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-Dat
  a-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?

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?

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.

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.

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.