Module 4: Regression Methods - Concepts and Applications

Introduction

The goal of these lab exercises is to use the cholesterol data set to explore relationships among the variables. 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/SISG2019/master/data/SISG-Data-cholesterol.csv", head=T)
```

Install R packages

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

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

Exercises

We will first explore the data set using descriptive statistics and use simiple linear regression to investigate bivariate associations. The objective of this initial analysis is to explore the relationship between triglycerides and BMI.

1. Use plots and descriptive statistics to explore the variables triglycerides and BMI individually as well as their relationship to each other. Based on your graphical summaries does there appear to be an association between triglycerides and BMI?

```
summary(TG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    47.0   114.8   156.5   177.4   234.0   586.0
```

```
summary(BMI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.40   22.90   24.60   25.00   26.73   38.80
```

```
group = 1*(BMI > 25)
group=factor(group,levels=c(0,1), labels=c("<=25",">25"))
table(group)
```

```
## group
## <=25  >25
##  224  176
```

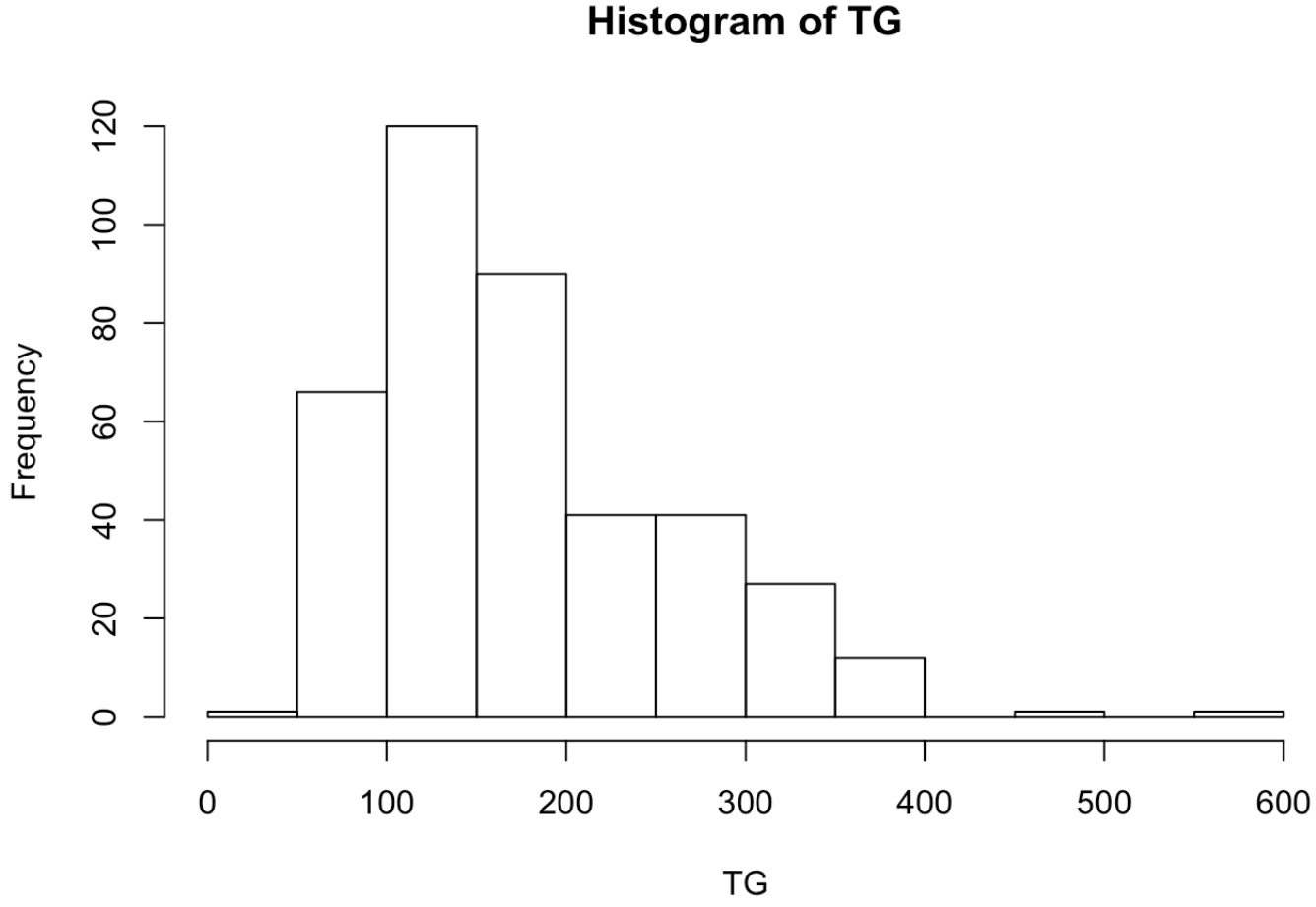
```
by(TG, group, mean)
```

```
## group: <=25
## [1] 147.3839
## ------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------
------------------------------
## group: >25
## [1] 215.6932
```

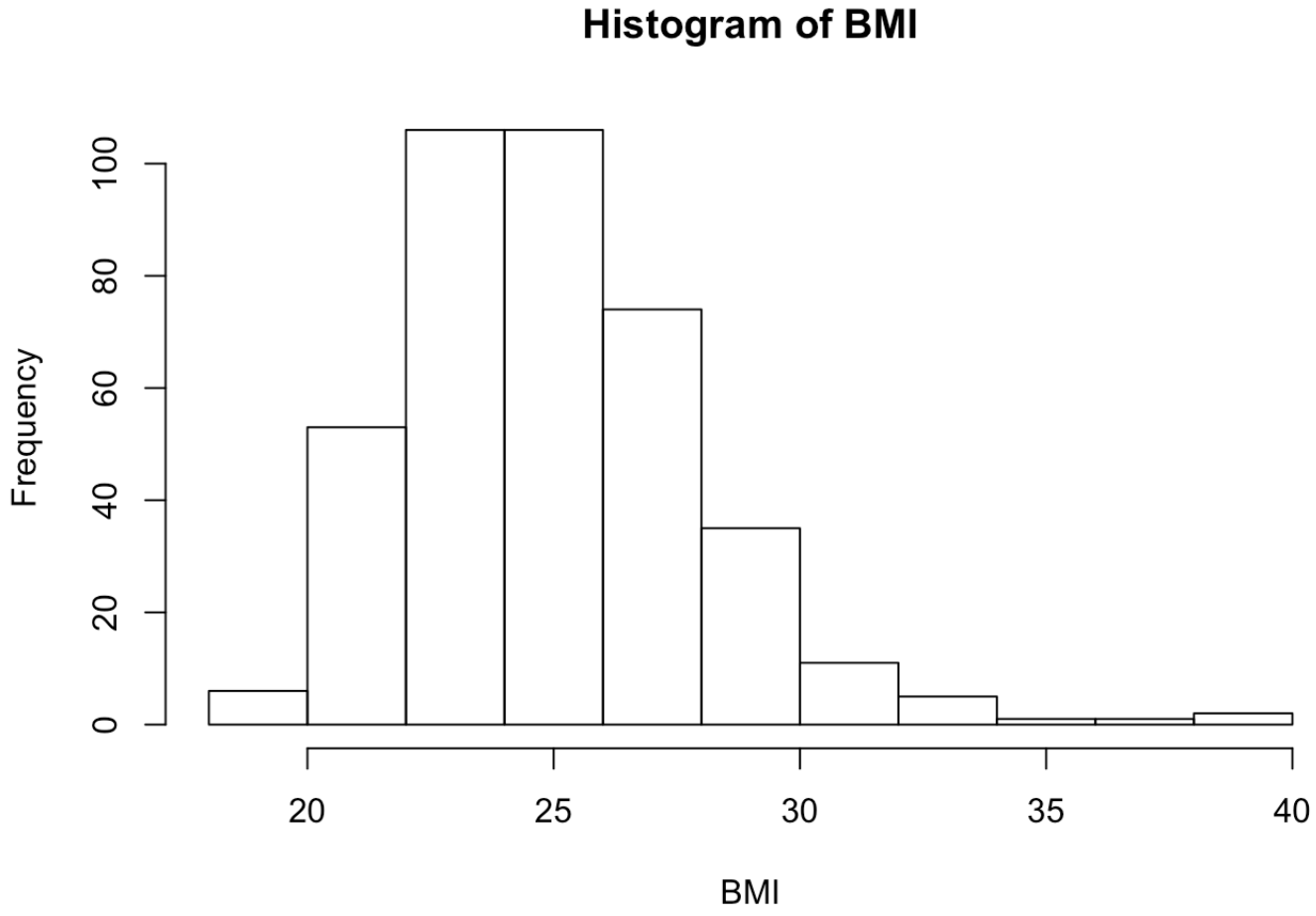
```
by(TG, group, sd)
```

