Atividade pontual 3 - Modelos para Dados Correlacionados

Verônica Stamatto Peres - 00304684

03/04/2021

Exercício Capítulo 5 - Applied Longitudinal Analysis

Problema 5.1

In the National Cooperative Gallstone Study (NCGS), one of the major interests was to study the safety of the drug chenodiol for the treatment of cholesterol gallstones (Schoenfield et al., 1981; Wei and Lachin, 1984). In this study, patients were randomly assigned to high-dose (750 mg per day), low-dose (375 mg per day), or placebo. We focus on a subset of data on patients who had floating gallstones and who were assigned to the high-dose and placebo groups.

In the NCGS it was suggested that chenodiol would dissolve gallstones but, in doing so, might increase levels of serum cholesterol. As a result serum cholesterol (mg/dL) was measured at baseline and at 6, 12, 20, and 24 months of follow-up. Many cholesterol measurements are missing because of missed visits, laboratory specimens were lost or inadequate, or patient follow-up was terminated.

The NCGS serum cholesterol data are stored in an external file: cholesterol.dat

Each row of the data set contains the following seven variables:

Group, ID, Y_1, Y_2, Y_3, Y_4, Y_5.

Note: The categorical variable Group is coded $1 = High - Dose$, $2 = Placebo$.

5.1.1

Read the data from the external file and keep it in a “multivariate” or “wide” format.

```
colesterol <- read_dta(file = "cholesterol.dta")
head(colesterol)
```

```
## # A tibble: 6 x 7
##   group    id    y1    y2    y3    y4    y5
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     1     1   178   246   295   228   274
## 2     1     2   254   260   278   245   340
## 3     1     3   185   232   215   220   292
## 4     1     4   219   268   241   260   320
## 5     1     5   205   232   265   242   230
## 6     1     6   182   213   173   200   193
```

Lendo os dados e mantendo em um formato multivariado temos as variáveis:

- Grupo, com valores 1 quando o paciente recebe uma alta dose da droga chenodiol e 2 quando o paciente recebe placebo.
- id, que representa um índice para cada pessoa.
- y_1, que representa a medida de colesterol sérico no início do estudo para cada indivíduo.
- y_2, que representa a medida de colesterol sérico após 6 meses do estudo para cada indivíduo.
- y_3, que representa a medida de colesterol sérico após 12 meses do estudo para cada indivíduo.
- y_4, que representa a medida de colesterol sérico após 20 meses do estudo para cada indivíduo.
- y_5, que representa a medida de colesterol sérico após 24 meses do estudo para cada indivíduo.

5.1.2

Calculate the sample means, standard deviations, and variances of the serum cholesterol levels at each occasion for each treatment group.

Médias para cada grupo em cada tempo

```
med1 = tapply(colesterol$y1,colesterol$group,median)
med2 = tapply(colesterol$y2,colesterol$group,median)
med3 = tapply(colesterol$y3,colesterol$group,median,na.rm=TRUE)
med4 = tapply(colesterol$y4,colesterol$group,median,na.rm=TRUE)
med5 = tapply(colesterol$y5,colesterol$group,median,na.rm=TRUE)

median1 = as.matrix(c(med1[1],med2[1],med3[1],med4[1],med5[1]))
median2 = as.matrix(c(med1[2],med2[2],med3[2],med4[2],med5[2]))
median1
```

```
##    [,1]
## 1 222.0
## 1 248.5
## 1 260.0
## 1 254.0
## 1 251.0
```

```
median2
```

```
##   [,1]
## 2  230
## 2  242
## 2  238
## 2  259
## 2  248
```

Temos para o grupo high-dose médias para os níveis de colesterol 222, 248.5, 260, 254 e 251 para os tempos 0, 6, 12, 20 e 24 respectivamente.

Temos para o grupo placebo médias para os níveis de colesterol 230, 242, 238, 259 e 248 para os tempos 0, 6, 12, 20 e 24 respectivamente.

