2009 Q2

November

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

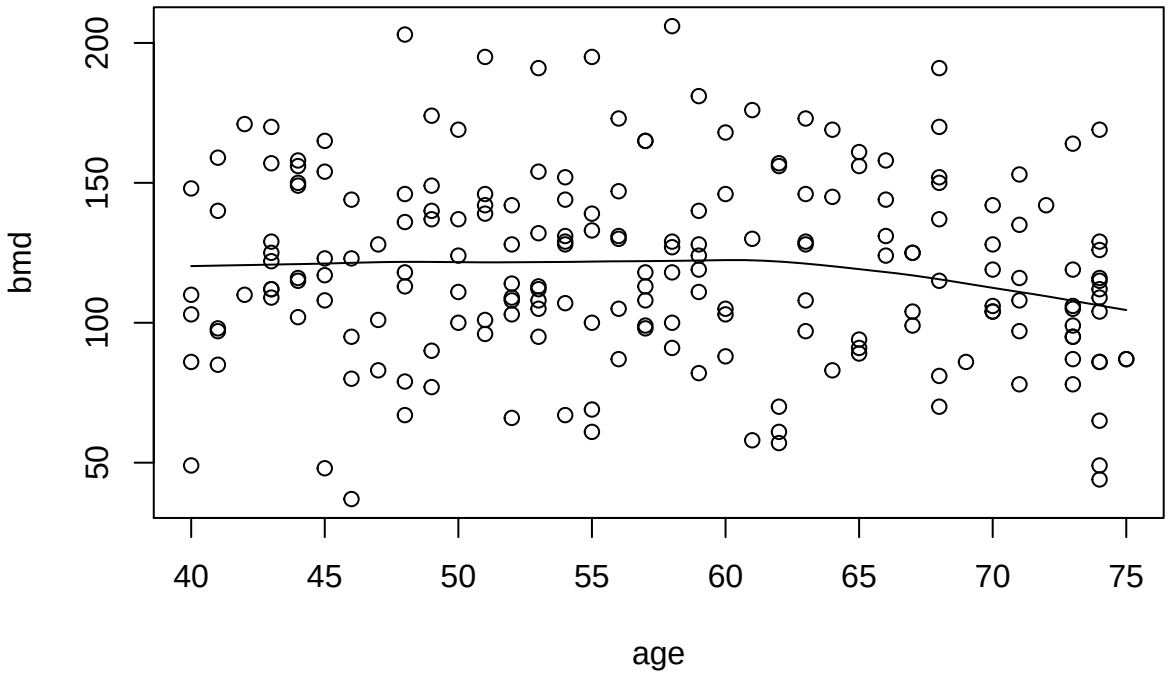
```
##        id    gender       age  diabetes   smoking      drug     visit       bmd
## "integer" "integer" "integer" "integer" "integer" "integer" "integer" "integer"
```

```
# 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
```