```
## group: <=25
## [1] 61.70787
## ------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------
------------------------------
## group: >25
## [1] 90.66584
```

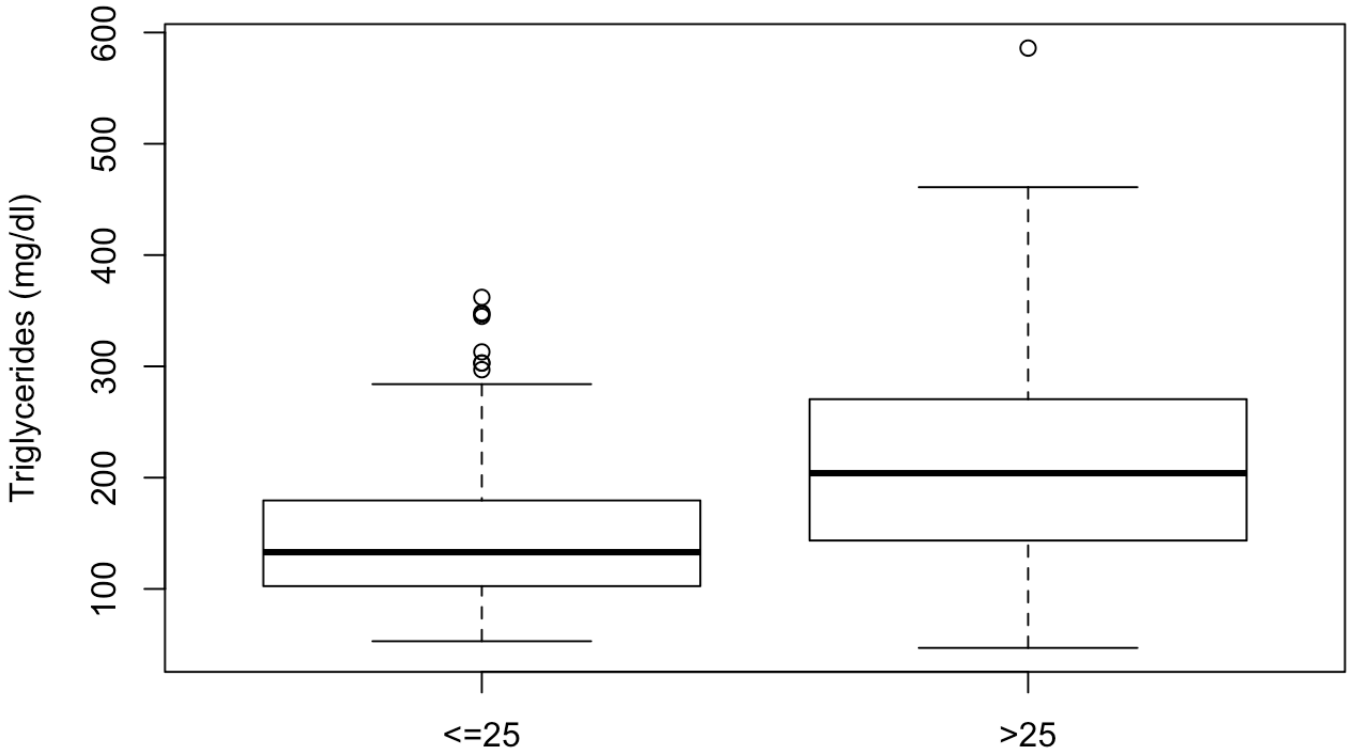
```
hist(TG)
```


Histogram of TG
Frequency
0
20
40
60
80
100
120
0
100
200
300
400
500
600
TG

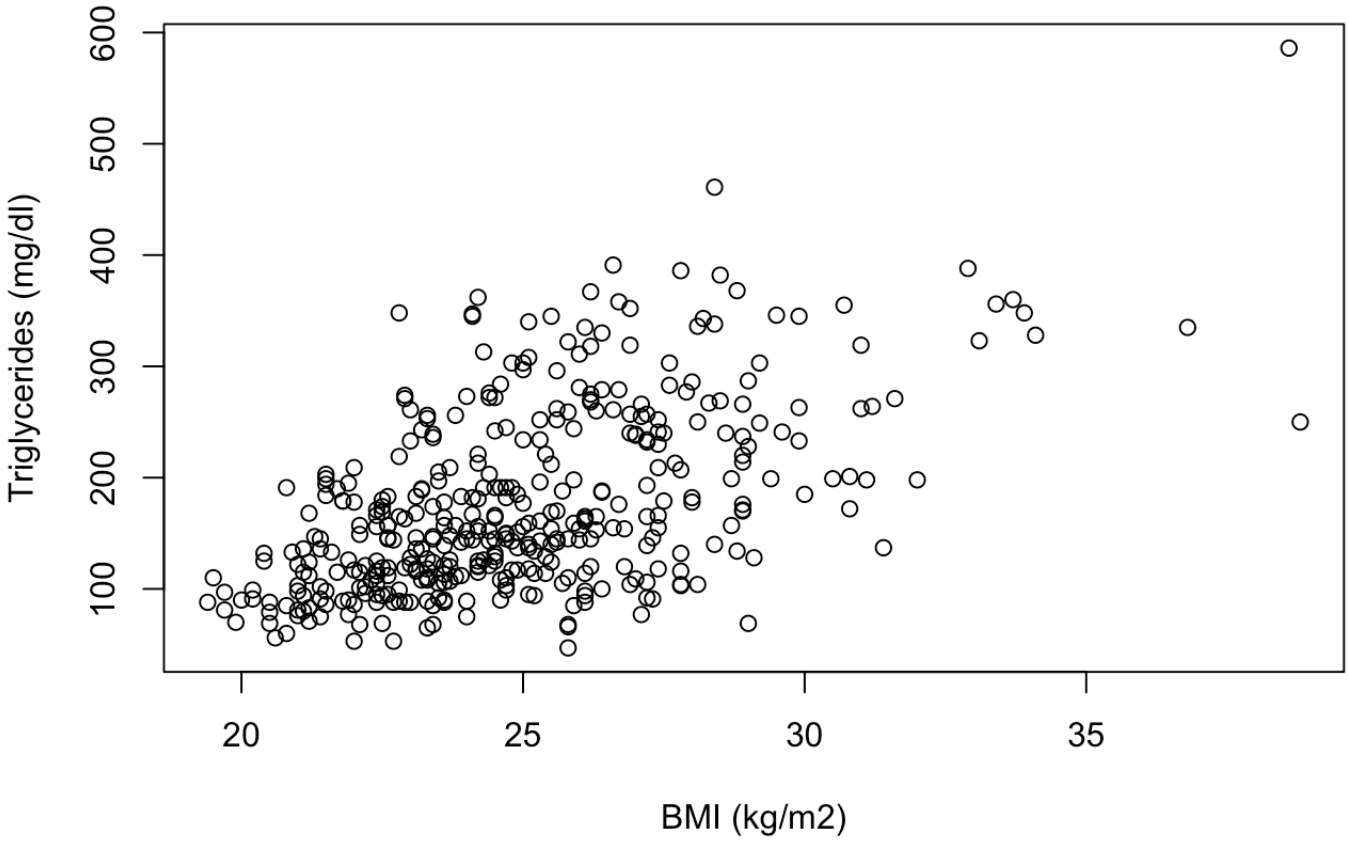
```
hist(BMI)
```




