Chapter 12.3a Bayesian Logistic Modeling

Jim Albert and Monika Hu

Chapter 12 Bayesian Multiple Regression and Logistic Models

Example: U.S. women labor participation

- The University of Michigan Panel Study of Income Dynamics (PSID) is the longest running longitudinal household survey in the world.
- Information on these individuals includes data covering employment, income, wealth, expenditures, health, marriage, childbearing, and child development.
- The PSID 1976 survey provides helpful self-reporting data sources for studies of married women's labor supply. A sample includes information on family income exclusive of wife's income (in $1000) and the wife's labor participation (yes or no).

The PSID Sample

This PSID sample contains 753 observations and two variables. Table provides the description of each variable in the PSID sample.

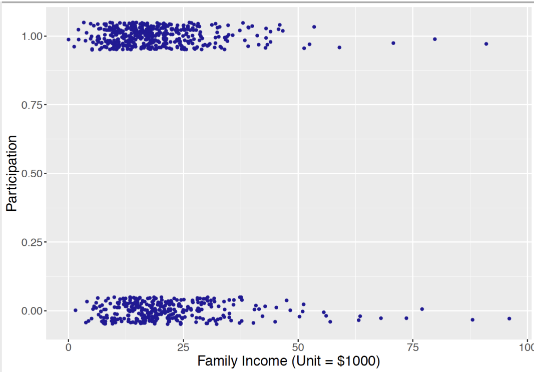
Variable	Description
LaborParticipation	Binary; the labor participation status of the wife 1 = yes, 0 = no
FamilyIncome	Continuous; the family income exclusive of wife' income, in $1000, 1975 U.S. dollars

A Prediction Problem

- Suppose one is interested in predicting a wife's labor participation status from the family income exclusive of her income.
- The response variable (labor participation) is not continuous, but binary – either the wife is working or she is not.
- One is interested in estimating the probability of a labor participation (yes) as a function of the predictor variable, family income exclusive of her income.
- Requires a new model that can express the probability of a yes as a function of the predictor variable.

Graph of the Data

- Figure displays a scatterplot of the family income against the labor participation status.
- We see that roughly half of the wives are working and it is difficult to see if the family income is predictive of the participation status.



Regression Modeling

- In Chapter 11, when one had a continuous-valued response variable and a single continuous predictor, the mean response μ_i was be expressed as a linear function:

$$\mu_i = \beta_0 + \beta_1 x_i.$$

- Moreover we assumed the response Y_i is Normally distributed with mean μ_i.
- However, such a Normal density setup is not sensible for a binary response Y_i since the response is not continuous-valued.

A logistic regression model

- Recall the definition of odds was introduced – an odds is the ratio of the probability of some event will take place over the probability of the event will not take place.
- In the PSID example, let p_i be the probability of labor participation of married woman i, and the corresponding odds of participation is $\frac{p_i}{1-p_i}$.
- If one applies a logarithm transformation on the odds, one obtains a quantity, called a log odds or logit, that is real-valued.
- One obtains a model for a binary response by writing the logit in terms of the linear predictor.

A logistic regression model

- The response Y_i is assumed to have a Bernoulli distribution with probability of success p_i.
- The logistic regression model writes that the logit of the probability p_i is a linear function of the predictor variable x_i:

$$\mathrm{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i.$$

Interpreting the model

- With the logit function, the regression coefficients β_0 and β_1 are directly related to the log odds $\log\left(\frac{p_i}{1-p_i}\right)$ instead of p_i.
- The intercept β_0 is the log odds when the predictor takes a value of 0. In the PSID example, it refers to the log odds of labor participation of a married woman, whose family has 0 family income exclusive of her income.
- The slope β_1 refers to the change in the log odds of labor participation of a married woman who has an additional $1000 family income exclusive of her own income.

Rewriting the Model

- By rearranging the logistic model, one expresses the regression as a nonlinear equation for the probability of success p_i:

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

$$\frac{p_i}{1-p_i} = \exp(\beta_0 + \beta_1 x_i)$$

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}.$$

- This equation shows that the logit function guarantees that the probability p_i lies in the interval (0, 1).