Interaction

Chapter 7, Lab 4: Solutions

OpenIntro Biostatistics

This lab introduces the concept of a statistical interaction, specifically in the case of an interaction between a categorical predictor and a numerical predictor.

The material in this lab corresponds to Section 7.7 in *OpenIntro Biostatistics*.

Introduction

An important implicit assumption in the multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

is that when one of the predictor variables x_j changes by 1 unit and the values of the other variables remain constant, the predicted response changes by β_j, regardless of the values of the other variables.

A statistical **interaction** occurs when this is assumption is not true, such that the effect of one explanatory variable x_j with the response depends on the particular value(s) of one or more other explanatory variables.

This course specifically examines interaction in a two-variable setting, where one of the predictors is categorical and the other is numerical. Interaction effects between two numerical variables and between more than two variables can be complicated to interpret. A more complete treatment of interaction is best left to a specialized regression course.

Interaction is best understood through considering a specific example. This lab introduces the concept of interaction using a sample from the NHANES data.[1]

[1] The NHANES data were introduced in Chapter 1, Lab 1 (Introduction to Data). The data can be treated as a simple random sample from the American population.

Interaction with NHANES

The NHANES collected information about various demographic and health variables for each participant, including total cholesterol level in mmol/L (`TotChol`), age in years (`Age`), and diabetes status (`Diabetes`, coded as either `No` or `Yes`).

The following set of questions step through exploring the association of total cholesterol with age and diabetes status, using `nhanes.samp.adult.500`, a sample of $n = 500$ adults from the larger NHANES dataset.

1. Load `nhanes.samp.adult.500` from the `oibiostat` package. Fit a linear model for predicting total cholesterol level from age and diabetes status.

```
#load the data
library(oibiostat)
data("nhanes.samp.adult.500")

#fit the model
model.TotCholvsAgeDiabetes = lm(TotChol ~ Age + Diabetes,
                                data = nhanes.samp.adult.500)
coef(model.TotCholvsAgeDiabetes)
```

```
##   (Intercept)           Age   DiabetesYes
##  4.800011340  0.007491805 -0.317665963
```

   a) Write the model equation in terms of the variable names.

$$\widehat{TotChol} = 4.80 + 0.0075(Age) - 0.32(DiabetesYes)$$

   b) Interpret the coefficients of the model, including the intercept.

   The coefficient for age indicates that a one year increase in age is associated with an increase in predicted mean total cholesterol of 0.0075 mmol/L, assuming diabetes status is held constant. The coefficient for diabetes indicates that diabetics have an average total cholesterol that is 0.32 mmol/L than non-diabetic individuals, assuming age is held constant.

   The intercept for the model represents an individual of age 0 years who is not diabetic; the intercept does not have interpretive value because the model specifically uses data for adults (minimum age of 21 years).

   c) Make predictions.

      i. How does the predicted mean total cholesterol for a 60-year-old individual compare to that of a 50-year-old individual, if both are diabetic?

      If both individuals have diabetes, then the change in predicted total cholesterol level can be determined directly from the coefficient for age. An increase in one year of age is associated with a 0.0075 increase in predicted mean total cholesterol; thus, an increase in ten years is associated with (10)(0.0075) = 0.075 mmol/L increase in predicted total cholesterol.

ii. How does the predicted mean total cholesterol for a 60-year-old individual compare to that of a 50-year-old individual, if both are not diabetic?

The calculation from part i. does not differ if both individuals are non-diabetic. According to the model, the relationship between age and total cholesterol remains the same regardless of diabetes status as long as diabetes status is held constant.

