Bayesian Methods for Data Science (DATS 6450 - 11)

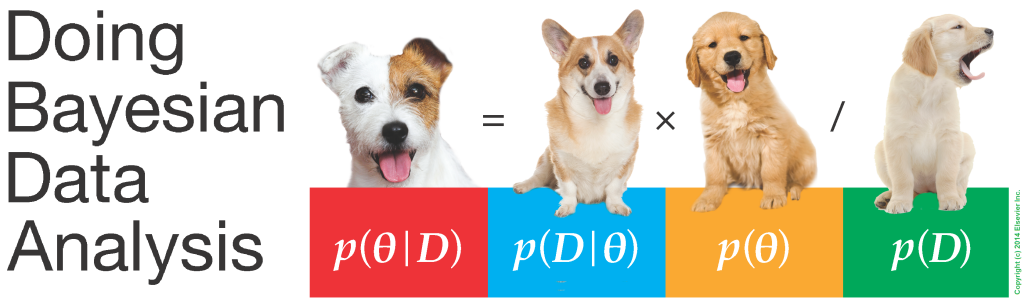
Dichotomous Predicted Variable

Yuxiao Huang

Data Science, Columbian College of Arts & Sciences
George Washington University
yuxiaohuang@gwu.edu

November 20, 2019

Reference



Picture courtesy of the book website

- This set of slices is an excerpt of the book by Professor John K. Kruschke, with some trivial changes by the creator of the slides
- Please find the reference to and website of the book below:
  - Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier
  - `https://sites.google.com/site/doingbayesiandataanalysis/`

Overview

1. Multiple metric predictors
2. Logistic Regression
3. Robust Logistic Regression

The Data

- Suppose we have a dataset measuring the height, weight, and gender (male or female) of a sample of full-grown adults
- The dataset is shown in Figure 21.1
- The data are plotted as 1's or 0's, with gender arbitrarily coded as male $= 1$ and female $= 0$
- All 0's are located on the bottom plane and all 1's are located on the top plane

Figure 21.1

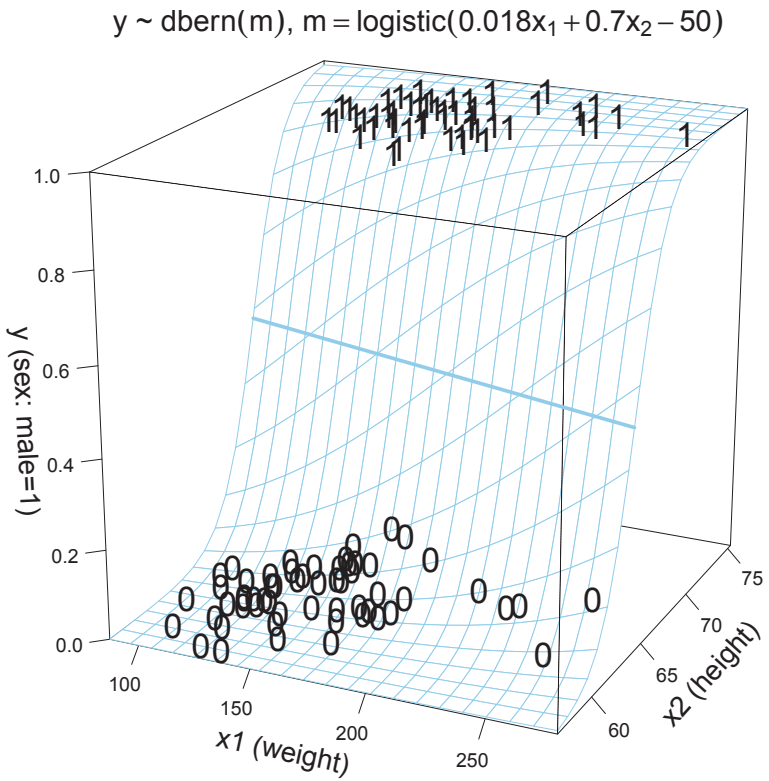


Figure 21.1: Data show gender (arbitrarily coded as male=1, female=0) as a function of weight (in pounds) and height (in inches). All 0's are located on the bottom plane of the cube, and all 1's are located on the top plane of the cube. Logistic surface shows maximum-likelihood estimate. Heavy line shows 50% threshold. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

The Goal

- We want to use the features (height and weight) to predict the target (gender)
- **Q:** Can we use GLM to do so?

