Introduction to Multiple Regression

Chapter 7, Lab 1: Solutions

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.

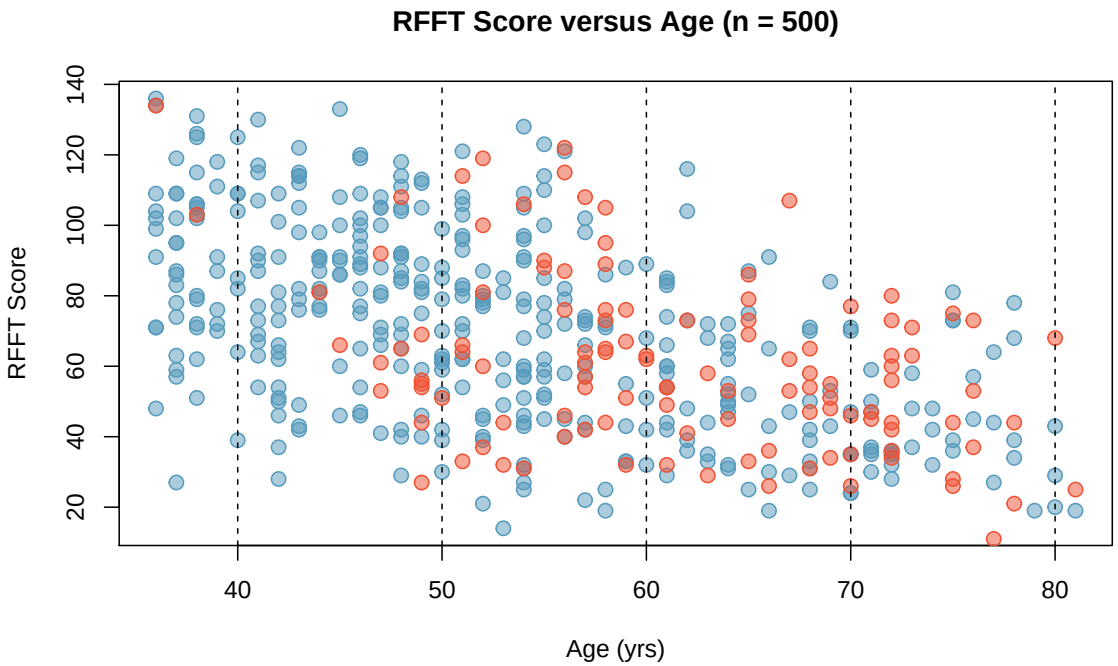
1. In the following scatterplot, statin users are represented with red points, while participants not using statins are shown as blue points.

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

   It is clear that age and statin use are associated, with statin use becoming more common as age increases; the red points are more prevalent on the right side of the plot than the left.

   It is also clear that age is associated with lower RFFT scores; ignoring the colors, the point cloud drifts down and to the right.

   A closer inspection of the plot suggests that for ages in relatively small ranges (e.g., ages 50-60 years), statin use may not be strongly associated with RFFT score—there are approximately as many red dots with low RFFT scores as with high RFFT scores in a given age range. In other words, for participants with similar ages, statin use may not be associated with RFFT.



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

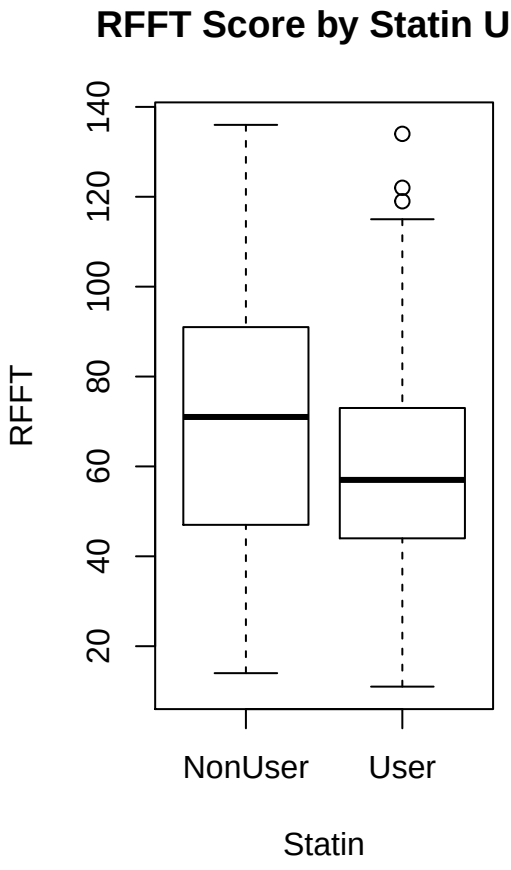
   a) Statin use is coded as an integer vector, where `0` represents a non-user and `1` represents a user. Convert the variable `Statin` into a factor variable, with levels `NonUser` and `User`.

```
#convert Statin to a factor
prevend.samp$Statin <- factor(prevend.samp$Statin, levels = c(0, 1),
                              labels = c("NonUser", "User"))
```