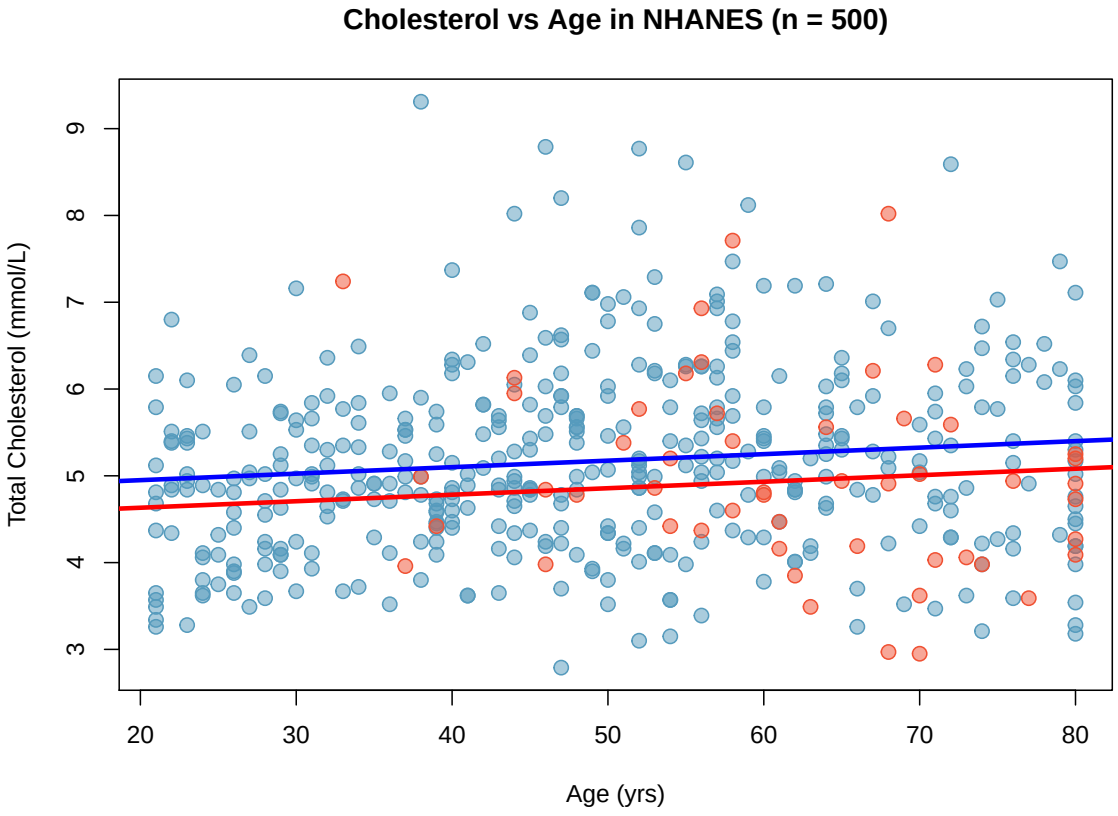
d) Based on the model equation from part a), write two separate model equations: one for diabetic individuals and one for non-diabetic individuals.

For non-diabetics (`DiabetesYes = 0`), the model equation is $\widehat{TotChol} = 4.80 + 0.0075(Age) - 0.32(0) = 4.80 + 0.0075(Age)$.

For diabetics (`DiabetesYes = 1`), the model equation is $\widehat{TotChol} = 4.80 + 0.0075(Age) - 0.32(1) = 4.48 + 0.0075(Age)$.

e) Make a scatterplot of total cholesterol versus age and plot the two models from part d). Describe what you see; compare the models.

The least-squares lines are parallel, with the same slope and different intercepts. While predicted mean total cholesterol is higher overall in non-diabetics (as indicated by the larger intercept), the rate of change in predicted mean total cholesterol by age is equal for diabetics and non-diabetics.