```
boxplot(TG~group,ylab="Triglycerides (mg/dl)")
```




```
plot(TG ~ BMI, xlab = "BMI (kg/m2)", ylab = "Triglycerides (mg/dl)")
```



2. Use linear regression to investigate the association between triglycerides and BMI. What do the linear regression model results tell you about the association? Make sure you can interpret the model coefficients and any hypothesis testing.

```
fit1 = lm(TG ~ BMI)
summary(fit1)
```

```
##
## Call:
## lm(formula = TG ~ BMI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -170.19  -45.10  -12.89   39.60  231.08
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -208.50      28.95  -7.203 2.97e-12 ***
## BMI            15.44       1.15  13.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.93 on 398 degrees of freedom
## Multiple R-squared:  0.3118, Adjusted R-squared:  0.3101
## F-statistic: 180.3 on 1 and 398 DF,  p-value: < 2.2e-16
```

3. Compute a prediction for the mean value of triglycerides at BMI = 23 as well as for a new individual with BMI = 23. How do these two intervals differ and why?

```
predict(fit1, newdata = data.frame(BMI = 23), interval = "confidence")
```

```
##        fit      lwr      upr
## 1 146.5612 138.4161 154.7062
```

```
predict(fit1, newdata = data.frame(BMI = 23), interval = "prediction")
```

```
##        fit      lwr      upr
## 1 146.5612 10.80972 282.3126
```

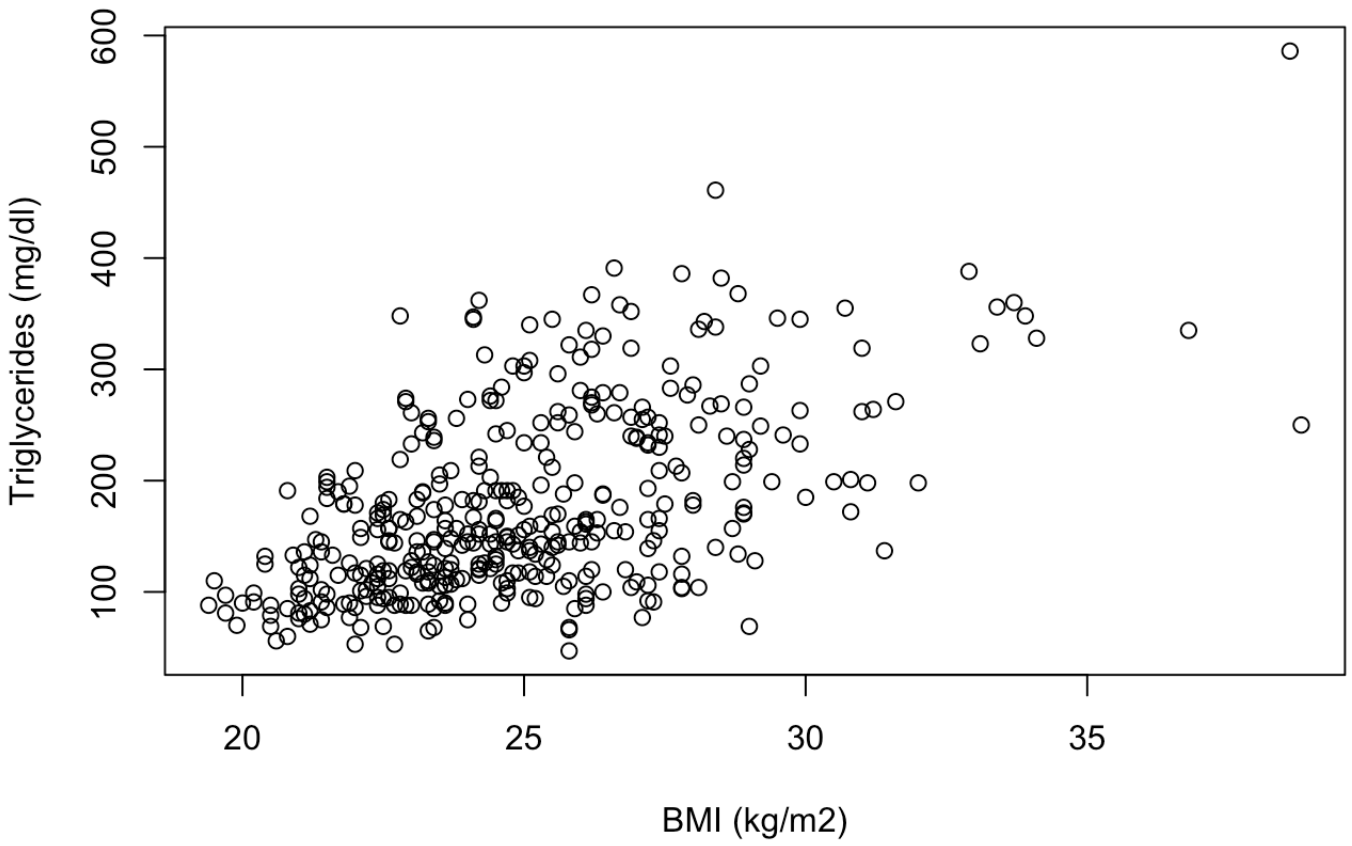
4. What is the R^2 value for the regression of triglycerides on BMI? What does this value tell you about the relationship between these two variables?

```
fit1 = lm(TG ~ BMI)
summary(fit1)$r.squared
```

```
## [1] 0.3118064
```

5. Based on a scatterplot of triglycerides versus BMI, are there any points that you suspect might have a large influence on the regression estimates? Compare linear regression results with and without the possibly influential points. Does it appear that these points had much influence on your results?

```
# Scatterplot of triglycerides vs BMI
plot(TG ~ BMI, xlab = "BMI (kg/m2)", ylab = "Triglycerides (mg/dl)")
```




```
# Identify observations with BMI <=37
bmi37 = which(BMI<=37)