The Goal

- We want to use the features (height and weight) to predict the target (gender)
- **Q:** Can we use GLM to do so?
- **A:** No, since the target here is discrete (binary, to be more specific)

The Model and Implementation in JAGS

- We will use a logistic function of a linear combination of the predictors
- The idea is that a linear combination of metric predictors is mapped to a probability value via the logistic function, and the predicted 0's and 1's are Bernoulli distributed around the probability:

$$y \sim \text{Bernoulli}(\mu), \tag{1}$$

$$\mu = \text{logistic}(\beta_0 + \beta_1 x_1 + \beta_2 x_2), \tag{2}$$

$$\text{logistic}(x) = \frac{1}{(1 + \exp(-x))}. \tag{3}$$

- A diagram of the model is presented in Figure 21.2
- See `Jags-Ydich-XmetMulti-Mlogistic.R` for details

Figure 21.2

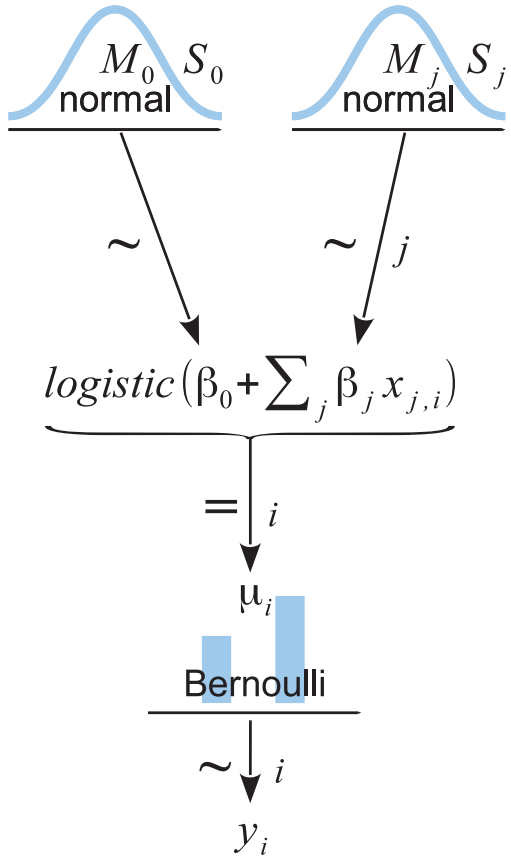


Figure 21.2: Dependency diagram for multiple logistic regression. Compare with the diagram for robust multiple linear regression in Figure 18.4 (p. 498). Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Example: Height, Weight, and Gender

- Figure 21.3 shows the results of predicting gender from weight alone
- Superimposed on the data are logistic curves that have parameter values from various steps in the MCMC chain
  - the spread of the logistic curves indicates the uncertainty of the estimate
  - the steepness of the logistic curves indicates the magnitude of the regression coefficient
  - the 50% probability threshold is marked by arrows that drop down from the logistic curve to the x-axis, near a weight of approximately 160 pounds
- The lower panels of Figure 21.3 show the marginal posterior distribution on the parameters
- In particular, the slope coefficient, β_1, has a mode larger than 0.03 and a 95% HDI that is well above zero (presumably by enough to exclude at least some non-zero ROPE)