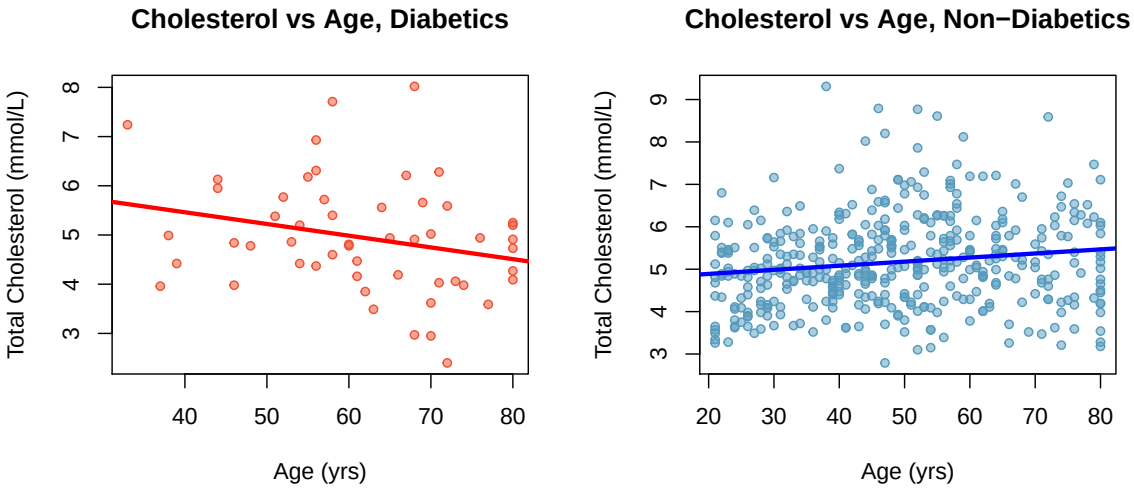


A model that assumes the relationship between total cholesterol and age does not depend on diabetes status might be overly simple and potentially misleading.

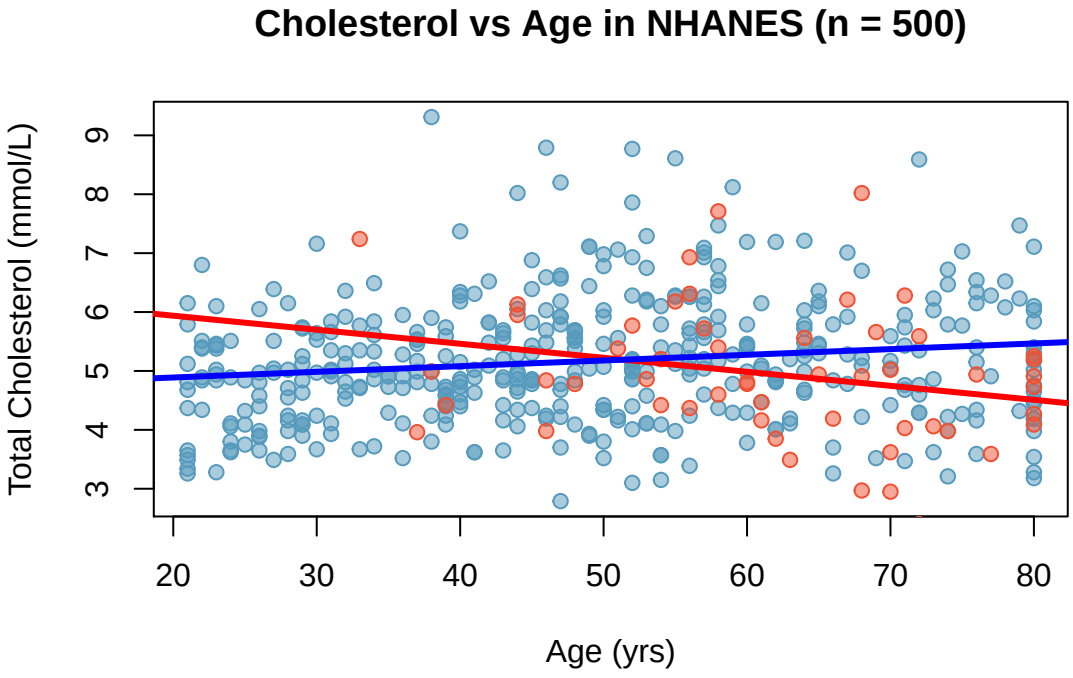
2. To explore this visually, fit two separate models for the relationship between total and cholesterol and age.

   a) Fit a model predicting total cholesterol from age in diabetic individuals. Create a plot specific to diabetic individuals and plot the least-squares line.

   b) Fit a model predicting total cholesterol from age in non-diabetic individuals. Create a plot specific to non-diabetic individuals and plot the least-squares line.



   c) Run the code in the template to create a single plot with data from all 500 individuals and the least-squares lines from parts a) and b).



d) Describe what you see in the plots. Does the association between total cholesterol level and age seem different between diabetics and non-diabetics?

