Lab 15: Logistic regression for wine quality

Today's forecast: sunny with a chance of wine

– American Proverb (credit: southernliving.com/culture/wine-quotes)

Logistic Regression

We are interested in fitting a model that can predict whether a wine is high quality or mediocre (in a subjective sense), using objective physical and chemical characteristics of the liquid. This work was originally presented in the paper "Modeling wine preferences by data mining from physicochemical properties" by Cortez et al (2009).

Part 1

1. *Why might winemakers want a physicochemical model of quality? What is one possible use of such a model? Is this use best characterized as description, prediction, inference to a population, or causal inference?
2. What was the response variable in the original study? What type of variable is this? What were the explanatory variables?
3. Sketch the head of the original red wine dataset (you do not need to include every explanatory variable). What is the unit of observation? What are the dimensions of the complete dataset?
4. Comment on the design of the study. Was random sampling involved? Was blinding involved? What is the population of interest? Was there a "treatment" variable that was manipulated? Is this an observational study or an experiment? Can we draw causal conclusions from the data?
5. The authors fit models for red and white wine separately. Why do you think they did this? How many observations did they have for each model?
6. *To fit our logistic regression we will dichotomize the response so that 6 or below is recorded as a "failure" (low quality wine) and 7 or above is a "success" (high quality wine). Pick an explanatory variable and sketch a plot of its (hypothetical) relationship with the dichotomized response.
7. Write out a logistic regression model that describes the probability of a success as a function of some of the explanatory variables.
8. Suppose we are given 5 wines along with predicted probabilities that each is high quality:

Wine	Probability high quality
1	.31
2	.19
3	.92
4	.45
5	.68

Which wines are high quality? What decision rule are you applying to make that choice?

9. *The sommelier at Chez Panisse (a fancy Berkeley restaurant) can sell a high quality wine for $50 and a low quality wine for only $30. As above, they can only access the predicted *probability* that a given

wine is high quality but they need to classify the wine in order to decide the price. If they sell a high quality wine as low quality, they lose $20 in potential sales. If they sell a low quality wine as high quality they will offend their patron and lose a lot of money (say >$100) in future business. Given this information, should they use the same decision rule as in 8 or should they use a higher/lower threshold when classifying a wine as high quality?

Part 2

Use the following code to load in the Wine data. We have excluded 50 of the wines from this dataset in order to use them later. The original red wine data is available from UC Irvine ("https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv").

```
#red wines
library(tidyverse)
wine_data <- read_csv("https://www.dropbox.com/s/tsbk5rznycces1y/train_wine_data.csv?dl=1")
```

10. What are the dimensions of the dataset? What is the unit of observation?
11. Create a new response variable by dichotomizing the original scores to 6 or below (low quality) and 7 or above (high quality). In this new column, a wine with 6 or below in `quality` should have a 0, while a wine with 7 or above should have a 1. What proportion of wines in the data are "high quality"?
12. Generate the plots you proposed in question 6. Describe the relationship. Does this variable seem to be a good predictor of wine quality?
13. *Fit a simple logistic regression of quality on sulphates. Interpret the coefficient and the p-value. Write out the equation for the logistic regression model with the estimated intercept and coefficient. What is the residual deviance? What is the AIC?
14. Fit a multiple logistic regression model including every explanatory variable. Interpret the coefficent for sulphates. Which coefficients have small p-values?What is the AIC? Does this model predict better or worse than the simple model?
15. *The below data contains 50 new red wines. Create a dichotomized response as you did for question 10, then predict the probability of being high quality using the full model. Round these probabilities (threshold the probabilities at 50%) to predict the dichotomized response. What proportion of these predictions are correct? How does this compare to the misclassification rate of the null model that always predicts the wine is poor quality?

```
test_wine_data <- read_csv("https://www.dropbox.com/s/mbq78m5f4guvw79/test_wine_data.csv?dl=1")
```

16. The original authors conclude with the following claim: *Furthermore, the relative importance of the inputs brought interesting insights regarding the impact of the analytical tests. Since some variables can be controlled in the production process this information can be used to improve the wine quality. For instance, alcohol concentration can be increased or decreased by monitoring the grape sugar concentration prior to the harvest.* Is this claim valid? Under what assumptions?