Bayesian Methods for Data Science (DATS 6450 - 11)

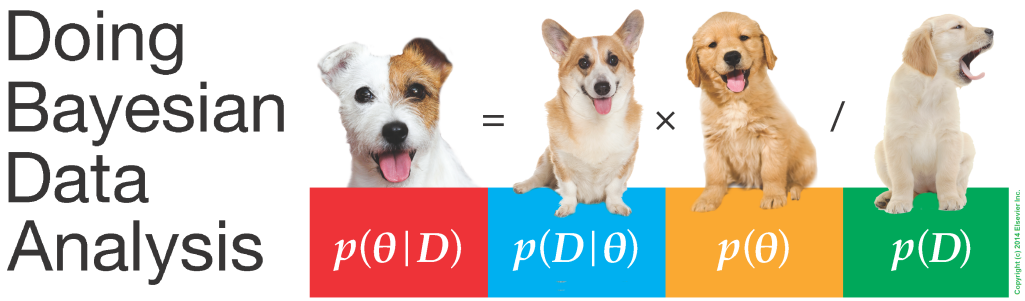
Metric Predicted Variable with Multiple Metric Predictor

Yuxiao Huang

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

November 6, 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 Linear Regression
2. Correlated Predictors
3. Multiplicative Interaction
4. Variable Selection

Simple Linear Regression

- Simple linear regression assumes the following data generation steps:
  1. generating μ based on the liner model determined by single predictor x and its weights β_0 and β_1

$$\mu = \beta_0 + \beta_1 x$$

  2. generating y based on a normal distribution determined by μ and σ

$$y \sim \text{norm}(\mu, \sigma)$$

- This is illustrated in Figure 17.1 (see next page)