Yes, the association between total cholesterol and age seems different between diabetics and non-diabetics. The lines fit separately are not parallel, and in fact, have slopes with different signs. The plots suggest that among non-diabetics, age is positively associated with total cholesterol. Among diabetics, however, age is negatively associated with total cholesterol.

With the addition of another parameter (commonly referred to as an interaction term), a linear regression model can be extended to allow the relationship of one explanatory variable with the response to vary based on the values of other variables in the model. Consider the model

$$E(TotChol) = \beta_0 + \beta_1(Age) + \beta_2(Diabetes) + \beta_3(Diabetes \times Age).$$

The term $(Diabetes \times Age)$ is the interaction term between diabetes status and age, and β_3 is the coefficient of the interaction term. Diabetes status and age, the main independent variables in the model, are sometimes referred to as "main effect variables" in the context of a model with an interaction term.

3. Use the code provided in the template to fit a model for predicting total cholesterol that includes age, diabetes, and the interaction term between age and diabetes status.

```
#fit model with interaction term
model.interact = lm(TotChol ~ Age*Diabetes, data = nhanes.samp.adult.500)
coef(model.interact)
```

```
##     (Intercept)             Age     DiabetesYes Age:DiabetesYes
##     4.695702513     0.009638183     1.718704342    -0.033451562
```

a) Write prediction equations.

i. Write the overall model equation.

$$\widehat{TotChol} = 4.70 + 0.0096(Age) + 1.72(DiabetesYes) - 0.033(Age \times DiabetesYes)$$

ii. Write the model equation for diabetics.

For diabetics (`DiabetesYes = 1`), the model equation is

$$\begin{aligned}\widehat{TotChol} &= 4.70 + 0.0096(Age) + 1.72(DiabetesYes) - 0.033(Age \times DiabetesYes) \\ &= 4.70 + 0.0096(Age) + 1.72(1) - 0.034(Age \times 1) \\ &= 4.70 + 1.72 + (0.0096 - 0.034)(Age) \\ &= 6.42 - 0.024(Age)\end{aligned}$$

iii. Write the model equation for non-diabetics.

For non-diabetics (`DiabeticsYes = 0`), the model equation is

$$\begin{aligned}\widehat{TotChol} &= 4.70 + 0.0096(Age) + 1.72(DiabetesYes) - 0.033(Age \times DiabetesYes) \\ &= 4.70 + 0.0096(Age) + 1.72(0) - 0.034(Age \times 0) \\ &= 4.70 + (0.0096)(Age)\end{aligned}$$

b) Interpret the model coefficients (of the overall equation), including the interaction term.

The intercept represents the predicted mean total cholesterol for a non-diabetic of age 0 years; as before, this term does not have a meaningful interpretation in context of the data.

The coefficient for age indicates that for non-diabetics, an increase in age of one year is associated with an increase in mean predicted total cholesterol of 0.0096 mmol/L.

The coefficient for diabetes represents the change in intercept value between the line of best fit for non-diabetics versus diabetics.

The interaction term indicates the difference in the slope coefficient of *Age* between diabetics and non-diabetics. For diabetics, an increase in 1 year of age is associated with a lower predicted mean total cholesterol of 0.034 mmol/L (0.0096 - 0.034 = -0.024). The difference is large enough that although total cholesterol and age is positively associated in non-diabetics, they are negatively associated in diabetics.

c) Make predictions.

i. How does the predicted mean total cholesterol for a 60-year-old individual compare to that of a 50-year-old individual, if both are diabetic?

If both individuals are diabetic, then the difference in predicted mean total cholesterol level is 10(0.024) = 0.24 mmol/L, with the older individual having the lower predicted mean total cholesterol.

ii. How does the predicted mean total cholesterol for a 60-year-old individual compare to that of a 50-year-old individual, if both are not diabetic?