Figure 21.3

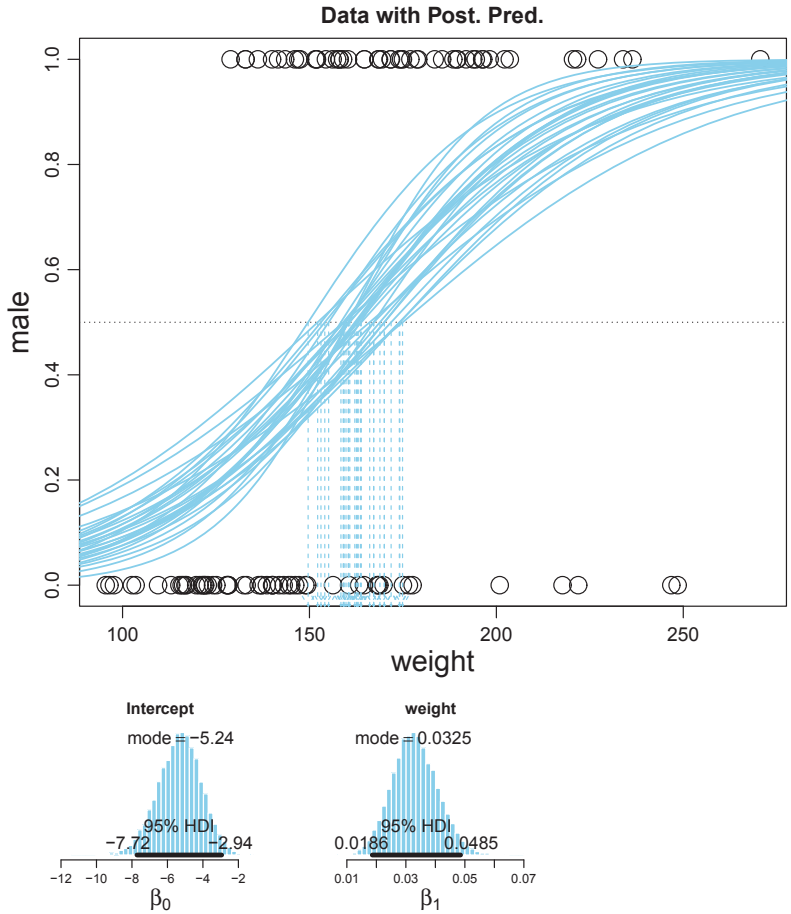


Figure 21.3: Predicting gender (arbitrarily coded as male=1, female=0) as a function of weight (in pounds), using logistic regression. Upper panel: Data are indicated by dots. Logistic curves are a random sample from the MCMC posterior. Descending arrows point to threshold weights at which the probability of male is 50%. Lower panels: Marginal posterior distribution on baseline and slope. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Example: Height, Weight, and Gender

- Figure 21.4 shows the results when using two predictors (height and weight)
- Superimposed on the data are credible level contours at which $p(\text{male}) = 50\%$
  - The 50% level contour is the set of x_1, x_2 values for which $\mu = 0.5$
  - the steepness of the logistic curves indicates the magnitude of the regression coefficient
  - the perpendicular to the level contour indicates the direction in which probability changes the fastest
  - the probability of being male increases rapidly as height goes up, but the probability of being male increases only a little as weight goes up
- The interpretation above is confirmed by the lower panels of Figure 21.4
- In particular, the regression coefficient on weight has a modal value less than 0.02 and its 95% HDI essentially touches zero

Figure 21.4

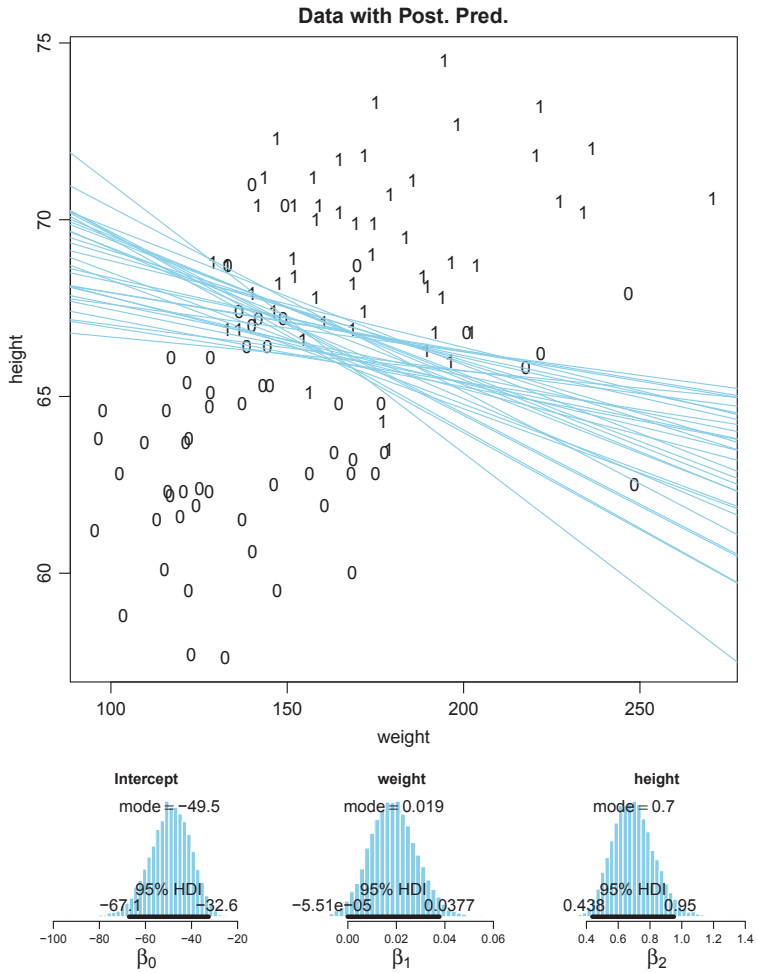


Figure 21.4: Predicting gender (arbitrarily coded as male=1, female=0) as a function of weight (in pounds) and height (in inches), using logistic regression. Upper panel: Data are indicated by 0's and 1's. Lines show a random sample from the MCMC posterior of thesholds at which the probability of male is 50%. Lower panels: Marginal posterior distribution on baseline and slopes. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Correlated Predictors

