Module 4: Regression Methods: Concepts and Applications

Lab 3: One-Way and Two-Way ANOVA

The goal of this lab is to answer the following scientific questions using the cholesterol dataset:

- Is rs4775401 associated with cholesterol levels?
- Are rs174548 and APOE associated with cholesterol levels?
- Does the effect of APOE on cholesterol levels depend on rs174548?

The cholesterol data set is available for download from the module Github repository and contains the following variables:

> ID: Subject ID
>
> sex: Sex: 0 = male, 1 = female
>
> age: Age in years
>
> chol: Serum total cholesterol, mg/dl
>
> BMI: Body-mass index, kg/m2
>
> TG: Serum triglycerides, mg/dl
>
> APOE: Apolipoprotein E genotype, with six genotypes coded 1-6: 1 = e2/e2, 2 = e2/e3, 3 = e2/e4, 4 = e3/e3, 5 = e3/e4, 6 = e4/e4
>
> rs174548: Candidate SNP 1 genotype, chromosome 11, physical position 61,327,924. Coded as the number of minor alleles: 0 = C/C, 1 = C/G, 2 = G/G.
>
> rs4775401: Candidate SNP 2 genotype, chromosome 15, physical position 59,476,915. Coded as the number of minor alleles: 0 = C/C, 1 = C/T, 2 = T/T.
>
> HTN: diagnosed hypertension: 0 = no, 1 = yes
>
> chd: diagnosis of coronary heart disease: 0 = no, 1 = yes

You can download the data file and read it into R as follows:

```
cholesterol = read.csv("https://raw.githubusercontent.com/rhubb/SISG2018/master/data/SISG-Data-cholesterol.csv", header=T)
```

Install R packages

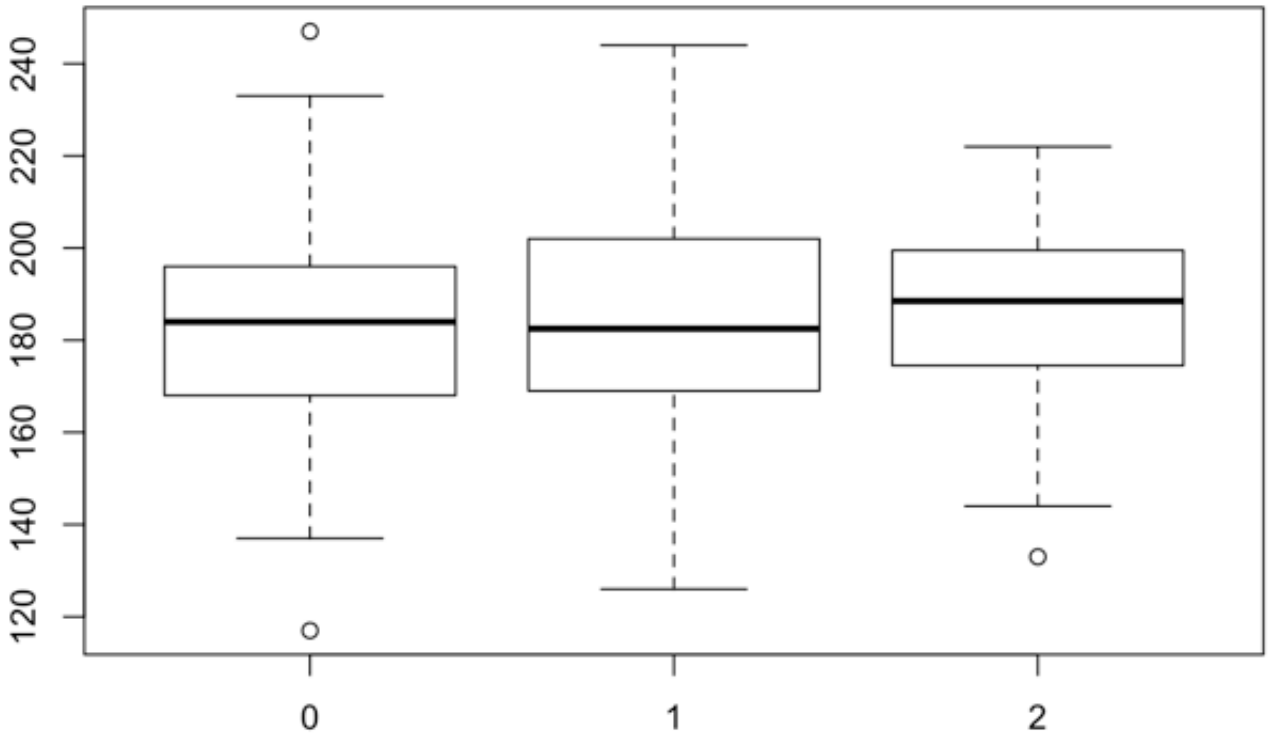
- For this lab you will need the *gee* and *multcomp* packages.
- If you have not already, install the packages first. You will then need to load the libraries each time you execute your R script.

```
install.packages("gee")
install.packages("multcomp")
library(gee)
library(multcomp)
```