Desvio padrão para cada grupo em cada tempo

```
#varpop = function(x){
#  var(x)*(length(x)-1)/length(x)
#}
sd1 = tapply(colesterol$y1,colesterol$group,sd)
sd2 = tapply(colesterol$y2,colesterol$group,sd)
sd3 = tapply(colesterol$y3,colesterol$group,sd,na.rm=TRUE)
sd4 = tapply(colesterol$y4,colesterol$group,sd,na.rm=TRUE)
sd5 = tapply(colesterol$y5,colesterol$group,sd,na.rm=TRUE)

sdg1 = as.matrix(c(sd1[1],sd2[1],sd3[1],sd4[1],sd5[1]))
sdg2 = as.matrix(c(sd1[2],sd2[2],sd3[2],sd4[2],sd5[2]))
sdg1
```

```
##       [,1]
## 1 39.66437
## 1 39.45228
## 1 38.32922
## 1 34.48935
## 1 49.96198
```

```
sdg2
```

```
##       [,1]
## 2 55.87459
## 2 49.23967
## 2 46.11058
## 2 51.14179
## 2 49.38817
```

Temos para o grupo high-dose o desvio padrão aproximado para os níveis de colesterol 39.66, 39.45, 38.33, 34.49 e 49.96 para os tempos 0, 6, 12, 20 e 24 respectivamente.

Temos para o grupo placebo o desvio padrão aproximado para os níveis de colesterol 55.88, 49.24, 46.11, 51.14 e 49.39 para os tempos 0, 6, 12, 20 e 24 respectivamente.

Variâncias para cada grupo em cada tempo

```
#varpop = function(x){
#  var(x)*(length(x)-1)/length(x)
#}
var1 = tapply(colesterol$y1,colesterol$group,var)
var2 = tapply(colesterol$y2,colesterol$group,var)
var3 = tapply(colesterol$y3,colesterol$group,var,na.rm=TRUE)
var4 = tapply(colesterol$y4,colesterol$group,var,na.rm=TRUE)
var5 = tapply(colesterol$y5,colesterol$group,var,na.rm=TRUE)

varg1 = as.matrix(c(var1[1],var2[1],var3[1],var4[1],var5[1]))
varg2 = as.matrix(c(var1[2],var2[2],var3[2],var4[2],var5[2]))
varg1
```

```
##       [,1]
## 1 1573.262
## 1 1556.483
## 1 1469.129
## 1 1189.515
## 1 2496.200
```

```
varg2
```

```
##       [,1]
## 2 3121.970
## 2 2424.545
## 2 2126.186
## 2 2615.482
## 2 2439.191
```

Temos para o grupo high-dose as variâncias aproximadas para os níveis de colesterol 1573.26, 1556.48, 1469.13, 1189.52 e 2496.2 para os tempos 0, 6, 12, 20 e 24 respectivamente.

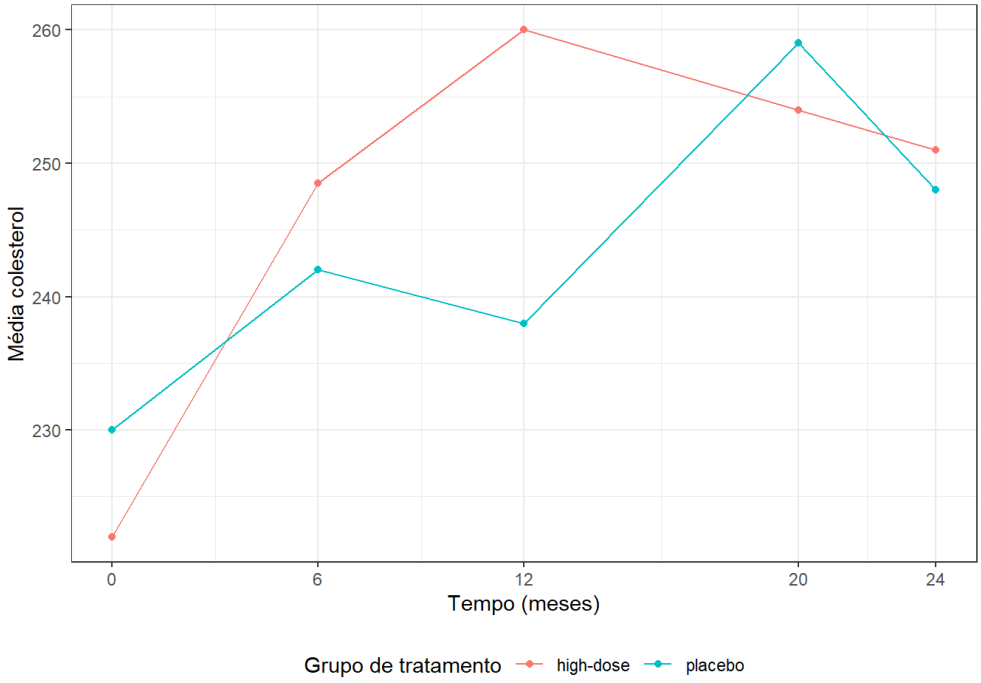
Temos para o grupo placebo as variâncias aproximadas para os níveis de colesterol 3121.97, 2424.54, 2126.19, 2615.48 e 2439.19 para os tempos 0, 6, 12, 20 e 24 respectivamente.