Figure 17.1

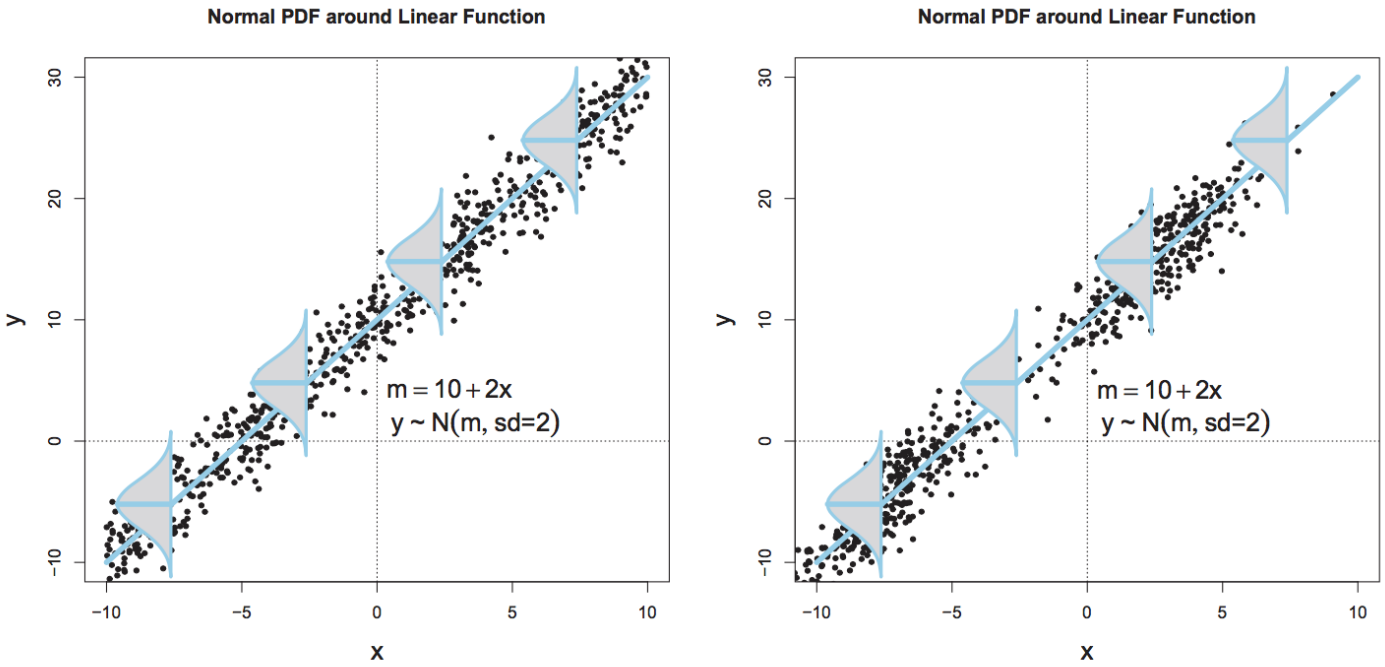


Figure 17.1: Examples of points normally distributed around a linear function. (The left panel repeats Figure 15.9, p. 406.) The model assumes that the data y are normally distributed vertically around the line, as shown. Moreover, the variance of y is the same at all values of x. The model puts no constraints on the distribution of x. The right panel shows a case in which x are distributed bimodally, whereas in the left panel the x are distributed uniformly. In both panels, there is homogeneity of variance. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Multiple Linear Regression

- We can extend simple linear regression to multiple linear regression simply by allowing multiple predictors
- That is, multiple linear regression assumes the following data generation steps
  1. generating μ based on the liner model determined by multiple predictors $x_1, x_2, \cdots, x_n$ and its weights β_0 and $\beta_1, \beta_2, \cdots, \beta_n$

$$\mu = \beta_0 + \sum_j \beta_j x_j$$

  2. generating y based on a normal distribution determined by μ and σ

$$y \sim \text{norm}(\mu, \sigma)$$

- This is illustrated in Figure 18.1 (see next page)

Figure 18.1

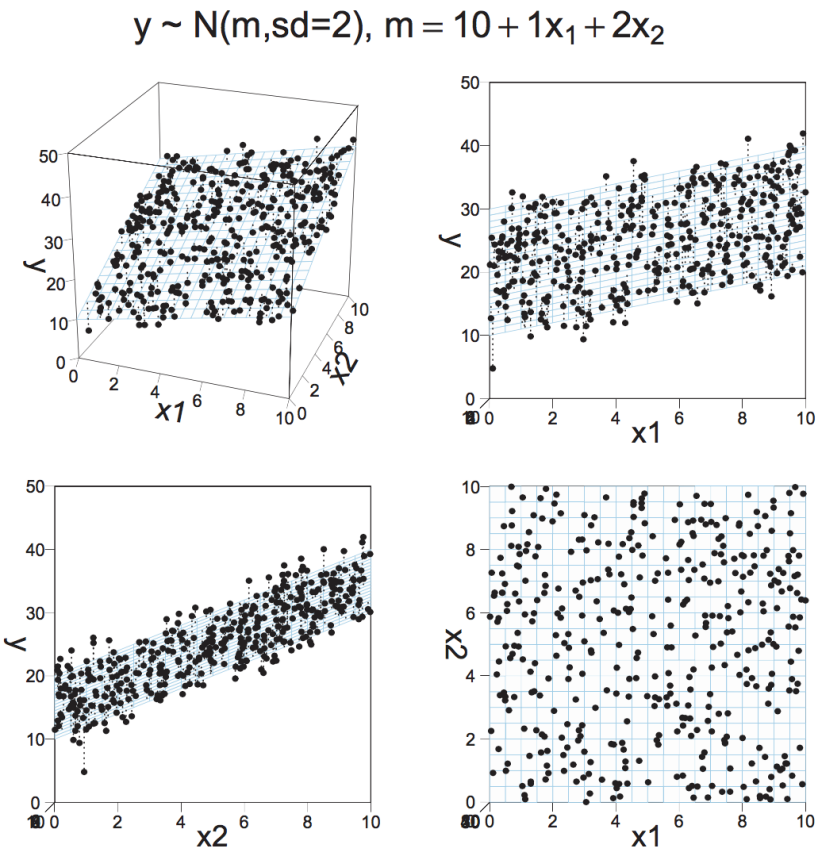


Figure 18.1: Data, y, that are normally distributed around the values in the plane. The $\langle x_1, x_2 \rangle$ values are independent of each other, as shown in the lower-right panel. The panels show different perspectives on the same plane and data. Compare with Figure 18.2. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Outliers and robust estimation: the t distribution