# Consider again the regression of TG on BMI
fit1=lm(TG~BMI)
summary(fit1)
```

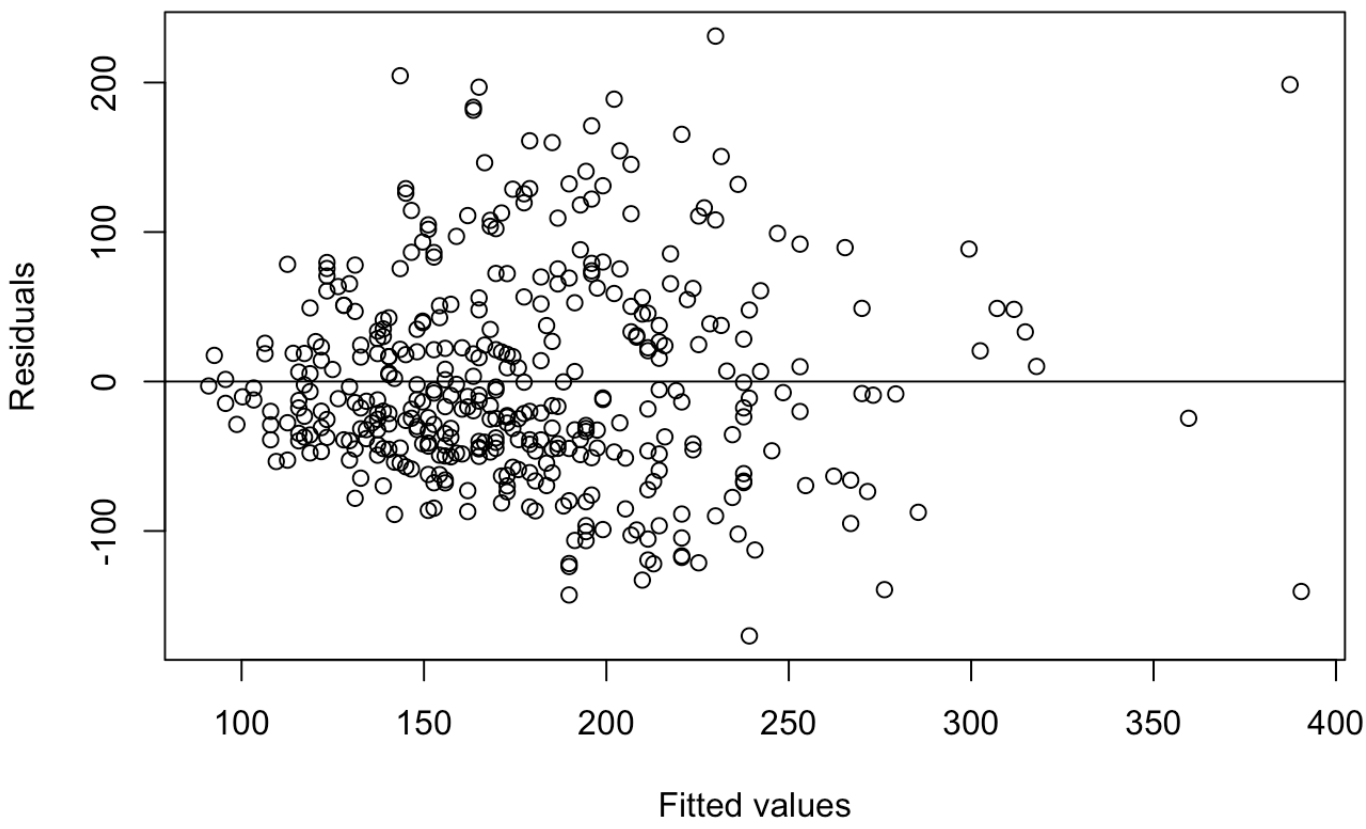
```
##
## Call:
## lm(formula = TG ~ BMI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -170.19  -45.10  -12.89   39.60  231.08
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -208.50      28.95  -7.203 2.97e-12 ***
## BMI            15.44       1.15  13.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.93 on 398 degrees of freedom
## Multiple R-squared:  0.3118, Adjusted R-squared:  0.3101
## F-statistic: 180.3 on 1 and 398 DF,  p-value: < 2.2e-16
```

```
# excluding subjects with BMI > 37
fit2 = lm(TG[bmi37] ~ BMI[bmi37])
summary(fit2)
```

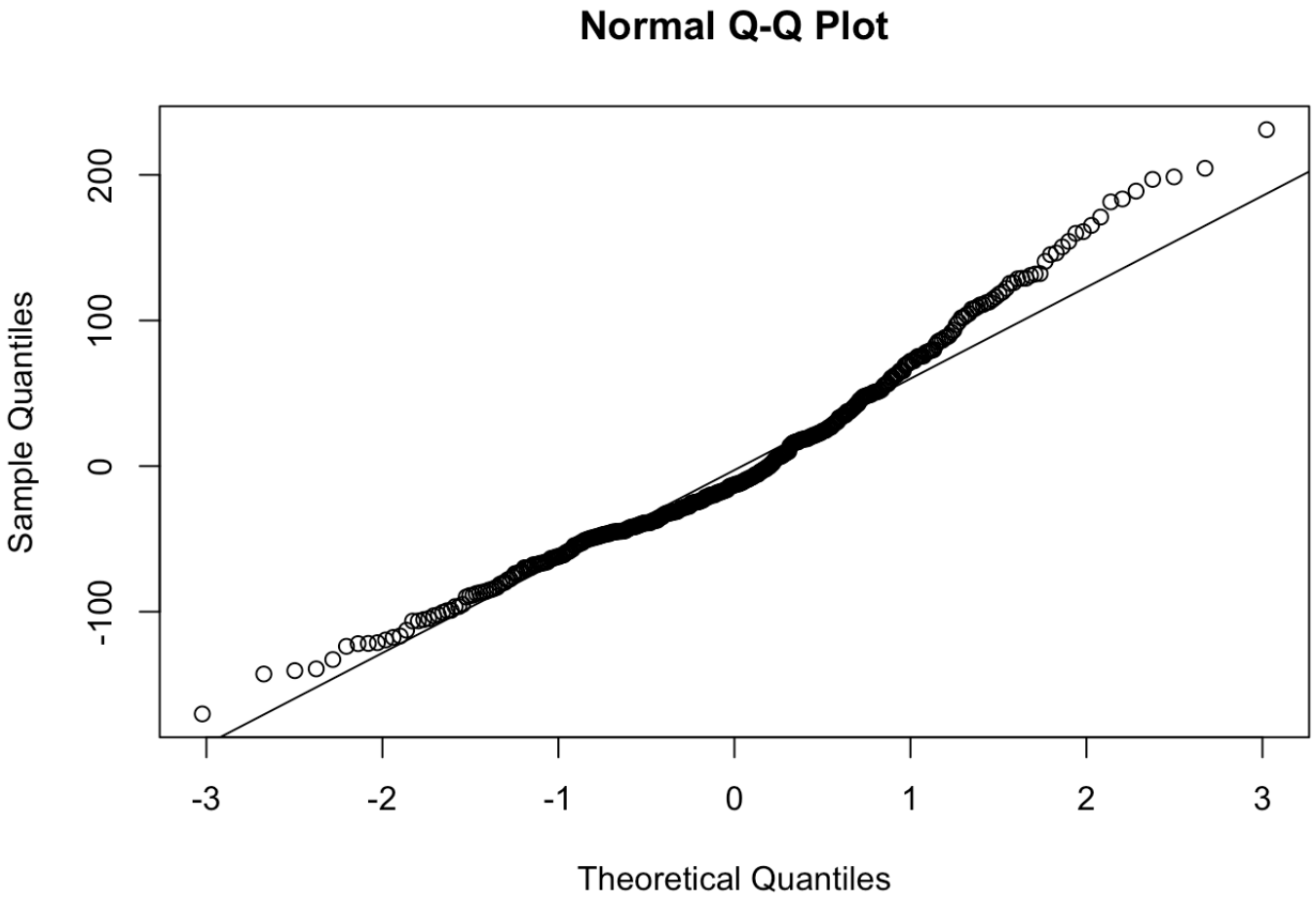
```
##
## Call:
## lm(formula = TG[bmi37] ~ BMI[bmi37])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -169.07  -44.87  -13.22   39.45  232.05
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -202.707     30.084  -6.738 5.68e-11 ***
## BMI[bmi37]    15.199      1.199  12.677  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.01 on 396 degrees of freedom
## Multiple R-squared:  0.2887, Adjusted R-squared:  0.2869
## F-statistic: 160.7 on 1 and 396 DF,  p-value: < 2.2e-16
```

6. Conduct a residuals analysis (using all data) to check the linear regression model assumptions. Do any modeling assumptions appear to be violated? How do model results change if you use robust standard errors?