5.1.3

On a single graph, construct a time plot that displays the mean serum cholesterol versus time (in months) for the two treatment group. Describe the general characteristics of the time trends for the two groups.

```
n <- 103
colesterol_resumo <- data.frame(tempo = c(0,6,12,20,24,0,6,12,20,24),
                                median = c(median1,median2),
                                group = c(rep("high-dose",5),rep("placebo",5)))

p <- ggplot(data = colesterol_resumo,
mapping = aes(x = tempo, y = median,
colour = group)) +
geom_point() + geom_line() +
scale_x_continuous(breaks = c(0,6,12,20,24)) +
labs(x = "Tempo (meses)",
y = expression("Média colesterol"),
colour = "Grupo de tratamento") +
theme_bw() + theme(legend.position = "bottom")
p
```



O gráfico contém as médias do colesterol sérico para cada grupo nos meses de acompanhamento. A linha rosa representa o grupo de indivíduo que recebeu uma dose alta de droga chenodiol e a linha azul representa o grupo de indivíduos placebo.

Podemos analisar que inicialmente a média do colesterol sérico para o grupo placebo está um pouco maior. No mês 6 do estudo, ambos os grupos tiveram um aumento dessa média, porém o grupo dos indivíduos que receberam alta dose da droga em análise teve um aumento bem maior. No mês 12, o grupo da alta dose continua crescendo e observamos um decréscimo na média do grupo placebo. No mês 20, o grupo placebo tem um crescimento grande na média do colesterol sérico e o grupo da alta dose tem uma diminuição na sua média. No último mês observado, ambos os grupos tiveram uma redução na média, com valores próximos.

O grupo dos indivíduos que receberam altas doses tem um rápido crescimento da média de colesterol sérico nos primeiros meses e depois tem uma leve queda nessa média. Já o grupo placebo tem um crescimento leve, depois uma leve queda da média no início do estudo e depois um crescimento grande da média, que volta a cair no último momento observado no estudo.

No final do estudo encontramos médias bem parecidas de colesterol sérico nos dois grupos, o que pode indicar que a droga chenodiol não afeta significativamente os níveis de colesterol sérico. Porém, no 12º mês do estudo percebemos uma grande diferença entre as médias dos dois grupos. Assim, podemos notar que em alguns meses de acompanhamento o chenodiol afeta o nível de colesterol, mas no final essa diferença não é tão grande entre os indivíduos que receberam uma alta dose e indivíduos placebos.

5.1.4

Next read the data from the external file and put the data in a “univariate” or “long” format, with five “records” per subject.

```
colesterol.longo <- gather(data = colesterol,
                       key = "tempo",
                       value = "colesterol", -id, -group)

colesterol.longo$tempo <- factor(colesterol.longo$tempo,
                      labels = c(0, 6, 12, 20,24))

colesterol.longo$tempo.num <- as.numeric(colesterol.longo$tempo,
                      labels = c(0, 6, 12, 20,24))

colesterol.longo$group <- factor(colesterol.longo$group,
                           labels = c("high-dose","Placebo"))

colesterol.longo$group = relevel(colesterol.longo$group, ref=2)

head(colesterol.longo)
```

```
## # A tibble: 6 x 5
##   group        id tempo colesterol tempo.num
##   <fct>     <dbl> <fct>      <dbl>     <dbl>
## 1 high-dose     1 0            178         1
## 2 high-dose     2 0            254         1
## 3 high-dose     3 0            185         1
## 4 high-dose     4 0            219         1
## 5 high-dose     5 0            205         1
## 6 high-dose     6 0            182         1
```

```
class(colesterol.longo$tempo)
```

```
## [1] "factor"
```

```
class(colesterol.longo$group)
```

```
## [1] "factor"
```

Transformando os dados em um formato longo ou univariado e transformando as variáveis tempo e grupo em fatores para seguir com a análise. O grupo de referência foi definido como o grupo placebo.

5.1.5

Assuming an unstructured covariance matrix, conduct an analysis of response profiles. Determine whether the patterns of change over time differ in the two treatment groups.

