Lab 1: Introduction to R and Simple Linear Regression

The data set SISG-Data-cholesterol.csv, available in the class Github directory (https://github.com/rhubb/SISG2017), contains the following variables:

Field Descriptions

`ID`: Subject ID
`sex`: Sex: 0 = male, 1 = female
`age`: Age in years
`chol`: Serum total cholesterol, mg/dl
`BMI`: Body-mass index, kg/m^2
`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

The goal of the regression labs will be to explore relationships among the variables in this data set using the statistical methods presented in the class sessions. The objective of this first lab will be

- Become familiar with R and RStudio
- Begin to explore the cholesterol dataset.
- Use graphical and descriptive methods to investigate the association between triglycerides and BMI
- Use simple linear regression to investigate the association between triglycerides and BMI

1. Open RStudio.

2. Create a new script file to record your R code. Open a script file by clicking on File -> New File -> R Script.

3. Load the cholesterol data set.

4. Familiarize yourself with the variables in this dataset.

5. Compute the sample mean, median and standard deviation of triglycerides.

6. Create a boxplot and histogram for triglycerides.

7. BMI is used to define overweight and obesity: overweight: $25 \leq BMI < 30$ kg/m^2 and obese: $BMI \geq 30$kg/m^2. Create a variable called "ibmi" that takes the value 1 if $BMI \geq 25$ and 0 if $BMI<25$.

8. Compute summary measures of triglycerides for the two groups of subjects defined by "ibmi".

9. Create boxplots for triglycerides separately for the two groups of subjects defined by "ibmi". Does there appear to be an association between BMI and triglycerides? Conduct a test of the null hypothesis that mean triglycerides do not differ between those with $BMI \geq 25$ and $BMI <25$.

10. Plot a scatterplot of triglycerides vs BMI. Based on this plot does there appear to be an association between BMI and triglycerides? What can you additionally say about the relationship between these variables that was not possible using the boxplot?

11. Use regression to investigate the association between triglycerides and BMI. What do the linear regression model results tell us about the association?

12. Compute the predicted value and its 95% confidence interval 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?

13. Check your script file. Make sure that all important commands that you have used and any output you want to save are included in here.

R code for Lab 1:
https://raw.githubusercontent.com/rhubb/SISG2017/master/Rcode/lab1_script.R

Solutions for Lab 1:
https://github.com/rhubb/SISG2017/raw/master/labs/SISG_Module_4_Lab_1_Solutions.pdf

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 the association between triglycerides and BMI modified by APOE e4 allele?

1) Load the `gee` package.

2) Construct again a scatterplot of triglycerides versus BMI. Are there any points that you suspect might have a large influence on the regression estimates?

3) Use regression to investigate the association between triglycerides and BMI with and without the observations with BMI > 37. Do the points with BMI>37 appear to affect your results? If so, how?

4) Use residuals analysis (using all data) to check the linear regression model assumptions. Create a scatterplot of residuals vs fitted values and a quantile-quantile plot of residuals and consider deletion diagnostics. Do any modeling assumptions appear to be violated? How do model results change if you use robust standard errors?

5) Investigate the association between triglycerides and BMI after log transforming triglycerides. Does this appear to correct violations of modeling assumptions? How does this impact the TG confidence and prediction intervals for BMI=23?

6) Summarize the variable APOE. Create a new binary variable indicating presence of the APOE e4 allele (APOE = 3, 5, or 6).

7) 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?

8) Fit a linear regression model that investigates whether the association between triglycerides and BMI is modified by the APOE4 allele. Is there an association between APOE e4 and triglycerides? Is there evidence of effect modification? Make sure that you can interpret the regression coefficients from the model with interaction.

R code for Lab 2:
https://raw.githubusercontent.com/rhubb/SISG2017/master/Rcode/lab2_script.R

Solutions for Lab 2:
https://github.com/rhubb/SISG2017/raw/master/labs/SISG_Module_4_Lab_2_Solutions.pdf

Lab 3: One-Way and Two-Way ANOVA

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

- Is rs4775401 associated with cholesterol levels?
- Is APOE genotype associated with cholesterol levels?
- Are rs174548 and APOE associated with cholesterol levels?
- Does the effect of APOE on cholesterol levels depend on rs174548?

1. Load packages `multcomp` and `gee`

2. Perform a descriptive analysis to investigate the scientific questions of interest using numeric and graphical methods.

3. Compare the mean cholesterol levels between genotype groups defined by rs4775401.
    a. Perform the one-way ANOVA using the regression approach.
    b. Compare the above results with those obtained when
        i. allowing for unequal variances
        ii. using robust standard errors
        iii. using a nonparametric test
    c. Is there evidence that mean cholesterol levels between genotype groups are different? If so, perform all pairwise multiple comparisons using Bonferroni's adjustment. Try out different adjustment methods.
    d. Interpret your results

4. Repeat the steps described in problem 4 to compare the mean cholesterol levels between genotype groups defined by APOE.

5. Obtain a cross-tabulation of the groups defined by rs174548 and APOE.

6. Perform a descriptive analysis to investigate the scientific questions of interest using numeric and graphical methods.

7. Fit a two-way ANOVA model with an interaction between rs174548 and APOE. Test the interaction. What do you conclude?

8. Fit a two-way ANOVA model without the interaction between rs174548 and APOE. Test the main effects of rs174548 and APOE. What do you conclude?

R code for Lab 3:
https://raw.githubusercontent.com/rhubb/SISG2017/master/Rcode/lab3_script.R

Solutions for Lab 3:
https://github.com/rhubb/SISG2017/raw/master/labs/SISG_Module_4_Lab_3_Solutions.pdf

Lab 4: ANCOVA and Logistic Regression

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

- Controlling for age, is APOE associated with cholesterol levels?
- Does age modify the association between APOE and cholesterol levels?
- Is hypertension associated with rs174548?
- Is hypertension associated with triglycerides?
- Is hypertension associated with rs174548 after adjusting for triglyceride levels?

1. Perform a descriptive analysis to investigate the relationships between age, APOE, and cholesterol using numeric and graphical methods.

2. Fit an ANCOVA model with an interaction between APOE and age. Test the interaction. What do you conclude?

3. Fit an ANCOVA model without an interaction between APOE and age. Compare the results with the one-way ANOVA model that compares mean cholesterol levels among genotypes defined by APOE. What can you say about the role of age? [Is it an effect modifier? Or is it a confounder? Or is it a precision variable?]

4. Using the ANCOVA model that includes main effects for APOE and age, what is the age-adjusted mean cholesterol level for each APOE genotype?

5. Is there an association between rs174548 and hypertension? Analyze this relationship using descriptive statistics, a chi-squared test and logistic regression. What additional information does logistic regression provide that the chi-squared test does not?

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

7. Revise your logistic regression model to include both triglycerides and rs174548. What does this model tell you about the association between rs174548 and hypertension? What role does triglycerides play in this analysis?

R code for Lab 4:
https://raw.githubusercontent.com/rhubb/SISG2017/master/Rcode/lab4_script.R

Solutions for Lab 4:
https://github.com/rhubb/SISG2017/raw/master/labs/SISG_Module_4_Lab_4_Solutions.pdf