- In reality, there could be outliers in the data
- Normal distribution may not be able to address the outliers due to its thin tails
- A more robust distribution is (student) t distribution, which has three parameters:
  - μ: the central tendency
  - σ: the standard deviation
  - ν (where $\nu \geq 1$): the normality parameter
- Figures 17.2 and 18.4 (see next two pages) show the robust simple and multiple linear regression using t distribution

Figure 17.2

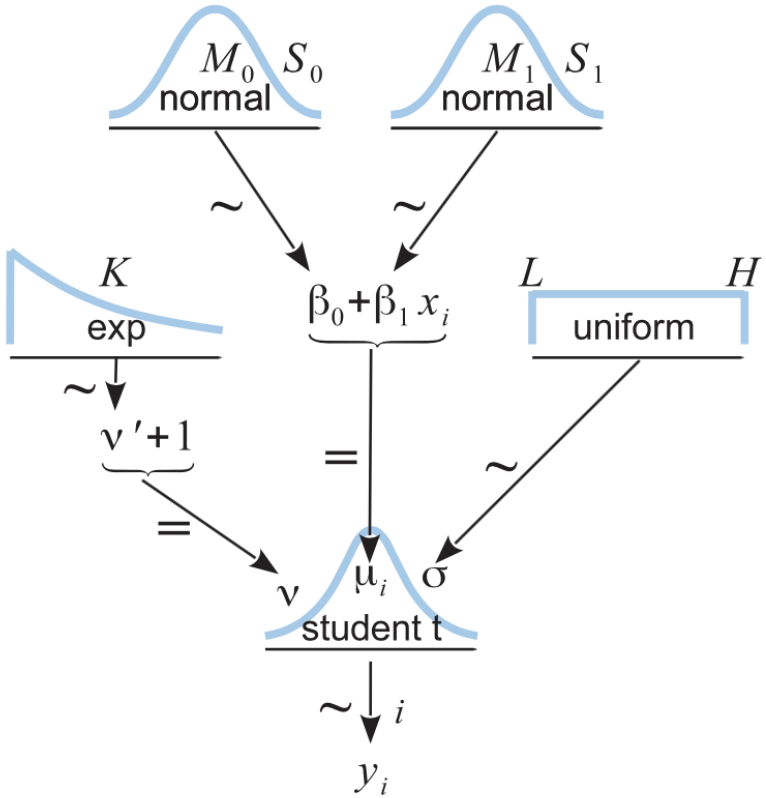


Figure 17.2: A model of dependencies for robust linear regression. The datum, y_i at the bottom of the diagram, is distributed around the central tendency μ_i, which is a linear function of x_i. Compare with Figure 16.11 on p. 437. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Figure 18.4