Foi utilizada a linha de base (mês 0) e o grupo placebo (grupo 2) como referência.

```
# modelo de perfis de respostas
# com matriz de covariância não estruturada
mod.pr <- gls(colesterol ~ group * tempo,      # argumento formula
              corr = corSymm(form = ~ 1 | id),
              weights = varIdent(form = ~ 1 | tempo), # matriz de covariancia nao estruturada
              method = "REML",  # metodo maxima verossimilhanca restrita
              data = colesterol.longo, # dados na forma longa
              na.action = na.omit)

summary(mod.pr)
```

```
## Generalized least squares fit by REML
##   Model: colesterol ~ group * tempo
##   Data: colesterol.longo
##        AIC      BIC    logLik
##   4314.588 4416.587 -2132.294
##
## Correlation Structure: General
##  Formula: ~1 | id
##  Parameter estimate(s):
##  Correlation:
##   1     2     3     4
## 2 0.770
## 3 0.732 0.773
## 4 0.738 0.800 0.726
## 5 0.586 0.665 0.678 0.625
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | tempo
##  Parameter estimates:
##         0         6        12        20        24
## 1.0000000 0.9320568 0.8791668 0.8974016 1.0300809
##
## Coefficients:
##                            Value Std.Error  t-value p-value
## (Intercept)            235.92683  7.305948 32.29243  0.0000
## grouphigh-dose          -9.67829  9.412956 -1.02819  0.3044
## tempo6                   7.24390  4.805425  1.50744  0.1324
## tempo12                  8.84620  5.207262  1.69882  0.0901
## tempo20                 23.10333  5.292171  4.36557  0.0000
## tempo24                 21.12230  7.398137  2.85508  0.0045
## grouphigh-dose:tempo6   12.21751  6.193407  1.97266  0.0492
## grouphigh-dose:tempo12  16.28893  6.738391  2.41733  0.0160
## grouphigh-dose:tempo20   4.75670  6.973253  0.68213  0.4955
## grouphigh-dose:tempo24   6.53598  9.763271  0.66945  0.5036
##
##  Correlation:
##                        (Intr) grphg- tempo6 temp12 temp20 temp24 grp-:6 gr-:12
## grouphigh-dose         -0.776
## tempo6                 -0.429  0.333
## tempo12                -0.500  0.388  0.581
## tempo20                -0.466  0.362  0.606  0.526
## tempo24                -0.392  0.304  0.476  0.522  0.438
## grouphigh-dose:tempo6   0.333 -0.429 -0.776 -0.451 -0.470 -0.369
## grouphigh-dose:tempo12  0.387 -0.497 -0.449 -0.773 -0.407 -0.404  0.578
## grouphigh-dose:tempo20  0.354 -0.456 -0.460 -0.400 -0.759 -0.332  0.592  0.513
## grouphigh-dose:tempo24  0.297 -0.378 -0.361 -0.396 -0.332 -0.758  0.463  0.503
##                        gr-:20
## grouphigh-dose
## tempo6
## tempo12
## tempo20
## tempo24
## grouphigh-dose:tempo6
## grouphigh-dose:tempo12
## grouphigh-dose:tempo20
## grouphigh-dose:tempo24  0.419
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.32029916 -0.68866948 -0.02685013  0.60855779  3.89204113
##
## Residual standard error: 46.7809
## Degrees of freedom: 447 total; 437 residual
```

```
knitr::kable(
Anova(mod.pr),
digits = c(0, 2, 4))
```

	Df	Chisq	Pr(>Chisq)
group	1	0.00	0.9516
tempo	4	65.48	0.0000
group:tempo	4	7.92	0.0947

Com esses resultados encontramos a matriz de correlação, a variância dos dados para cada tempo, os coeficientes estimados de β, estatísticas descritivas (ex: mínimo, mediana, máximo).

Podemos concluir que parece existir diferença entre os dois grupos. Os primeiros dois tempos do estudo (tempo6 e tempo12) temos um aumento maior da média do grupo da high-dose e nos dois últimos tempos do estudo (tempo20 e tempo24) temos um aumento maior da média do grupo placebo.

Percebemos que a interação entre grupo e tempo é importante ao realizar a anova, pois temos valores significativos à 10% para essa interação, de acordo com o teste qui-quadrado.

5.1.6

Display the estimated 5 x 5 covariance and correlation matrices for the five repeated measurements of serum cholesterol.