```
# Plot residuals vs fitted values
plot(fit1$fitted, fit1$residuals,xlab="Fitted values",ylab="Residuals")
abline(0,0)
```




```
# Quantile-quantile plot
qqnorm(fit1$residuals)
qqline(fit1$residuals)
```




```
# Deletion diagnostics
dfb=dfbeta(fit1)
index=order(abs(dfb[,2]),decreasing=T)
cbind(dfb[index[1:15],],BMI[index[1:15]],TG[index[1:15]])
```

```
##     (Intercept)        BMI
## 266  -19.330846  0.7942199 38.6 586
## 152   13.901014 -0.5709072 38.8 250
## 42     5.931197 -0.2513651 31.4 137
## 105   -4.913771  0.2197891 28.4 461
## 182   -4.740550  0.1986603 32.9 388
## 269    4.338551 -0.1906778 29.0  69
## 41     4.106832 -0.1731648 32.0 198
## 278    3.636316 -0.1550624 30.8 172
## 354   -3.306959  0.1474196 28.5 382
## 232   -3.365307  0.1436724 30.7 355
## 94    -3.176430  0.1403325 28.8 368
## 345    2.953435 -0.1294906 29.1 128
## 85    -2.819085  0.1293738 27.8 386
## 102    2.976553 -0.1265171 31.1 198
## 306   -2.929242  0.1264456 29.9 345
```

```
# fit a linear regression model with robust standard errors
fit.gee = gee(TG ~ BMI, id = seq(1,length(TG)))
```

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

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

```
## (Intercept)         BMI
##  -208.50096    15.43748
```

```
summary(fit.gee)
```

```
##
##  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 = TG ~ BMI, id = seq(1, length(TG)))
##
## Summary of Residuals:
##        Min          1Q     Median          3Q         Max
## -170.18608  -45.09554  -12.88618   39.60133  231.07641
##
##
## Coefficients:
##               Estimate Naive S.E.   Naive z Robust S.E.  Robust z
## (Intercept) -208.50096  28.946250 -7.203039   32.021396 -6.511301
## BMI           15.43748   1.149603 13.428538    1.322308 11.674646
##
## Estimated Scale Parameter:  4750.958
## Number of Iterations:  1
##
## Working Correlation
##      [,1]
## [1,]    1
```

```
# calculate p-values for robust regression
z = abs(fit.gee$coef/sqrt(diag(fit.gee$robust)))
2*(1-pnorm(z))
```

```
##  (Intercept)          BMI
## 7.450263e-11 0.000000e+00
```

7. Summarize the variable APOE. Create a new binary variable indicating presence of the APOE e4 allele (APOE = 3, 5, or 6). Investigate the association between triglycerides and BMI adjusting for presence of the APOE e4 allele. What do the linear regression model results tell you about the adjusted association? Make sure you can interpret the model

coefficients and any hypothesis testing.

```
# Summarize the variable APOE
table_APOE=table(APOE)
table_APOE
```

```
## APOE
##   1   2   3   4   5   6
##   2  51   5 267  65  10
```

```
prop.table(table_APOE)
```

```
## APOE
##      1      2      3      4      5      6
## 0.0050 0.1275 0.0125 0.6675 0.1625 0.0250
```

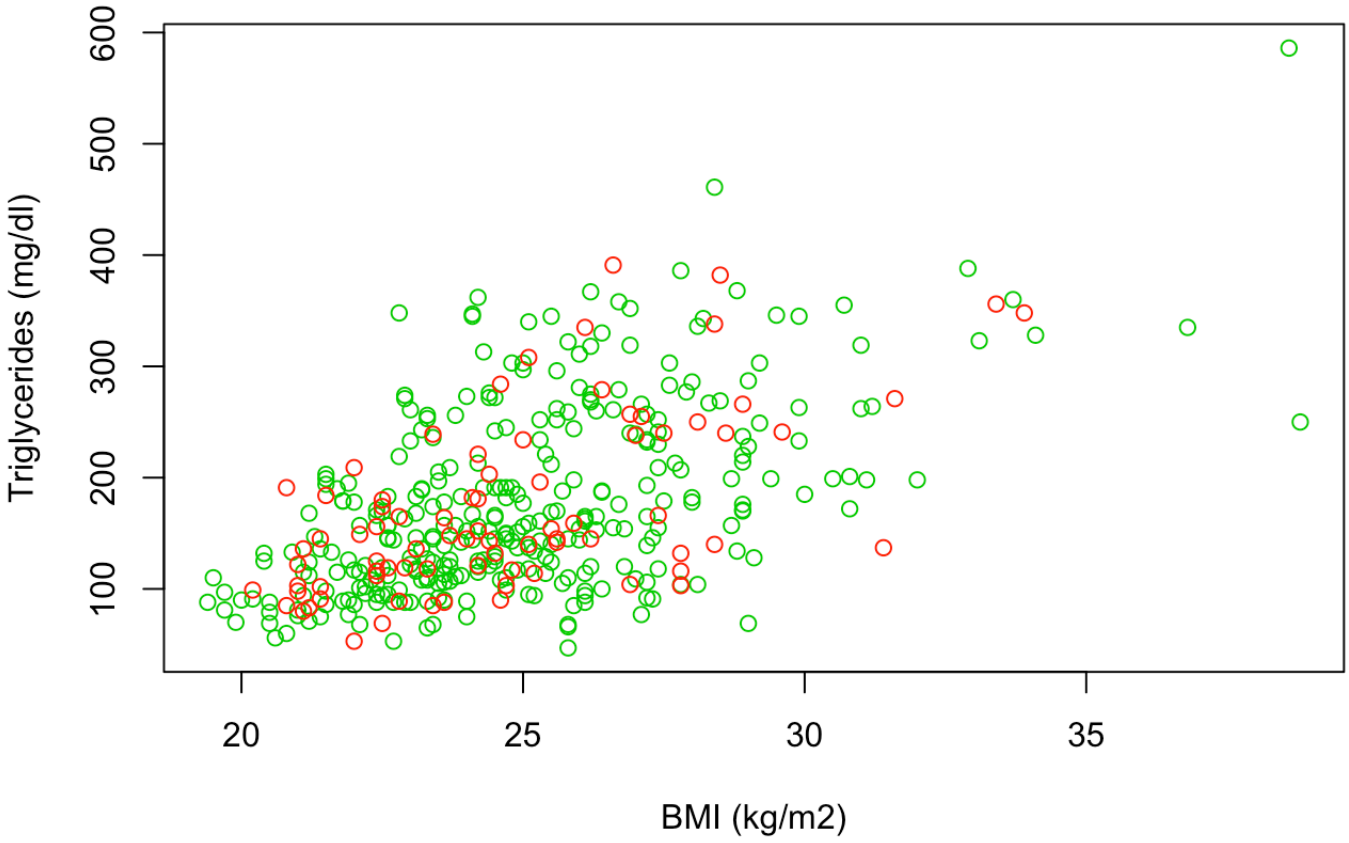
```
# binary variable indicating presence of APOE4
APOE4 = ifelse(APOE %in% c(3,5,6), 1, 0)

## Linear regression analyses for association of APOE4 and BMI with TG  ----------
# multiple linear regression of triglycerides on BMI and APOE4
fit3=lm(TG~BMI+APOE4)
summary(fit3)
```

```
##
## Call:
## lm(formula = TG ~ BMI + APOE4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -170.62  -45.59  -12.70   39.09  230.64
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -207.674     29.129  -7.130 4.79e-12 ***
## BMI           15.424      1.152  13.389  < 2e-16 ***
## APOE4         -2.427      8.634  -0.281    0.779
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.01 on 397 degrees of freedom
## Multiple R-squared:  0.3119, Adjusted R-squared:  0.3085
## F-statistic: 89.99 on 2 and 397 DF,  p-value: < 2.2e-16
```

8. Plot separate scatterplots for triglycerides vs BMI for subjects in the two groups defined by presence of the APOE e4 allele. Do these plots suggest effect modification? Fit a linear regression model that investigates whether the association between triglycerides and BMI is modified by the APOE4 allele. Is there evidence of effect modification? Make sure that you can interpret the regression coefficients from this model as well as any hypothesis tests.

