# Stats112 HW2

Viraj Vijaywargiya

2023-04-26

```
library(lattice)
library(nlme)
library(lme4)
```

```
## Loading required package: Matrix

##
## Attaching package: 'lme4'

## The following object is masked from 'package:nlme':
##
##     lmList
```

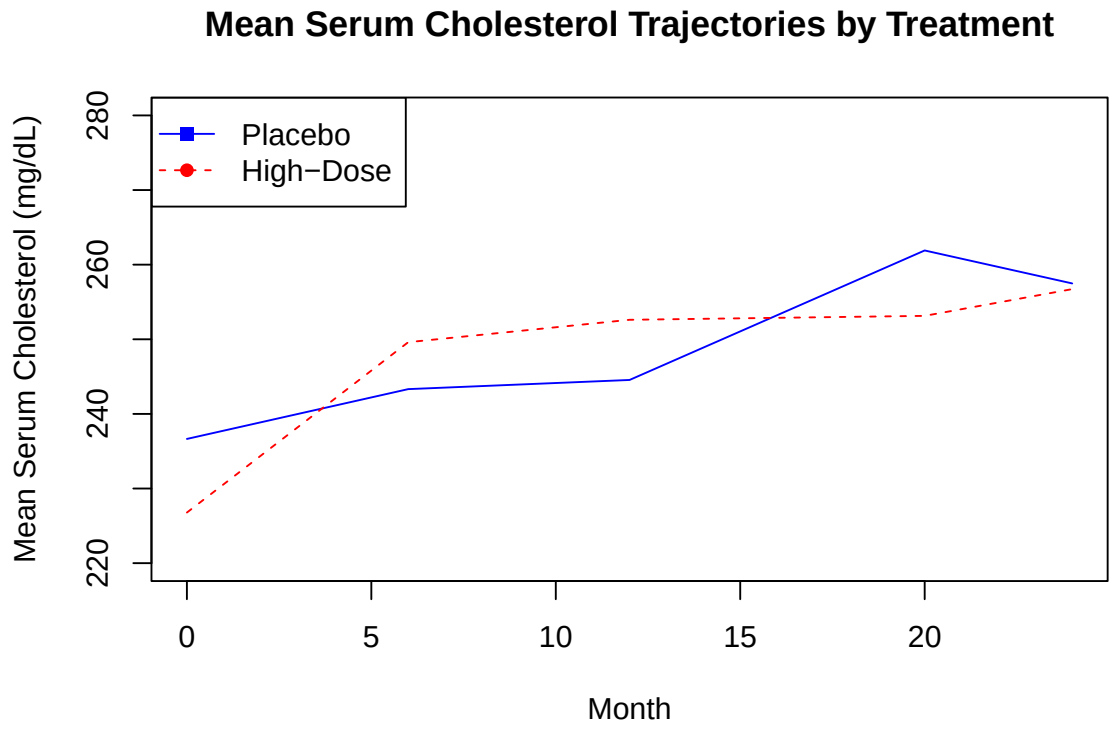
```
library(survival)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::collapse() masks nlme::collapse()
## x tidyr::expand()   masks Matrix::expand()
## x dplyr::filter()   masks stats::filter()
## x dplyr::lag()      masks stats::lag()
## x tidyr::pack()     masks Matrix::pack()
## x tidyr::unpack()   masks Matrix::unpack()
```

1. 
```
NCGS = read.table("/Users/virajvijaywargiya/Downloads/cholesterol-data.txt", na.strings=".")

NCGS = NCGS[complete.cases(NCGS), ]

names(NCGS) = c("Trt", "ID", "M0", "M6", "M12", "M20", "M24")

NCGS$Trt = factor(NCGS$Trt, levels=c(2,1), labels=c("Placebo","HighDose"))

NCGS.long = reshape(NCGS, varying=list(3:7), idvar="ID", timevar="Month", times=c(0,6,12,20,24), v

NCGS.long$Time = as.numeric(factor(NCGS.long$Month))
```

1a)

```
means = tapply(NCGS.long$Chol, list(NCGS.long$Month, NCGS.long$Trt), mean)
times = c(0,6,12,20,24)
plot(times, means[,1], type="l", xlab="Month",
    ylab="Mean Serum Cholesterol (mg/dL)",
    ylim=c(220,280), main="Mean Serum Cholesterol Trajectories by Treatment",
    col="blue", lty=1, pch=15)
points(times, means[,2], type="l",
    col="red", lty=2, pch=16)
legend("topleft",c("Placebo","High-Dose"),
    col=c("blue","red"), lty=c(1,2), pch=c(15,16))
```