If both individuals are not diabetic, then the difference in predicted mean total cholesterol level is 10(0.0096) = 0.096 mmol/L, with the younger individual having the lower predicted mean total cholesterol.

iii. Compare the predictions made in parts i. and ii. to those made in Question 1 using the model without an interaction term. How does fitting an interaction term change the model?

For the model in Question 1, the predicted difference in mean cholesterol level between two individuals was the same regardless of whether the two individuals were both diabetic or both non-diabetic. Fitting an interaction term allows for the association between total cholesterol and age to be different between diabetics and non-diabetics.

d) Speculate as to what might explain a positive association between age and cholesterol for non-diabetics, but a negative association between age and cholesterol for diabetics.

Cholesterol levels tend to increase in age, and keeping cholesterol levels in a healthy range eventually becomes a concern for all individuals. However, diabetics are particularly at risk for having elevated cholesterol levels, and so may be prescribed cholesterol-lowering medication more so than non-diabetic individuals. The observed interaction between age and diabetes may be the result of more frequent cholesterol-lowering medication in diabetic individuals.

The estimated equations for non-diabetic and diabetic individuals from the model with the interaction term, fit to all individuals, show the same behavior as seen when two separate models were fit to diabetics and non-diabetics.

In practice, it is more efficient to model the data using a single model with an interaction term than working with subsets of the data. The subset approach shown at the beginning of this lab was used to demonstrate the logic behind interaction.

4. Using a single model allows for a formal test of whether there is significant evidence of an interaction.

```
#print summary of the model
summary(model.interact)
```

```
##
## Call:
## lm(formula = TotChol ~ Age * Diabetes, data = nhanes.samp.adult.500)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3587 -0.7448 -0.0845  0.6307  4.2480
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.695703   0.159691  29.405  < 2e-16 ***
## Age              0.009638   0.003108   3.101  0.00205 **
## DiabetesYes      1.718704   0.763905   2.250  0.02492 *
## Age:DiabetesYes -0.033452   0.012272  -2.726  0.00665 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.061 on 469 degrees of freedom
##   (27 observations deleted due to missingness)
## Multiple R-squared:  0.03229,    Adjusted R-squared:  0.0261
## F-statistic: 5.216 on 3 and 469 DF,  p-value: 0.001498
```

a) Is there evidence that the interaction term between age and diabetes status is statistically significant at $\alpha = 0.05$?

The p-value of the interaction coefficient is 0.0067, which is less than $\alpha = 0.05$. There is statistically significant evidence of an interaction at the 0.05 level.

b) Based on adjusted R^2, is the model with the interaction term an improvement over the model with only the main effect variables?

The adjusted R^2 of the initial model with only age and diabetes status is 0.013. The model with the interaction has adjusted R^2 of 0.026; the higher value indicates the model is an improvement.

Note that the R^2 of the model with the interaction is only 0.032; the model explains very little of the observed variability in total cholesterol. In the setting of a large study to examine factors associated with cholesterol level in adults, a model like this one is typically a starting point for building a more refined model.

```
summary(model.TotCholvsAgeDiabetes)$adj.r.squared
```

```
## [1] 0.01277467
```

Interaction with PREVEND

The following set of questions step through taking a closer look at the association of RFFT score with age and statin with `prevend.samp`, a sample of $n = 500$ individuals from the PREVEND data.

5. Run the code in the template to load `prevend.samp` from the `oibiostat` package and convert `Statin` to a factor variable. Fit a model for predicting RFFT score from age, statin use, and the interaction term between age and statin use.

```
#load the data
library(oibiostat)
data("prevend.samp")

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

#fit the model
model.RFFT.interact = lm(RFFT ~ Age*Statin, data = prevend.samp)
coef(model.RFFT.interact)
```

```
##     (Intercept)             Age     StatinUser Age:StatinUser
##     140.2031114      -1.3149119     -13.9720216       0.2474466
```

a) Write prediction equations.

i. Write the overall model equation.

