Interaction

Chapter 7, Lab 4

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.

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

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

   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?

      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?

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

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

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?

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.

   a) Write prediction equations.

      i. Write the overall model equation.

      ii. Write the model equation for diabetics.

      iii. Write the model equation for non-diabetics.

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

   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?

      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?

      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?

   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.

The estimated equations for non-diabetic and diabetic individuals from the model with the interaction term, fit to all individuals, show the same qualitative behavior as seen when two separate models were fit to diabetics and non-diabetics. Note, however, that the values of the estimated coefficients are not the same between the two approaches.

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.

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

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

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.

   a) Write prediction equations.

      i. Write the overall model equation.

      ii. Write the model equation for statin users.

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

   b) Interpret the model coefficients.

   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?

      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?

      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?

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