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

   Median RFFT score is higher in statin non-users than statin-users, by about 10 points. RFFT score is more variable in non-users than users; the IQR for non-users spans about 50 - 90 points, while the IQR for users spans about 50 - 70 points.

```
#create a plot
plot(RFFT ~ Statin, data = prevend.samp,
 main = "RFFT Score by Statin Use")
```



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

   The slope coefficient is -10.05; statin users have a mean RFFT score that is 10 points lower than that of non-users.

```r
#fit a model and print the model coefficients
lm(RFFT ~ Statin, data = prevend.samp)$coef
```

```
## (Intercept)  StatinUser
##    70.71429   -10.05342
```

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

The model is not sufficient for understanding whether statin use is associated with decreased cognitive ability, because it does not account for age as a potential confounder. The observed association may simply reflect the underlying association between age and statin use (statin users tend to be older) and between age and RFFT score (older individuals tend to have lower RFFT scores).

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

```r
#create subsets
youngest = prevend.samp[prevend.samp$Age < 50, ]
younger = prevend.samp[prevend.samp$Age >= 50 & prevend.samp$Age < 60, ]
older = prevend.samp[prevend.samp$Age >= 60, ]
```

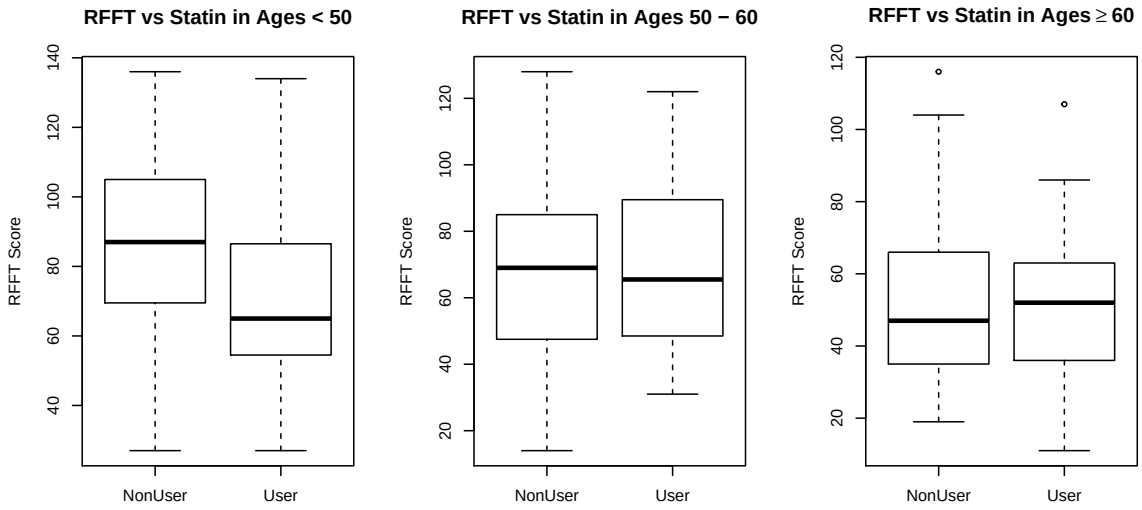
b) For each age cohort, create a plot showing the association between RFFT score and statin use. Compare these plots to each other and to the plot from Question 2, part b). Does the nature of the association between RFFT score and statin use seem to differ depending on age?

Yes, the nature of the association between RFFT score and statin use seem to differ depending on age. In the `younger` and `older` age cohorts, median RFFT score between statin users and non-users seem fairly similar. In the `older` cohort, median RFFT is higher in the statin users, while in the `younger` cohort, median RFFT is higher in the non-users.

The plot for the `youngest` group looks most similar to the plot from part b), which showed a larger difference between median RFFT score. This more pronounced difference is a consequence of the small number of statin users under age 50; there are only 15 statin users to 171 non-users in the `youngest` group. From the initial scatterplot, about a third of statin users under age 50 have RFFT scores above 70. It is likely that if more observations on statin users were available in this age range, the difference in median RFFT score would look more like that observed in the other two age cohorts.[2]

[2] Not shown in these solutions, but this is indeed the case if `prevend` rather than `prevend.samp` is subsetted.

```r
#create plots
par(mfrow = c(1, 3))
boxplot(RFFT ~ Statin, data = youngest,
    ylab = "RFFT Score", main = "RFFT vs Statin in Ages < 50")
boxplot(RFFT ~ Statin, data = younger,
    ylab = "RFFT Score", main = "RFFT vs Statin in Ages 50 - 60")
boxplot(RFFT ~ Statin, data = older,
    ylab = "RFFT Score", main = expression(bold("RFFT vs Statin in Ages">= "60")))
```




```r
#examine group size
table(youngest$Statin)
```

```
##
## NonUser    User
##     171      15
```

```r
table(younger$Statin)
```

```
##
## NonUser    User
##     108      40
```

```r
table(older$Statin)
```

```
##
## NonUser    User
##     106      60
```

