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?

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.

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?

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?

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?

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?

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.

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.

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.

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?

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?

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?

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.

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.

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?

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?

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?