571HW2

Coco_Luo

2023-01-21

Question 1

1).

Based on lecture notes LMM3, we know that

$$l(Y,b) = \frac{-1}{2}(Y - X\beta - Zb)^T R^{-1}(Y - X\beta - Zb) - \frac{1}{2}b^T Db + C$$

Since $R = R(\theta)$ and $D = D(\theta)$,

$$L(b|Y,\hat{\beta},\theta) \propto e^{-1/2}[b^T(Z^T R^{-1} Z + D^{-1})b - 2(Y - X\beta)^T R^{-1} Zb]$$

$$b|Y,\hat{\beta},\theta \sim N[(Z^T R^{-1} Z + D^{-1})^{-1} Z^T R^{-1}(Y - X\beta), (Z^T R^{-1} Z + D^{-1})^{-1}]$$

Now we want to show that $(Z^T R^{-1} Z + D^{-1})^{-1} Z^T R^{-1}(Y - X\beta) = DZ^T V^{-1}(Y - X\beta)$ where $V = ZDZ^T + R$

Observe that this is equivalent as showing:

$$\begin{gathered}
(Z^T R^{-1} Z + D^{-1})^{-1} Z^T R^{-1} = DZ^T V^{-1} \\
Z^T R^{-1} V = (Z^T R^{-1} Z + D^{-1}) DZ^T \\
Z^T R^{-1}(ZDZ^T + R) = Z^T R^{-1} ZDZ^T + Z^T \\
Z^T R^{-1} ZDZ^T + Z^T = Z^T R^{-1} ZDZ^T + Z^T
\end{gathered}$$

As a result, we have completed the proof:

$$E[b|y,\hat{\beta},\theta] = (Z^T R^{-1} Z + D^{-1})^{-1} Z^T R^{-1}(Y - X\beta) = DZ^T V^{-1}(Y - X\beta)$$

2).

$$l = -\frac{1}{2}(Y - X\beta - Zb)^T R^{-1}(Y - X\beta - Zb) - \frac{1}{2}b^T D^{-1} b$$

$$l = -\frac{1}{2}(Y - [X \quad Z]\begin{bmatrix}\beta \\ b\end{bmatrix})^T R^{-1}(Y - [X \quad Z]\begin{bmatrix}\beta \\ b\end{bmatrix}) - \frac{1}{2}[\beta^T \quad b^T]\begin{bmatrix}0 & \\ & D^{-1}\end{bmatrix}\begin{bmatrix}\beta \\ b\end{bmatrix}$$

$$\frac{dl}{d(\beta,b)} = \begin{bmatrix}X^T \\ Z^T\end{bmatrix} R^{-1}[X \quad Z]\begin{bmatrix}\hat{\beta} \\ \hat{b}\end{bmatrix} + \begin{bmatrix}0 & \\ & D^{-1}\end{bmatrix}\begin{bmatrix}\hat{\beta} \\ \hat{b}\end{bmatrix}$$

$$= \begin{bmatrix}X^T \\ Z^T\end{bmatrix} R^{-1} Y$$

$$\begin{bmatrix}X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + D^{-1}\end{bmatrix}\begin{bmatrix}\hat{\beta} \\ \hat{b}\end{bmatrix} = \begin{bmatrix}X^T R^{-1} Y \\ Z^T R^{-1} Y\end{bmatrix}$$

$$* X^T R^{-1} X\hat{\beta} + X^T R^{-1} Z\hat{b} = X^T R^{-1} Y$$

$$**\, Z^TR^{-1}X\hat{\beta}+(Z^TR^{-1}Z+D^{-1})\hat{b}=Z^TR^{-1}Y$$

From $*$ and $**$ we can get $\hat{b}=(Z^TR^{-1}Z+D^{-1})^{-1}Z^TR^{-1}(Y-X\hat{\beta})$, put it back into the previous equation, we get

$$X^TR^{-1}X\hat{\beta}+X^TR^{-1}Z(Z^TR^{-1}Z+D^{-1})^{-1}(Z^TR^{-1}(Y-X\hat{\beta}))=X^TR^{-1}Y$$

$$X^T(R^{-1}X-R^{-1}Z(Z^TR^{-1}Z+D^{-1})^{-1}Z^TR^{-1})X\hat{\beta}=X^T(R^{-1}-R^{-1}Z(Z^TR^{-1}Z+D^{-1})^{-1}Z^TR^{-1})Y$$

Now, I would like to prove that $V^{-1}=R^{-1}-R^{-1}Z(Z^TR^{-1}Z+D^{-1})^{-1}Z^TR^{-1}$ by VV^{-1}

$$VV^{-1}=ZDZ^TR^{-1}+I-ZDZ^TR^{-1}Z(Z^TR^{-1}Z+D^{-1})^{-1}Z^TR^{-1}-Z(Z^TR^{-1}Z+D^{-1})^{-1}Z^TR^{-1}$$

Assuming $W=(Z^TR^{-1}Z+D^{-1})^{-1}$,

$$\begin{aligned}&ZDZ^TR^{-1}-ZDZ^TR^{-1}ZWZ^TR^{-1}-ZWZ^TR^{-1}+I\\&=ZDZ^TR^{-1}-ZDW^{-1}WZ^TR^{-1}+I\\&=ZDZ^TR^{-1}-ZDZ^TR^{-1}+I\\&=I\end{aligned}$$

