Exploration 4.3: Predicting blood glucose

Learning Goals:

- Adjust the relationship between two quantitative variables based on a categorical variable
- Create indicator variables in order to include binary categorical variables in the regression model
- Evaluate the validity of the regression model

Blood glucose levels are the primary indicator of diabetes, with higher levels associated with individuals having diabetes or pre-diabetes. The Framingham Heart study is a well-known ongoing study of cardiovascular health that that started in 1948 and is now on its third generation of participants.

The dataset **GlucoseData** contains information on the blood glucose levels (measured in milligrams per deciliter or mg/dL), body mass index or BMI (measured in kg/m^2), whether the respondent has been previously diagnosed with Type II diabetes, and whether person is taking cholesterol medication regularly, for a large sample from this third generation of participants (surveyed between 2005 and 2008). The average age of these participants is about 66 years, with standard deviation SD = 9 years.

Note: There were 17 participants who had missing values for at least one of the four variables in this data set: Blood glucose level, BMI, Diabetic status, and whether taking cholesterol medications. These participants were removed and the following questions use the cleaned data.

STEP 1: Ask a research question.

1. What do you predict about the association between *bloodglucose* and *BMI*?

STEP 2: Design a study and collect data.

2. Do the data arise from an observational study or a randomized experiment? How are you deciding?

Open the **GlucoseData** datafile.

3. What is the sample size in the dataset?

STEP 3: Explore the data.

Open the Multiple Variables applet. Copy and paste the data into the Sample data window. Drag the *BloodGlucose* variable into the Response box. Check the **Show descriptive** box.

4. Describe the shape, center, and variation in blood glucose levels. Are there any outliers or other unusual characteristics?

Obesity and hence higher BMI is believed to be associated with higher blood glucose levels. Let's see whether *BMI* explains variation in *BloodGlucose*.

Drag *BMI* into the **Explanatory** box. Check the **Show Equation** box.

5. Describe the association between *BloodGlucose* and *BMI*. Is this the association direction that you expected? Are there any unusual observations? If so, what do you suggest?

Perform any necessary data cleaning steps. You might wish to use a spreadsheet program to do this and consider outliers and other potential issues with your data, and then reload into the applet.

6. What data cleaning, if any, did you do? Why?

7. What percentage of variation in *BloodGlucose* is explained by *BMI*? (*Hint:* You can check the **R-squared** box.) Also record the *SSModel* and *SSError* values. Also, report and interpret the *SE residuals* value (*Hint*: You can check the **Regression SE** box or the **Show residuals** box to find *SE residuals*). What does the SE residuals value tell you about the expected accuracy of your prediction of the person's blood glucose based on their BMI?

8. Provide an interpretation of the slope coefficient of *BMI*, in context.

9. Provide an interpretation of the intercept of the model, in context. Is this an extrapolation? Why or why not?

10. Why would it not make sense to fit a "separate means model" for each BMI value?

Can we improve the model?

Some cholesterol drugs fall under the category of statins, a type of medication that has been found to be associated with an increased risk of Type II diabetes (see study about statins and Type II diabetes). We expect that whether a person is taking cholesterol medications may also explain variation in blood glucose levels. But, if we already know the person's *BMI*, is knowing whether the person takes cholesterol medications also helpful in predicting their blood glucose levels?

11. Make a prediction: If we have our statistical model predict the *BloodGlucose* of a person from both the person's *BMI* and *cholesterolmeds*, do you think the amount of explained variation will increase over a model predicting blood glucose levels from BMI alone? Explain.

First, verify that *cholesterolmeds* is related to *BloodGlucose*.

Remove *BMI* from the Explanatory box and move *cholesterolmeds* into the Explanatory box. Check the **Show Equation** box. Note, the stack on the right are the "yes" values for cholesterol medication.

12. What do you learn from the graph and means? How do you interpret the coefficient of *cholesterolmeds* in the prediction equation? What percentage of variation in *BloodGlucose* is explained by *cholesterolmeds*? Does this seem like a substantial amount?

But if we already have the person's *BMI*, how worthwhile is it to also know whether the person takes cholesterol medication?

