic
##         (Intercept)  Residual
## StdDev:  0.01399878 0.1238445
...
```

```
intervals(m5.lme2)$reStruct
```

```
...
## [1] "Random Effects:"
NA
NA
NA
...
```

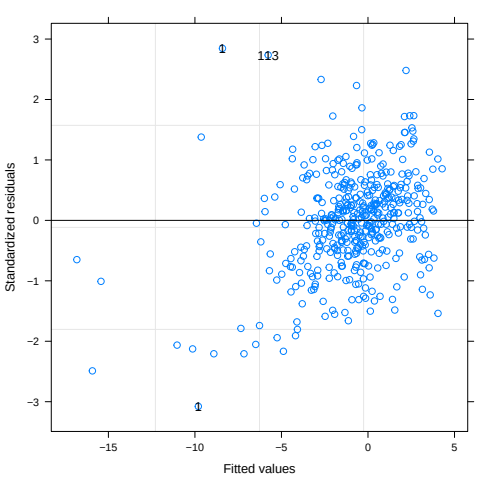
```
(2.504619e-06)^2 / ((2.504619e-06)^2 + 0.1223433^2)
```

```
...
NA
NA
NA
NA
...
```

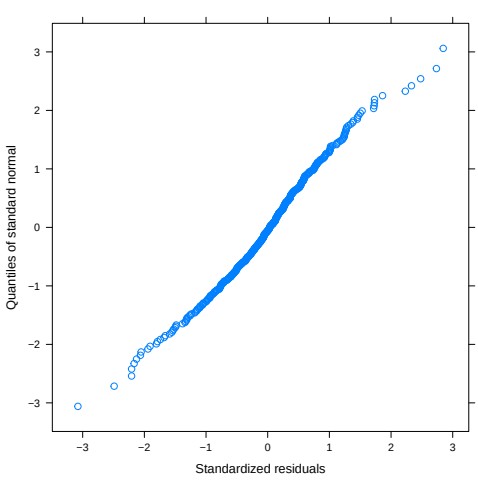
The variability in BMD that is due to the clinic random effect is 4.191051e-10, very small.

Appendix

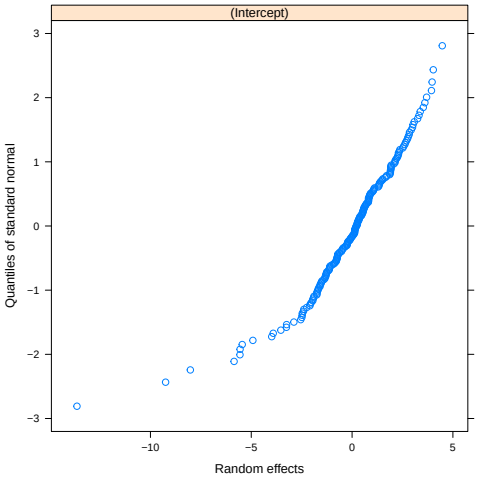
Figures



(a) residual plot of m.lme4



(b) qqnorm of residuals



(c) qqnorm of random effects

