Stats 112 Homework 2

Owen Lin

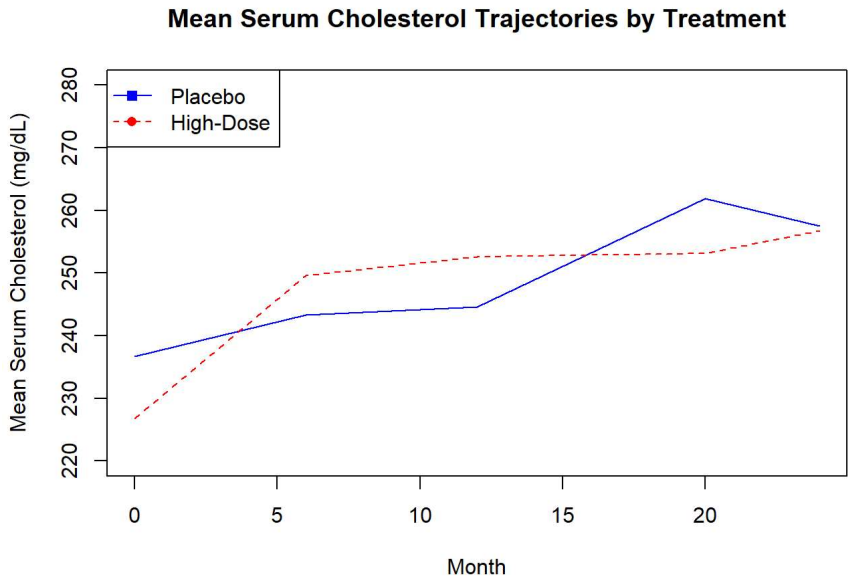
```
library(nlme)
library(lme4)
```

Problem 1

```
setwd("D:\\Coding\\Stats112\\Data")
NCGS = read.table("cholesterol-data.txt", na.strings=".")
# remove observations with na entries
NCGS = NCGS[complete.cases(NCGS), ]
# Give col names
names(NCGS) = c("Trt", "ID", "M0", "M6", "M12", "M20", "M24")
# label Treatment
NCGS$Trt = factor(NCGS$Trt, levels=c(2,1), labels=c("Placebo","HighDose"))
# convert to long form
NCGS.long = reshape(NCGS, varying=list(3:7), idvar="ID", timevar="Month", times=c(0,6,12,20,24), v.names="Chol", direction
="long")
# change month to time 1 2 3 4
NCGS.long$Time = as.numeric(factor(NCGS.long$Month))
```

1a:
For both High Dose and Placebo group, the mean cholesterol level generally goes up as time progresses.

```
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))
```



1b:
$Y_{ij} = \beta_1 + \beta_2 * Trt_{ij} + \beta_3 * I(Month = 6)_{ij} + \beta_4 * I(Month = 12)_{ij} + \beta_5 * I(Month = 20)_{ij} + \beta_6 * I(Month = 24)_{ij} + \beta_7 * TrtI(Month =$
where $Trt_{ij} = 1$ for HighDose and = 0 for Placebo group

```
mod.unst = gls(Chol ~ Trt+factor(Month)+Trt*factor(Month), data=NCGS.long, weight=varIdent(form = ~ 1 | Time),corr=corSymm(,
form = ~ Time | ID))
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 0.694 0.704 0.650
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | Time
##  Parameter estimates:
##         1         2         3         4         5
## 1.0000000 0.9411481 0.8717237 0.8636889 1.0036558
##
## Coefficients:
##                                Value Std.Error   t-value p-value
## (Intercept)                236.64516  8.941329 26.466441  0.0000
## TrtHighDose                 -9.86738 12.197989 -0.808935  0.4191
## factor(Month)6               6.67742  5.981019  1.116435  0.2651
## factor(Month)12              7.90323  6.036859  1.309162  0.1914
## factor(Month)20             25.25806  5.907502  4.275592  0.0000
## factor(Month)24             20.83871  7.955291  2.619478  0.0092
## TrtHighDose:factor(Month)6  16.15591  8.159458  1.980023  0.0485
## TrtHighDose:factor(Month)12 17.93011  8.235637  2.177137  0.0302
## TrtHighDose:factor(Month)20  1.10305  8.059164  0.136869  0.8912
## TrtHighDose:factor(Month)24  9.10573 10.852810  0.839021  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
## -2.32348164 -0.70392520 -0.06752126  0.60899296  3.64289136
##
## Residual standard error: 49.78321
## Degrees of freedom: 335 total; 325 residual
```

1c:
The estimated correlation between time 1 and time 2 is 0.764, whereas the estimated correlation between time 1 and time 5 is 0.606.

1d:

```
mod.ar = gls(Chol ~ Trt+factor(Month)+Trt*factor(Month), data=NCGS.long,corr=corAR1(, form = ~ Time | ID))
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:
The estimated correlation between time 1 and time 2 is 0.755, whereas the estimated correlation between time 1 and time 5 is $0.755^4 = 0.325$.

1f:
I would go with the model in part b (the unstructured model) because it has a lower AIC.

1g:
H_0: unstructured model is better.
H_a: ar(1) model is better.
p-value: <0.0001
Conclusion: We reject null and conclude that ar(1) model is better, which does not agree with the conclusion in part 1f.