As $\hat{b}=DZ^TV^{-1}(Y-X\hat{\beta})=DZ^T(R^{-1}-R^{-1}Z(Z^TR^{-1}Z+D^{-1})^{-1}ZR^{-1})(Y-X\hat{\beta})$, assuming $W=(Z^TR^{-1}Z+D^{-1})^{-1}$,

$$\begin{aligned}&DZ^T(R^{-1}-R^{-1}ZWZR^{-1})(Y-X\hat{\beta})\\&=D(W^{-1}-Z^TR^{-1}Z)WZ^TR^{-1}(Y-X\hat{\beta})\\&=D(Z^TR^{-1}Z+D^{-1}-Z^TR^{-1}Z)WZ^TR^{-1}(Y-X\hat{\beta})\\&=WZ^TR^{-1}(Y-X\hat{\beta})\\&=(Z^TR^{-1}Z+D^{-1})^{-1}Z^TR^{-1}(Y-X\hat{\beta})\end{aligned}$$

which is equivalent to the previous results of b we obtained from $*$ and $**$.

3).

$$L(\beta,\theta)=\int L(Y,b|\beta,\theta)db=\int L(Y|b)L(b)db$$

where $L(Y|b)\sim N(X\beta+Zb,R)$ and $L(b)\sim N(0,D)$. As both $L(Y|b)$ and $L(b)$ fllows normal distributions, $L(Y|\beta,\theta)$ also follows a normal distribution

$$E[Y|\beta,\theta]=E[E[Y|\beta,\theta,b]]=E[X\beta+Zb]=X\beta$$

$$\begin{aligned}Var(Y|\beta,\theta)&=E[Var(Y|\beta,\theta,b)]+Var(E[Y|\beta,\theta,b])\\&=E[R]+Var(X\beta+Zb)=R+Z^TDZ\end{aligned}$$

Thus:

$$L(Y|\beta,\theta)\sim N(X\beta,R+Z^TDZ)$$

$$V=R+Z^TDZ$$

$$l(\beta,\theta)=logL(Y|\beta,\theta)=-\frac{1}{2}ln|V|-\frac{1}{2}(Y-X\beta)^TV^{-1}(Y-X\beta)$$

4).

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij} \\ &= \begin{bmatrix} 1 & t_{ij} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 1 & t_{ij} \end{bmatrix} \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} + \epsilon_{ij} \end{aligned}$$

Which is same as the form $Y = X\beta + Zb + \epsilon$, thus $Y|b \sim N(X\beta + Zb, R)$, follow from that, we have

$$\mu_i(t) = E[Y_i(t)|b_i] = \frac{1}{m} \sum_{j=1}^{m} (\beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij})$$

$$Var(\hat{\mu}_i(t)) = E(Var(\hat{\mu}_i(t)|b_i)) + Var(E(\hat{\mu}_i(t)|b_i)) = R + Z^T D Z = R + \begin{bmatrix} 1 \\ t_{ij} \end{bmatrix} D \begin{bmatrix} 1 & t_{ij} \end{bmatrix}$$

Question 2

1).

We know that $R = \sigma^2 I_n$ and $X = \begin{bmatrix} 1 & X_1 \\ ... & ... \\ 1 & X_n \end{bmatrix}$, D = 0, Z=0, $V = R + Z^T D Z = \sigma^2 I_n$.

$$l_R(\theta) = \frac{-1}{2} ln[X^T V^{-1} X] - \frac{1}{2} ln|V| - \frac{1}{2}(Y - X\hat{\beta})^T V^{-1} (Y - X\hat{\beta})$$

$$\frac{-1}{2} ln[(\sigma^2)^{-p} |X^T X|] - \frac{1}{2} ln[(\sigma^2)^n] - \frac{1}{2\sigma^2}(Y - X\hat{\beta})^T V^{-1} (Y - X\hat{\beta})$$

$$\frac{dl_R(\theta)}{d\theta} = \frac{1}{2\sigma^2} - \frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(Y - X\hat{\beta})^T (Y - X\hat{\beta}) = 0$$

$$\hat{\sigma}^2_{REML} = \frac{1}{n-p}(Y - X\hat{\beta})^T (Y - X\hat{\beta})$$

$$\hat{\sigma}^2_{REML} = \frac{1}{n-p} \sum_{i=1}^{n} (Y_i - X_i^T \hat{\beta})^2$$

2).