- An important cause of parameter uncertainty is correlated predictors
- Figure 21.6 shows a case of data with strongly correlated predictors
- The 50% level contours (the threshold lines) are extremely ambiguous, with many different possible angles
- The same ambiguity can arise for correlations of multiple predictors, but higher-dimensional correlations are difficult to graph
- As in the case of linear regression, it is important to consider the correlations of the predictors when interpreting the parameters of logistic regression

Figure 21.6

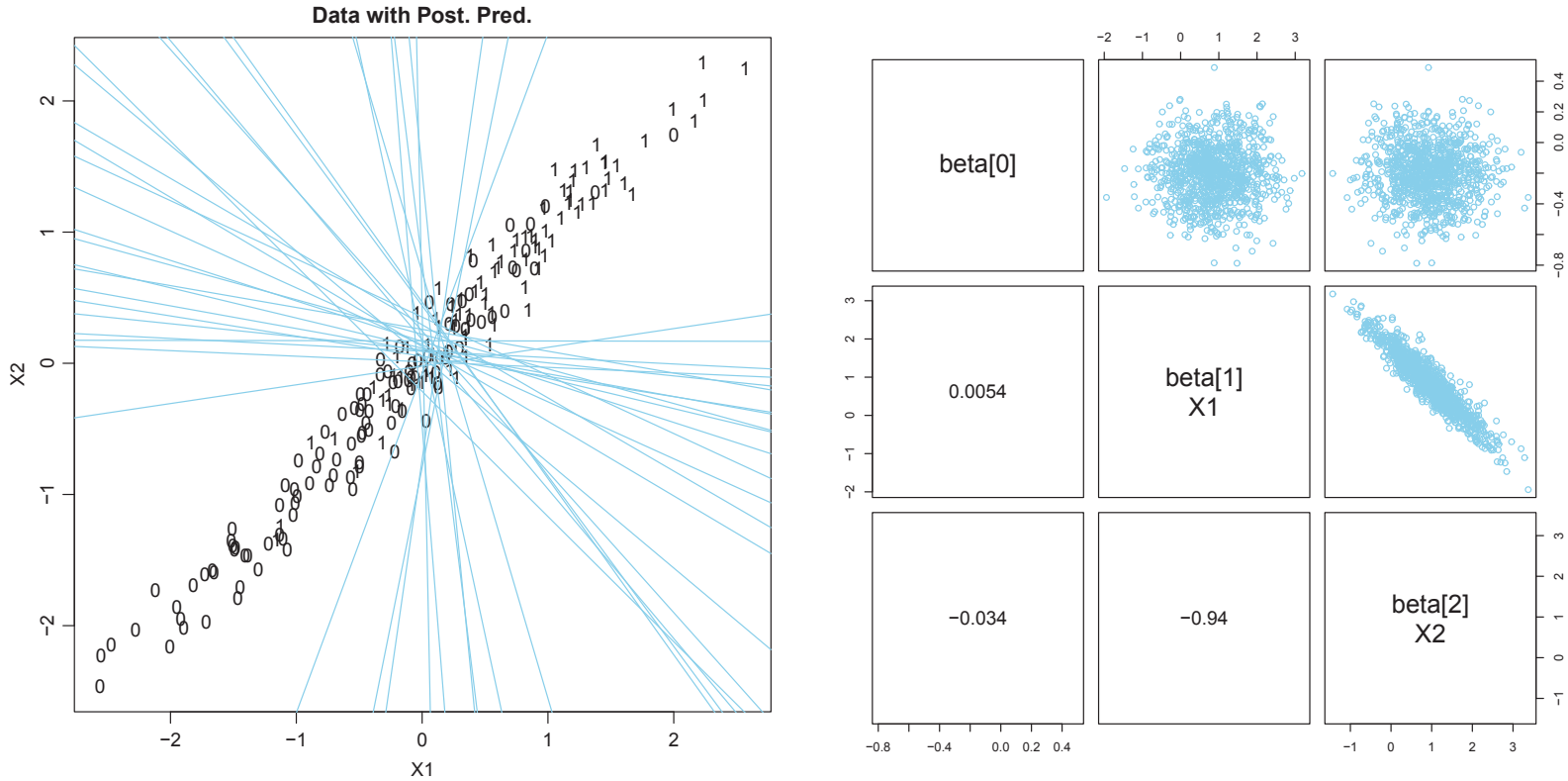


Figure 21.6: Estimates of slope parameters trade off when the predictors are correlated. Left panel shows credible 50% level contours superimposed on data. Right panel shows strong anti-correlation of credible β_1 and β_2 values. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Interaction of metric predictors

- There may be applications in which it is meaningful to consider a multiplicative interaction of metric predictors
