STAT 608, Spring 2022 - Assignment 6
Solutions

1. For logistic regression with one predictor, we use the model

$$\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 x$$

   (a) Show that solving for the probability of success for a given value of the predictor, $\theta(x)$, gives

$$\theta(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

$$\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 x$$

$$\frac{\theta(x)}{1-\theta(x)} = \exp(\beta_0 + \beta_1 x)$$

$$\theta(x) = \exp(\beta_0 + \beta_1 x) - \theta(x)\exp(\beta_0 + \beta_1 x)$$

$$\theta(x) + \theta(x)\exp(\beta_0 + \beta_1 x) = \exp(\beta_0 + \beta_1 x)$$

$$\theta(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

   (b) and

$$\theta(x) = \frac{1}{1 + \exp(-\{\beta_0 + \beta_1 x\})}$$

$$\begin{aligned}
\theta(x) &= \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \\
&= \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \left[\frac{\exp(-(\beta_0 + \beta_1 x))}{\exp(-(\beta_0 + \beta_1 x))}\right] \\
&= \frac{1}{\exp(-(\beta_0 + \beta_1 x)) + 1} \\
&= \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x))}
\end{aligned}$$

2. On page 285 of the text, it says "When X is a dummy variable, it can be shown that the log odds are also a linear function of x." Suppose that X is a dummy variable, taking the value 1 with probability π_j, $j = 0, 1$, conditional on $Y = 0, 1$.

(a) Show that the log odds are a linear function of x.

FIRST, DEFINE THE BERNOULLI PROBABILITY $P(X|Y=j) = \pi_j^x(1-\pi_j)^{1-x}$. THEN NOTE THAT

$$\frac{\theta(x)}{1-\theta(x)} = \frac{P(Y=1|X)}{P(Y=0|X)} = \frac{P(Y=1, X=x)P(X=x)}{P(Y=0, X=x)P(X=x)} = \frac{P(Y=1)P(X=x|Y=1)}{P(Y=0)P(X=x|Y=0)}$$

FINALLY, WE HAVE

$$\begin{aligned}
\log\left(\frac{\theta(x)}{1-\theta(x)}\right) &= \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + \log\left(\frac{P(X=x|Y=1)}{P(X=x|Y=0)}\right) \\
&= \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + \log\left(\frac{\pi_1^x(1-\pi_1)^{1-x}}{\pi_0^x(1-\pi_0)^{1-x}}\right) \\
&= \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + x\log\left(\frac{\pi_1}{\pi_0}\right) + (1-x)\log\left(\frac{1-\pi_1}{1-\pi_0}\right) \\
&= \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + \log\left(\frac{1-\pi_1}{1-\pi_0}\right) + x\left(\log\left(\frac{\pi_1}{\pi_0}\right) - \log\left(\frac{1-\pi_1}{1-\pi_0}\right)\right) \\
&= \log\left(\frac{P(Y=1)P(X=0|Y=1)}{P(Y=0)P(X=0|Y=0)}\right) + x\log\left(\frac{\pi_1/(1-\pi_1)}{pi_0/(1-\pi_0)}\right) \\
&= a + bx
\end{aligned}$$

(b) Define the slope and intercept for the linear function.

THE INTERCEPT IS

$$\log\left(\frac{P(Y=1)}{P(Y=0)}\right) + \log\left(\frac{1-\pi_1}{1-\pi_0}\right)$$

AND THE SLOPE IS

$$\log\left(\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}\right)$$

3. On page 284 of the text, the author quotes Cook and Weisberg: "When conducting a binary regression with a skewed predictor, it is often easiest to assess the need for x and $\log(x)$ by including them both in the model so that their relative contributions can be assessed directly." Show that indeed the log odds are a function of x and $\log(x)$ for the gamma distribution.

THE TEXT GIVES THE FOLLOWING, SO WE HANDLE ONLY THE SECOND TERM IN THE SUM, AS THE FIRST TERM IS A CONSTANT WITH RESPECT TO x:

$$\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + \log\left(\frac{f(x|Y=1)}{f(x|Y=0)}\right)$$

WE'LL USE THE PARAMETRIZATION OF THE GAMMA DISTRIBUTION AS FOLLOWS:

$$f(x|\alpha,\beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}$$

Then we can start rearranging:

$$\begin{aligned}\log\left(\frac{f(x|Y=1)}{f(x|Y=0)}\right) &= \log\left(\frac{x^{\alpha_1-1}e^{-x/\beta_1}\Gamma(\alpha_0)\beta_0^{\alpha_0}}{x^{\alpha_0-1}e^{-x/\beta_0}\Gamma(\alpha_1)\beta_1^{\alpha_1}}\right)\\ &= \log\left(\frac{\Gamma(\alpha_0)\beta_0^{\alpha_0}}{\Gamma(\alpha_1)\beta_1^{\alpha_1}}\right) + (\alpha_1-\alpha_0)\log(x) + \left(\frac{1}{\beta_0}+\frac{1}{\beta_1}\right)x\end{aligned}$$

Finally, we see that the log odds is a function of x and $\log(x)$ when x has the gamma distribution.

4. Chapter 8, Question 4

   (a) No, model (8.6) is not a valid model: we see from the marginal model plots that our model does not match a nonparametric fit to the data.

   (b) The real answer here is that the predictors may very well have nonlinear associations and we may be missing some important predictors like diet and exercise; however, based on the kernel density estimates, we first observe that x_1 and x_4 are both right-skewed. As shown above, this may indicate that adding the log transformations of both of these variables may be important. To a smaller degree, it appears variable x_1 has unequal variances when heart disease = yes and no, which may indicate adding a quadratic term if we think x_1 is approximately normally distributed. The fact that the plot is much wider than it is narrow may be exaggerating the skewness to our eyes.

   (c) This is a much-improved model. We seem to still be having trouble modeling the relationship between x_1 and the odds of a heart attack at very small values of x_1, but otherwise, the model seems to match a nonparametric fit pretty well. We don't have the ability to conduct a goodness of fit test for these binary data.

   (d) Holding systolic blood pressure, cholesterol, obesity, and age constant, when a patient has a family history of heart disease, our model predicts their odds of a heart attack to be $e^{0.941056} = 2.56$ times as large as when a patient does not have a family history of heart disease.