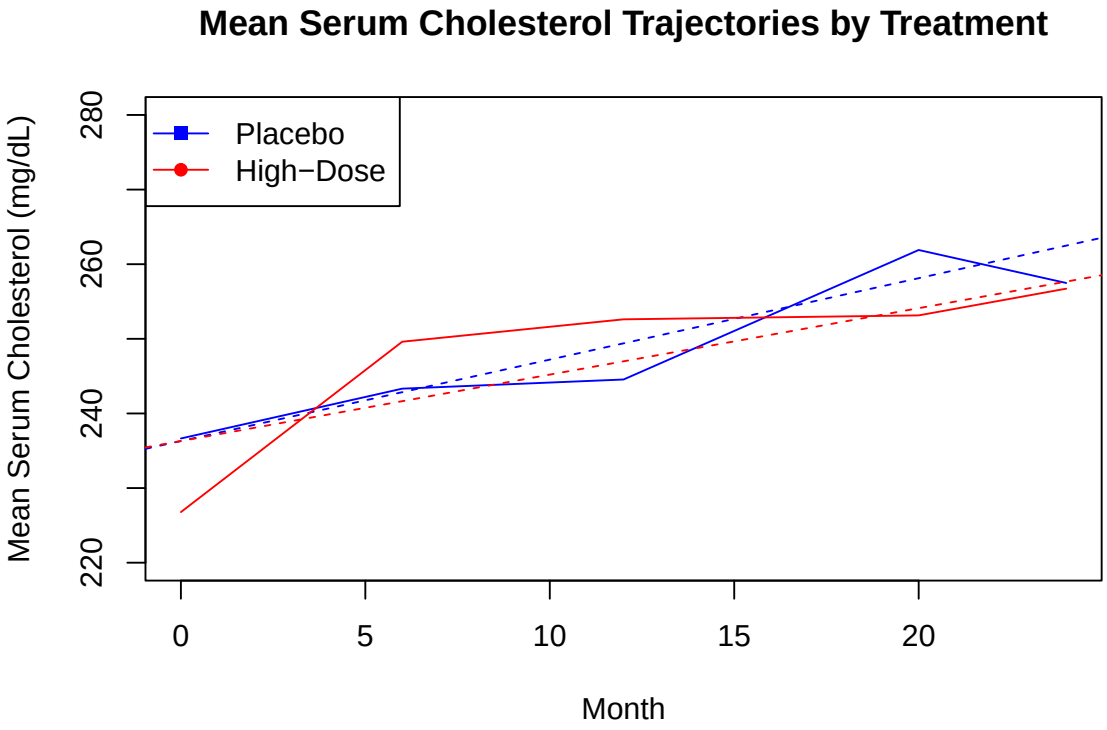



```
mod = gls(Chol ~  Month +Trt:Month , data=NCGS.long, weight=varIdent(form = ~ 1 | Time),corr=corSym
summary(mod)
```

```
## Generalized least squares fit by REML
##   Model: Chol ~ Month + Trt:Month
##   Data: NCGS.long
##        AIC      BIC    logLik
##   3284.969 3353.461 -1624.484
##
## Correlation Structure: General
##  Formula: ~Time | ID
##  Parameter estimate(s):
##  Correlation:
##   1     2     3     4
```

```
## 2 0.736
## 3 0.723 0.808
## 4 0.751 0.812 0.732
## 5 0.600 0.686 0.703 0.648
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | Time
##  Parameter estimates:
##         1         2         3         4         5
## 1.0000000 0.9391971 0.8703265 0.8585931 0.9968276
##
## Coefficients:
##                       Value Std.Error  t-value p-value
## (Intercept)       236.30142  5.609715 42.12360  0.0000
## Month               1.09116  0.212904  5.12510  0.0000
## Month:TrtHighDose  -0.20088  0.270289 -0.74322  0.4579
##
##  Correlation:
##                   (Intr) Month
## Month             -0.366
## Month:TrtHighDose  0.000 -0.682
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.2459962 -0.7227297 -0.0335675  0.6623791  3.6451359
##
## Residual standard error: 49.84686
## Degrees of freedom: 335 total; 332 residual
```

```
b = coef(mod)

plot(times, means[,1], type="l", xlab="Month",
    ylab="Mean Serum Cholesterol (mg/dL)",
    ylim=c(220,280), main="Mean Serum Cholesterol Trajectories by Treatment",
    col="blue", lty=1, pch=15)

abline(a=b[1], b=b[2] , col="blue" , lty=2)
points(times, means[,2], type="l",
    col="red", lty=1, pch=16)
abline(a=b[1], b=b[2]+b[3] , col="red" , lty=2)
legend("topleft",c("Placebo","High-Dose"),
    col=c("blue","red"), lty=c(1,1), pch=c(15,16))
```




From the dotted lines above for the two groups, we can see that the mean cholesterol level tends to increase as time moves on.