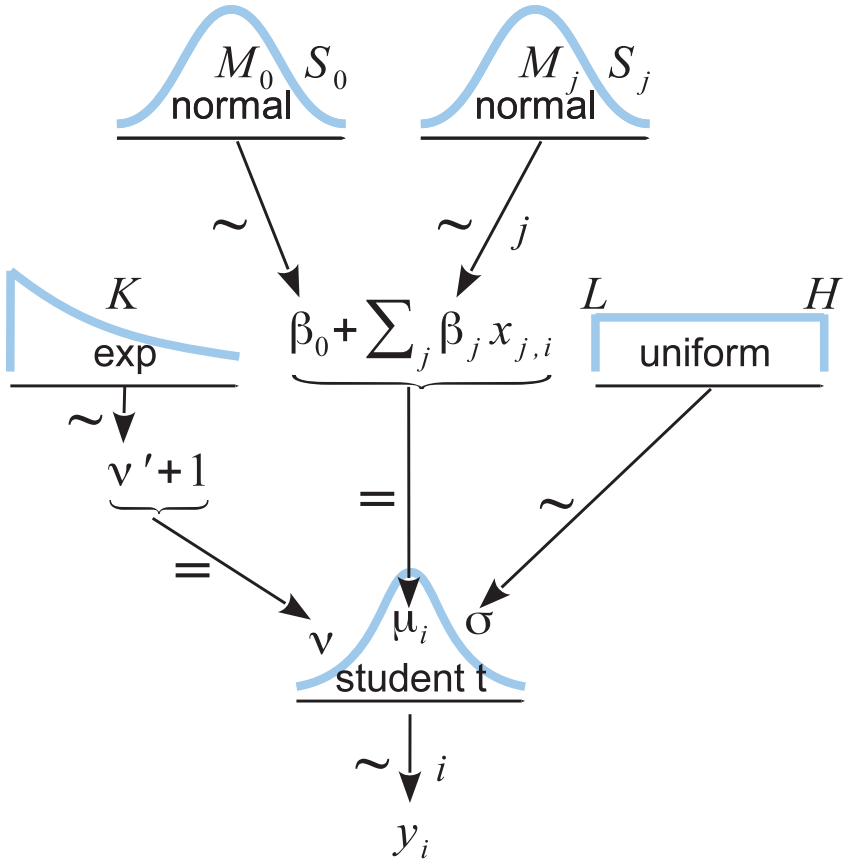


Figure 18.4: Hierarchical diagram for multiple linear regression. Compare with Figure 17.2 (p. 463). Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Independent predictors

- Consider data generated by a multiple linear regression model:

$$y \sim \text{norm}(\mu, 2) \quad \text{where} \quad \mu = 10 + x_1 + 2x_2$$

- Here the predictors, x_1 and x_2, are distributed independently
- Figure 18.1 (see next page) shows different perspectives on the data