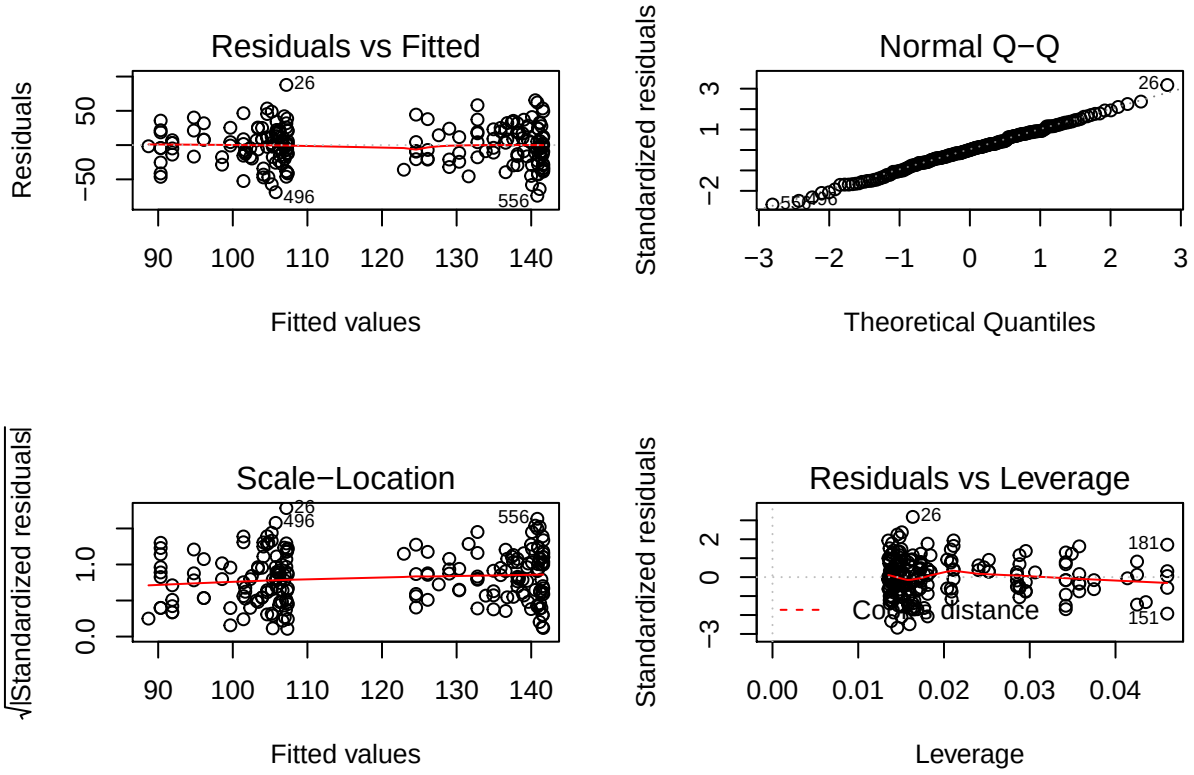
```
anova(m, m2)
```

```
## 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
```