In order to show that they are equivalent, we want to show that the max value of equation (1):

$$(1) - \frac{1}{2}(y - X\hat{\beta} - Z\hat{b})^T R^{-1} (y - X\hat{\beta} - Z\hat{b}) - \frac{1}{2}\hat{b} D^{-1} \hat{b}$$

$$(2) = -\frac{1}{2}(y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta})$$

$$(1) = -\frac{1}{2}[\hat{b}^T Z^T R^{-1} Z\hat{b} + \hat{b} D^{-1} \hat{b} - 2(Y - X\hat{\beta})^T R^{-1} Z\hat{b} + (Y - X\hat{\beta})^T R^{-1} (Y - X\hat{\beta})]$$

Let $A = Z^T R^{-1} Z + D^{-1}$, we have:

$$(1) = -\frac{1}{2}[(\hat{b} - A^{-1} Z^T R^{-1} (Y - X\hat{\beta}))^T A (\hat{b} - A^{-1} Z^T R^{-1} (Y - X\hat{\beta}))]$$

$$= -\frac{1}{2}[-(y - X\hat{\beta}^T R^{-1} Z A^{-1} R^{-1} (Y - X\hat{\beta})) + (Y - X\hat{\beta})^T R^{-1} (Y - X\hat{\beta})]$$

Thus, the above equation (1) is maximized when $\hat{b} = A^{-1} Z^T R^{-1} (Y - X\hat{\beta})$, The max quantity of (1) is:

$$-\frac{1}{2}(Y - X\hat{\beta})^T [R^{-1} - R^{-1} Z A^{-1} R^{-1}] (Y - X\hat{\beta})$$

Then, we want to show that $R^{-1} - R^{-1}ZA^{-1}Z^TR^{-1} = V^{-1}$ and thus (1) == (2). Since $V = R + ZDZ^T$, we only need to prove that $V(R^{-1} - R^{-1}ZA^{-1}Z^TR^{-1}) = I$

$$(R + ZDZ^T)(R^{-1} - R^{-1}ZA^{-1}Z^TR^{-1}) = I$$

$$I + ZDZ^TR^{-1} - ZA^{-1}Z^TR^{-1} - ZDZ^TR^{-1}ZA^{-1}Z^TR^{-1} = I$$
$$I + Z[D - A^{-1} - DZ^TR^{-1}ZA^{-1}]Z^TR^{-1} = I$$
$$I + Z[D - (I - DZ^TR^{-1}Z)A^{-1}]Z^TR^{-1} = I$$
$$I + Z[D - D(D^{-1} - Z^TR^{-1}Z)A^{-1}]Z^TR^{-1} = I$$

Since $A = D^{-1} - Z^TR^{-1}Z, I = I$, $V(R^{-1} - R^{-1}ZA^{-1}Z^TR^{-1}) = I$ and $R^{-1} - R^{-1}ZA^{-1}Z^TR^{-1} = V^{-1}$ the max value of,

$$(1) - \frac{1}{2}(y - X\hat{\beta} - Z\hat{b})^TR^{-1}(y - X\hat{\beta} - Z\hat{b}) - \frac{1}{2}\hat{b}D^{-1}\hat{b} = (2) - \frac{1}{2}(y - X\hat{\beta})^TV^{-1}(y - X\hat{\beta})$$

3).

$$L_{REML}(\theta) = \int L(Y; \beta, \theta)d\beta$$

$$= e^{-1/2ln|V|}\int e^{-1/2[\beta^TX^TV^{-1}X\beta - 2Y^TV^{-1}X\beta + Y^TV^{-1}Y]}d\beta$$

$$= e^{-1/2ln|V|}\int e^{-1/2[\beta-(X^TVX^{-1}X^TV^{-1}Y)^T(X^TV^{-1}X)(\beta-(X^TVX)^{-1}X^TV^{-1}Y)]}e^{-1/2[Y^TV^{-1}Y]-Y^TV^{-1}X(X^TV^{-1}X)^{-1}XV^{-1}Y}d\beta$$

Since $\hat{\beta} = (X^TV^{-1}X)^{-1}X^TV^{-1}Y$,

$$L_{REML}(\theta) = e^{-1/2ln|V|}e^{-1/2[Y^TV^{-1}Y-\hat{\beta}^TX^TV^{-1}X\hat{\beta}]}\int e^{-1/2[\beta-(X^TVX^{-1}X^TV^{-1}Y)^T(X^TV^{-1}X)(\beta-(X^TVX)^{-1}X^TV^{-1}Y)]}d\beta$$

$$= e^{-1/2ln|V|}e^{-1/2ln|X^TV^{-1}X|}e^{-1/2[(Y-X\hat{\beta})^TV^{-1}(Y-X\hat{\beta})]}$$
$$= e^{-1/2ln|X^TV^{-1}X|-1/2ln|V|-1/2[(Y-X\hat{\beta})^TV^{-1}(Y-X\hat{\beta})]}$$

We can use numerical methods like Laplace approximatoin or Markov Chain to compute the integral over β. And we can maximize the REML likelihood function with respect to θ to get the REML estimate of θ. Therefore, we have shown that the REML likelihood can be obtained by the Bayesian model assuming a flat prior for β.