- For example, it might be that only the combination of tall and heavy is a good indicator of being male, while being tall but not heavy, or heavy but not tall, both indicate being female
- This sort of conjunctive combination of predictors can be expressed by their multiplication (or other ways)
- Figure 21.7 shows examples of logistic surfaces with multiplicative interaction of predictors
- The left column shows examples without interaction
- The right column shows the same examples from the left column but with a multiplicative interaction included

Figure 21.7

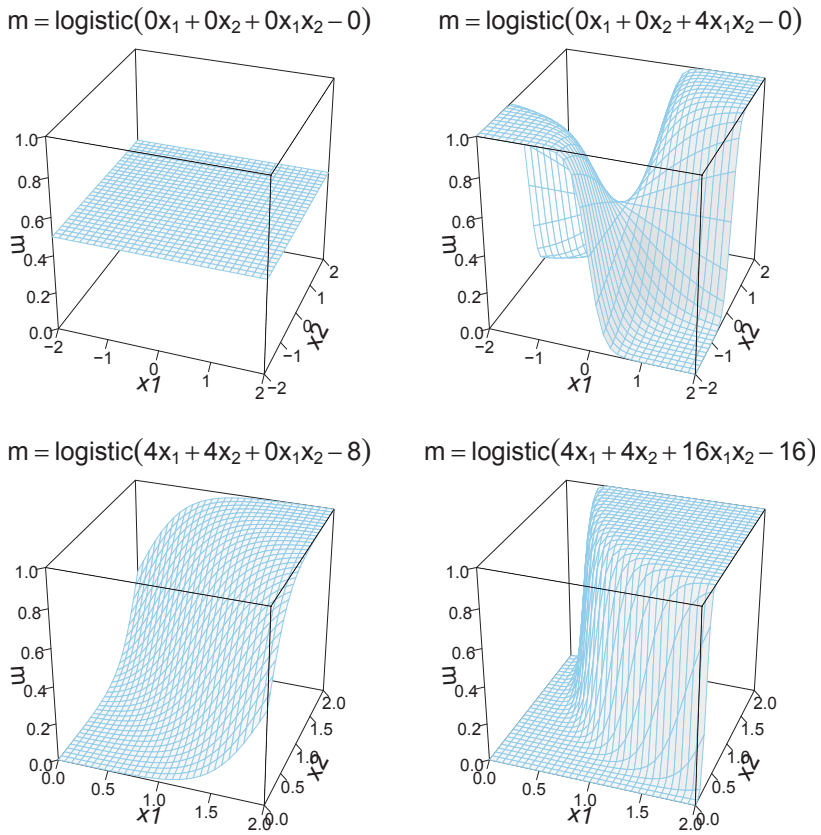


Figure 21.7: Multiplicative interaction of metric predictors in logistic regression. Left column shows examples of no interaction. Right column shows corresponding logistic surfaces with interaction. Title of each plot shows the coefficients on the predictors. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Figure 21.3