Exercises

1. Perform a descriptive analysis to investigate the scientific questions of interest using numeric and graphical methods.

```
# graphical display: boxplot
boxplot(chol ~ factor(rs4775401))
```




```
# numeric descriptives
tapply(chol, factor(rs4775401), mean)
```

```
##        0        1        2
## 183.4505 184.2882 185.0000
```

```
tapply(chol, factor(rs4775401), sd)
```

```
##        0        1        2
## 20.70619 23.85693 21.70851
```

2. Conduct an analysis of differences in mean cholesterol levels across genotype groups defined by rs4775401. Is there evidence that mean cholesterol levels differ across genotypes? If so, perform all pairwise multiple comparisons using Bonferroni's adjustment. Try out different adjustment methods.

```r
# ANOVA for cholesterol and rs4775401
fit1 = lm(chol ~ factor(rs4775401))
summary(fit1)
```

```
##
## Call:
## lm(formula = chol ~ factor(rs4775401))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.450 -15.450  -0.288  15.550  63.550
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         183.4505     1.5597 117.618   <2e-16 ***
## factor(rs4775401)1   0.8377     2.3072   0.363    0.717
## factor(rs4775401)2   1.5495     4.4702   0.347    0.729
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.17 on 397 degrees of freedom
## Multiple R-squared:  0.0005135,  Adjusted R-squared:  -0.004522
## F-statistic: 0.102 on 2 and 397 DF,  p-value: 0.9031
```

```r
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: chol
##                     Df Sum Sq Mean Sq F value Pr(>F)
## factor(rs4775401)   2    100   50.11   0.102 0.9031
## Residuals         397 195089  491.41
```

```r
# construct contrasts for all pairwise comparisons
M2 = contrMat(table(rs4775401), type="Tukey")
fit2 = lm(chol ~ -1 + factor(rs4775401))

# explore options to correct for multiple comparisons
mc2 = glht(fit2, linfct =M2)
summary(mc2, test=adjusted("none"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = chol ~ -1 + factor(rs4775401))
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## 1 - 0 == 0   0.8377     2.3072   0.363    0.717
## 2 - 0 == 0   1.5495     4.4702   0.347    0.729
## 2 - 1 == 0   0.7118     4.5212   0.157    0.875
## (Adjusted p values reported -- none method)
```

```r
summary(mc2, test=adjusted("bonferroni"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = chol ~ -1 + factor(rs4775401))
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## 1 - 0 == 0   0.8377     2.3072   0.363        1
## 2 - 0 == 0   1.5495     4.4702   0.347        1
## 2 - 1 == 0   0.7118     4.5212   0.157        1
## (Adjusted p values reported -- bonferroni method)
```

```
summary(mc2, test=adjusted("holm"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = chol ~ -1 + factor(rs4775401))
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## 1 - 0 == 0   0.8377     2.3072   0.363        1
## 2 - 0 == 0   1.5495     4.4702   0.347        1
## 2 - 1 == 0   0.7118     4.5212   0.157        1
## (Adjusted p values reported -- holm method)
```

```
summary(mc2, test=adjusted("hochberg"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = chol ~ -1 + factor(rs4775401))
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## 1 - 0 == 0   0.8377     2.3072   0.363    0.875
## 2 - 0 == 0   1.5495     4.4702   0.347    0.875
## 2 - 1 == 0   0.7118     4.5212   0.157    0.875
## (Adjusted p values reported -- hochberg method)
```

```
summary(mc2, test=adjusted("hommel"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = chol ~ -1 + factor(rs4775401))
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## 1 - 0 == 0   0.8377     2.3072   0.363    0.875
## 2 - 0 == 0   1.5495     4.4702   0.347    0.875
## 2 - 1 == 0   0.7118     4.5212   0.157    0.875
## (Adjusted p values reported -- hommel method)
```

```
summary(mc2, test=adjusted("BH"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = chol ~ -1 + factor(rs4775401))
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## 1 - 0 == 0   0.8377     2.3072   0.363    0.875
## 2 - 0 == 0   1.5495     4.4702   0.347    0.875
## 2 - 1 == 0   0.7118     4.5212   0.157    0.875
## (Adjusted p values reported -- BH method)
```

```
summary(mc2, test=adjusted("BY"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = chol ~ -1 + factor(rs4775401))
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## 1 - 0 == 0   0.8377     2.3072   0.363        1
## 2 - 0 == 0   1.5495     4.4702   0.347        1
## 2 - 1 == 0   0.7118     4.5212   0.157        1
## (Adjusted p values reported -- BY method)
```

```
summary(mc2, test=adjusted("fdr"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = chol ~ -1 + factor(rs4775401))
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## 1 - 0 == 0   0.8377     2.3072   0.363    0.875
## 2 - 0 == 0   1.5495     4.4702   0.347    0.875
## 2 - 1 == 0   0.7118     4.5212   0.157    0.875
## (Adjusted p values reported -- fdr method)
```