Figure 18.1

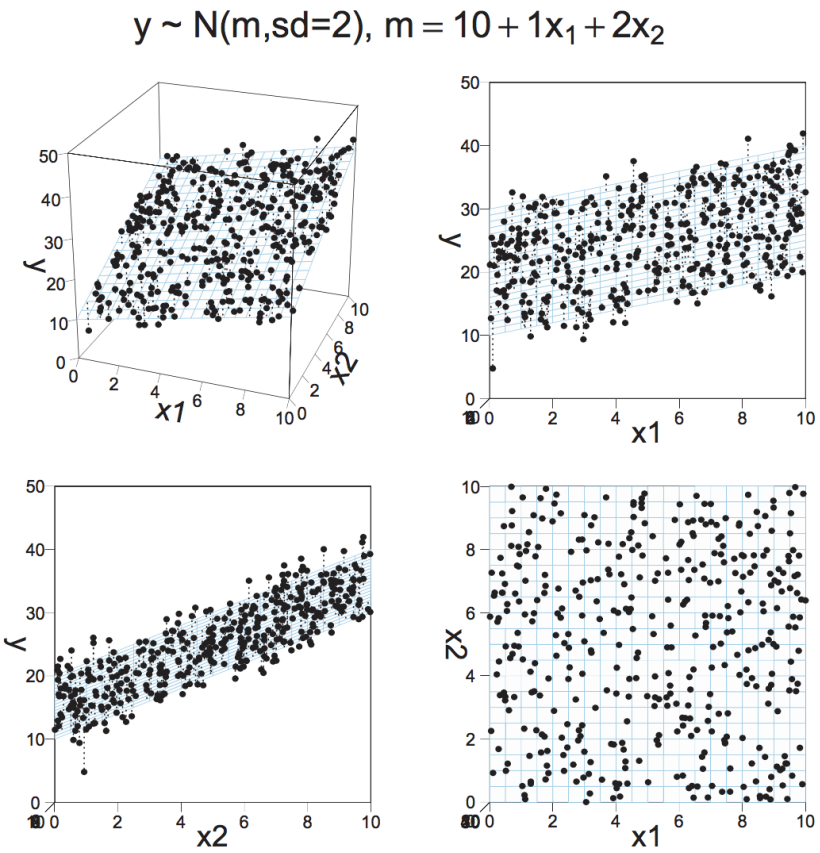


Figure 18.1: Data, y, that are normally distributed around the values in the plane. The $\langle x_1, x_2 \rangle$ values are independent of each other, as shown in the lower-right panel. The panels show different perspectives on the same plane and data. Compare with Figure 18.2. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Correlated predictors

- Now consider data generated by the exact same multiple linear regression model:

$$y \sim \text{norm}(\mu, 2) \quad \text{where} \quad \mu = 10 + x_1 + 2x_2$$

- However the predictors, x_1 and x_2, are (negatively) correlated
- Figure 18.2 (see next page) shows different perspectives on the data
- **Q:** Can you find anything wrong?

Figure 18.2

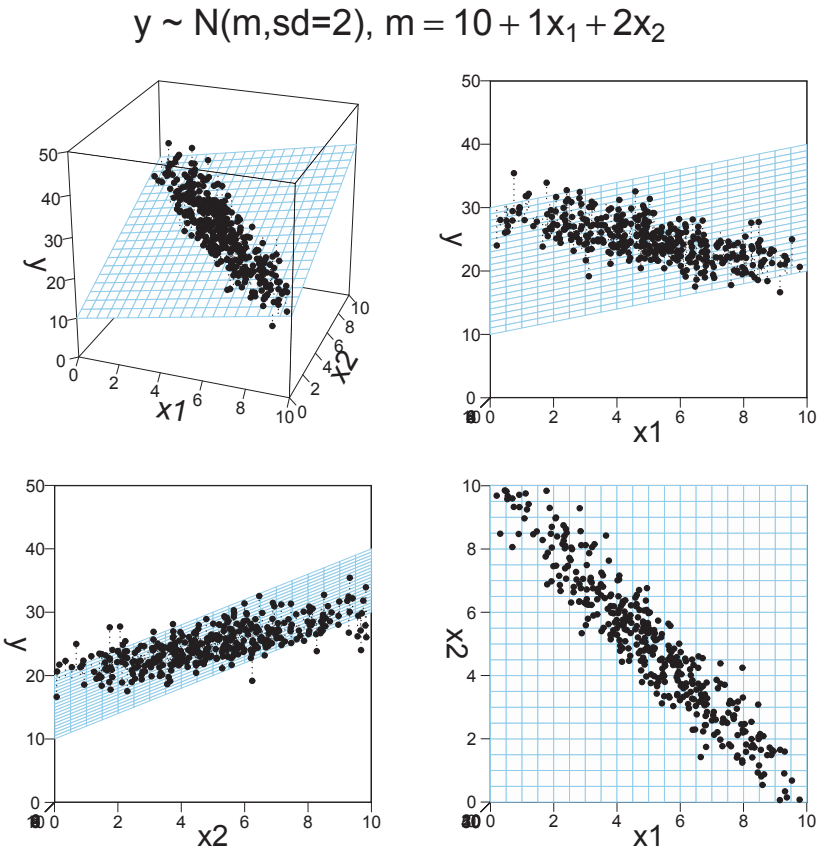


Figure 18.2: Data, y, that are normally distributed around the values in the plane. The $\langle x_1, x_2 \rangle$ values are (anti-)correlated, as shown in the lower-right panel. The panels show different perspectives on the same plane and data. Compare with Figure 18.1. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

The perils of correlated predictors

- Interpreting the relationship between individual predictor and the target can be misleading, when the predictors are correlated
  - top right panel in Figure 18.2: reversed relationship between y and x_1
  - bottom left panel in Figure 18.2: not so accurate relationship between y and x_2

A real example

- Predicting a state's average high-school SAT score based on the amount of money the state spends per pupil
- The top right panel in Figure 18.3 (see next page) shows the relationship
- **Q:** Is it a positive or negative relationship?
- **Q:** What could be a natural conclusion if the finding were real?
- **Q:** What is the reason for the spurious finding?