4).

5).

Question 3

Prompt: We want to conduct a study where we put N individuals to two groups of the same size and assign treatments to each group. If we measure a numerical variable on each individual for several times. What is the better approach, average all measurement for each individual, Linear regression model or a mixed model with random intercept?

Answer: It depends on the specific research question and the nature of the data. If the goal of the study is to compare the overall treatment effects across the two groups, then averaging the 10 measurements for each individual and comparing the means of the two groups using a t-test or ANOVA would be appropriate. If the goal is to examine the individual differences in response to the treatment, then a linear regression model with an interaction term between treatment group and time would be appropriate. In this model, the slope of the line representing the change in the outcome over time would be different for the two groups, and the intercept would represent the baseline level of the outcome for each group. If the goal is to examine the individual differences in response to the treatment and control for the clustering effect, then a mixed model with random intercept would be appropriate. This model would account for the dependence among the observations within the same group, and allow for the estimation of both fixed effects (treatment) and random effects (intercept). It's important to note that before deciding which model to use, it's important to check the assumptions of the model and the data, such as normality and independence of errors, as well as the estimation method.

Prompt: What are the pros and cons of using the mixed model with random intercept?

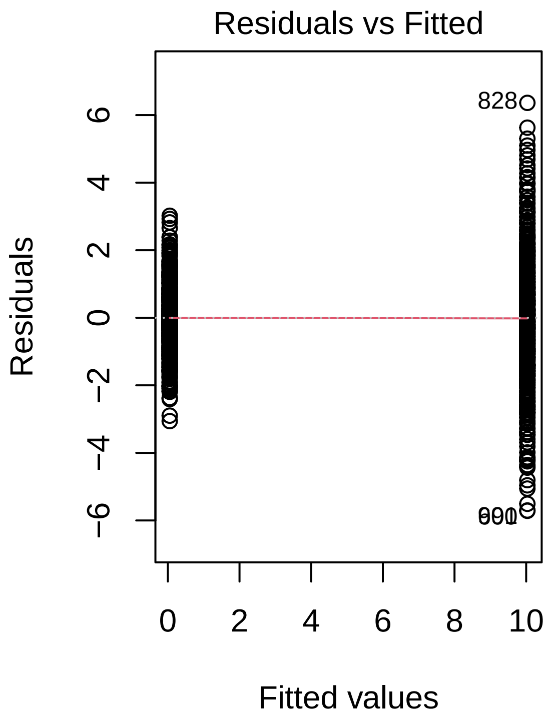
Answer: The pros and cons of using a mixed model with a random intercept include: Pros: They can account for the non-independence of observations that occur in clustered or hierarchical data structures, such as when observations are nested within subjects or within groups. They can increase the precision of the estimates by accounting for the additional variation due to the clustered structure of the data. They can account for the potential confounding effect of clustering or grouping on the outcome variable. They can be used to estimate both fixed and random effects, which allows for more complex and informative models. Cons: They can be computationally intensive and may require specialized software and expertise to fit. They may not be appropriate if the data does not have a clear clustered or hierarchical structure. They require a large enough sample size to estimate the random effects. They need to be interpreted with care, as the presence of random effects can change the interpretation of fixed effects. Additionally, it's important to keep in mind that any mixed model, including a mixed model with a random intercept, relies on the assumptions of the model being met, such as normality and constant variance of the residuals. It's important to check these assumptions and address any violation of them.

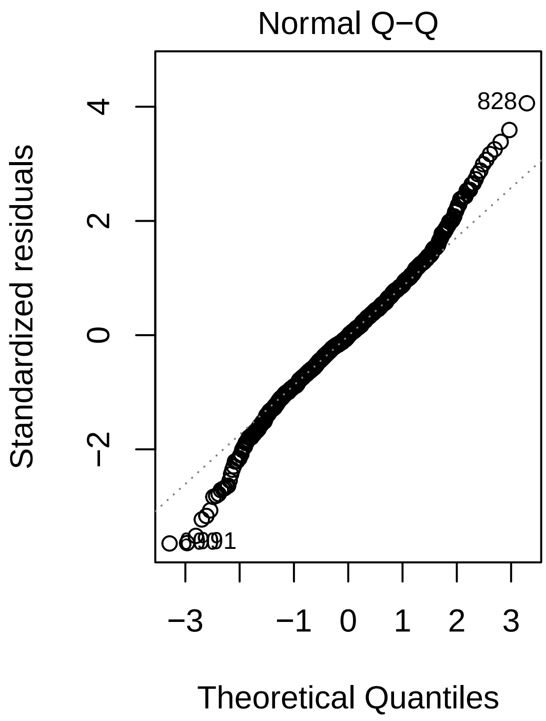
I think the answer given by the openchat AI looks great, different measurements or models have their pros and cons based on certain situations; I will run a stimulation study and see if any of the methods will provide a statistically significant result.