Figure 1: Model Diagnostics

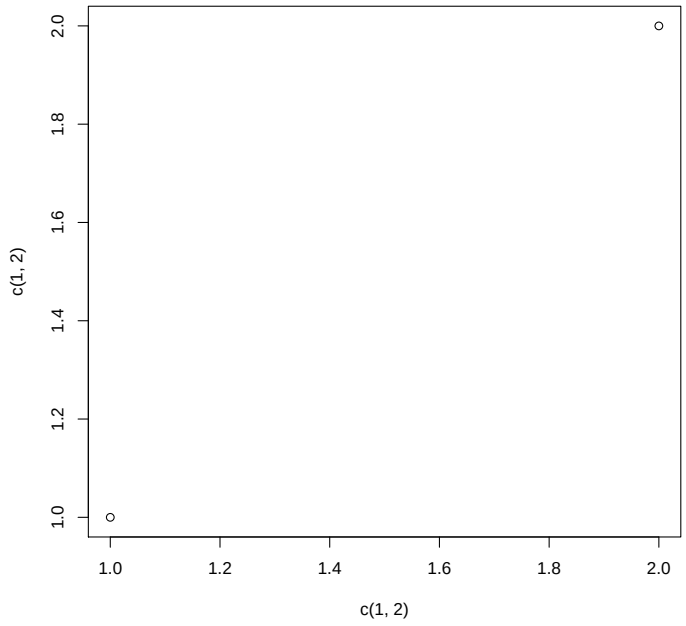


Figure 2: Scatter Plot Age vs. np.chg

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)
```

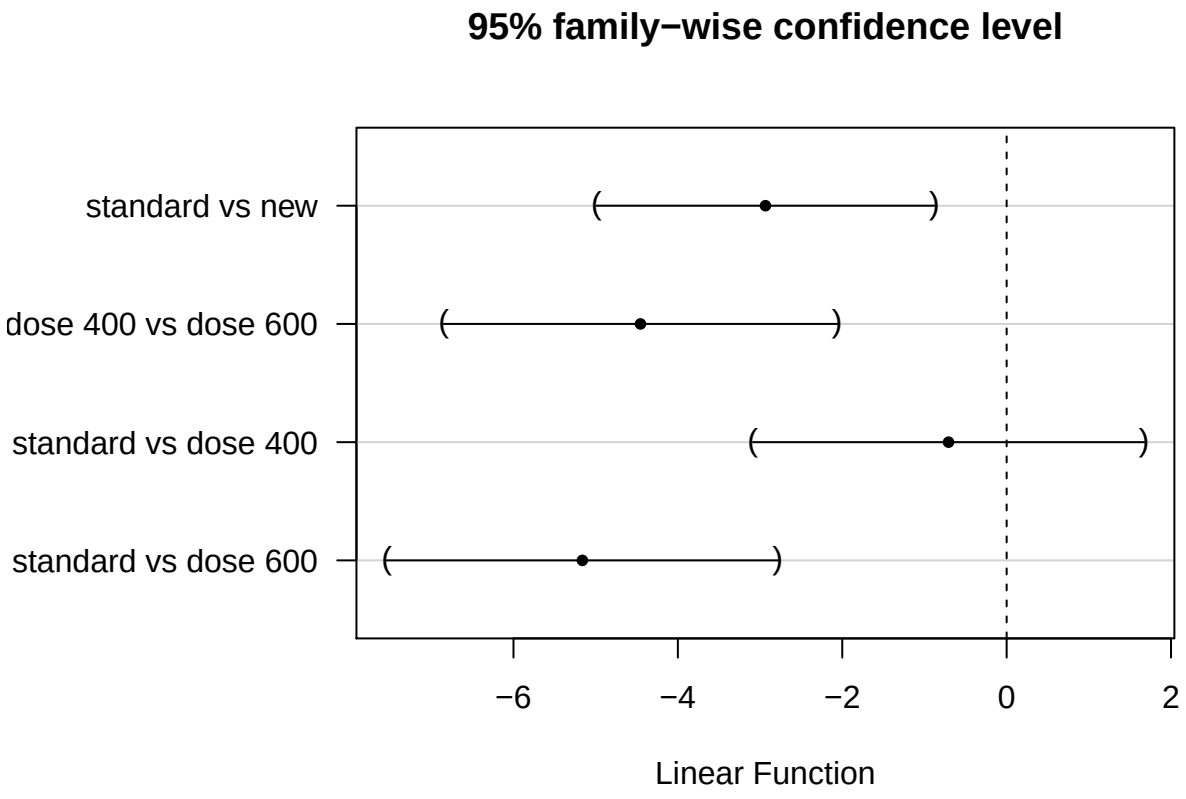
```
ma.lme2 <- update(ma.lme1, method = "ML")
ma.lme3 <- update(ma.lme2, fixed = . ~ drug * statin + factor(visit) +
                    ldl_0 + gender)
ma.lme4 <- update(ma.lme2, fixed = . ~ drug + statin + factor(visit) +
                    ldl_0 + gender)
anova(ma.lme2, ma.lme3, ma.lme4)
```

```
        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 ***
...