Figure 18.3

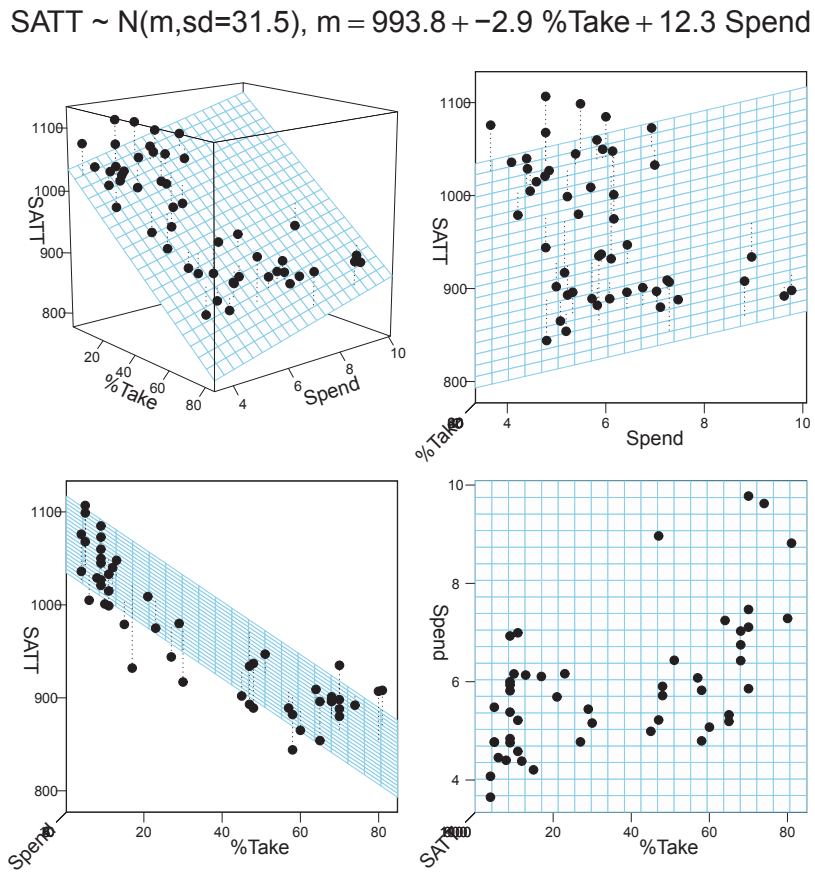


Figure 18.3: The data (Guber, 1999) are plotted as dots, and the grid shows the best fitting plane. "SATT" is the average total SAT score in a state. "%Take" is the percentage of students in the state who took the SAT. "Spend" is the spending per pupil, in thousands of dollars.

The implementation and posterior distributions

- See `Jags-Ymet-XmetMulti-Mrobust-Example.R` for details
- Figure 18.5 (see next page) shows the posterior distributions

