BANA 7042: Assignment 4 (25 points)

Multinomial and Ordinal Outcomes

Note: Please be sure to properly label all figures and include a caption for each!

Before answering the following questions, please do the following:

- Download the wine quality data from our course Canvas page (the file is called `winequality.csv`). They're also available from the UCI Machine Learning Repository here; please read the background information on the webpage carefully.
- Search for "wine quality" on YouTube and watch some videos to gain some basic knowledge about wine. Here's an example video.
- Read the paper "Modeling wine preferences by data mining from physicochemical properties" (2009); you can access the full article through the UC library or by downloading the PDF attached to this assignment. You don't have to understand how they did the analysis. Instead, try to understand the background of the project, the reasons they for the analyses, and what they want to achieve. Specifically, read the following sections carefully:
  - Abstract
  - Section 1: Introduction
  - Section 2.1: Wine data
  - Section 2.2: Data mining approach and evaluation
  - Section 4: Conclusions and implications

Questions

The `winequality.csv` data set includes data on both red and white wines; for this analysis, you'll only be focusing on the red wines. The response variable of interest is `quality`.

1) Subset the data to only include observations on red wines. Describe the distribution of the response. Based on the nature of the response variable, what type of regression model would you suggest starting with (and why)?

2) Explore the data using summary statistics and graphics. Do any of the variables appear to have an association with the overal quality score? If so, which? Describe the nature of these associations.

3) Construct a binary response according to the following rule

$$Y = \begin{cases} 1, & \texttt{quality} \geq 7 \\ 0, & \texttt{quality} < 7 \end{cases}$$

Fit a logistic regression to the data using all possible main effects (i.e., include each variable, but no interaction effects). Assess the performance of this model. Does the model seem *well-calibrated*? Discuss and provide a plot of a calibration curve.

4) Interpret the effect of the predictor `alcohol` on the odds that `quality >= 7`. Construct an effect plot visualizing the effect of `alcohol` on the probability that `quality >= 7` and describe the relationship. Does this plot look linear or nonlinear? If nonlinear, discuss how this is possible.

5) Discuss reasons why the modeling approach used in 3) is ill-advised for modeling these data.

6) Fit an ordinal regression model to the data using the original response (i.e., `quality`) using the `orm()` function from R package **rms**. Construct an effect plot for each predictor showing the effect on the predicted probability that `quality >= 7`. From these, try to determine the top three predictors solely in terms of their effect on the predicted probability that `quality >= 7`.

7) Consider a single observation x_0 (i.e., a single red wine) with the following characteristics:

```
fixed.acidity         7.3000
volatile.acidity      0.6500
citric.acid           0.0000
residual.sugar        1.2000
chlorides             0.0650
free.sulfur.dioxide  15.0000
total.sulfur.dioxide 21.0000
density               0.9946
pH                    3.3900
sulphates             0.4700
alcohol              10.0000
```

Based on your fitted model from 6), provide estimates for the following quantities:

- $\mathrm{P}(\texttt{quality} = 7|x_0)$
- $\mathrm{P}(\texttt{quality} \geq 7|x_0)$
- $\mathrm{P}(\texttt{quality} = 9|x_0)$
- $\mathrm{P}(\texttt{quality} \leq 9|x_0)$

8) You're asked to use your model from part 6) to provide predictions for the white wines included in the original sample. Discuss whether or not you think this is a reasonable request and why. What would you do in practice (e.g., what if this was your boss asking)?