3. Compare results obtained using classical ANOVA to those based on ANOVA allowing for unequal variances, using robust standard errors, and using a nonparametric test.

```
# One-way ANOVA (not assuming equal variances)
oneway.test(chol ~ factor(rs4775401))
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  chol and factor(rs4775401)
## F = 0.10457, num df = 2.000, denom df = 75.608, p-value = 0.9008
```

```
# Using robust standard errors
summary(gee(chol ~ factor(rs4775401), id=seq(1,length(chol))))
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
##        (Intercept) factor(rs4775401)1 factor(rs4775401)2
##        183.4504950          0.8377402          1.5495050
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     Independent
##
## Call:
## gee(formula = chol ~ factor(rs4775401), id = seq(1, length(chol)))
##
## Summary of Residuals:
##         Min          1Q      Median          3Q         Max
## -66.4504950 -15.4504950  -0.2882353  15.5495050  63.5495050
##
##
## Coefficients:
##                       Estimate Naive S.E.     Naive z Robust S.E.    Robust z
## (Intercept)        183.4504950   1.559715 117.6179395    1.453272 126.2327489
## factor(rs4775401)1   0.8377402   2.307238   0.3630923    2.332437   0.3591694
## factor(rs4775401)2   1.5495050   4.470234   0.3466273    4.282708   0.3618049
##
## Estimated Scale Parameter:  491.4078
## Number of Iterations:  1
##
## Working Correlation
##      [,1]
## [1,]    1
```

```
# Non-parametric ANOVA
kruskal.test(chol ~ factor(rs4775401))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  chol by factor(rs4775401)
## Kruskal-Wallis chi-squared = 0.57611, df = 2, p-value = 0.7497
```

4. Perform a descriptive analysis to investigate the relationships between cholesterol, APOE and rs174548. Conduct an analysis to investigate the association between cholesterol, APOE and rs174548, with and without an interaction between APOE and rs174548. Is there evidence of an interaction between APOE and rs174548?

```
# exploratory data analysis
table(rs174548, APOE)
```

```
##          APOE
## rs174548   1   2   3   4   5   6
##         0   2  33   2 144  40   6
##         1   0  17   3  99  24   4
##         2   0   1   0  24   1   0
```