```

```
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
...
```

```
# 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.

```
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.

```
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%.

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

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

```
exp(-0.67473)
```

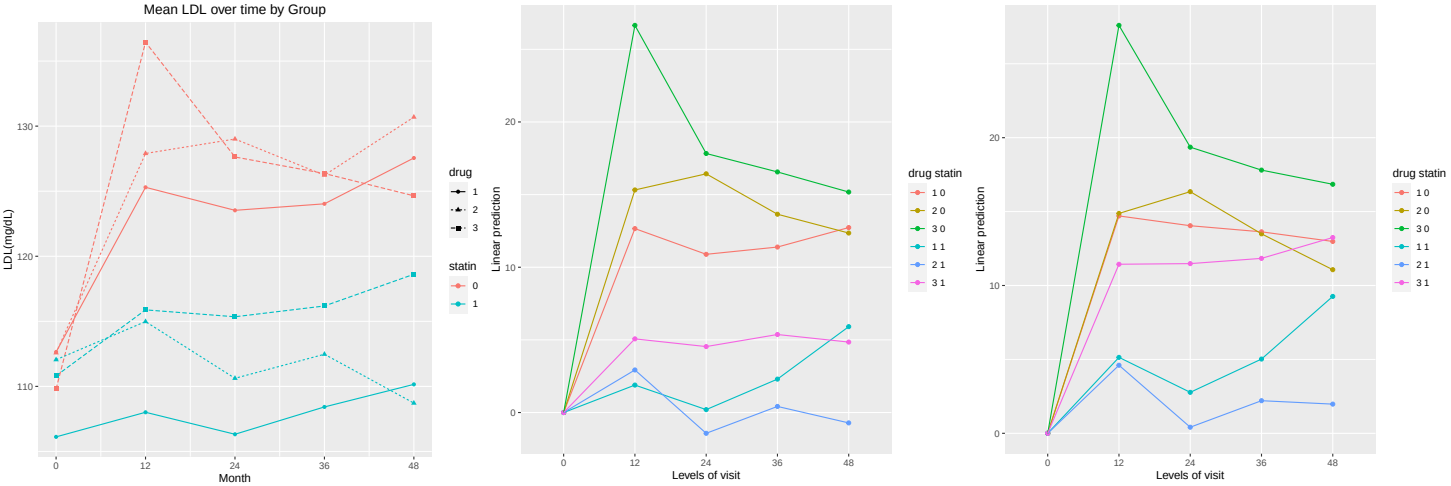
```
[1] 0.509
```

Appendix

Packages

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

Figures

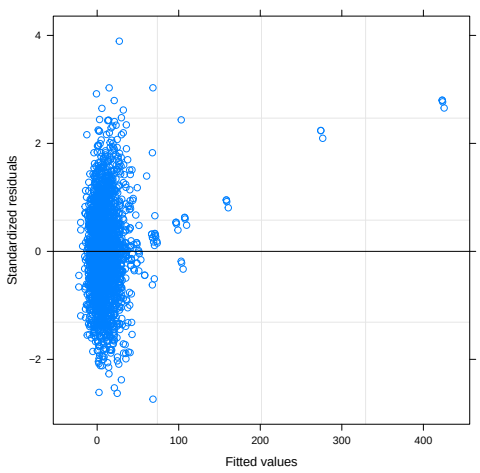


