Module 4: Regression Methods: Concepts and Applications

Lab 2: Model Checking and Multiple Linear Regression

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

- Are triglyceride levels associated with BMI?
- Are linear regression model assumptions satisfied for this relationship?
- Is there an association between triglycerides and BMI after adjusting for the APOE e4 allele?
- Is the association between triglycerides and BMI modified by the APOE e4 allele?

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

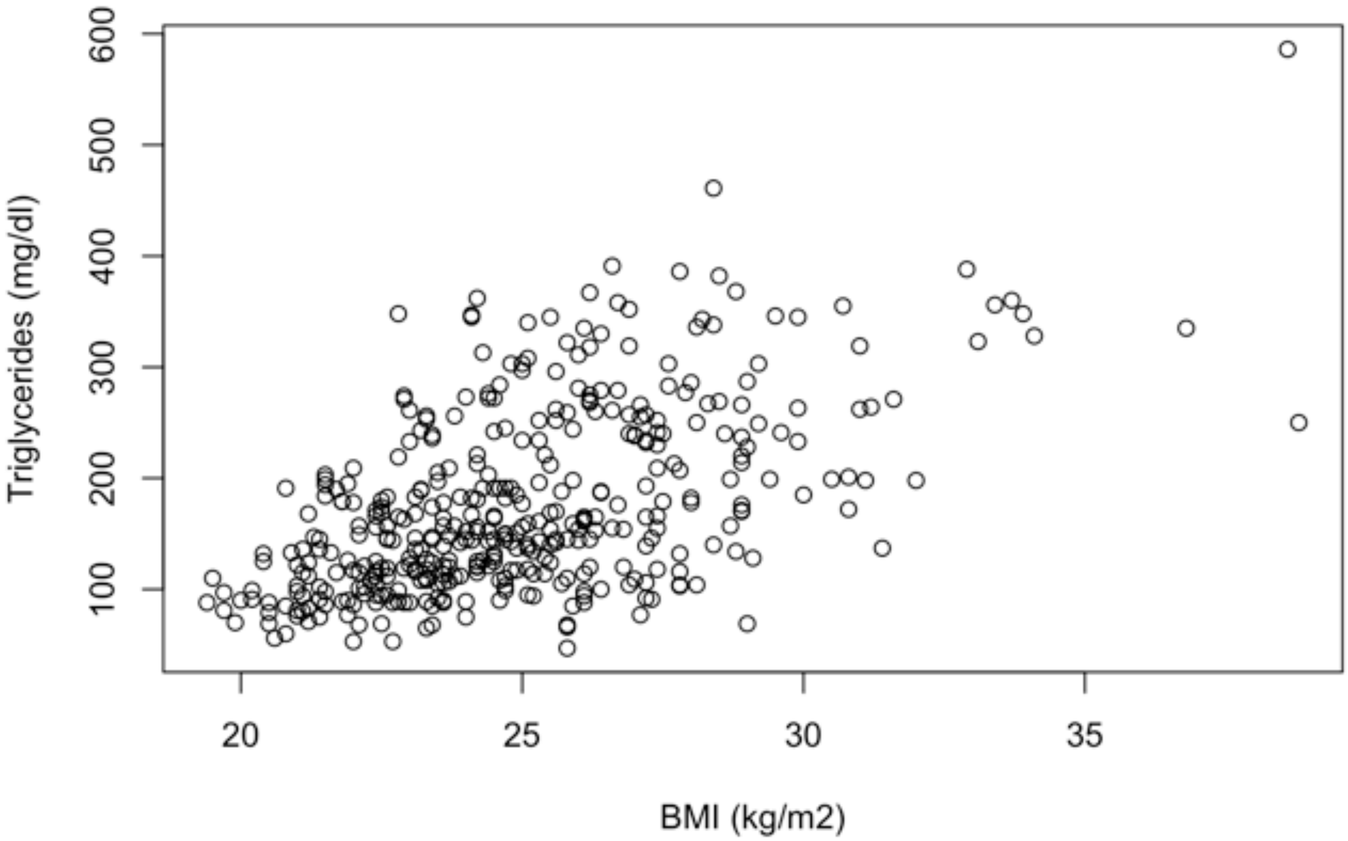
- For this lab you will need the *gee* package.
- If you have not already, install the package first. You will then need to load the library each time you execute your R script.