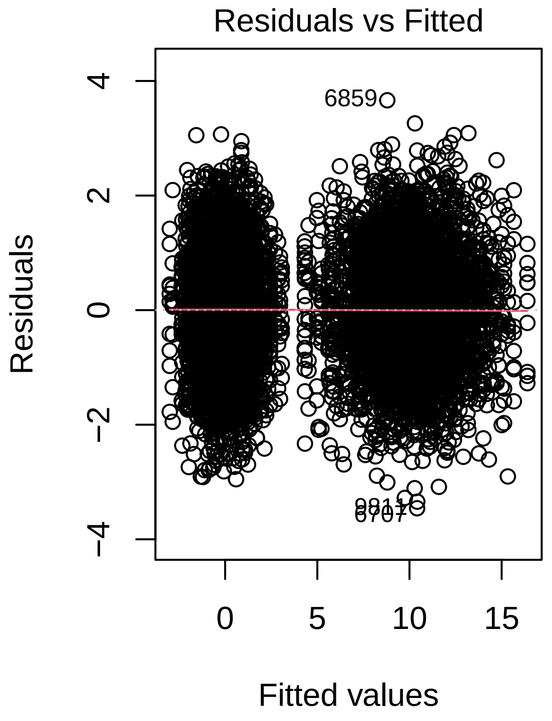
Assuming $n = 1000$, the two treatment groups are labeled as A and B. Also assume that within each treatment group, the 10 repeated measurements on the same person are i.i.d distributed. The error terms $\epsilon \sim N(0, \sigma^2)$. The true measurements in group A $Y_A \sim N(\mu_A, \sigma_A^2 I)$ and the true measurements in group B $Y_b \sim N(\mu_B, \sigma_B^2 I)$. I assigned $\mu_A = 0, \mu_B = 10, \sigma = 1, \sigma_A = 1, \sigma_B = 2$.

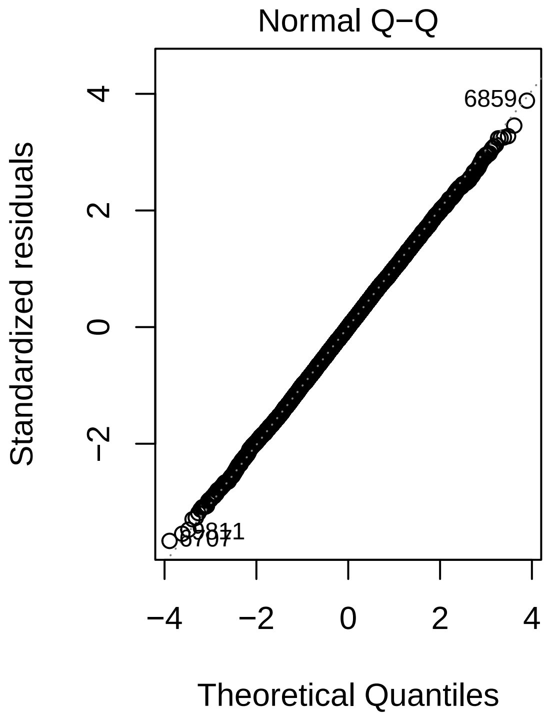
More details about the code will be shown in the appendix. All simulation methods have generated models with statistically significant p-values.

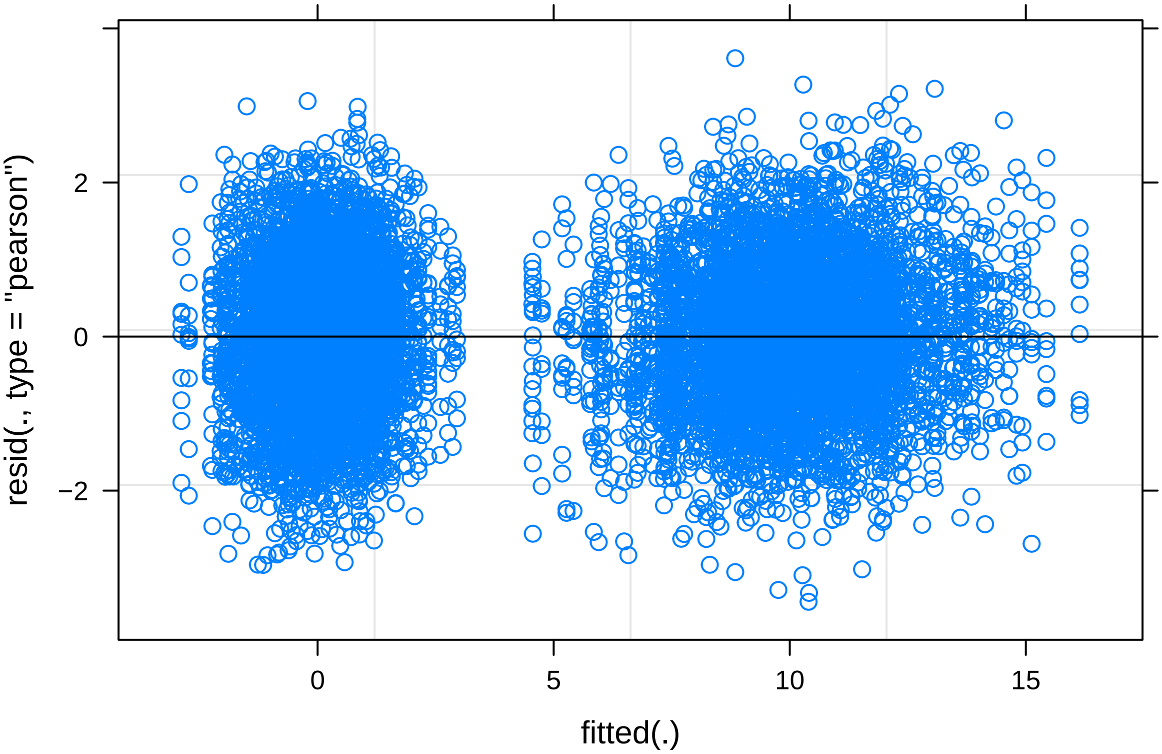
I ploted the residual plot and normal QQ plot to see how well the models fit.