Now drag the *BMI* variable to the top of the Explanatory box (drop it above *cholesterolmeds*). Your scatterplot of *BloodGlucose vs. cholesterolmeds* should now be color coded by BMI (darker colors (e.g., red, black) indicating smaller BMI). (*Hint*: You can check the Show 2-variable graphs box.)

13. Based on this graph, who tends to have larger BMI, those who take cholesterol meds or those who don't? How can you tell?

14. If we adjust the *BloodGlucose* values based on the *BMI* values, which observations should move up and which should move down? Predict how this will change the (adjusted) group means for the cholesterol groups.

Check the **Adjust values** box to check your prediction. The scatterplot now shows the relationship between *BloodGlucose* and *cholesterolmeds* after adjusting for *BMI*. From the pie chart, report the *SSprev* value and the *SScholesterolmeds* value.

15. How have the two group means changed? How has this impacted the difference between them?

16. What is the adjusted slope coefficient of *cholesterolmeds*? Is it larger or smaller than the unadjusted coefficient?

17. What does *SSprev* represent? (*Hint*: Where have you seen this value before?)

18. How much additional variation in blood glucose levels is explained by adding *cholesterolmeds* to a model that already includes *BMI*? (*Hint: SScholesterolmeds,* R^2) Is the (adjusted) association between *cholesterolmeds* and *BloodGlucose* stronger or weaker if we first adjust for *BMI*? Does this make sense in context? Explain.

We can also explore the association between *BloodGlucose* and *BMI* after adjusting for *cholesterolmeds*.

Uncheck the **Show Equation** box. In the applet. Change the order of the *BMI* and *cholesterolmeds* variables in the Explanatory box so that *cholesterolmeds* is listed first and *BMI* is listed second. You should see a scatterplot of *BloodGlucose* and *BMI* color coded for yes and no on whether they take cholesterol medication.

19. Explain how/whether the association between blood glucose levels and BMI differs for those who take cholesterol medication and those who don't.

20. If we adjust the blood glucose levels for taking cholesterol medication, will the red observations move up or down? The blue observations? Why?

Then check the **Adjust *y* values** to check your prediction.

With a quantitative variable, we will go one step further and also adjust the *BMI* values for *cholesterolmeds*. This adjustment will isolate the information in the *BMI* variable that is not related to *cholesterolmeds*.

21. Suppose we adjust the BMIs for taking cholesterol medication, how do you expect the red and blue observations to move?

Check the **Adjust *x* values** box. This creates an ***added variable plot***.

Definition: An ***added variable plot*** shows the association between a response variable and the explanatory variable that is last entered into the model (usually the primary explanatory of interest), after adjusting for any other variables already in the model (often explanatory variables of less interest). The graph shows the residuals between the response and the explanatory variables of less interest versus the residuals between the explanatory variable of interest and explanatory variables of less interest. Added variables plots are most useful when you want to visually explore whether a new explanatory variable explains additional variation in the response variable, after adjusting for other variable(s).

Check the **Show Equation** box.

22. Was your prediction correct? What does the applet report for the slope between the adjusted *BMIs* and these adjusted blood glucose levels? Is this slope larger or smaller than the unadjusted slope?

Using Categorical Variables in the Statistical Model

The least squares estimation method can again be used to estimate the slope coefficients for both explanatory variables simultaneously.

In the applet, check the **Statistical model** box.

23. Provide interpretations of the intercept and of the slope coefficients, keeping in mind where you have previously seen these values.

24. Based on this output, write out two separate prediction equations, one for those who take cholesterol medication and one for those who don't.

Check your work by unchecking both the **Adjust *y*** and **Adjust *x*** boxes and check the **Separate Lines** box. (The scatterplot should show the two lines and their equations.)

Another way to represent this model, especially when we have a binary categorical variable, uses an ***indicator variable***.

Definition: An ***indicator variable*** converts a binary categorical variable to a (0,1) variable where 1 indicates that the observation is in that category and 0 indicates that the observation is not in that category.