1b)

```
mod.unst = gls(Chol ~ Trt+factor(Month)+Trt*factor(Month), data=NCGS.long, weight=varIdent(form = ~

summary(mod.unst)
```

```
## Generalized least squares fit by REML
##   Model: Chol ~ Trt + factor(Month) + Trt * factor(Month)
##   Data: NCGS.long
##        AIC      BIC    logLik
##   3234.935 3329.531 -1592.468
##
## Correlation Structure: General
##  Formula: ~Time | ID
##  Parameter estimate(s):
##  Correlation:
##   1     2     3     4
## 2 0.764
## 3 0.748 0.807
## 4 0.758 0.822 0.741
## 5 0.606 0.695 0.704 0.650
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | Time
##  Parameter estimates:
```

```
##         1         2         3         4         5
## 1.0000000 0.9411488 0.8717255 0.8636942 1.0036302
##
## Coefficients:
##                                Value Std.Error   t-value p-value
## (Intercept)                236.64516  8.941450 26.466082  0.0000
## TrtHighDose                 -9.86738 12.198154 -0.808924  0.4192
## factor(Month)6               6.67742  5.981007  1.116437  0.2651
## factor(Month)12              7.90323  6.036858  1.309162  0.1914
## factor(Month)20             25.25806  5.907533  4.275569  0.0000
## factor(Month)24             20.83871  7.955226  2.619500  0.0092
## TrtHighDose:factor(Month)6  16.15591  8.159442  1.980027  0.0485
## TrtHighDose:factor(Month)12 17.93011  8.235636  2.177137  0.0302
## TrtHighDose:factor(Month)20  1.10305  8.059206  0.136868  0.8912
## TrtHighDose:factor(Month)24  9.10573 10.852721  0.839028  0.4021
##
##  Correlation:
##                            (Intr) TrtHgD fc(M)6 f(M)12 f(M)20 f(M)24 THD:(M)6
## TrtHighDose                -0.733
## factor(Month)6             -0.420  0.308
## factor(Month)12            -0.515  0.378  0.644
## factor(Month)20            -0.523  0.383  0.665  0.564
## factor(Month)24            -0.441  0.323  0.553  0.592  0.511
## TrtHighDose:factor(Month)6  0.308 -0.420 -0.733 -0.472 -0.487 -0.405
## TrtHighDose:factor(Month)12 0.378 -0.515 -0.472 -0.733 -0.413 -0.434  0.644
## TrtHighDose:factor(Month)20 0.383 -0.523 -0.487 -0.413 -0.733 -0.375  0.665
## TrtHighDose:factor(Month)24 0.323 -0.441 -0.405 -0.434 -0.375 -0.733  0.553
##                            THD:(M)1 THD:(M)20
## TrtHighDose
## factor(Month)6
## factor(Month)12
## factor(Month)20
## factor(Month)24
## TrtHighDose:factor(Month)6
## TrtHighDose:factor(Month)12
## TrtHighDose:factor(Month)20  0.564
## TrtHighDose:factor(Month)24  0.592    0.511
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.32343593 -0.70391415 -0.06751993  0.60899199  3.64284197
##
## Residual standard error: 49.78389
## Degrees of freedom: 335 total; 325 residual
```

Y(ij) = 236.64 - 9.87 Trt(ij) + 6.68 I(Month=6)(ij) + 7.90 I(Month=12)(ij) + 25.26 I(Month=20)(ij) + 20.84 I(Month=24)(ij) + 16.16 TrtI(Month=6)(ij) + 17.93 TrtI(Month=12)(ij) + 1.10 TrtI(Month=20)(ij) + 9.11 TrtI(Month=24)(ij)

1c) Estimated correlation between time 1 and time 2: 0.764.
Estimated correlation between time 1 and time 5: 0.606.