Residuals vs Fitted
Residuals
Fitted values
6
4
2
0
−2
−4
−6
0
2
4
6
8
10
828
691


Normal Q−Q
Standardized residuals
Theoretical Quantiles
4
2
0
−2
−3
−1
0
1
2
3
828
691


Residuals vs Fitted
Residuals
Fitted values
4
2
0
−2
−4
0
5
10
15
6859
9811
6707

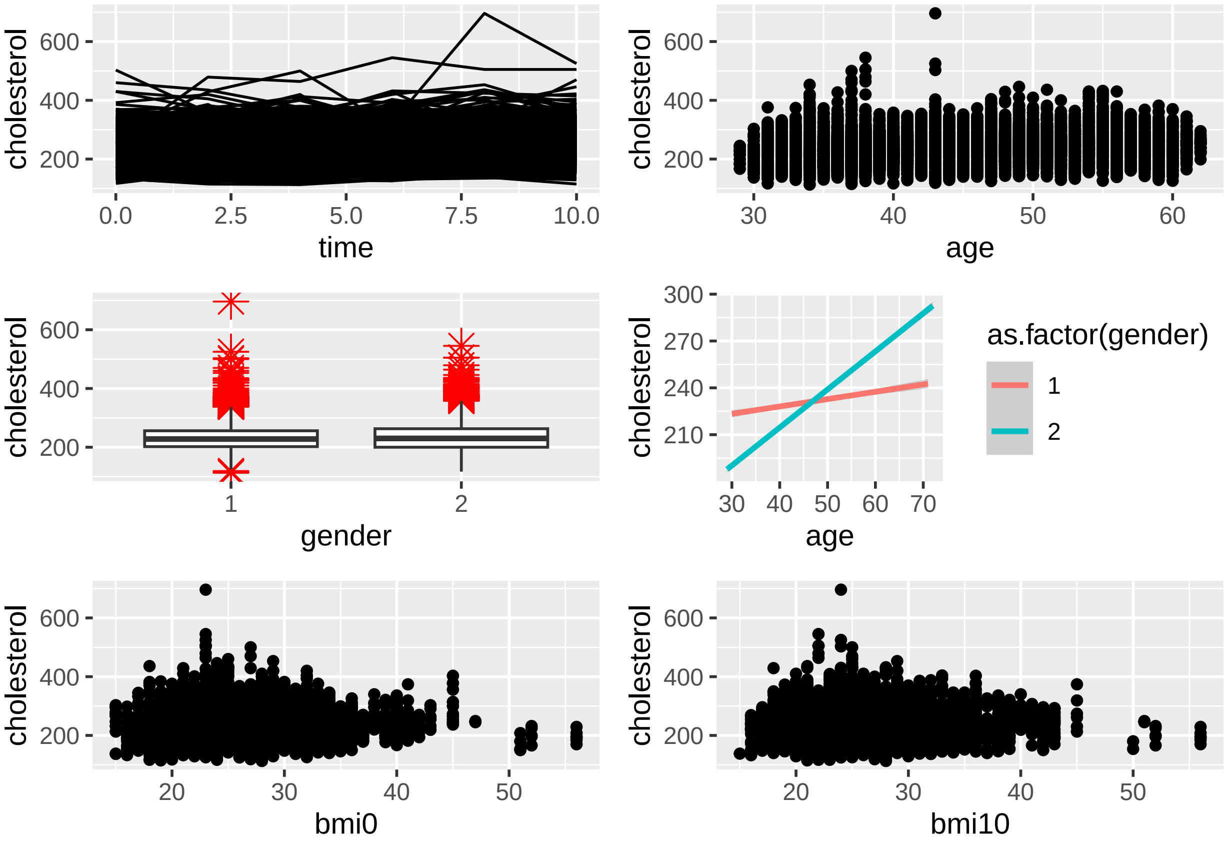

Normal Q−Q
Standardized residuals
Theoretical Quantiles
4
2
0
−2
−4
−2
0
2
4
6859
9811
6707