For example, we can define an indicator variable for taking cholesterol medication by

$$cmed\_ind = \begin{cases} 0, & if\ no \\ 1, & if\ yes \end{cases}$$

It doesn't matter which outcome is coded as zero and which as 1, but the one coded as zero is called the *reference group*.

Use the pull-down menu to change from **Effect coding** to **Indicator coding.**

25. Use the output to construct a single prediction equation using the *BMI* and the *cmed_ind* indicator variable. Did the coefficient of *BMI* change?

26. Provide an interpretation of the intercept of this model in context. (*Hint*: What does it mean for "all explanatory terms to be equal to zero"?)

27. Determine the prediction equation for those who do NOT take cholesterol medication (substitute *cmed_ind* = 0 into the above equation and simplify):

28. Determine the prediction equation for those who take cholesterol medication (substitute *cmed_ind* = 1 into the above equation and simplify):

29. Provide an interpretation of the *cmed_ind* coefficient from the single prediction equation in context. (*Hint*: How are the two equations from the two previous questions related? How does this correspond to the scatterplot with the two lines?)

Notice that the resulting two regression equations for those who take cholesterol medication and those who don't that were derived using indicator coding are identical to the equations obtained from the effects coding. Software packages will generally choose which of these two coding options to use. Most of the time there is an option to ask for the other. You can also use the indicator variable (as a quantitative variable) directly.

Key Idea: Statistical software packages typically automatically use either effects coding or indicator variable coding to change categorical variables into quantitative variables in a multiple regression model. Check the documentation to know which approach the software package you choose is using.

Key Idea: Two ways to represent a binary variable include

- ***Indicator coding***: the binary variable is coded (0, 1) and the slope coefficient represents the difference in group means (non-reference – reference category);
- ***Effect coding***: the binary variable is coded (-1, 1) and the slope coefficient represents the difference in the group means from an "overall average" (least squares mean); this coefficient is also half the size of the difference in group means.

STEP 4: Draw inferences beyond the data.

Make sure that you still are predicting blood glucose by *cholesterolmeds* and BMI (with BMI listed second). Check **Show residuals**.

30. Assess the fit of this model using the residual plots. Be sure to address all four validity conditions.

From the pie chart, report the (adjusted) *SSBMI* value.

31. How has the *SSBMI* changed from the one-variable model using *BMI* alone to the two-variable model using *BMI* and *cholesterolmeds*? Why?

32. What is the value of *SSprev* in the pie chart? What proportion of variation is explained by *SSprev*? Explain what *SSprev* represents.

Key Idea: The adjusted sums of squares report the decrease in the *SSError* when the last variable in the list is added to the model. The ANOVA table reports only the adjusted sum of squares for each variable, the difference between the sum of these values and *SSModel* represents the covariation among the explanatory variables.

Check the **ANVOA Table** box.

33. Confirm the adjusted SS values. What is *SSModel*? What does this represent? How does it compare to the sum of the SS values? What does *SSModel – SScholesterolmeds – SSBMI* represent?

34. What is *SSError* for the two-variable model? Is it smaller than the *SSError* for the one-variable models? By a lot? What does this tell you?

35. Is the overall model significant? Are the individual variables, after adjusting for the other, significant? Explain any discrepancy in your answers.

36. Toggle between Effect and Indicator coding. Does this change the ANOVA table?

STEP 5: Formulate conclusions.

37. Write a paragraph summarizing the results of your analysis. Be sure to address the following:

- Include a possible Sources of Variation diagram. Be sure to include the covariation between *cholesterolmeds* and *BMI*.
- Is there a statistically significant linear association between the *BloodGlucose* and the *BMI* of a subject, after adjusting for *cholesterolmeds*? What is your evidence?
- If so, what is the nature of that association?
- Which variable, *cholesterolmeds* or *BMI*, explains more variation in *BloodGlucose*?
- Explain how you would use that information to predict someone blood glucose, as if to a health care provider. Also provide a measure of how accurate you expect this prediction to be.
- To what population would you generalize these conclusions?

STEP 6: Look back and ahead.

38. What do you suggest next in analyzing these data?