(a) residual plot of ma.lme0

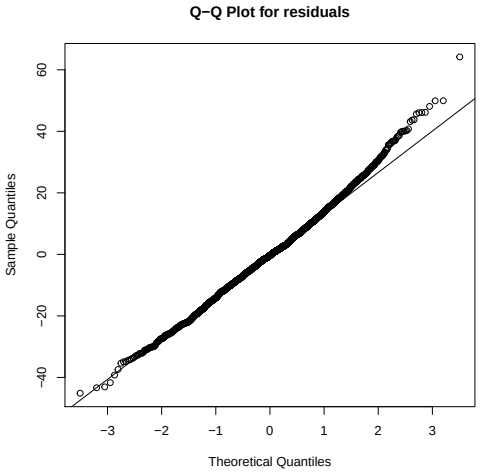
(b) qqnorm of residuals

(c) qqnorm of random effects of ma.lme0

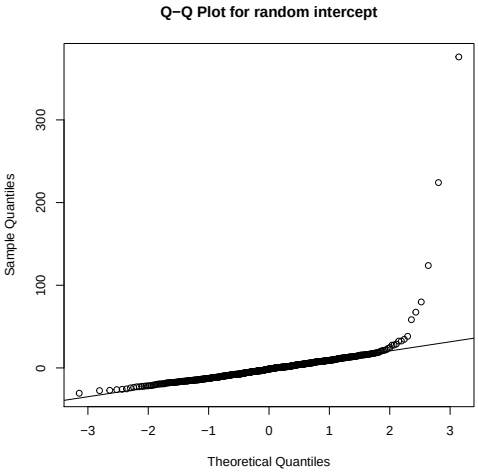
Figure 1: Model Diagnostics



(a) residual plot of ma.lme0

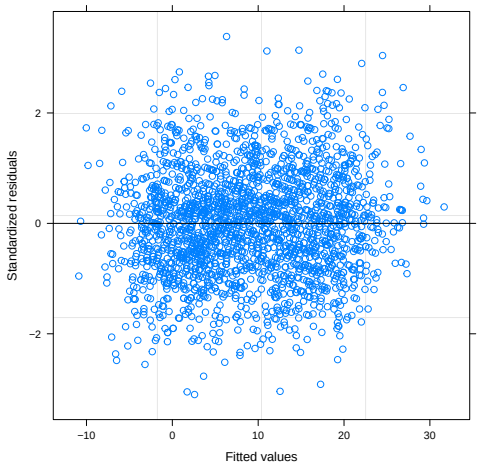


(b) qqnorm of residuals

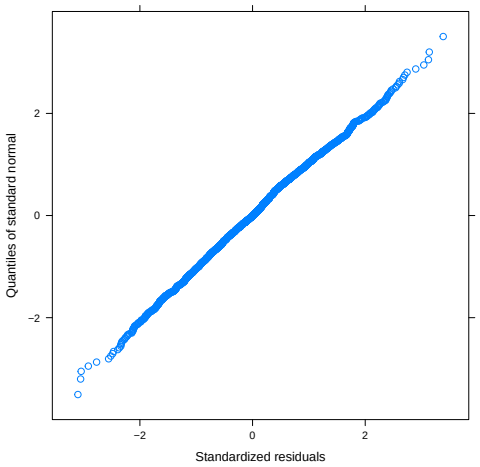


(c) qqnorm of random effects of ma.lme0

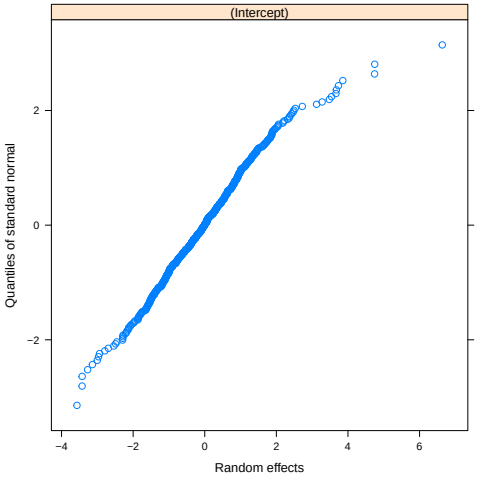
Figure 2: Model Diagnostics



(a) residual plot of ma.lme1



(b) qqnorm of residuals



(c) qqnorm of random effects of ma.lme1

Figure 3: Model Diagnostics