1d)

```
mod.ar = gls(Chol ~ Trt+factor(Month)+Trt*factor(Month), data=NCGS.long,corr=corAR1(, form = ~ Time
summary(mod.ar)
```

```
## Generalized least squares fit by REML
##   Model: Chol ~ Trt + factor(Month) + Trt * factor(Month)
##   Data: NCGS.long
##        AIC     BIC    logLik
##   3275.724 3321.13 -1625.862
##
## Correlation Structure: AR(1)
##  Formula: ~Time | ID
##  Parameter estimate(s):
##       Phi
## 0.7550983
##
## Coefficients:
##                                 Value Std.Error   t-value p-value
## (Intercept)                236.64516  8.589057 27.551938  0.0000
## TrtHighDose                 -9.86738 11.717410 -0.842113  0.4003
## factor(Month)6               6.67742  6.011133  1.110842  0.2675
## factor(Month)12              7.90323  7.963557  0.992424  0.3217
## factor(Month)20             25.25806  9.166282  2.755541  0.0062
## factor(Month)24             20.83871  9.978859  2.088286  0.0376
## TrtHighDose:factor(Month)6  16.15591  8.200541  1.970103  0.0497
## TrtHighDose:factor(Month)12 17.93011 10.864087  1.650402  0.0998
## TrtHighDose:factor(Month)20  1.10305 12.504875  0.088209  0.9298
## TrtHighDose:factor(Month)24  9.10573 13.613414  0.668880  0.5040
##
##  Correlation:
##                             (Intr) TrtHgD fc(M)6 f(M)12 f(M)20 f(M)24 THD:(M)6
## TrtHighDose                 -0.733
## factor(Month)6              -0.350  0.257
## factor(Month)12             -0.464  0.340  0.662
## factor(Month)20             -0.534  0.391  0.515  0.762
## factor(Month)24             -0.581  0.426  0.431  0.627  0.806
## TrtHighDose:factor(Month)6   0.257 -0.350 -0.733 -0.486 -0.377 -0.316
## TrtHighDose:factor(Month)12  0.340 -0.464 -0.486 -0.733 -0.559 -0.459  0.662
## TrtHighDose:factor(Month)20  0.391 -0.534 -0.377 -0.559 -0.733 -0.591  0.515
## TrtHighDose:factor(Month)24  0.426 -0.581 -0.316 -0.459 -0.591 -0.733  0.431
##                             THD:(M)1 THD:(M)20
## TrtHighDose
## factor(Month)6
## factor(Month)12
## factor(Month)20
## factor(Month)24
## TrtHighDose:factor(Month)6
## TrtHighDose:factor(Month)12
## TrtHighDose:factor(Month)20  0.762
## TrtHighDose:factor(Month)24  0.627    0.806
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.1187510 -0.6658150 -0.0607092  0.6090313  3.7923012
```

```
##
## Residual standard error: 47.82185
## Degrees of freedom: 335 total; 325 residual
```

1e)

1f) Model 1 (mod.ar) AIC: 3275.724; Model 2 (mod.unst) AIC: 3234.935.

Based on the AIC values provided, we can see that Model 2 has a lower AIC value (3234.935) compared to Model 1 (3275.724). This suggests that Model 2 (mod.unst) is a better fit to the data compared to Model 1 (mod.ar).

In general, a lower AIC value indicates a better fit of the model to the data. AIC takes into account the complexity of the model as well as the goodness of fit. Thus, a model with a lower AIC value is preferred over a model with a higher AIC value.

1g)

```
anova(mod.ar, mod.unst)
```

```
##          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## mod.ar       1 12 3275.724 3321.130 -1625.862
## mod.unst     2 25 3234.935 3329.531 -1592.468 1 vs 2 66.78882  <.0001
```

The likelihood ratio test can be used to compare the goodness of fit of two models. The null hypothesis is that the more restricted model (AR(1) covariance structure) is a better fit, and the alternative hypothesis is that the less restricted model (unstructured covariance) is a better fit.

In this case, the output from R shows that the likelihood ratio test statistic (L.Ratio) is 66.78882 with a corresponding p-value of less than 0.0001. This means that the unstructured covariance model is a significantly better fit than the AR(1) covariance model.

Therefore, we reject the null hypothesis that the AR(1) covariance structure is a better fit than the unstructured covariance structure in favor of the alternative hypothesis that the unstructured covariance structure is a better fit. This conclusion is based on the p-value which is less than 0.05 (or any reasonable significance level), indicating strong evidence against the null hypothesis.