Fitting and interpreting a model

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

```
#fit the multiple regression model
lm(RFFT ~ Statin + Age, data = prevend.samp)
```

```
##
## Call:
## lm(formula = RFFT ~ Statin + Age, data = prevend.samp)
##
## Coefficients:
## (Intercept)   StatinUser          Age
##    137.8822       0.8509      -1.2710
```

   a) Write the equation of the linear model.

$$\widehat{RFFT} = 137.88 + 0.85(StatinUser) - 1.27(Age)$$

   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.

   The use of statins is associated with a mean RFFT score that is higher by 0.85 points, when age is held constant; i.e., comparing two individuals of the same age. In the simple regression model, the slope coefficient was -10.05, indicating a negative association between RFFT score and statin use.

   c) Interpret the slope coefficient for age.

   A one year increase in age is associated with a decrease in mean RFFT score of 1.27 points, when statin use is held constant; i.e., in terms of comparing two individuals who are either both statin users or both statin non-users.

   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?

      If two individuals are the same age, then the difference in their predicted mean RFFT score is directly given by the value of the slope coefficient for statin use. According to the model, the statin user has a predicted mean RFFT score higher than the non-user by 0.85 points.

      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?

      For two individuals that have the same statin use status (either both users or both non-users), the difference in their predicted mean RFFT score is indicated by the slope coefficient for age. A one year increase in age is associated with a lower mean RFFT score by 1.27 points; thus, a ten-year difference in age is associated with a $10(1.27) = 12.7$ difference in predicted RFFT score, with the predicted mean score of the 60-year-old being 12.7 points lower than the 50-year-old's.

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?

The predicted mean RFFT score for a 70-year-old individual who uses statins is lower (49.76) than that for a 50-year-old individual who does not use statins (75.18).

```
model.RFFTvsStatinAge = lm(RFFT ~ Statin + Age, data = prevend.samp)

#use predict( )
predict(model.RFFTvsStatinAge, newdata = data.frame(Statin = "User", Age = 70))
```

```
##        1
## 49.76347
```

```
predict(model.RFFTvsStatinAge, newdata = data.frame(Statin = "NonUser", Age = 50))
```

```
##        1
## 74.33249
```

```
#alternatively, use r as a calculator...
b0 = 137.8822
b1 = 0.8509
b2 = -1.271

x1 = c(1, 0)
x2 = c(70, 50)

y = b0 + (b1*x1) + (b2*x2); y
```

```
## [1] 49.7631 74.3322
```

e) As in simple linear regression, inferences can be made about the slope parameters estimated by the model slope coefficients.[3] Based on the multiple regression model, is there a statistically significant association between RFFT score and statin use?

The p-value associated with the slope coefficient for statin use in the model adjusting for age is 0.743. The large p-value indicates the data are consistent with the null hypothesis of no association between RFFT score and statin use after adjusting for age.

```
summary(model.RFFTvsStatinAge)
```

```
##
## Call:
## lm(formula = RFFT ~ Statin + Age, data = prevend.samp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.855 -16.860  -1.178  15.730  58.751
##
## Coefficients:
```

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

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 137.8822     5.1221  26.919   <2e-16 ***
## StatinUser    0.8509     2.5957   0.328    0.743
## Age          -1.2710     0.0943 -13.478   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.21 on 497 degrees of freedom
## Multiple R-squared:  0.2852, Adjusted R-squared:  0.2823
## F-statistic: 99.13 on 2 and 497 DF,  p-value: < 2.2e-16
```

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.

Although the use of statins appeared to be associated with lower RFFT score when no adjustment was made for possible confounders, statin use is not significantly associated with RFFT score in a model that adjusts for age. After adjusting for age, the estimated difference in mean RFFT score between statin users and non-users is 0.85 points; there is a 74% chance of observing such a difference if there is no difference between mean RFFT score in the population of statin users and non-users.

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

No, it would be inappropriate to conclude from this study that as one ages, one's cognitive function declines. The study was a cross-sectional study, in which all variables were measured at a single point in time. While the results of the study support the conclusion that older individuals tend to have lower RFFT scores, they cannot be used to conclude that scores decline with age because there were no repeated measurements of RFFT as individuals aged.

For example, there may be a 'cohort effect'. It may be the case that younger participants have more post-secondary school education or better health practices, and that this cohort effect is driving the observed association between RFFT score and age.