Exploration 6.3: Predicting breast cancer

Learning Goals:

- Utilize a logistic regression model using multiple categorical and/or quantitative explanatory variables

Background: Researchers Patricio et al. (*BMC Cancer*, 2018) looked at predicting the incidence of breast cancer from patient characteristics such as age and body mass index (BMI), and from routine blood analyses looking at health markers such as levels of glucose and resistin.

You may wish to read the article. It can be found online at <https://bmccancer.biomedcentral.com/articles/10.1186/s12885-017-3877-1>.

Sixty-four women newly diagnosed with breast cancer were recruited from the Gynecology Department of the University Hospital Centre of Coimbra (Portugal) between 2009 and 2013, and 52 healthy female volunteers were selected and enrolled in the study as controls. None of the patients had prior treatment for cancer, and all were free from any other infections or diseases at the time of enrollment in the study.

During the first consultation, the same research physician collected data on variables such as age, weight, and height for all participants. BMI was calculated from the weight and height measurements. Blood samples were collected from all 116 women at the same time of the day after an overnight fasting, and all samples (2500 g) were first centrifuged at 4 °C and then stored at −80 °C before being tested for levels of glucose (mmol/L) among other things. The same equipment and protocols were used each time. The data are available in the file ***BreastCancerData***.

Note: To be mindful of how the authors of the study described the participants, we have described the participants as women, but we recognize that this language is not as inclusive as we would like.

STEP 1: Ask a research question.
Can we predict whether or not a woman has breast cancer from her age, BMI, and level of glucose and resistin in her blood?

STEP 2: Design a study and collect data.

1. Identify the observational/experimental units.

2. Was this an experiment or an observational study? How are you deciding?

3. Did the study use random sampling, random assignment, both, or neither? Why is this information relevant?

4. What is the response variable? Is it quantitative or categorical? If quantitative, are high values or low values desirable for the response? If categorical, what are the possible categories?

STEP 3: Explore the data.

Fasting blood glucose levels of 100 mg/dL or higher are believed to be indicative of prediabetes or diabetes. The following two-way table shows the women cross classified by whether they had normal (below 100 mg/dl) or high (100 mg/dl or more) fasting blood glucose levels, and whether or not they had breast cancer.

	Normal blood glucose level	**High blood glucose level**	**Total**
Diagnosed with breast cancer	34	30	64
Healthy (control)	44	8	52
Total	78	38	116

5. Calculate and report the odds ratio of having breast cancer, comparing women with high fasting blood glucose levels to those with normal blood glucose levels. Also, interpret this odds ratio in the context of the study.

6. Use software to fit a logistic regression model to predict the outcome for *diseasestatus* from *glucoselevel*. Confirm that the model produces the same odds ratio of breast cancer as the sample odds ratio. Is the association statistically significant?

7. Now, investigate whether there is evidence of an association between *glucoselevel* and either *BMI* or *age*. Anticipate how the relationship of the glucose level and having breast cancer might change if we add age and BMI to the model.

STEP 4: Draw inferences beyond the data.

8. Use statistical software to fit a logistic regression model predicting disease status from *glucoselevel, BMI*, and *age*. Fill in the "Regression table" and the "ANOVA/Deviance" table.

Term	**Coeff**	**SE**	***z*-value**	χ^2	**p-value**
Intercept					
Glucose.level					
BMI					
Age					

Source	**DF**	**Deviance**	χ^2 **p-value**
Model			
Glucose.level			

BMI			
Age			

The model deviance compares this current model to the model with only an intercept. The deviances of the individual rows assume that term was the last one added to the model, that is, after adjusting for the other terms. When the sample size is large, the deviance statistic follows a chi-square distribution with df = number of parameters. Most software packages will report individual regression coefficients and their corresponding standard errors. Dividing the coefficients by the standard error gives a *z*-statistic, which, when the sample size is large, will follow a standard normal distribution. Squaring a *z*-statistic corresponds to a chi-square statistic with df = 1. These chi-square statistics are not quite equivalent to the ones from the deviance table (here is where our analogy to *t* and *F*-statistics breaks down), but, when the sample size is large, should be very similar.

9. Determine and interpret in context a 95% confidence interval for the *adjusted* odds ratio of breast cancer from whether or not the respondent has high or normal fasting blood glucose levels from this model (which also includes BMI and age).

10. Did the odds ratio (as reported in the previous answer) change much from before we adjusted for age and BMI? Would you consider this a practically significant (large) change?

11. This model ignores the potential for interactions between pairs of explanatory variables. Add the interactions between *age* and *glucose level* and between *BMI* and *glucose level*. Are either of the interactions statistically significant additions to the model? If so, write a detailed interpretation in context of the significance interaction(s).

Measuring Model Performance

Although statistical significance is important, we might also ask about overall model performance. With a quantitative response, we considered R^2, the proportion of variation in the response variable explained by the model. Recall that R^2 = correlation($\hat{y}$, y)^2. In this data set, the y values are 0 or 1, so does a correlation coefficient make sense?

Save the predicted probabilities for the model with no interactions and calculate the correlation coefficient between these values and the observed *disease.status* outcomes (as a 0/1 numeric variable).

12. What does this tell you about the model's performance?

Note: We cannot square this value and interpret it as the percentage of variation in *disease.status* explained by the model. The 0/1 variable also prevents these two variables from having a linear relationship. Researchers have proposed several other alternative measures of model performance, and we consider one more here that compares the actual success/failure outcome with the predicted outcome based on the model.

13. Use software to store the predicted probability of disease status for each individual in your data set based on the model with all three variables but no interactions. Produce a graph (e.g., histogram or dotplot) of these predicted probabilities. What is the range of these predicted values?

How do we convert these predicted probabilities to predicted outcomes for comparison? We could predict any observation with a predicted probability larger than 0.50 as a "predicted breast cancer" and any observation with a predicted probability less than 0.50 a "predicted healthy."

Use your cut-off value to predict the outcome for each person in the dataset. To do this, create a new variable in your data set that codes each person as either a (predicted) "success" or "failure" based on your cut-off value. Then, create a table of the observed outcomes vs. these predicted outcomes.

14. What percentage of the outcomes did you predict correctly? Based on this value, do you think the model is doing well?

	Predicted breast cancer	**Predicted healthy**	**Total**
Actual breast cancer			
Actual healthy			
Total			

Definition: The ***correct classification rate*** compares the predicted outcomes of the binary response variable to the actual outcomes and sees how many outcomes the model correctly predicts in the dataset.

Like R^2, there is no global agreement on how large the correct classification rate needs to be in order to be "doing well." However, it can also be useful for comparing models. For example, you could calculate the correct classification rate for the model that includes the interaction terms to see whether they help appreciably to improve the predictive performance of the model.

15. Another possible cut-off value is the overall proportion of successes. How does this value compare to the cutoff used in the previous two questions? How does the correct classification rate change, if at all?

STEP 5: Formulate conclusions.

16. Is there evidence that individuals with high blood glucose levels are more likely to develop breast cancer after adjusting for age and BMI? Can we draw a cause-and-effect conclusion? To whom do these findings apply?

STEP 6: Look back and ahead.

17. What cautions or concerns do you have about this study? What changes would you make if you were to conduct a similar study?