1h)

```
mod.unst.numeric = gls(Chol ~ Trt+Month+Trt*Month, data=NCGS.long, weight=varIdent(form = ~ 1 | Tim
summary(mod.unst.numeric)
```

```
## Generalized least squares fit by maximum likelihood
##   Model: Chol ~ Trt + Month + Trt * Month
##   Data: NCGS.long
##        AIC      BIC    logLik
##   3289.222 3361.691 -1625.611
##
## Correlation Structure: General
##  Formula: ~Time | ID
##  Parameter estimate(s):
##  Correlation:
##   1     2     3     4
## 2 0.733
## 3 0.721 0.805
## 4 0.753 0.812 0.728
## 5 0.600 0.685 0.700 0.643
## Variance function:
```

```
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | Time
##  Parameter estimates:
##         1         2         3         4         5
## 1.0000000 0.9346026 0.8645659 0.8553460 0.9925752
##
## Coefficients:
##                        Value Std.Error   t-value p-value
## (Intercept)       235.01567  8.231693 28.550100  0.0000
## TrtHighDose         2.55246 11.229885  0.227292  0.8203
## Month               1.10902  0.227204  4.881149  0.0000
## TrtHighDose:Month  -0.23634  0.309958 -0.762477  0.4463
##
##  Correlation:
##                   (Intr) TrtHgD Month
## TrtHighDose       -0.733
## Month             -0.503  0.369
## TrtHighDose:Month  0.369 -0.503 -0.733
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.24019187 -0.72053352 -0.04057204  0.66651742  3.68317295
##
## Residual standard error: 49.68117
## Degrees of freedom: 335 total; 331 residual
```

1i)

```
mod.unst.quad = gls(Chol ~ Trt+Month+Trt*Month+I(Month^2)+Trt*I(Month^2), data=NCGS.long, weight=va
summary(mod.unst.quad)
```

```
## Generalized least squares fit by maximum likelihood
##   Model: Chol ~ Trt + Month + Trt * Month + I(Month^2) + Trt * I(Month^2)
##   Data: NCGS.long
##        AIC      BIC    logLik
##   3285.172 3365.269 -1621.586
##
## Correlation Structure: General
##  Formula: ~Time | ID
##  Parameter estimate(s):
##  Correlation:
##   1     2     3     4
## 2 0.757
## 3 0.746 0.801
## 4 0.757 0.817 0.735
## 5 0.596 0.693 0.701 0.637
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | Time
##  Parameter estimates:
##         1         2         3         4         5
## 1.0000000 0.9438417 0.8721044 0.8653080 1.0095773
##
## Coefficients:
```

```
##                              Value Std.Error   t-value p-value
## (Intercept)             235.78493  8.747582 26.954297  0.0000
## TrtHighDose              -6.18577 11.933674 -0.518346  0.6046
## Month                     0.90435  0.801812  1.127880  0.2602
## I(Month^2)                0.00857  0.032181  0.266241  0.7902
## TrtHighDose:Month         2.08861  1.093853  1.909407  0.0571
## TrtHighDose:I(Month^2)   -0.09733  0.043902 -2.216895  0.0273
##
##  Correlation:
##                          (Intr) TrtHgD Month  I(M^2) TrHD:M
## TrtHighDose              -0.733
## Month                    -0.452  0.331
## I(Month^2)                0.330 -0.242 -0.959
## TrtHighDose:Month         0.331 -0.452 -0.733  0.703
## TrtHighDose:I(Month^2) -0.242  0.330  0.703 -0.733 -0.959
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.24386428 -0.69095535 -0.02248774  0.67814551  3.71247054
##
## Residual standard error: 49.08189
## Degrees of freedom: 335 total; 329 residual
```

1j)

```
anova(mod.unst.numeric, mod.unst.quad)
```

```
##                   Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## mod.unst.numeric      1 19 3289.222 3361.691 -1625.611
## mod.unst.quad         2 21 3285.172 3365.269 -1621.586 1 vs 2 8.050406  0.0179
```

Model 2 (mod.unst.quad) fits the data better because it has a lower AIC value (3285.172) compared to that of Model 1 (mod.unst.numeric).

2. Linear mixed effects model: Yij = B0 + B1 tij + B2 Xij + b0i + b1i tij + eij.

   2a) The expression for the marginal mean cholesterol level for a male subject on day 20 is: E(Yij|tij=20, Xi=1) = B0 + B1 tij + B2 Xij + E(b0i) + E(b1i) tij.
   Since the random effects are assumed to be normally distributed with mean 0 and variances sigmab0^2 and sigmab1^2, respectively, we have: E(b0i) = 0 and E(b1i) = 0.
   Therefore, the expression simplifies to: E(Yij|tij=20, Xi=1) = B0 + B1(20) + B2(1).