Covariância

```
knitr::kable(
getVarCov2(mod.pr)$Omega,
digits = 3)
```

	0	6	12	20	24
0	2188.452	1571.425	1407.913	1449.214	1320.705
6	1571.425	1901.174	1387.080	1463.678	1397.530
12	1407.913	1387.080	1691.530	1254.365	1343.293
20	1449.214	1463.678	1254.365	1762.425	1264.163
24	1320.705	1397.530	1343.293	1264.163	2322.094

Podemos perceber que a variância no início do estudo, ou seja, na linha de base é um pouco menor que no final do estudo, mas são próximas. Nota-se que a variância nos meses 6, 12 e 20 são menores.

Correlação

```
mod.pr$modelStruct$corStruct
```

```
## Correlation structure of class corSymm representing
##  Correlation:
##   1     2     3     4
## 2 0.770
## 3 0.732 0.773
## 4 0.738 0.800 0.726
## 5 0.586 0.665 0.678 0.625
```

Podemos notar que as correlações, na maioria das vezes, diminuem à medida que a separação tempo entre as medidas repetidas aumenta.

5.1.7

With baseline (month 0) and the placebo group (group 2) as the reference group, write out the regression model for mean serum cholesterol that corresponds to the analysis ofresponse profiles in Problem 5.1.5.

O modelo de regressão para o colesterol sérico médio que corresponde à análise de perfis de resposta do problema 5.1.5 pode ser escrita como:

$$Y = \beta_0 + \beta_1 group + \beta_2 tempo + \beta_3 tempo * gruop$$

Onde Y é a variável resposta que representa o nível de colesterol. Estamos analisando essa variável levando em consideração os dois grupos (placebo e high-dose), os tempos dos estudo (0, 6, 12, 20 e 24 meses) e a interação entre tempo e grupo, ou seja, como cada grupo é afetado ao longo do tempo.

Com os valores dos coeficientes encontrados na questão 5.1.5

```
mod.pr$coefficients
```

```
##           (Intercept)         grouphigh-dose                 tempo6
##             235.926829              -9.678291               7.243902
##                tempo12                tempo20                tempo24
##               8.846205              23.103334              21.122304
##  grouphigh-dose:tempo6 grouphigh-dose:tempo12 grouphigh-dose:tempo20
##              12.217511              16.288928               4.756696
## grouphigh-dose:tempo24
##               6.535983
```

Temos que o modelo de regressão para o colesterol sérico médio pode ser escrito da seguinte forma:

$$Y = 235.927 - 9.678hd + 7.244t_6 + 8.846t_{12} + 23.103t_{20} + 21.122t_{24} + 12.218hd * t_6 + 16.289hd * t_{12} + 4.757hd * t_{20} + 6.536hd * t$$

Onde hd(high-dose) e p(placebo), assim como t_0, t_6, t_{12}, t_{20} e t_{24} recebem 1 quando a observação faz parte do grupo e tempo determinado e 0 se não faz parte. Dessa forma, uma observação de um indivíduo do grupo high-dose no tempo 20 meses pode ter o modelo reduzido para:

$$Y = 235.927 - 9.678hd + 4.757hd * t_{20} = 235.927 - 9.678 + 4.757$$

5.1.8

Let L denote a matrix of known weights and β the vector of linear regression parameters from the model assumed in Problem 5.1.7. The null hypothesis that the patterns of change over time do not differ in the two treatment groups can be expressed as H0 : $L\beta = 0$. Describe an appropriate weight matrix L for this null hypothesis.

Vamos considerar uma matriz L de pesos conhecidos e β o vetor de parâmetros de regressão linear assumido no problema 5.1.7. Queremos descobrir L para a hipótese nula de que os padrões de mudança ao longo do tempo não diferem nos dois grupos de tratamento, ou seja, H0: $L\beta = 0$

De acordo com os estimadores de β:

```
l0 <- c(1,0)
l1 <- c(24.37691,0)
l2 <- c(1.686593,0)
l3 <- c(1.841346,0)
l4 <- c(0,-1)
l5 <- c(0,-1)
l6 <- c(-1,0)
l7 <- c(-1,0)
l8 <- c(0,4.857013)
l9 <- c(0,3.231695)

L <- matrix(c(l0,l1,l2,l3,l4,l5,l6,l7,l8,l9), ncol=10)
L
```

```
##      [,1]     [,2]     [,3]     [,4] [,5] [,6] [,7] [,8]     [,9]    [,10]
## [1,]    1 24.37691 1.686593 1.841346    0    0   -1   -1 0.000000 0.000000
## [2,]    0  0.00000 0.000000 0.000000   -1   -1    0    0 4.857013 3.231695
```

