Module 4: Regression Methods: Concepts and Applications

Lab 4: Logistic Regression and GLMs

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

- Is hypertension associated with rs174548?
- Is hypertension associated with triglycerides?
- Is hypertension associated with rs174548 after adjusting for triglyceride levels?

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 *lmtest* 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("lmtest")
library(gee)
library(lmtest)
```

Exercises

1. 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
```

```
# 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
```

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

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

```
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
```

2. Use logistic regression to investigate the association between triglycerides and hypertension. Interpret the results of this model. 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
```

3. 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
```

4. 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
```

5. 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?

```
# relative risk 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
```

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.