BANA 7042: Assignment 2

Logistic regression (part I)

Part I: multiple choice (1 point each)

Please clearly circle (or highlight) your multiple choice answers

Question 1: True or false. In the LR model, it is assumed that there is a linear relationship between the predictor variables and the mean response (i.e., probability of success or $Y = 1$). If false, write the correct statement.

Question 2 Why is ordinary linear regression not appropriate for binary outcomes?

a. It only has one weight per feature
b. It only has one output value
c. The bias parameter skews the output value
d. Its predictions are not restricted to values between 0 and 1

Question 3 Large values of the log-likelihood statistic indicates that:

a. There is potential multicollinearity in the data
b. The associated model fits the data well
c. As the predictor variables increase in value, the likelihood of the outcome occurring decreases
d. The associated model is a poor fit of the data

Question 4 LR is used to predict the probability of a _____?

a. Binary independent variable
b. Binary dependent variable.
c. Continuous dependent variable.
d. Continuous independent variable.

Question 5 In a logistic regression, if the predicted logit is 0, what's the transformed probability?

a. 0

b. 1
c. 0.5
d. 0.05

Part II: short answer response (3 points each)

Question 6. For ease of notation, let $\eta = x^\top \beta = \beta_0 + \beta_1 x_1 + \beta_2 x_1 + \cdots + \beta_p x_p$ be the linear predictor. For the LR model, where

$$\text{logit}\,(p) = x^\top \beta,$$

show that

$$p = \frac{1}{1 + \exp\,(-\eta)}$$

Question 7: (2 pts.) Using historical data from past offers sent to customers, your company uses a logistic regression model to predict the likelihood that a customer will redeem a particular offer that was sent to them (e.g., a coupon for milk). The model is to be used on a new set of $n = 500$ customers from the same population that the original training sample came from. However, your stakeholder asks another question. In addition to predicting the probability that each of the 500 new customers will redeem the offer sent to them, you're asked to also provide an estimate of the total number of offers that will be redeeemd. Discuss how you would approach this problem. (No wrong answer here so don't Google it, I just want to hear your group's solution in detail.)

For the next few questions, you'll be using the blood donation data available from the UC Irvine Machine Learning Repository; you can download the data from here.

Question 8 Import the data into R and fit a simple logistic regression model using `Donated_Blood` as the binary response and `Recency` as the predictor. Interpret the coefficient of `Recency` (be sure to use full sentences and include the units). How would you interpret the estimated intercept in this model?

Question 9 Plot the fitted LR model from **Question 8** on the probability scale (i.e., `Recency` on the x-axis and the predicted probability of donating on the y-axis). The donation center asks you to estimate the `Recency` associated with a predicted likelihood of donating of $p = 0.5$. (In Pharmacology, this is referred to as the median effective dose.) Provide an estimate using your fitted model and describe how you arrived at this value.

Question 10 Fit an LR model using all four predictors and provide an interpretation for each coefficient estimate. Without "Googling" or "ChatGPT-ing", come up with a way of measuring the relative "importance" of each variable. (There's a million ways to do this, none of which are perfect, so please use your own intuition and expertise to devise a way to accomplish this.) Which variable seems most important in predicting `Donate_Blood`? Provide a table with the ranking of each variable. Describe how your group arrived at these rankings. What are some potential drawbacks of your approach?