L está definido dessa forma para que $\beta_0 = \beta_1$, $\beta_2 = \beta_6$, $\beta_3 = \beta_7$, $\beta_4 = \beta_8$ e $\beta_5 = \beta_9$, já que essas igualdades de β mantém os tempos constantes mudando apenas o grupo.

```
coef <- matrix(c(235.926829,-9.678291,7.243902,8.846205,23.103334,21.122304,12.217511,16.288928,4.756696,6.535983),ncol=1)
d <- L%*%coef
knitr::kable(d,digits = 1)
```

0
0
Multiplicando a matriz L com os coeficientes temos uma matriz de 0, ou seja, $L\beta = 0$ é válida com esse L.

5.1.9

Show how the estimated regression coefficients from an analysis of response profiles can be used to construct the time-specific means in the two groups. Compare these estimated means with the sample means obtained in Problem 5.1.2.

```
coefplac <- matrix(c(235.926829,7.243902,8.846205,23.103334,21.122304),ncol=1)
mean_plac <- getVarCov2(mod.pr)$Omega%*%coefplac
mean_plac
```

```
##        [,1]
## 0  601530.5
## 6  460118.5
## 12 414529.3
## 20 431027.5
## 24 411850.8
```

```
coefhd <- c(-9.678291,12.217511,16.288928,4.756696,6.535983)
mean_hd <- getVarCov(mod.pr)%*%coefhd
mean_hd
```

```
##          [,1]
## [1,] 35763.55
## [2,] 44141.20
## [3,] 44989.95
## [4,] 40269.83
## [5,] 46728.53
```

Os resultados apresentados anteriormente estão muito distantes dos verdadeiros valores da média. Isso indica um erro nos códigos acima.

5.1.10

With baseline (month 0) and the placebo group (group 2) as the reference group, provide an interpretation for each of the estimated regression coefficients in terms of the effect of the treatments on the patterns of change in mean serum cholesterol.

```
knitr::kable(summary(mod.pr)$tTable[,-4],
             digits = c(3, 3, 2),
col.names = c("Estimativa", "EP", "Z"))
```

	Estimativa	EP	Z
(Intercept)	235.927	7.306	32.29
grouphigh-dose	-9.678	9.413	-1.03
tempo6	7.244	4.805	1.51
tempo12	8.846	5.207	1.70
tempo20	23.103	5.292	4.37
tempo24	21.122	7.398	2.86
grouphigh-dose:tempo6	12.218	6.193	1.97
grouphigh-dose:tempo12	16.289	6.738	2.42
grouphigh-dose:tempo20	4.757	6.973	0.68
grouphigh-dose:tempo24	6.536	9.763	0.67

A linha de base (mês 0) é escolhido como o nível de referência para o tempo e o grupo placebo é escolhido como nível de referência para o grupo de tratamento.

Vamos considerar as estimativas de β e seus erros padrões da tabela anterior para interpretar os coeficientes em relação ao efeito dos tratamentos sobre os padrões de mudança no colesterol sérico médio

Os resultados indicam que os indivíduos que receberam altas doses da droga apresentaram um aumento maior da média de colesterol sérico nas duas primeiras ocasiões, quando comparados com os indivíduos que receberam placebo.

Quando comparado ao grupo placebo, o grupo de indivíduos que receberam altas doses da droga tem um aumento adicional de 12.218 nos níveis médios de colesterol desde o início do estudo até o mês 6. Fazendo essa comparação novamente, o grupo “high-dose” tem um aumento adicional de 16.289 nos níveis médios de colesterol do início do estudo até o mês 12. Um aumento adicional de 4.757 nos níveis médios de colesterol do início do estudo até o mês 20 e um aumento adicional de 6.536 nos níveis médios de colesterol do início do estudo até o mês 24.

Podemos fazer uma análise parecida para o grupo placebo. O grupo placebo tem um aumento adicional de 7.244 nos níveis médios de colesterol desde o início do estudo até o mês 6. Um aumento adicional de 8.846 nos níveis médios de colesterol do início do estudo até o mês 12. Um aumento adicional de 23.103 até o mês 20 e um aumento de 21.122 até o mês 24. Tudo isso comparado com o grupo placebo no início do estudo, ou seja, tempo 0, representado pelo intercepto com nível de colesterol de 235.927.