   2b) The conditional mean cholesterol level for a male subject on day 20, given the individual-specific random effects, is: E(Yij|tij=20, Xi=1, b0i, b1i) = B0 + B1 tij + B2 Xij + b0i + b1i tij. Setting tij=20 & Xi=1, we get: E(Yij|tij=20, Xi=1, b0i, b1i) = B0 + B1(20) + B2(1) + b0i + b1i(20).
   Using the formula for the conditional expectation for an LME model, we can obtain the conditional mean cholesterol level for a male subject on day 20 as:
   E(Yij|tij=20, Xi=1) = B0 + B1(20) + B2(1) + E(b0i|tij=20, Xi=1) + E(b1i|tij=20, Xi=1)(20).

   2c) The expression for the marginal mean cholesterol level for a female subject on day 20 is: E(Yij|tij=20, Xi=0) = B0 + B1(20) + B2(0).
   Subtracting this expression from the expression for the marginal mean cholesterol level for a male subject on day 20, we get: E(Yij|tij=20, Xi=1) - E(Yij|tij=20, Xi=0) = (B0 + B1(20) + B2(1)) - (B0 + B1(20) + B2(0)).
   Simplifying this expression, we get: E(Yij|tij=20, Xi=1) - E(Yij|tij=20, Xi=0) = B2.

Therefore, the difference in marginal mean cholesterol levels between a male subject on day 20 and a female subject on day 20 is equal to the fixed effect B2.

2d) The expression for the conditional mean cholesterol level for a female subject at day j, given the individual-specific random effects, is: E(Yij|tij=j, Xi=0, b0i, b1i) = B0 + B1 tij + B2 Xij + b0i + b1i tij.
Setting Xi=0 for a female subject, between day 10 and day 40, we get:
E(Yij|tij=10, Xi=0, b0i, b1i) = B0 + B1(10) + b0i + b1i(10)
E(Yij|tij=40, Xi=0, b0i, b1i) = B0 + B1(40) + b0i + b1i(40)
The difference in conditional mean cholesterol levels between day 10 and day 40 for a female subject is: E(Yij|tij=40, Xi=0) - E(Yij|tij=10, Xi=0) = (B0 + B1(40) + b0i + b1i(40)) - (B0 + B1(10) + b0i + b1i(10)) => E(Yij|tij=40, Xi=0) - E(Yij|tij=10, Xi=0) = B1(40-10).
Therefore, the difference in conditional mean cholesterol levels between day 10 and day 40 for a female subject is equal to the fixed effect B1 multiplied by the time difference of 30 days.

2e) In the linear mixed effects model for cholesterol, there are two sources of random variation: the random intercept (b0i) and the random slope (b1i).
The random intercept represents the variability in the baseline cholesterol level for each individual, which is not accounted for by the fixed effects in the model. In other words, the random intercept captures the fact that different individuals may have different average cholesterol levels even if they have the same sex and are measured at the same time points. For example, some individuals may have a higher baseline cholesterol level due to genetic factors or lifestyle habits. The random intercept is assumed to be normally distributed with a mean of 0 and a variance of sigmab0^2.
The random slope represents the variability in the rate of change in cholesterol levels over time for each individual, which is not accounted for by the fixed effects in the model. In other words, the random slope captures the fact that different individuals may have different rates of change in their cholesterol levels over time, even if they have the same sex. For example, some individuals may experience a faster or slower increase in cholesterol levels over time due to differences in their metabolism or diet. The random slope is assumed to be normally distributed with a mean of 0 and a variance of sigmab1^2.
The random intercept and random slope are assumed to be independent of each other and independent of the error term (eij), which represents the residual variation not explained by the fixed or random effects. The random effects capture the variation between individuals that is not explained by the fixed effects, and they are an important source of variability in the model. By including the random intercept and random slope in the model, we can account for this variability and obtain more accurate estimates of the fixed effects.

3. 3a) E(Yi bar|X) = (B0 + B1 x1ij + B2 x2ij)/ni = B0/ni + x1i bar + x2i bar

   3b) This is done to keep the variances same across all i instructors.
   Var(Yi bar) = sigma^2/ni => ni*Var(Yi bar) = sigma^2. Therefore, the variance of instructors' average rating is sigma^2.