```
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
```

1h:

```
mod.h = gls(Chol ~ Trt + as.numeric(Month) + Trt:as.numeric(Month), data=NCGS.long, weight=varIdent(form = ~ 1 | Time),corr=
corSymm(, form = ~ Time | ID), method = "ML")
summary(mod.h)
```

```
## Generalized least squares fit by maximum likelihood
##   Model: Chol ~ Trt + as.numeric(Month) + Trt:as.numeric(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 0.9346032 0.8645644 0.8553447 0.9925740
##
## Coefficients:
##                                 Value Std.Error   t-value p-value
## (Intercept)                  235.01569  8.231686 28.550127  0.0000
## TrtHighDose                    2.55243 11.229876  0.227289  0.8203
## as.numeric(Month)              1.10902  0.227205  4.881125  0.0000
## TrtHighDose:as.numeric(Month)  -0.23633  0.309959 -0.762467  0.4463
##
##  Correlation:
##                                (Intr) TrtHgD as.(M)
## TrtHighDose                    -0.733
## as.numeric(Month)              -0.503  0.369
## TrtHighDose:as.numeric(Month)  0.369 -0.503 -0.733
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.24019615 -0.72053361 -0.04057246  0.66651870  3.68317427
##
## Residual standard error: 49.68114
## Degrees of freedom: 335 total; 331 residual
```

1i:

$$Y_{ij} = \beta_0 + \beta_1 * Trt_{ij} + \beta_2 * t_{ij} + \beta_3 * t_{ij}^2 + \beta_4 * Trt_{ij}t_{ij} + \beta_5 * Trt_{ij}t_{ij}^2 + \epsilon_i$$

```
Month2 = as.numeric(NCGS.long$Month)^2
mod.exp = gls(Chol ~ Trt + as.numeric(Month) + Trt*as.numeric(Month) + Month2 + Trt*Month2, data=NCGS.long, weight=varIdent
(form = ~ 1 | Time),corr=corSymm(, form = ~ Time | ID), method="ML")
summary(mod.exp)
```

```
## Generalized least squares fit by maximum likelihood
##   Model: Chol ~ Trt + as.numeric(Month) + Trt * as.numeric(Month) + Month2 +      Trt * Month2
##   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
## as.numeric(Month)              0.90435  0.801812  1.127880  0.2602
## Month2                         0.00857  0.032181  0.266241  0.7902
## TrtHighDose:as.numeric(Month)  2.08861  1.093853  1.909407  0.0571
## TrtHighDose:Month2            -0.09733  0.043902 -2.216895  0.0273
##
##  Correlation:
##                               (Intr) TrtHgD as.(M) Month2 THD:.(
## TrtHighDose                   -0.733
## as.numeric(Month)             -0.452  0.331
## Month2                         0.330 -0.242 -0.959
## TrtHighDose:as.numeric(Month)  0.331 -0.452 -0.733  0.703
## TrtHighDose:Month2            -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:
The model in part i (with quadratic terms) fit the data better based on likelihood ratio test at 95% confident level.

```
anova(mod.h, mod.exp)
```

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

Problem 2

2a:
The marginal mean (population expectation) cholesterol level for a male subject on day 20 is
$\beta_0 + \beta_1 * 20 + \beta_2 * 1$

2b:
The conditional mean (conditional expectation) cholesterol level for a male subject on day 20 is $\beta_0 + \beta_1 * 20 + \beta_2 * 1 + b_{0,i} + b_{1,i} * 20$

2c:
The difference in marginal mean cholesterol levels between a male subject on day 20 and a female subject on day 20 is
$\beta_2 * 1$

2d:
The difference in conditional mean cholesterol levels between a male subject on day 20 and a female subject on day 20 is
$\beta_2 * 1 + b_{0,i1} + b_{1,i1} * 20 - b_{0,i2} - b_{1,i2} * 20$
where i1 represent a male subject and i2 represent a female subject

2e:
For different individual, they could start with a different level of cholesterol, this is captured in random intercept ($b_{0,i}$). Similarly, the change in cholesterol level varies for different individual, this is captured in random slope ($b_{1,i}$).

Problem 3

3a:

$E(\bar{Y_i}|X) = \beta_0 + \beta_1 * \bar{x_{1i}} + \beta_2 * \bar{x_{2i}}$

3b:

We should weight the variance of each instructor's average rating by the number of raters because $var(\bar{Y_i}) = \frac{1}{n_i} * var(Y_{ij})$

```
library(alr4)
data(Rateprof)
gls = gls(quality ~ easiness + helpfulness, data = Rateprof)
coef(gls)
```

```
## (Intercept)    easiness helpfulness
## 0.009125009 0.019386786 0.965371707
```