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

Exercises

1. Based on the 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)
```

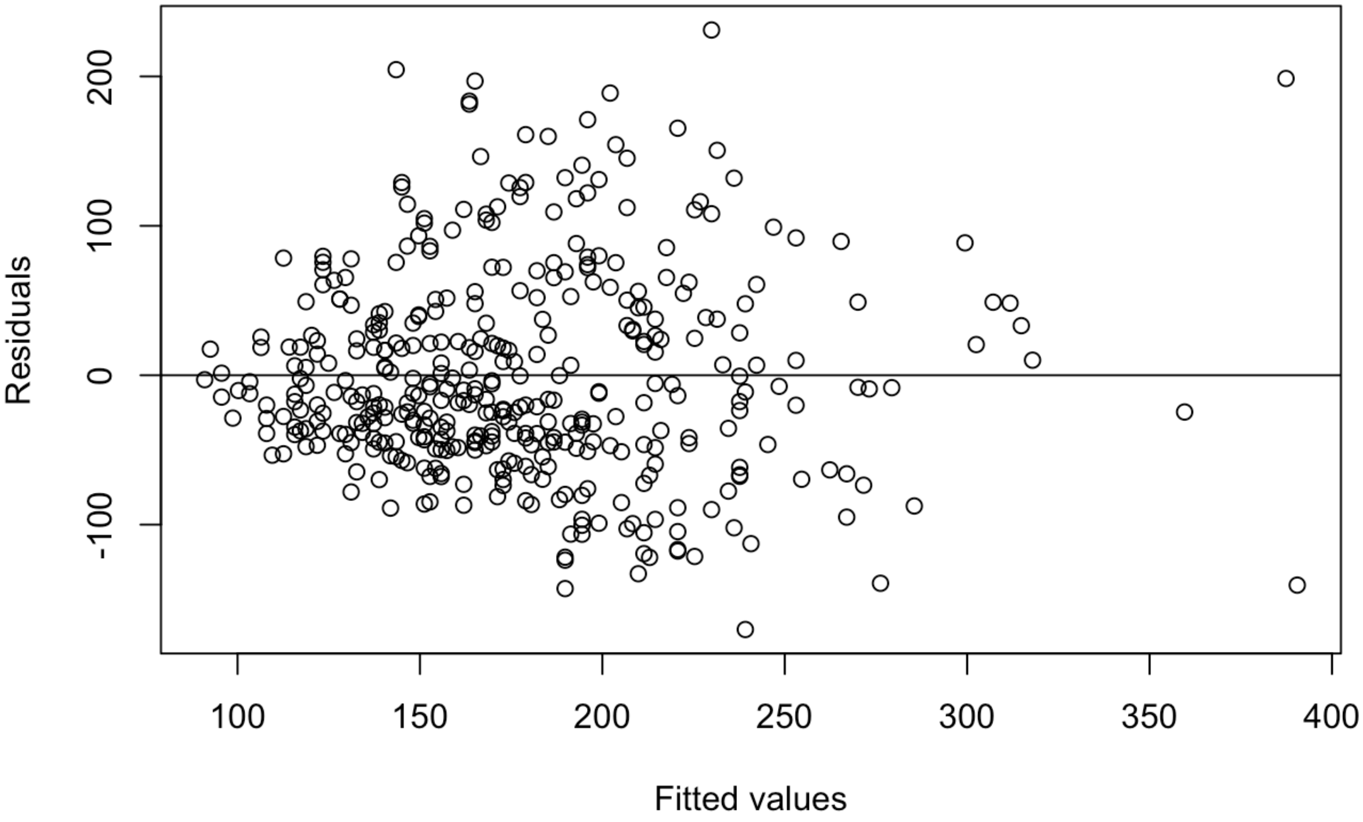
```
##
## Call:
## lm(formula = TG ~ BMI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -170.19  -45.10  -12.89   39.60  231.08
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -208.50      28.95  -7.203 2.97e-12 ***
## BMI            15.44       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)
```