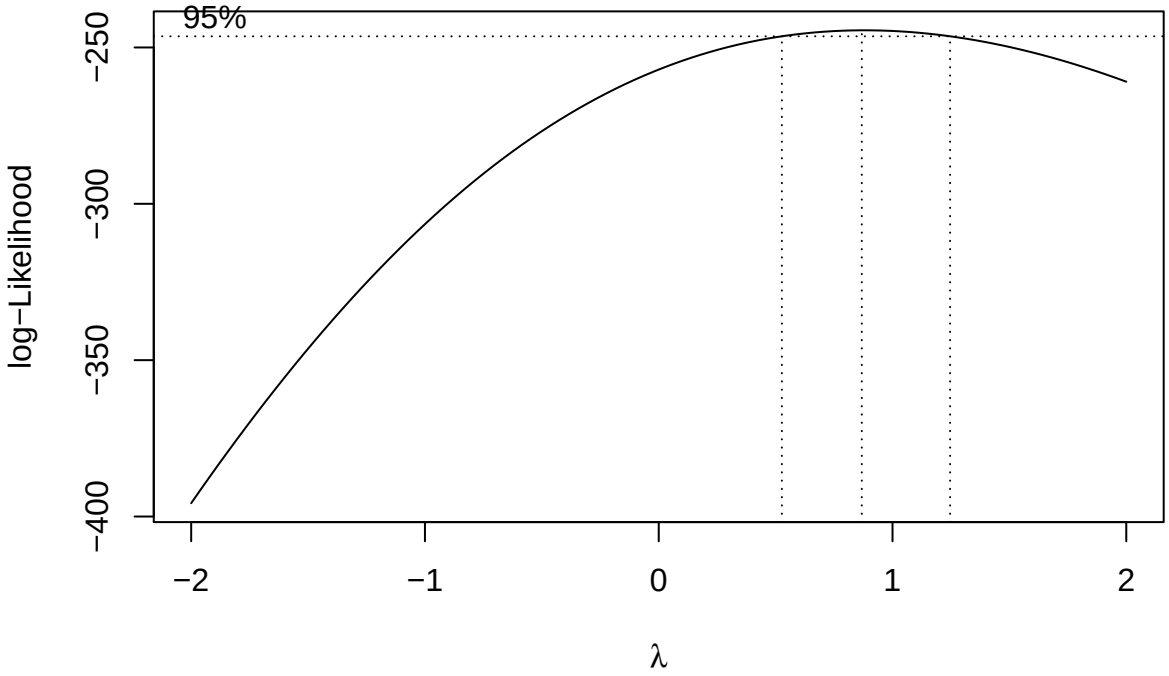
```
m3 <- lm(bmd ~ gender + I(age-57) + I((age-57)^2), data = d0)
m4 <- lm(bmd ~ gender + I(age-57), data = d0)
anova(m4, m3, m2)
```

```
## 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.

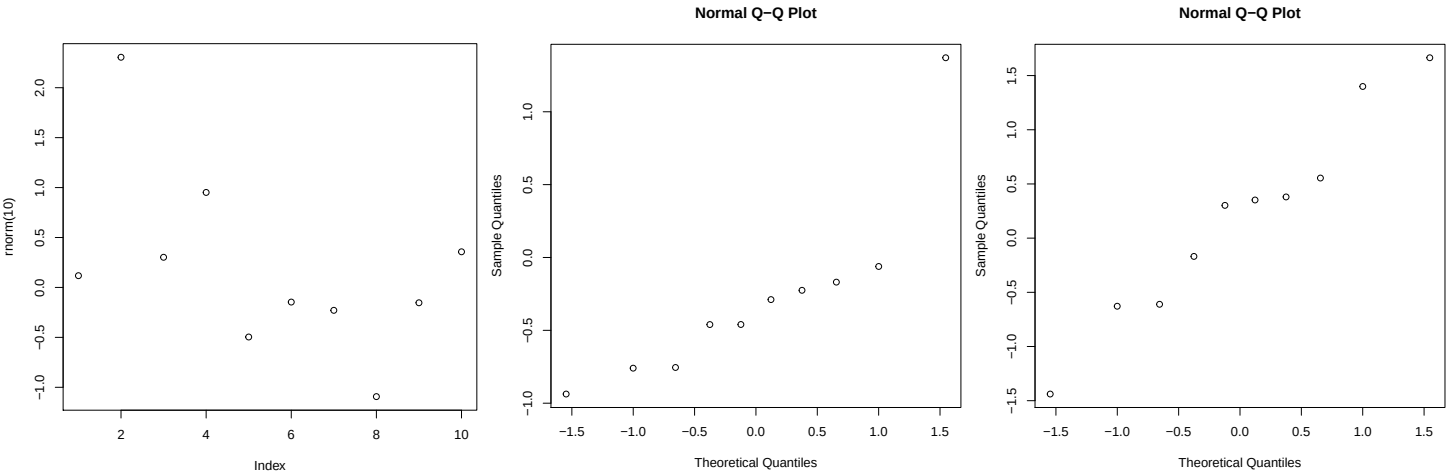
Packages

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

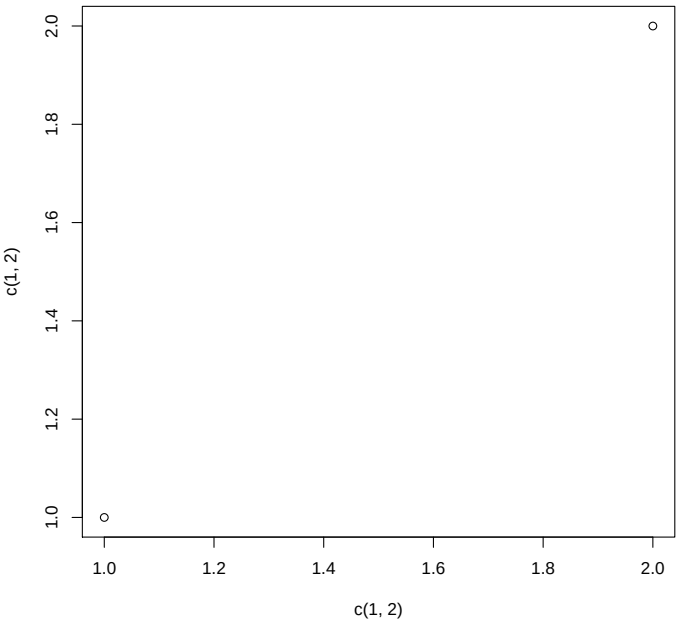


Figure 2: Scatter Plot Age vs. np.chg