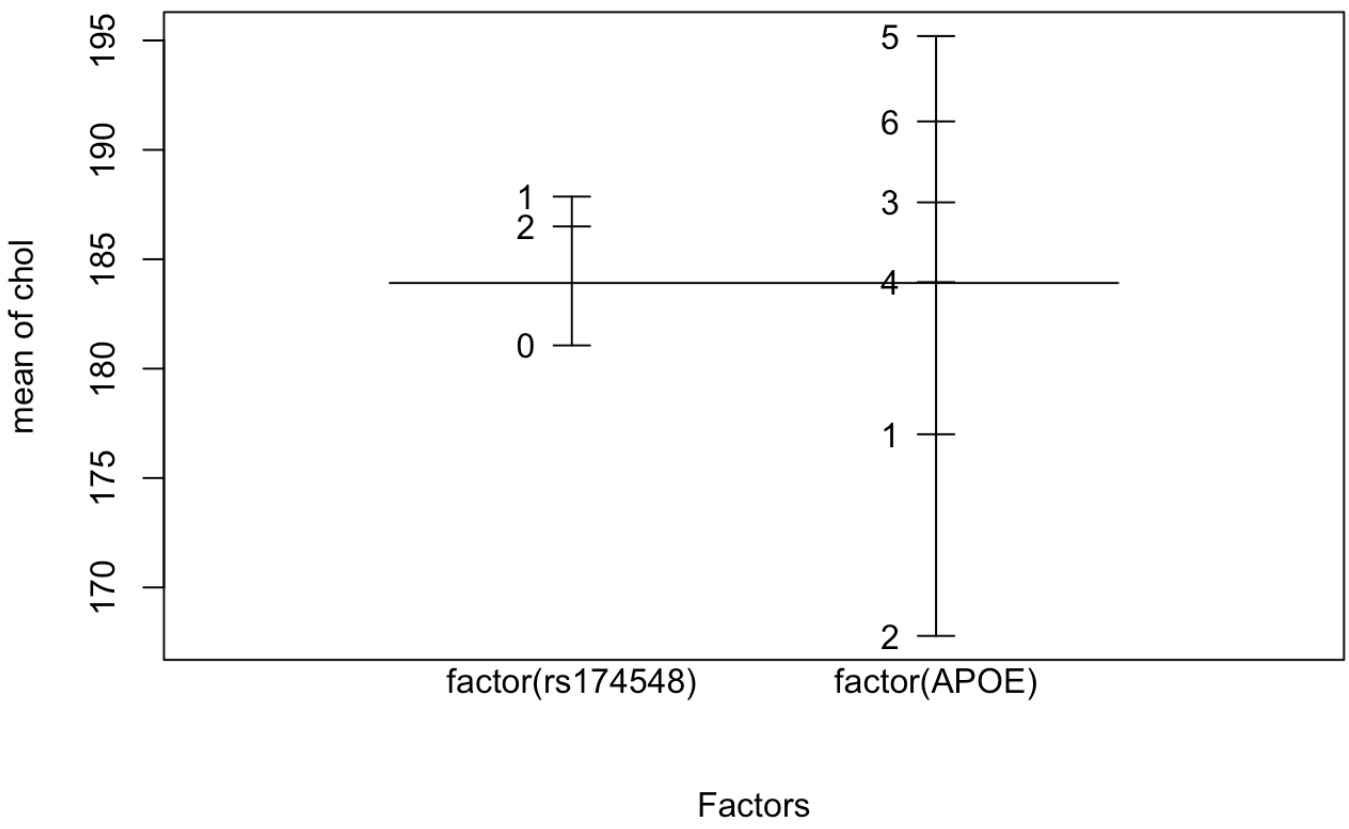
```
tapply(chol, list(factor(rs174548), factor(APOE)), mean)
```

```
##     1        2        3        4        5        6
## 0 177 168.0909 192.0000 180.4653 193.6250 180.6667
## 1  NA 167.7059 184.6667 187.9192 199.0833 207.2500
## 2  NA 159.0000       NA 188.5417 165.0000       NA
```

```
tapply(chol, list(factor(rs174548), factor(APOE)), sd)
```

```
##          1        2        3        4        5        6
## 0 16.97056 17.39318 18.38478 21.00646 18.07773 23.04488
## 1       NA 12.65783 37.85939 24.03810 18.82856 14.68276
## 2       NA       NA       NA 16.46598       NA       NA
```

```
par(mfrow = c(1,1))
plot.design(chol ~ factor(rs174548) + factor(APOE))
```




```
# model with interaction
fit1 = lm(chol ~ factor(rs174548)*factor(APOE))
summary(fit1)
```

```
##
## Call:
## lm(formula = chol ~ factor(rs174548) * factor(APOE))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.465 -13.021  -0.042  13.671  56.081
##
## Coefficients: (4 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     177.000     14.659  12.074   <2e-16 ***
## factor(rs174548)1                26.583     13.382   1.986   0.0477 *
## factor(rs174548)2               -28.625     20.989  -1.364   0.1734
## factor(APOE)2                    -8.909     15.097  -0.590   0.5555
## factor(APOE)3                    15.000     20.732   0.724   0.4698
## factor(APOE)4                     3.465     14.761   0.235   0.8145
## factor(APOE)5                    16.625     15.022   1.107   0.2691
## factor(APOE)6                     3.667     16.927   0.217   0.8286
## factor(rs174548)1:factor(APOE)2  -26.968     14.744  -1.829   0.0682 .
## factor(rs174548)2:factor(APOE)2   19.534     29.722   0.657   0.5114
## factor(rs174548)1:factor(APOE)3  -33.917     23.179  -1.463   0.1442
## factor(rs174548)2:factor(APOE)3       NA         NA      NA       NA
## factor(rs174548)1:factor(APOE)4  -19.129     13.653  -1.401   0.1620
## factor(rs174548)2:factor(APOE)4   36.701     21.481   1.709   0.0883 .