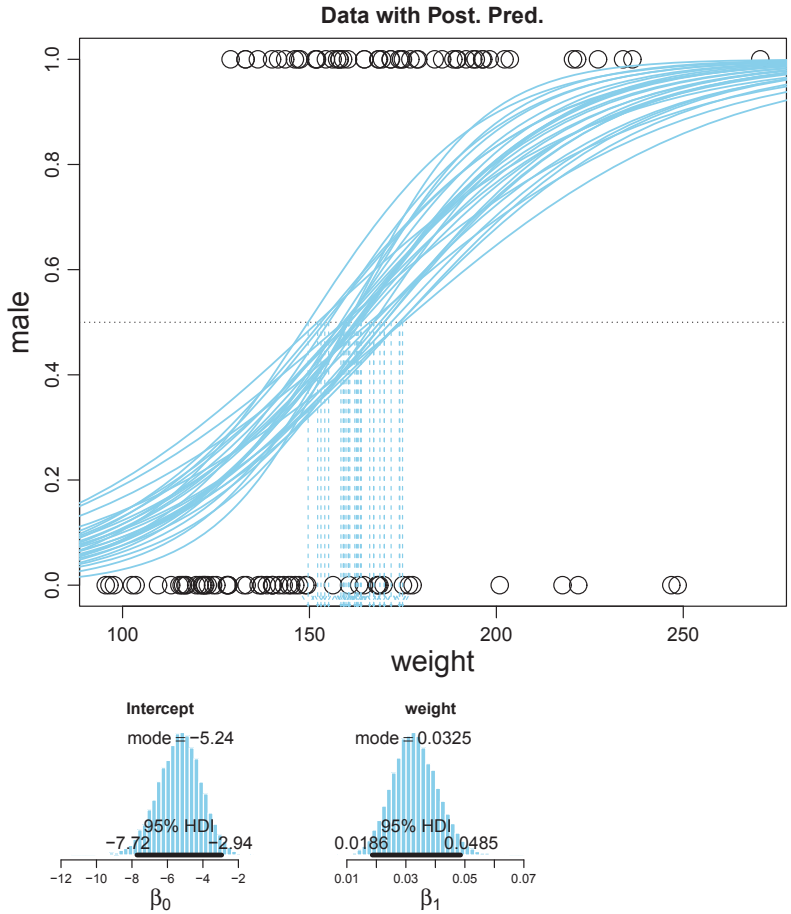


Figure 21.3: Predicting gender (arbitrarily coded as male=1, female=0) as a function of weight (in pounds), using logistic regression. Upper panel: Data are indicated by dots. Logistic curves are a random sample from the MCMC posterior. Descending arrows point to threshold weights at which the probability of male is 50%. Lower panels: Marginal posterior distribution on baseline and slope. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Robust Logistic Regression

- Figure 21.3 shows gender as a function of weight alone
- Notice that there are several unusual data points in the lower right that represent heavy females
- For these data points to be accommodated by a logistic function, its slope must not be too extreme. **Q:** Why?

Robust Logistic Regression

- Figure 21.3 shows gender as a function of weight alone
- Notice that there are several unusual data points in the lower right that represent heavy females
- For these data points to be accommodated by a logistic function, its slope must not be too extreme. **Q:** Why?
- **A:** Because if the slope is extreme, then the logistic function gets close to its asymptote at $y = 1$ for heavy weights, which then makes the probability of data points with $y = 0$ essentially nil
- We use a model that incorporates a description of outliers as such
- Because this sort of model provides parameter estimates that are relatively stable in the presence of outliers, it is called robust against outliers

Robust Logistic Regression

- We will describe the data as being a mixture of two different sources
- One source is the logistic function of the predictor(s)
- The other source is sheer randomness or "guessing" whereby the y value comes from the flip of a fair coin:

$$y \sim \text{Bernoulli}(\mu = \frac{1}{2}). \tag{4}$$

- We suppose that every data point has a small chance, α, of being generated by the guessing process, but usually, with probability $1 - \alpha$, the y value comes from the logistic function of the predictor

Robust Logistic Regression

- With the two sources combined, the predicted probability that $y = 1$ is

$$\mu = \alpha \cdot \frac{1}{2} + (1 - \alpha) \cdot \text{logistic}\left(\beta_0 + \sum_j \beta_j x_j\right). \tag{5}$$

- When the guessing coefficient is zero, then the conventional logistic model is completely recovered
- When the guessing coefficient is one, then the y values are completely random
- In most applications we would expect the proportion of random outliers in the data to be small, and therefore the prior should emphasize small values of α
- See `Jags-Ydich-XmetMulti-Mlogistic Robust.R` for details

Robust Logistic Regression

- Figure 21.8 shows the fit of the robust logistic regression model for predicting gender as a function of weight alone
- The superimposed curves show μ, which asymptote at levels away from 0 and 1, unlike the ordinary logistic curves in Figure 21.3
- The modal estimate of the guessing parameter is almost 0.2
- This implies that the asymptotes are around $y = 0.1$ and $y = 0.9$
- Especially for heavy weights, the non-1 asymptote lets the model accommodate outlying data while still having a steep slope at the threshold
- Notice that the slope of the logistic is larger than in Figure 21.3