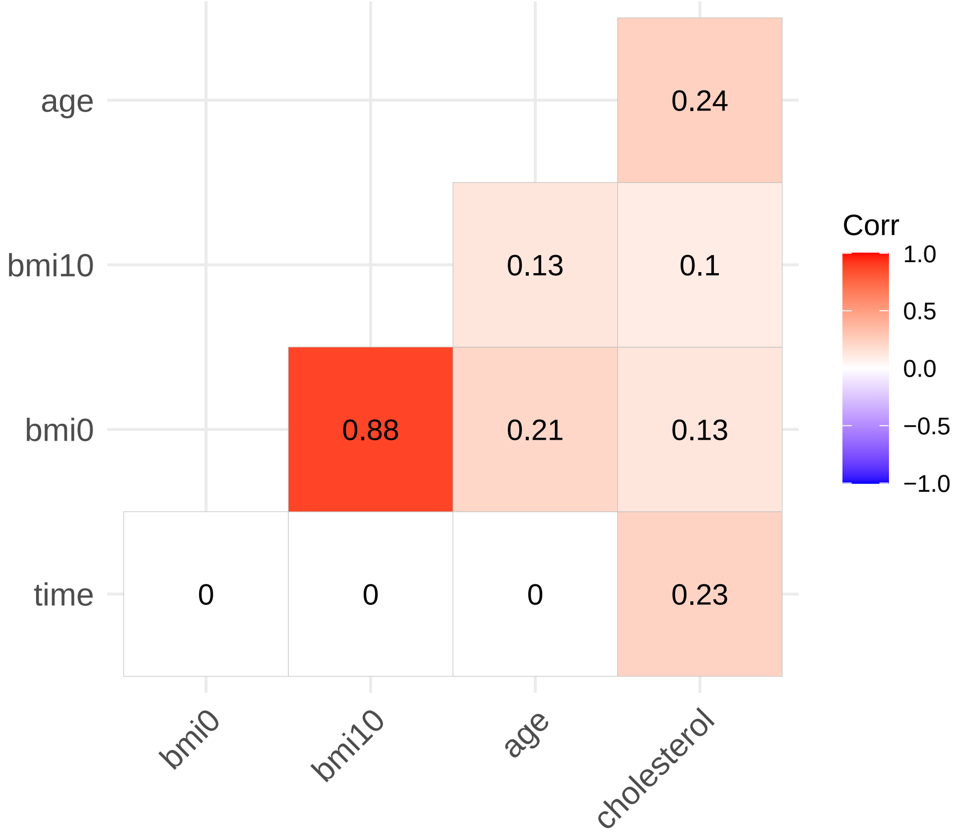


From the above plots, I conclude that the linear regression and mixed effect models perform better than the first approach. From a time-efficient perspective, however, linear regression and mixed effect models will cost a lot more but also reduce biases and possible errors in exchange.

Question 4

Before analysis, I rename the columns to be what they originally represent and adjusted the cholesterol levels into one column. Then, I convert -9 to NA. To investigate how cholesterol has changed over time and its relationships with age, gender, and BMI, I first used a line plot to visualize cholesterol level over time, we can observe that different subjects may have a different cholesterol level. Thus, I would like to add a random intercept and a random slope related to time when fitting our model. Looking at the scatterplots between cholesterol and age, we can see that cholesterol becomes higher when people becomes older, and it seems that cholesterol levels increases a lot in the females group as they become older, while the cholesterol level in males does not change a lot over age. If we only look at the boxplot of cholesterol across two genders, we cannot see much differences in the distribution. Males have a wider tails than females, and more outliers. Finally, from the last scatterplots of the relationship between BMI and cholesterol, I cannot see a strong relationship between those variables. In addition, from the correlation plots, bmi0 and bmi10 has a correlation score as large as 0.88 where bmi0 has a higher correlation with cholesterol, therefore I will only include age, gender, time, bmi0 and the interaction between gender and time in my model.





My fitted model is $E[Y_{ij}|b_i] = \beta_0 + \beta_1 \cdot age_i + \beta_2 \cdot gender_i + \beta_3 \cdot time_j + \beta_4 \cdot gender_i : time_j + \beta_5 \cdot bmi0_i +$

$b_{0i} + b_{1i} \cdot time_j$, and the estimated coefficients are shown below. We can see from the estimates that `gender` and `time` have the largest impact on the cholesterol level, followed by the age, the interaction term and finally bmi0. One example of interpreating the above estimates is for each unit increase in age, the cholesterol level will on average increase by 1.31 while holding all other variables constant. In general, we notice that an older male with a higher baseline BMI tends to have higher cholesterol level. And the cholesterol level tends to increase over time for all gender, especially for females. Since `gender` is marked as 1 for male and 2 for female, we do expect that females will have a relatively higher cholesterol level compared to males holding all the other factors constant.

