STAT 608 HW 7

Jack Cunningham (jgavc@tamu.edu)

11/22/24

1)

We have the logistic regression model with one predictor:

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

a)

We apply e^x to both sides:

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

Multiply both sides by $1-\theta(x)$:

$$\theta(x) = (1-\theta(x))e^{\beta_0+\beta_1 x}$$

$$\theta(x) = e^{\beta_0+\beta_1 x} - e^{\beta_0+\beta_1 x}\theta(x)$$

$$\theta(x) + e^{\beta_0+\beta_1 x}\theta(x) = e^{\beta_0+\beta_1 x}$$

On left side we factor out $\theta(x)$:

$$\theta(x)(1 + e^{\beta_0+\beta_1 x}) = e^{\beta_0+\beta_1 x}$$

We are left with the desired result:

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

b)

Starting from our result in a, we multiply by $e^{-(\beta_0+\beta_1 x)}/e^{-(\beta_0+\beta_1 x)}$:

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

With $e^{-(\beta_0+\beta_1 x)}e^{\beta_0+\beta_1 x} = 1$, we distribute through and are left with the desired result:

$$\theta(x) = \frac{1}{1+e^{-(\beta_0+\beta_x)}}$$

2)

We can use the Bernoulli distribution for the conditional distribution of $X|Y=1$ and $X|Y=0$. So:

$$P(X=x|Y=0) = \pi_0^x(1-\pi_0)^{1-x}$$

$$P(X=x|Y=1) = \pi_1^x(1-\pi_1)^{1-x}$$

This will help with the next two parts.

a)

On page 283 it is derived that when X is a discrete random variable we have:

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

Using our conditional distributions from earlier we have:

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

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

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

We can see that this the log odds of x are a linear function of x.

b)

The slope β_1 and intercept β_0 are:

$$\beta_1 = \log\left(\frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)}\right), \beta_0 = \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + \log\left(\frac{1-\pi_1}{1-\pi_0}\right)$$

3)

The gamma distribution has density function:

$$f(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta} \quad x,\alpha,\beta > 0$$

On page 283 it is shown that when X is a continuous variable the log odds function is:

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

Each conditional distribution will have its own α, β. So we can say $f(x|Y=j), j=0,1$ is a gamma density with α_j and β_j, $j=0,1$. Also for the sake of efficiency I will use $\xi_j = \frac{1}{\Gamma(\alpha_j)\beta_j^{\alpha_j}}$ to reduce the clutter of the equations. We then have:

$$\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + \log\left(\frac{\xi_1 x^{\alpha_1-1}e^{-x/\beta_1}}{\xi_0 x^{\alpha_0-1}e^{-x/\beta_0}}\right)$$

After simplifying we have:

$$\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + \log\left(\frac{\xi_1}{\xi_0}\right) + (\alpha_1-1)\log(x) - (\alpha_0-1)\log(x) + (-x/\beta_1) - (-x/\beta_0)$$

$$\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + \log\left(\frac{\xi_1}{\xi_0}\right) + (\alpha_1-\alpha_0)\log(x) + \left(\frac{1}{\beta_0} - \frac{1}{\beta_1}\right)x$$

We can clearly see that the log odds are a function of x and $\log(x)$ for the gamma distribution.

4)

a)

Model 8.6 is not a valid model for the data. We can see in the marginal model plots for X_1 and X_4 that there is significant deviation between the loess estimate of $E(Y|X_1), E(Y|X_4)$ and $E(\hat{Y}|X_1), E(\hat{Y}|X_4)$.

The two predictors X_1, blood pressure, and X_4, obesity, are intuitively important variables to predict heart disease yet are not statistically significant according to Model 8.6. These predictors should be reviewed.

After looking at the density estimates of X_1 and X_4 we can see that both have right skewed distributions. This would indicate that log odds could depend on both X_1 and $\log X_1$, and similarly for X_4. Model 8.6 has not included the log transformations of X_1 and X_4.

b)

In line with out discussion at the end of part a we should add $\log(X_1)$ and $\log(X_4)$ due to the skewed densities of both X_1 and X_4.

c)

Model 8.7 is a valid model for the data. The marginal model plots show close tracking between the loess estimates of $E(Y|X_i)$ and $E(\hat{Y}|X_i)$ for each predictor, particularly a large improvement in X_1 and X_4 which now have p-values significantly lower than before. To fully review the model we would need to see plots that show us leverage points, but in absence of that we can assume model 8.7 is a valid model.

d)

The predictor variable X_3 is a dummy variable equaling one for patients with a family history. The coefficient from model 8.7 is 0.941056. We interpret as follows: with all other variables remaining unchanged, a patient having a family history of heart disease increases the log-odds of heart disease by 0.941056 on average.