```
# scatterplot with subjects stratified by APOE4
par(mfrow = c(1,1))
plot(BMI[APOE4 == 0], TG[APOE4 == 0], pch = 1, col=75,xlab = "BMI (kg/m2)", ylab = "Triglyc
erides (mg/dl)")
points(BMI[APOE4 == 1], TG[APOE4 == 1], pch = 1, col=34)
```




```
# multiple linear regression of triglycerides on BMI, APOE4, and interaction
fit4 = lm(TG ~ BMI*APOE4)
summary(fit4)
```

```
##
## Call:
## lm(formula = TG ~ BMI * APOE4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -170.04  -45.72  -13.03   38.88  231.12
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -204.0193    32.4558  -6.286  8.6e-10 ***
## BMI           15.2780     1.2857  11.883  < 2e-16 ***
## APOE4        -20.9439    72.6801  -0.288    0.773
## BMI:APOE4      0.7464     2.9088   0.257    0.798
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.09 on 396 degrees of freedom
## Multiple R-squared:  0.3121, Adjusted R-squared:  0.3068
## F-statistic: 59.88 on 3 and 396 DF,  p-value: < 2.2e-16
```

```
# Compare the models with and without interaction
anova(fit3,fit4)
```

```
## Analysis of Variance Table
##
## Model 1: TG ~ BMI + APOE4
## Model 2: TG ~ BMI * APOE4
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    397 1890505
## 2    396 1890191  1    314.27 0.0658 0.7976
```

```
# Compare with the model without APOE4
anova(fit1,fit4)
```

```
## Analysis of Variance Table
##
## Model 1: TG ~ BMI
## Model 2: TG ~ BMI * APOE4
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    398 1890881
## 2    396 1890191  2    690.59 0.0723 0.9302
```

Next we will investigate the association between a set of categorical predictors and a continuous outcome. For these exercises, we will study the relationship between several genotypes included in the data set and total cholesterol level.

9. Perform a descriptive analysis to explore the variables for total cholesterol and rs4775401 as well as the relationship between them using numeric and graphical methods.

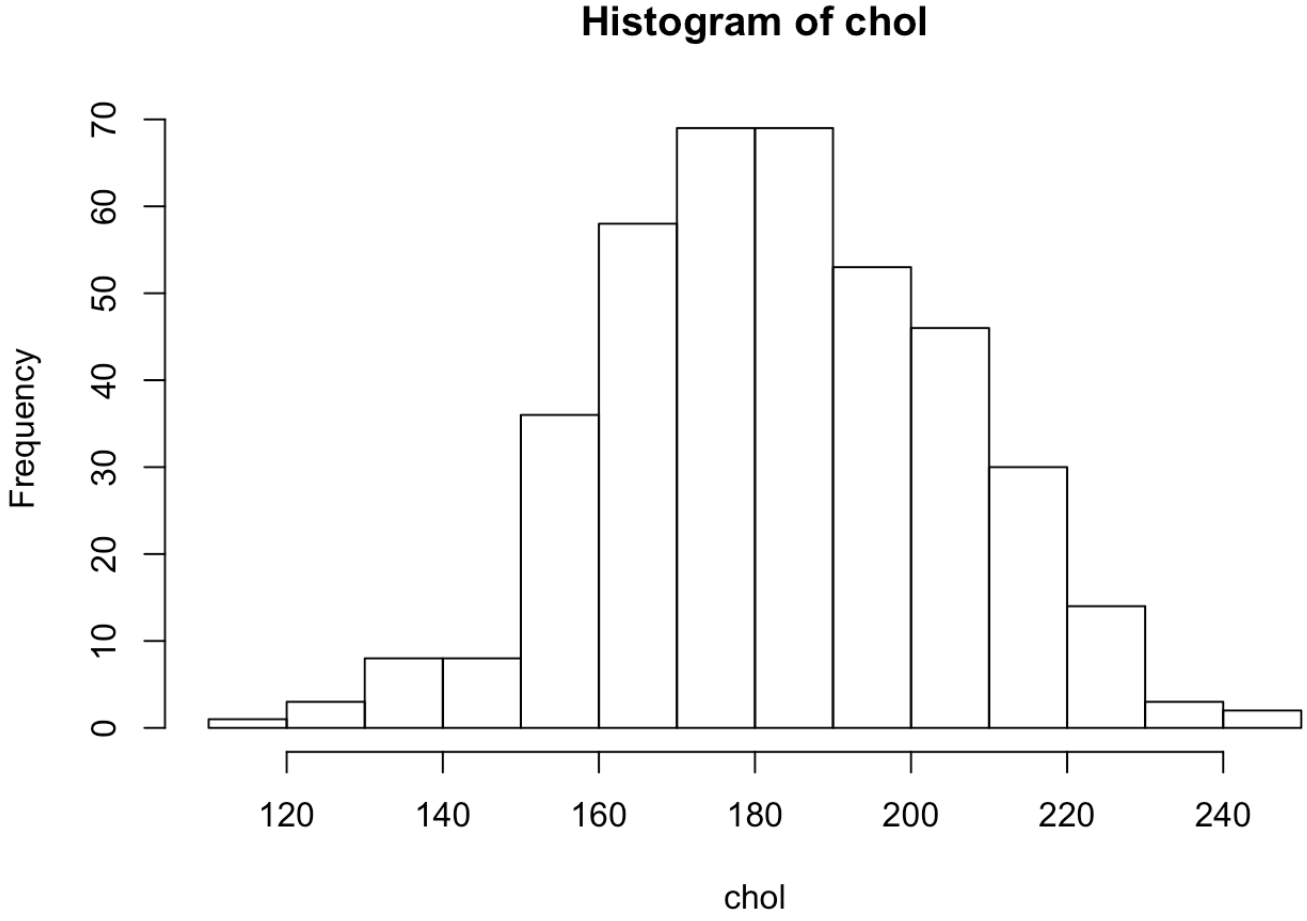
```
# descriptive statistics
summary(chol)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   117.0   168.0   184.0   183.9   199.2   247.0
```

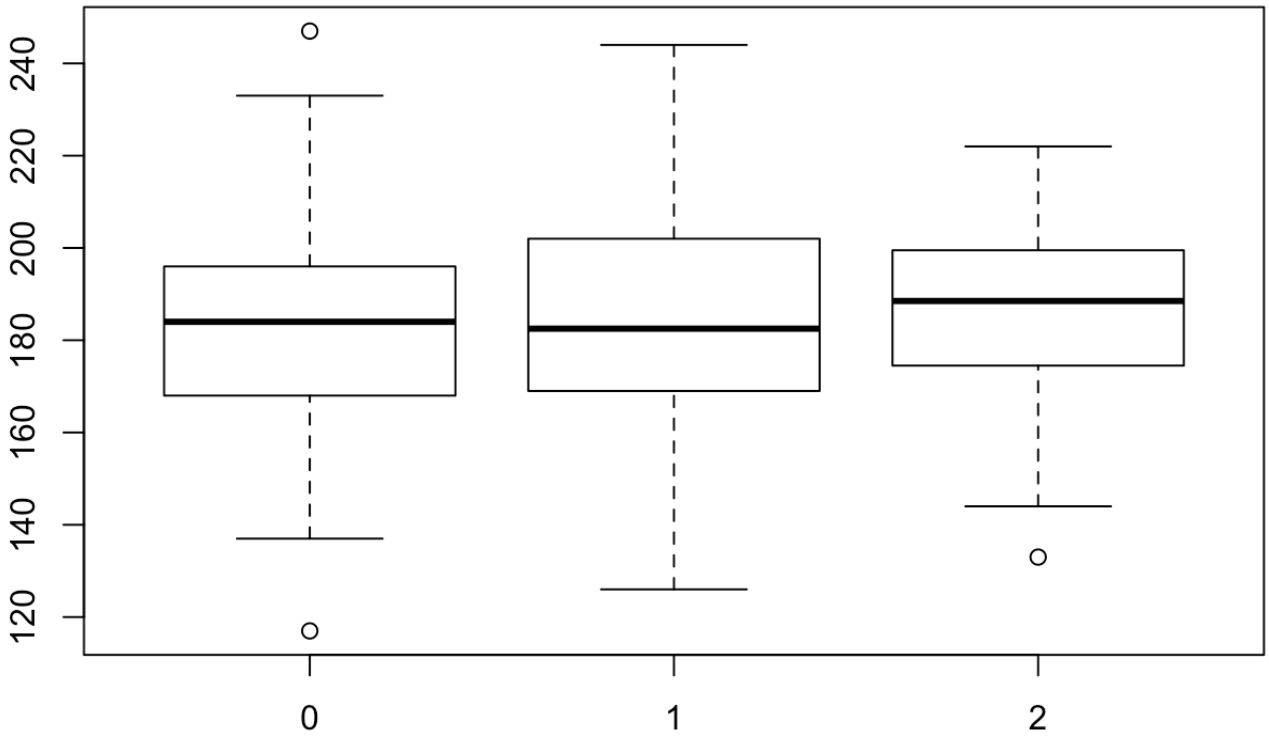
```
table(rs4775401)
```