Figure 21.8

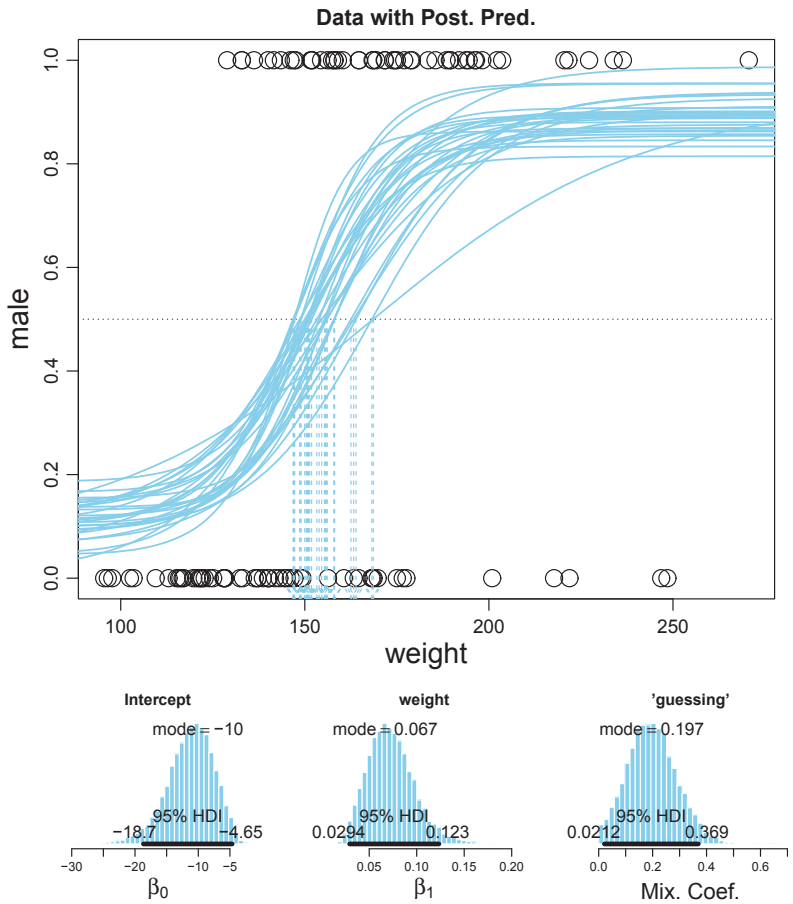


Figure 21.8: Predicting gender (arbitrarily coded as male=1, female=0) as a function of weight (in pounds), using *robust* logistic regression. Upper panel: Data are indicated by dots, the same as in Figure 21.3. Curves are a random sample from the MCMC posterior; notice asymptotes away from 0,1 limits. Descending arrows point to threshold weights at which the probability of male is 50%. Lower panels: Marginal posterior distribution on baseline, slope, and guessing coefficient. Pairwise plots are shown in Figure 21.9.

