Jack Rodoni
STAT 608
Assignment 06
4/20/21 (:

1) For logistic Regression w/ 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 $\theta(x) = \dfrac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$

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

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

$$\iff \theta(x)\left(1+\exp(\beta_0+\beta_1 x)\right) = \exp(\beta_0+\beta_1 x)$$

$$\iff \theta(x) = \exp(\beta_0+\beta_1 x) / \left(1+\exp(\beta_0+\beta_1 x)\right)$$

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

$$\theta(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \cdot \frac{e^{-(\beta_0+\beta_1 x)}}{e^{-(\beta_0+\beta_1 x)}} = \frac{e^{\beta_0-\beta_0+\beta_1 x-\beta_1 x}}{e^{-(\beta_0+\beta_1 x)} + e^{0}} = \boxed{\frac{1}{1+e^{-(\beta_0+\beta_1 x)}} = \theta(x)}$$

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" & that X is a dummy, taking the value 1 w/ probability π_j, $j=0,1$ conditional on $Y=0,1$.

(See H.O.S slide 67) → (a) Show that the log odds are a linear function x.

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

- Taking log of both sides:

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

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

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

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

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

[Note: $P(X=0|Y=1) = 1-\pi_1$ & $P(X=0|Y=0) = 1-\pi_0$]

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

$$\boxed{\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = a + bx}$$

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

<u>intercept</u>: $\log\left(\frac{P(y=1)\,P(x=0|y=1)}{P(y=0)\,P(x=0|y=0)}\right)$

<u>slope</u>: $\log\left(\frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)}\right)$

3.) On pg 284 of the text, the author quotes Cook and Weisberg: "When conducting a binary regression w/ 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 relative contributions can be assessed directly". Show that, indeed, the log odds are a function of x & log(x) for the gamma distribution.

(see slide 292 of text book) →

• Recall (PDF for gamma) $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$

• Note that the $\frac{\beta^\alpha}{\Gamma(\alpha)}$ is constant wrt x.

• $\frac{\theta}{1-\theta} = \frac{x_1^{\alpha-1} e^{-\beta x_1}}{x_0^{\alpha-1} e^{-\beta x_0}} \Rightarrow \log\left(\frac{\theta}{1-\theta}\right) = \log\left(\frac{x_1^{\alpha-1} e^{-\beta x_1}}{x_0^{\alpha-1} e^{-\beta x_0}}\right)$

$= \log\left(\frac{x_1^{\alpha-1}}{x_0^{\alpha-1}}\right) + \log\left(\frac{e^{-\beta x_1}}{e^{-\beta x_0}}\right)$

$$\boxed{\log\left(\frac{\theta}{1-\theta}\right) = \log\left(x_1/x_0\right) + \frac{x_1}{x_0}}$$

4.) Chapter 8, Question 4

(a) Is model (8.6) a valid model for the data? Give reasons to support your answers.

- No, this model doesn't seem to be a valid model. Looking at the marginal model plots we can see that the marginal models for x_1 and x_4 do not match up well w/ the nonparametric model.

(b) What extra predictor term or terms would you reccommend be added to model (8.6) in order to improve it.

- I would suggest adding a log transformed version of both x_1 & x_4 to the model. Both x_1 & x_4 seem to have distributions that are skewed right, applying the log transform might help normalize them.

c.) Following your advice in (b) extra predictor terms were added to the model. We shall denote these predictors $f_1(x_1)$ & $f_2(x_4)$. Marginal model plots for the new model are shown. Is the new model valid?

- Yes the model seems to be valid. Looking at the marginal model plots we see that the the marginal models for each one of our variables matches up well w/ the nonparametric curves.

(See H.O. 8, slide 26) → (d) Interpret the estimated coefficient of x_3 in the model.

$$e^{0.941056} = 2.56268619$$

- Our model predicts the odds of a patient having heart given they have a family history is about 2.56268619 times higher than the odds of a patient having heart disease, who doesn't have a family history, holding all else constant.