$$\widehat{RFFT} = 140.20 - 1.31(Age) - 13.97(StatinUser) + 0.25(Age \times StatinUser)$$

ii. Write the model equation for statin users.

$$\begin{aligned}\widehat{RFFT} &= 140.20 - 1.31(Age) - 13.97(StatinUser) + 0.25(Age \times StatinUser)\\ &= 140.20 - 1.31(Age) - 13.97(1) + 0.25(Age \times 1)\\ &= 140.20 - 13.97 + (-1.31 + 0.25)(Age)\\ &= 126.23 - 1.06(Age)\end{aligned}$$

iii. Write the model equation for statin non-users.

$$\begin{aligned}\widehat{RFFT} &= 140.20 - 1.31(Age) - 13.97(StatinUser) + 0.25(Age \times StatinUser)\\ &= 140.20 - 1.31(Age) - 13.97(0) + 0.25(Age \times 0)\\ &= 140.20 - 1.31(Age)\end{aligned}$$

b) Interpret the model coefficients.

The intercept represents the predicted mean RFFT score for an individual at age 0 years who is not using statins; this does not have a meaningful interpretation in context of the data since the RFFT cannot be administered to a newborn.

The coefficient for age indicates that for individuals not using statins, an increase in age of 1 year is associated with a predicted mean RFFT score that is 1.31 points lower.

The coefficient for statin use represents the change in intercept value between the line of best fit for statin non-users versus statin users.

The interaction term indicates the difference in the slope coefficient of *Age* between statin users and non-users. For statin users, a increase in age of 1 year is associated with a predicted mean RFFT score is associated with a predicted mean RFFT score that is 1.06 points lower. The slope for statin users is 0.25 points higher than the slope for non-users.

c) Make predictions.

i. How does the predicted mean RFFT score for a 55-year-old individual compare to that of a 65-year-old individual, if both are using statins?

If both individuals are using statins, the difference in predicted mean RFFT score between individuals that are 10 years apart in age is $(10)(1.06) = 10.06$ points, with the older individual having the lower predicted mean RFFT score.

ii. How does the predicted mean RFFT score for a 55-year-old individual compare to that of a 65-year-old individual, if both are not using statins?

If both individuals are not using statins, the difference in predicted mean RFFT score between individuals that are 10 years apart in age is $(10)(1.31) = 13.1$ points, with the older individual having the lower predicted mean RFFT score.

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

The predicted mean RFFT score for a 70-year-old individual using statins is 51.51 points. The predicted mean RFFT score for a 50-year-old individual not using statins is 74.46 points.

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

```
##        1
## 51.50851
```

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

```
##        1
## 74.45752
```

```
#alternatively, use r as a calculator...
b0 = coef(model.RFFT.interact)[1]
b1 = coef(model.RFFT.interact)[2]
b2 = coef(model.RFFT.interact)[3]
b3 = coef(model.RFFT.interact)[4]

#70 year old statin user
age = 70; user = 1

y = b0 + b1*age + b2*user + b3*(age*user)
y
```

```
## (Intercept)
##    51.50851
```

```
#50 year old statin non-user
age = 50; user = 0

y = b0 + b1*age + b2*user + b3*(age*user)
y
```

```
## (Intercept)
##    74.45752
```

d) Is there evidence of a statistically significant interaction between age and statin use?

No, there is not evidence of a statistically significant interaction between age and statin use. The p-value associated with the interaction term is 0.317.

```
summary(model.RFFT.interact)
```

```
##
## Call:
## lm(formula = RFFT ~ Age * Statin, data = prevend.samp)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.551 -16.963  -1.179  15.764  58.802
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    140.2031     5.6209  24.943   <2e-16 ***
## Age             -1.3149     0.1040 -12.646   <2e-16 ***
## StatinUser     -13.9720    15.0113  -0.931    0.352
## Age:StatinUser   0.2474     0.2468   1.003    0.317
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.21 on 496 degrees of freedom
## Multiple R-squared:  0.2866, Adjusted R-squared:  0.2823
## F-statistic: 66.42 on 3 and 496 DF,  p-value: < 2.2e-16
```