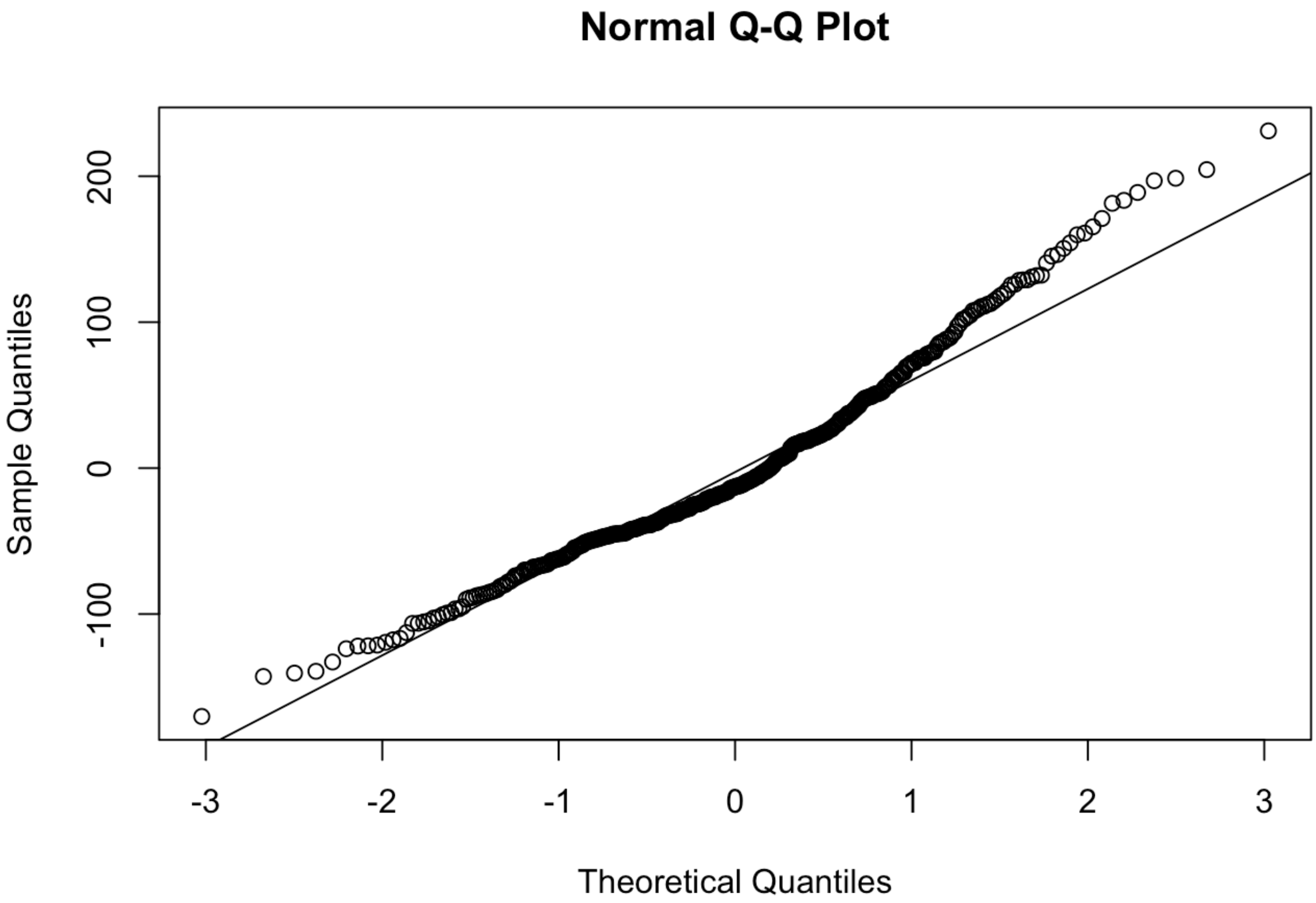
```
## 
## Call:
## lm(formula = TG[bmi37] ~ BMI[bmi37])
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -169.07  -44.87  -13.22   39.45  232.05 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -202.707     30.084  -6.738 5.68e-11 ***
## BMI[bmi37]    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
```

2. 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
```