```
## rs4775401
##   0   1   2
## 202 170  28
```

```
hist(chol)
```




```
# 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
```

10. 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? Compare results obtained using classical ANOVA to those based on ANOVA allowing for unequal variances, using robust standard errors, and using a nonparametric test. How do your results differ? Which approach do you prefer and why?

```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
# 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
```

11. Carry out all pairwise comparisons between rs4775401 genotypes and cholesterol using an adjustment method of your choice to address the issue of multiple comparisons. What do you conclude about differences in cholesterol between the genotypes?

```
# 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)
```

```
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("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("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)
```

12. Perform a descriptive analysis to investigate the relationships between cholesterol, APOE and rs174548. Use ANOVA 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
```

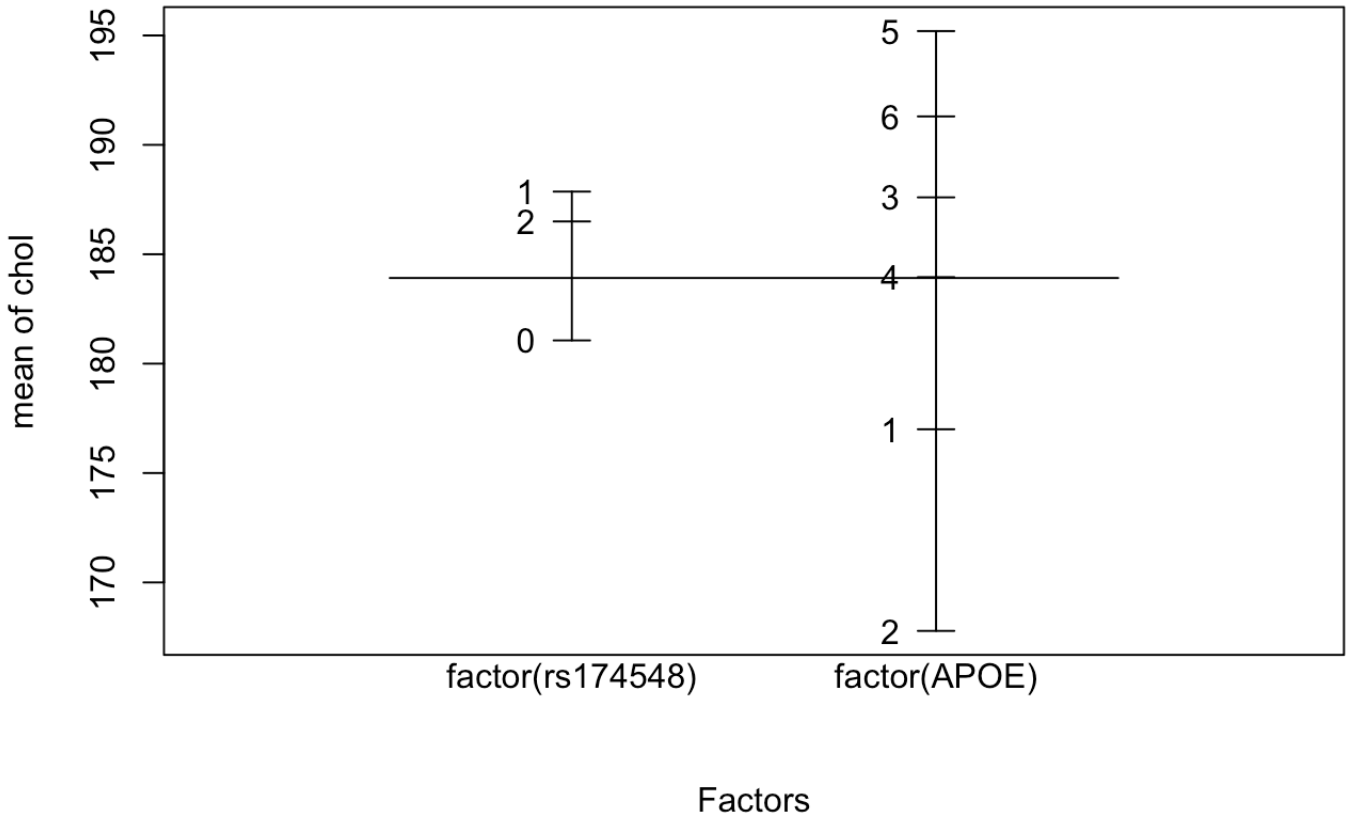
```
tapply(chol, list(factor(rs174548), factor(APOE)), mean)
```

```
##     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
```

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




```r
# 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
```

For the final set of exercises we will study the relationship between genotype, clinical characteristics, and the binary outcome hypertension

13. Is there an association between rs174548 and hypertension? Analyze this relationship using descriptive statistics as well as a logistic regression analysis.

```
# Descriptive statistics for hypertension
table(HTN)
```

```
## HTN
##   0   1
##  85 315
```

```
table(HTN,rs174548)
```

```
##    rs174548
## HTN   0   1   2
##   0  61  21   3
##   1 166 126  23
```

```
chisq.test(HTN,rs174548)
```

```
##
##  Pearson's Chi-squared test
##
## data:  HTN and rs174548
## X-squared = 10.014, df = 2, p-value = 0.006692
```

```
by(TG,HTN,mean)
```

```
## HTN: 0
## [1] 160.3412
## ------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------
------------------------------
## HTN: 1
## [1] 182.054
```

```r
# Logistic regression analysis for the association between rs174548 and hypertension
glm.mod1 <- glm(HTN ~ factor(rs174548), family = "binomial")
summary(glm.mod1)
```

```
##
## Call:
## glm(formula = HTN ~ factor(rs174548), family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0782   0.4952   0.5553   0.7912   0.7912
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.0011     0.1497   6.686 2.29e-11 ***
## factor(rs174548)1   0.7906     0.2792   2.831  0.00463 **
## factor(rs174548)2   1.0358     0.6318   1.639  0.10115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 413.80  on 399  degrees of freedom
## Residual deviance: 403.39  on 397  degrees of freedom
## AIC: 409.39
##
## Number of Fisher Scoring iterations: 4
```

```r
exp(glm.mod1$coef)
```

```
##       (Intercept) factor(rs174548)1 factor(rs174548)2
##          2.721311          2.204819          2.817269
```

```r
exp(confint(glm.mod1))
```

```
## Waiting for profiling to be done...
```

```
##                         2.5 %    97.5 %
## (Intercept)       2.0416424  3.675895
## factor(rs174548)1 1.2935601  3.883015
## factor(rs174548)2 0.9375188 12.174163
```