	Estimate	Std. Error	t value
(Intercept)	140.432	5.273	26.634
age	1.309	0.090	14.595
gender2	-2.627	1.527	-1.720
time	2.341	0.101	23.253
bmi0	0.938	0.173	5.407
gender2:time	1.106	0.137	8.092

2).

(3). Suppose you have time to develop your own program, what is a better and more systematic approach (or model) would you like to propose to answer the question in (2)? From what we got in the first problem, I would like to add `rate` into my model, which is the estimated coefficient for time plus the estimated coefficient for gender times gender and plus b1i. The logistic regression I am going to fit for p_i is (the probability of death rate after 30-years for subject i):

$$log\frac{p_i}{1-p_i} = \beta_0 + \beta_1 \cdot age_i + \beta_2 \cdot gender_i + \beta_3 \cdot cholesterol0_j + \beta_4 \cdot bmi0_i + \beta_4 \cdot rate_i$$

And we got the estimates for the logistic regression as below:

	Estimates	Odds Ratio	lower CI	upper CI	p-value
age	0.1070	1.1129	1.0982	1.1279	0.0000
gender2(Female)	-0.7873	0.4551	0.3604	0.5746	0.0000
bmi0	0.0175	1.0176	0.9933	1.0425	0.1562
cholesterol0	0.0018	1.0018	0.9994	1.0041	0.1374
rate	-0.0149	0.9852	0.9037	1.0741	0.7356

3).

I noticed the missing values take up about one third of the entire dataset. By dropping these values, we lost a large amount of useful information. If I am allowed to have more time for the analysis, I would try to impute these missing values, such as taking their median or mean or use prediction methods to generate similar figures. In addition, I noticed the categorical variable `gender` is labeled with 1 and 2. This could have introduced unwanted biases in the above linear and logistic regression. I would probably try to perform one-hot encoding on the `gender` variable to avoid such issues. In addition, we fit the model in problem 4(2) with limited variables. If I have time to develop my program, I may explore more about the dataset to see if there are any potential confounders. If I do, I will refit my model with those confounders and determine if those variables are statically significant. Last but not the least, I would consider other models that might give me a better performance. Those models include the parametric survival model, the accelerated failure time model, Bayesian survival analysis and etc. The last one could deal with uncertainty in estimation, handle missing data, and solve time-dependent covariates.

Appendix

```
# problem 3
# data generation
set.seed(571)
n = 1000
eps_A = rnorm(10*n/2)
eps_B = rnorm(10*n/2)
y_A <- rep(rnorm(n/2, 0, 1), each=10) + eps_A
y_B <- rep(rnorm(n/2, 10, 2), each=10) + eps_B


# data selection
y = c(y_A, y_B)
index_range = c(seq(1, 10*n, 10), 10*n+1)
y_random  = y_average = c()


for (i in 1:n) {
  i_range = index_range[i]:(index_range[i+1]-1)
  i_index = sample(i_range, 1)
  y_random = c(y_random, y[i_index]) # data selected randomly
  y_average = c(y_average, mean(y[i_range])) # data selected by averaging
}

# averages
df_average = data.frame(id = 1:n, y = y_average, x = rep(c("A","B"), each = n/2))
lm2 <- lm(y~x, data = df_average)

# linear regression
library(lme4)
df_d  = cbind(y, dummy(rep(1:n, each = 10)))
colnames(df_d) = c("y", paste0("x_", 2:n))
df_d = as.data.frame(df_d)
lm3 = lm(y~.,data = df_d)

# random intercept
df_lmm = data.frame(id = rep(1:n, each = 10), y = y, x = rep(c("A","B"), each = 10*n/2))
lm4 = lmer(y ~ x+(1|id), data = df_lmm)
```

```
## problem 4
dat <- read.table("~/Desktop/571/HW2/framingham.dat", quote="\"", comment.char="")
```

```
library(reshape2)
library(dplyr)
library(lme4)



colnames(dat) = c("age", "gender", "bmi_baseline","bmi_year10", "cigarattes",
                   "cholesterol_year0",  "cholesterol_year2",  "cholesterol_year4",
                   "cholesterol_year6",  "cholesterol_year8", "cholesterol_year10",
                   "dead")
dat["individuals"] = 1:nrow(dat)
```

```
dat[dat==-9]<-NA

Framingham = data.frame(id = rep(1:nrow(dat),6),
                        age = rep(dat[,1], 6),
                        gender = as.factor(rep(dat[,2],6)),
                        bmi0 = rep(dat[,3], 6),
                        bmi10 = rep(dat[,4], 6),
                        cigarattes = rep(dat[,5], 6),
                        cholesterol = c(dat[,6], dat[,7], dat[,8], dat[,9], dat[,10],
                                        dat[,11]),
                        time = rep(c(0,2,4,6,8,10),each = nrow(dat)),
                        dead = rep(dat[,12],6))

df = Framingham[complete.cases(Framingham),]
dat = dat[complete.cases(dat),]
```