Introduction to Multiple Regression

Chapter 7, Lab 1

OpenIntro Biostatistics

Topics

- Adjusting for a potential confounder
- Fitting and interpreting a model

In most practical settings, more than one explanatory variable is likely to be associated with a response. Multiple linear regression is an extension of simple linear regression that allows for more than one predictor variable in a linear model. As with simple linear regression, the response variable must be numerical, but the predictor variables can be either numerical or categorical.

The statistical model estimating the linear relationship between a response variable y and predictors $x_1, x_2, \ldots, x_p$ is based on

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p.$$

There are several applications of multiple regression. One of the most common applications in a clinical setting is estimating an association between a response variable and primary predictor of interest while adjusting for possible confounding variables.

This lab introduces the multiple regression model by examining the possible association between cognitive function and the use of statins after adjusting for a potential confounder.

The material in this lab corresponds to Sections 7.1 and 7.2 of *OpenIntro Biostatistics*.

Background information

Statins are a class of drug widely used to lower cholesterol. Research suggests that adults with elevated low density lipoprotein (LDL) cholesterol may be at risk for adverse cardiovascular events. A set of guidelines released in 2013 recommended statin therapy in individuals who are at high risk of adverse cardiovascular events, including individuals with Type II diabetes and moderately high LDL and non-diabetic individuals with atherosclerotic cardiovascular disease and high LDL. If these guidelines were to be followed, almost half of Americans ages 40 to 75 and nearly all men over 60 would be prescribed a statin.

However, some physicians have raised the question of whether treatment with a statin might be associated with an increased risk of cognitive decline.

The goal of this lab is to examine the association between cognitive decline and statin use, after adjusting for a potential confounder.

This lab uses data from the Prevention of REnal and Vascular END-stage Disease (PREVEND) study.[1] Clinical and demographic data for 4,095 individuals are stored in the `prevend` dataset in the `oibiostat` package.

[1] These data were introduced in Chapter 6, Lab 1 (Examining Scatterplots).

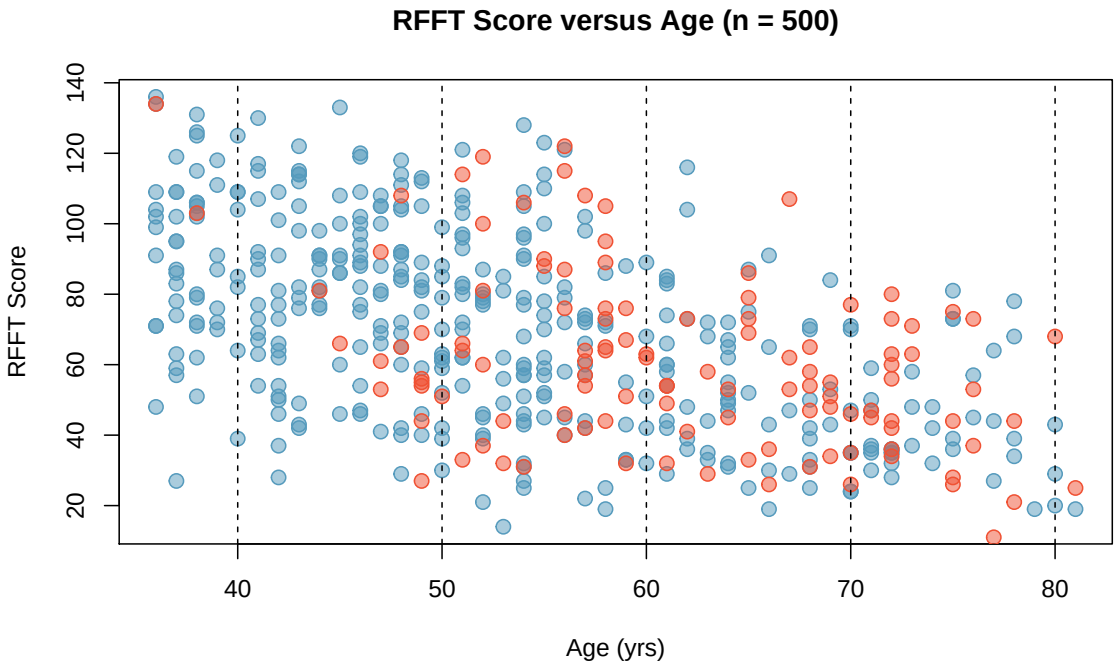
Adjusting for a confounder

Recall that the Unit 6 labs explored the association between cognitive function and age. Cognitive function in the PREVEND study was measured with the Ruff Figural Fluency Test (RFFT). Scores on the RFFT range from 0 to 175 points, where higher scores are indicative of better cognitive function. An analysis of the relationship between age and RFFT score showed evidence of a negative association between cognitive function and age; older individuals tend to have lower mean RFFT score than younger individuals.

The questions in this lab use data from a random sample of $n = 500$ individuals from the `prevend` dataset; the sample is stored as `prevend.samp` in the `oibiostat` package.

1. In the following scatterplot, statin users are represented with red points, while participants not using statins are shown as blue points.

   Examine the scatterplot and describe what you see regarding the relationship between RFFT score, age, and statin use.



2. Explore the relationship between RFFT score and statin use with the data in `prevend.samp`.

   a) Statin use is coded as an integer vector, where `0` represents a non-user and `1` represents a user. Convert the variable `Statin` into a factor variable, with levels `NonUser` and `User`.

   b) Create a plot showing the association between RFFT score and statin use. Describe what you see.

   c) Fit a simple regression model and interpret the slope coefficient.

   d) Discuss whether the model from part c) is sufficient for understanding whether statin use is associated with decreased cognitive ability.

3. Age is a potential confounder for the relationship between statin use and cognitive function. If older participants tend to use statins, and higher age is associated with lower cognitive ability, perhaps the observed negative association between cognitive ability and statin use is primarily driven by age.

    a) Subset the participants in `prevend.samp` by age to create three age cohorts:

        - `youngest`: individuals with age < 50 years
        - `younger`: individuals with age ≥ 50 years and < 60 years
        - `older`: individuals with age ≥ 60 years

    b) For each age cohort, create a plot showing the association between RFFT score and statin use. Compare these plots to each other and to the plot from Question 2, part b). Does the nature of the association between RFFT score and statin use seem to differ depending on age?

Fitting and interpreting a model

4. Fit a multiple regression model for predicting RFFT score from statin use and age.

    a) Write the equation of the linear model.

    b) Interpret the slope coefficient for statin use. Compare the coefficient to the one from the simple regression model between RFFT score and statin use.

    c) Interpret the slope coefficient for age.

    d) Make predictions.

        i. How does the predicted mean RFFT score for a 65-year-old individual using statins compare to that of an individual of the same age who is not using statins?

        ii. How does the predicted mean RFFT score for a 50-year-old individual compare to that of a 60-year-old individual, if they both use statins?

        iii. How does the predicted mean RFFT score for a 70-year-old individual who uses statins compare to that of a 50-year-old individual who does not use statins?

    e) As in simple linear regression, inferences can be made about the slope parameters estimated by the model slope coefficients.[2] Based on the multiple regression model, is there a statistically significant association between RFFT score and statin use?

    f) In a clinical setting, the interpretive focus lies on reporting the nature of the association between the primary predictor and the response, while specifying which potential confounders have been adjusted for. Briefly respond to a clinician who is concerned about a possible association between statin use and decreased cognitive function, based on the analyses conducted in this lab.

    g) Can the results of this study be used to conclude that as one ages, one's cognitive function (as measured by RFFT) declines?

[2] Inference in multiple regression will be introduced formally in Chapter 7, Lab 3.