Figure 18.5

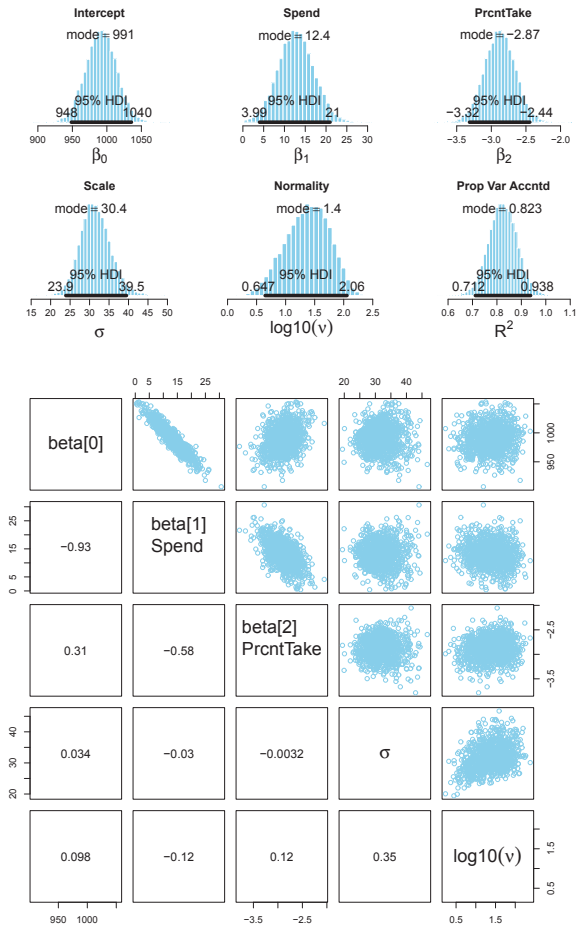


Figure 18.5: Posterior distribution for data in Figure 18.3 and model in Figure 18.4. Scatterplots reveal correlations among credible parameter values; in particular, the coefficient on Spending ("Spend") trades off with the coefficient on Percentage taking the exam ("PrcntTake"), because those predictors are correlated in the data. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

The R^2 statistic

- The R^2 statistic, a.k.a., the proportion of variance accounted for, is

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

- Here:
  - y_i is the real value of y in sample i
  - $\hat{y}_i$ is the linearly predicted value of y_i
  - $\bar{y}$ is the average of y across all the samples

Redundant predictors

- Suppose we have only two samples:

x_1	x_2	y
1	1	1
2	2	2

- **Q:** Which model wins?

$$y = \beta_1 x_1 + \beta_2 x_2$$
$$y = \beta_1 x_1$$
$$y = \beta_2 x_2$$

Redundant predictors

- Suppose we have only two samples:

x_1	x_2	y
1	1	1
2	2	2

- **Q:** Which model wins?

$$y = \beta_1 x_1 + \beta_2 x_2$$
$$y = \beta_1 x_1$$
$$y = \beta_2 x_2$$

- **A:** The second or the third (no difference)

Back to the SAT example

- Let us create a new predictor, PropNotTake, such that:

$$\text{PropNotTake} = \frac{100 - \text{PrcntTake}}{100}$$

- Here:
  - PrcntTake is the percentage of students taking SAT
  - PropNotTake is the proportion of students not taking SAT
- Thus PropNotTake is completely redundant because of PrcntTake
- Figure 18.6 (see next page) shows the posterior distributions with the redundant predictor
- **Q:** Can you find any sign of redundancy?