3. 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 us 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
```

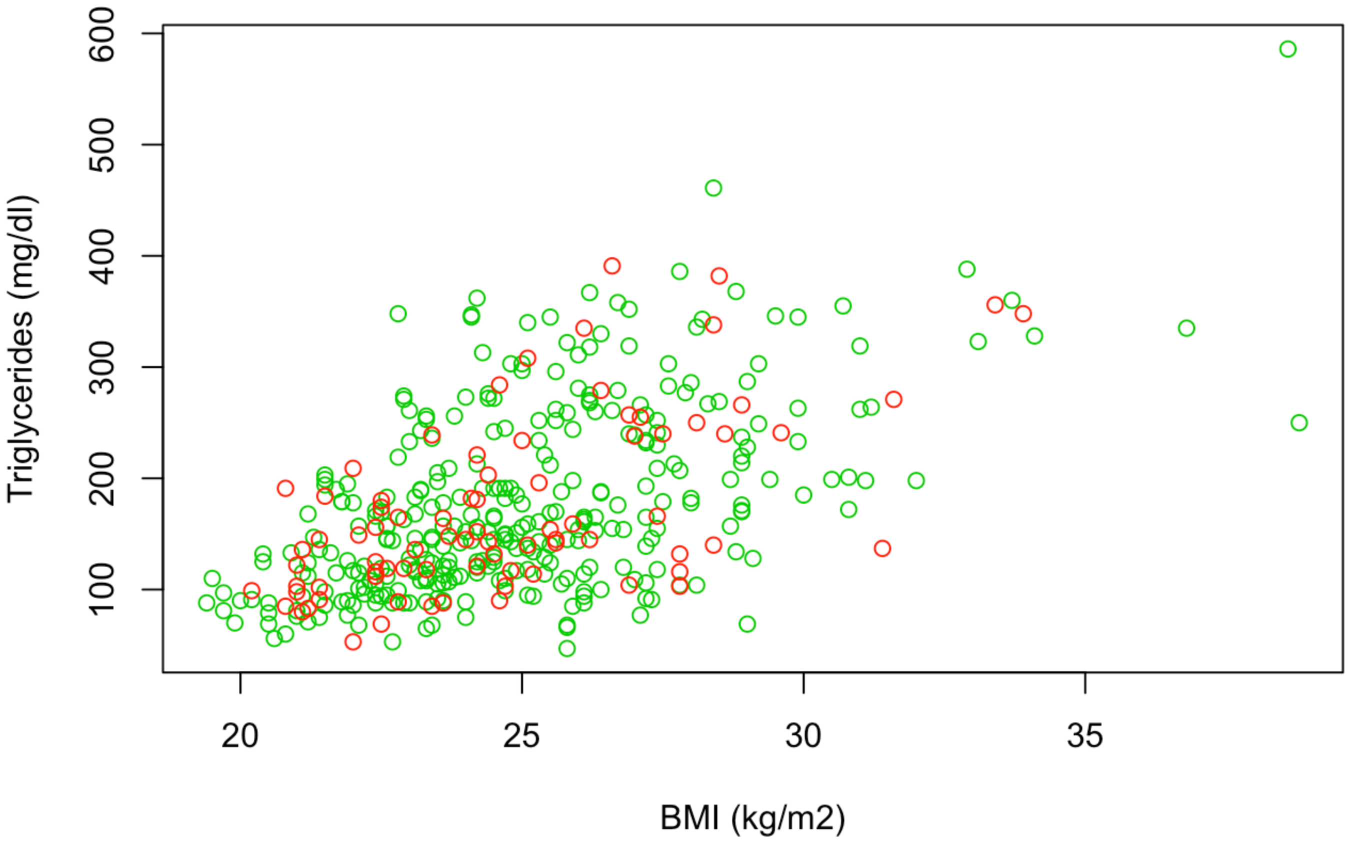
```
# binary variable indicating presence of APOE4
APOE4 = ifelse(APOE %in% c(3,5,6), 1, 0)

## 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
```

4. 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 the model with interaction 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
```

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.