Figure 21.3

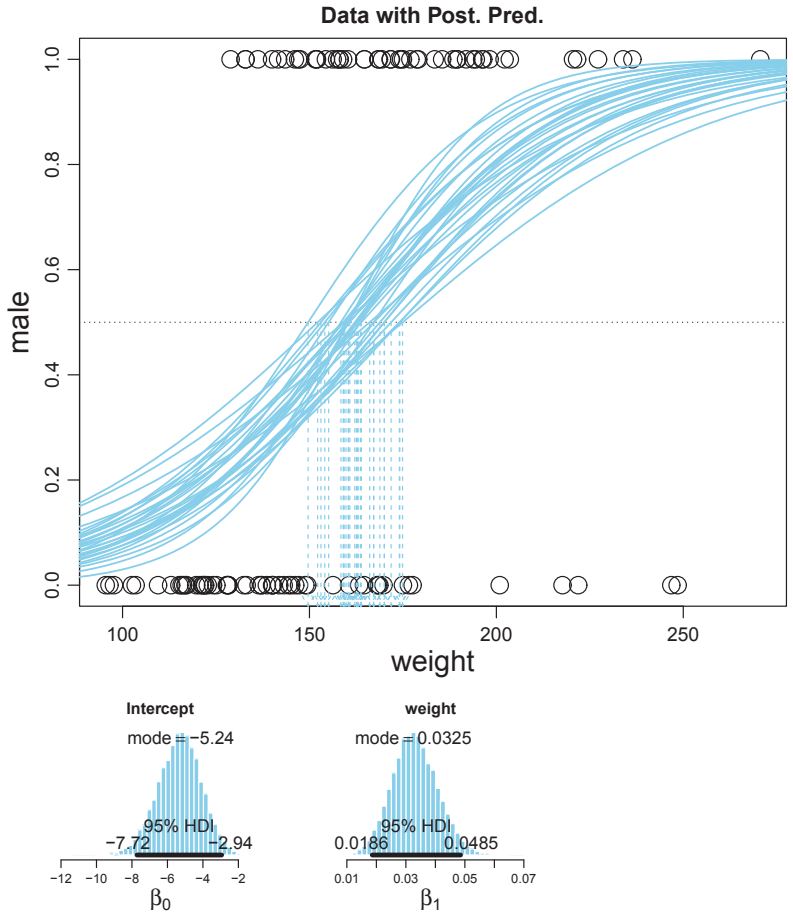


Figure 21.3: Predicting gender (arbitrarily coded as male=1, female=0) as a function of weight (in pounds), using logistic regression. Upper panel: Data are indicated by dots. Logistic curves are a random sample from the MCMC posterior. Descending arrows point to threshold weights at which the probability of male is 50%. Lower panels: Marginal posterior distribution on baseline and slope. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Robust Logistic Regression

- Figure 21.9 shows pairwise posterior parameters
- In particular, consider the panel that plots the value of β_1 (i.e., the slope on weight) against the value of the guessing parameter
- Notice the strong positive correlation between those two parameters
- This means that as the guessing parameter gets larger, credible values of the slope get larger also

Figure 21.9

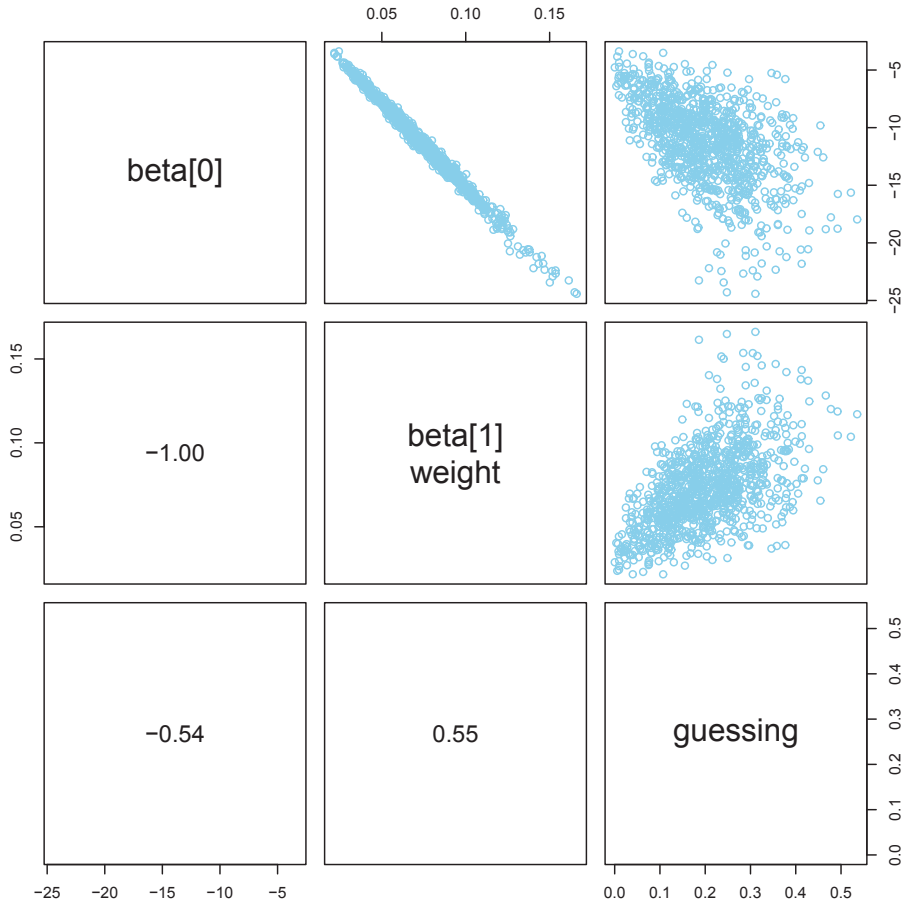


Figure 21.9: Pairwise plots for posterior distribution in Figure 21.8. Notice here that the guessing coefficient is correlated with the slope, beta[1]. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.