## factor(rs174548)1:factor(APOE)5  -21.125     14.413  -1.466   0.1435
## factor(rs174548)2:factor(APOE)5       NA         NA      NA       NA
## factor(rs174548)1:factor(APOE)6       NA         NA      NA       NA
## factor(rs174548)2:factor(APOE)6       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.73 on 386 degrees of freedom
## Multiple R-squared:   0.15,  Adjusted R-squared:  0.1214
## F-statistic: 5.241 on 13 and 386 DF,  p-value: 1.169e-08
```

```
# model without interaction
fit2 = lm(chol ~ factor(rs174548) + factor(APOE))
summary(fit2)
```

```
##
## Call:
## lm(formula = chol ~ factor(rs174548) + factor(APOE))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.074 -13.074  -0.328  14.390  56.507
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        177.000     14.685  12.053  < 2e-16 ***
## factor(rs174548)1    6.419      2.208   2.907  0.00385 **
## factor(rs174548)2    5.575      4.348   1.282  0.20060
## factor(APOE)2      -11.465     14.990  -0.765  0.44483
## factor(APOE)3        6.749     17.426   0.387  0.69876
## factor(APOE)4        4.074     14.772   0.276  0.78286
## factor(APOE)5       15.744     14.933   1.054  0.29237
## factor(APOE)6       11.733     16.111   0.728  0.46691
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.77 on 392 degrees of freedom
## Multiple R-squared:  0.1338, Adjusted R-squared:  0.1183
## F-statistic:  8.65 on 7 and 392 DF,  p-value: 6.989e-10
```

```
# compare models with and without interaction
anova(fit2,fit1)
```

```
## Analysis of Variance Table
##
## Model 1: chol ~ factor(rs174548) + factor(APOE)
## Model 2: chol ~ factor(rs174548) * factor(APOE)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    392 169074
## 2    386 165903  6    3170.5 1.2294 0.2901
```

Once your group has completed the lab exercises, please submit your R script file to the class Github repository:

https://github.com/rhubb/SISG2018/tree/master/submit (https://github.com/rhubb/SISG2018/tree/master/submit)

Sign in using the class username and password. Then click upload files to save your R script file to the repository.