14. Use logistic regression to investigate the association between triglycerides and hypertension. What can you conclude about the relationship based on these results? Make sure that you can interpret the model coefficients and hypothesis testing.

```
# Logistic regression analysis for the association between triglycerides and hypertension
glm.mod2 <- glm(HTN ~ TG, family = "binomial")
summary(glm.mod2)
```

```
##
## Call:
## glm(formula = HTN ~ TG, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0433   0.5219   0.6697   0.7417   0.8333
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.715580   0.295441   2.422   0.0154 *
## TG          0.003482   0.001637   2.127   0.0334 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 413.80  on 399  degrees of freedom
## Residual deviance: 408.92  on 398  degrees of freedom
## AIC: 412.92
##
## Number of Fisher Scoring iterations: 4
```

```
exp(glm.mod2$coef)
```

```
## (Intercept)          TG
##    2.045374    1.003488
```

```
exp(confint(glm.mod2))
```

```
## Waiting for profiling to be done...
```

```
##                2.5 %   97.5 %
## (Intercept) 1.144445 3.651986
## TG          1.000382 1.006839
```

15. Analyze the association between hypertension and rs174548 adjusted for triglycerides using logistic regression. What does this model tell you about the association between rs174548 and hypertension? What role does triglycerides play in this analysis?

```
# logistic regression analysis for the association between rs174548 and hypertension
# adjusting for triglycerides
glm.mod3 <- glm(HTN ~ TG+factor(rs174548), family = "binomial")
summary(glm.mod3)
```

```
##
## Call:
## glm(formula = HTN ~ TG + factor(rs174548), family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1280   0.4335   0.5995   0.7758   0.9378
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.436636   0.310955   1.404  0.16027
## TG                0.003339   0.001658   2.013  0.04411 *
## factor(rs174548)1 0.786461   0.280547   2.803  0.00506 **
## factor(rs174548)2 0.963842   0.634925   1.518  0.12900
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 413.80  on 399  degrees of freedom
## Residual deviance: 399.05  on 396  degrees of freedom
## AIC: 407.05
##
## Number of Fisher Scoring iterations: 4
```

```
exp(glm.mod3$coef)
```

```
##       (Intercept)                TG factor(rs174548)1 factor(rs174548)2
##          1.547492          1.003344          2.195611          2.621751
```

```
exp(confint(glm.mod3))
```

```
## Waiting for profiling to be done...
```

```
##                         2.5 %    97.5 %
## (Intercept)        0.8383655  2.843689
## TG                 1.0001933  1.006736
## factor(rs174548)1 1.2847081  3.876255
## factor(rs174548)2 0.8652782 11.375999
```

```
lrtest(glm.mod2,glm.mod3)
```

```
## Likelihood ratio test
##
## Model 1: HTN ~ TG
## Model 2: HTN ~ TG + factor(rs174548)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   2 -204.46
## 2   4 -199.52  2 9.8682   0.007197 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

16. Use a GLM to estimate the relative risk of hypertension for patients with different rs174548 genotypes, adjusting for triglyceries. Make sure you can interpret the coefficients. How do these results compare to the results of the logistic regression analysis?

```
# relative risk regression for the association between rs174548 and hypertension
# adjusting for triglycerides
glm.mod4 <- gee(HTN ~ TG+factor(rs174548), family = "poisson", id = seq(1,nrow(cholesterol)
))
```

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

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

```
##       (Intercept)                TG factor(rs174548)1 factor(rs174548)2
##      -0.419615759       0.000605558       0.155797546       0.175538367
```

```
summary(glm.mod4)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logarithm
##  Variance to Mean Relation: Poisson
##  Correlation Structure:     Independent
##
## Call:
## gee(formula = HTN ~ TG + factor(rs174548), id = seq(1, nrow(cholesterol)),
##     family = "poisson")
##
## Summary of Residuals:
##         Min           1Q      Median           3Q          Max
## -0.90949342  0.06820756  0.17449240  0.26578251  0.32372436
##
##
## Coefficients:
##                       Estimate   Naive S.E.   Naive z Robust S.E.  Robust z
## (Intercept)       -0.419615759 0.0654482041 -6.411417 0.065735698 -6.383377
## TG                 0.000605558 0.0003069945  1.972537 0.000262569  2.306282
## factor(rs174548)1  0.155797546 0.0547601891  2.845088 0.052279059  2.980114
## factor(rs174548)2  0.175538367 0.1033407933  1.698636 0.080279415  2.186593
##
## Estimated Scale Parameter:  0.2146029
## Number of Iterations:  1
##
## Working Correlation
##      [,1]
## [1,]    1
```

```
exp(glm.mod4$coef)
```

```
##       (Intercept)                TG factor(rs174548)1 factor(rs174548)2
##         0.6572993         1.0006057         1.1685896         1.1918877
```

```
p <- 2*(1-pnorm(abs(glm.mod4$coef)/sqrt(diag(glm.mod4$robust.variance))))
p
```

```
##       (Intercept)                TG factor(rs174548)1 factor(rs174548)2
##      1.732243e-10      2.109491e-02      2.881413e-03      2.877229e-02
```

17. Use a GLM to estimate the risk difference for hypertension according to rs174548 genotypes, adjusting for triglyceries. Make sure you can interpret the coefficients. How do these results compare to the results of the logistic regression and relative risk regression analyses?

```
# risk difference regression for the association between rs174548 and hypertension
# adjusting for triglycerides
glm.mod5 <- gee(HTN ~ TG+factor(rs174548), id = seq(1,nrow(cholesterol)))
```

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

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

```
##       (Intercept)                TG factor(rs174548)1 factor(rs174548)2
##      0.6456470422      0.0004917309      0.1235863772      0.1412652004
```

```
summary(glm.mod5)
```

```
##
##  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 = HTN ~ TG + factor(rs174548), id = seq(1, nrow(cholesterol)))
##
## Summary of Residuals:
##         Min           1Q      Median           3Q         Max
## -0.90642633  0.07354151  0.17225061  0.26448914  0.33124161
##
##
## Coefficients:
##                       Estimate   Naive S.E.   Naive z  Robust S.E.  Robust z
## (Intercept)       0.6456470422 0.0502859906 12.839501 0.0498961114 12.939827
## TG                0.0004917309 0.0002443104  2.012730 0.0002161362  2.275098
## factor(rs174548)1 0.1235863772 0.0427748139  2.889232 0.0410937749  3.007423
## factor(rs174548)2 0.1412652004 0.0838391168  1.684956 0.0683354838  2.067231
##
## Estimated Scale Parameter:  0.1631336
## Number of Iterations:  1
##
## Working Correlation
##      [,1]
## [1,]    1
```

```
p <- 2*(1-pnorm(abs(glm.mod5$coef)/sqrt(diag(glm.mod5$robust.variance))))
p
```

```
##       (Intercept)                TG factor(rs174548)1 factor(rs174548)2
##       0.000000000       0.022900079       0.002634726       0.038712434
```