Figure 18.6

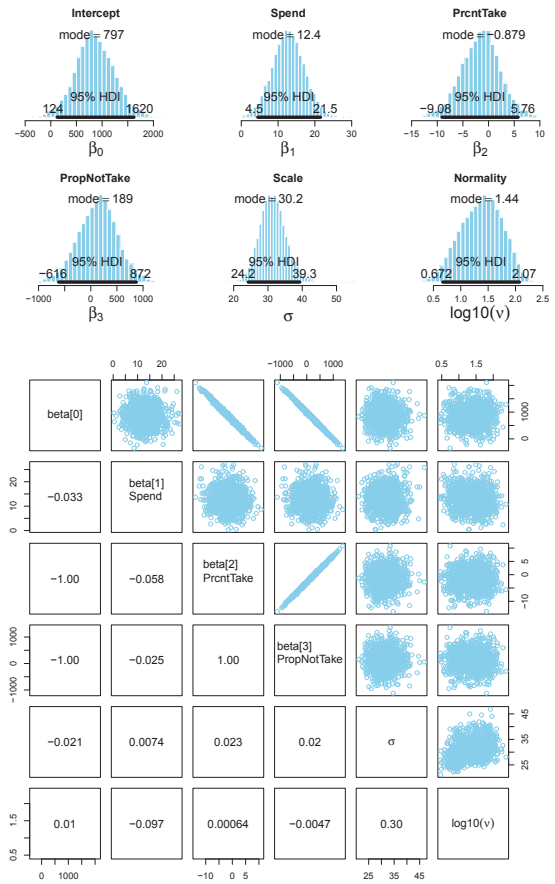


Figure 18.6: Posterior distribution for data in Figure 18.3 with a redundant predictor, the proportion of students not taking the exam. Compare with the result without a redundant predictor in Figure 18.5. Notice the perfect correlation between credible values of the regression coefficients on percentage taking the exam (PrcntTake) and proportion not taking the exam (PropNotTake). The posterior on the redundant predictors is strongly reflective of the prior distribution, which is shown in Figure 18.7. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Frequentist VS Bayesian approach

- When predictors are strongly correlated, traditional multiple linear regression models may break down
- Conversely, Bayesian multiple linear regression can still produce posterior distributions, which are largely determined by the prior
- Compare the posteriors in Figure 18.6 (see previous page) with the priors in Figure 18.7 (see next page) to see the resemblance

Figure 18.7

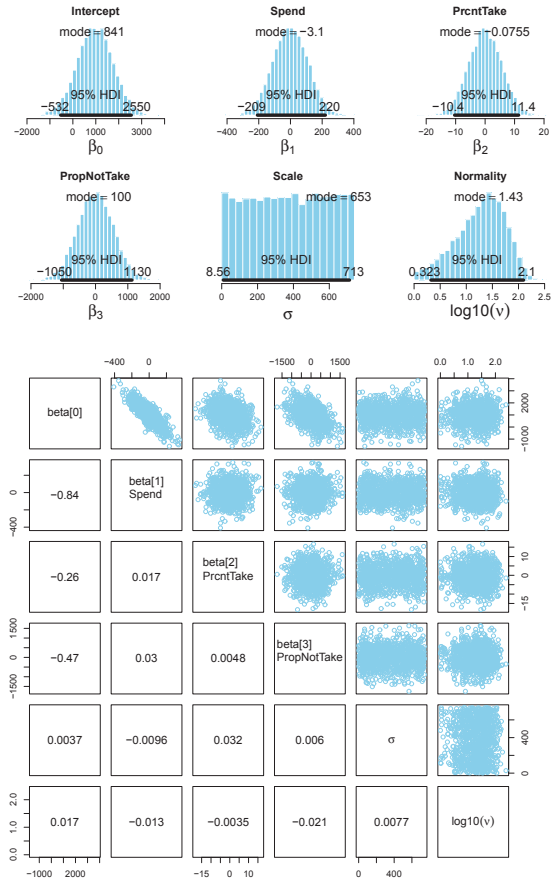


Figure 18.7: The prior distribution for the posterior distribution in Figure 18.6. Notice that the marginal posterior distributions of the redundant predictors (in Figure 18.6) is only a little narrower than the priors shown here. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Dealing with redundant predictors

- When predictors are perfectly correlated, we can simply drop all but one
- When predictors are not perfectly correlated, we can:
  - retain the top k predictors based on metrics such as variable importance
  - create a new predictor averaging the correlated ones
  - compress the correlated predictors into one using Principal Components Analysis (PCA)
  - estimate an underlying common factor using factor analysis or Structural Equation Modeling (SEM)

Informative priors, sparse data, and correlated predictors

- Informative priors are particularly useful when we have sparse data and correlated predictors
- Example
  - one flip of a coin
  - height and weight as shown in Figure 17.3 (see next page)
  - the two samples example

Figure 17.3

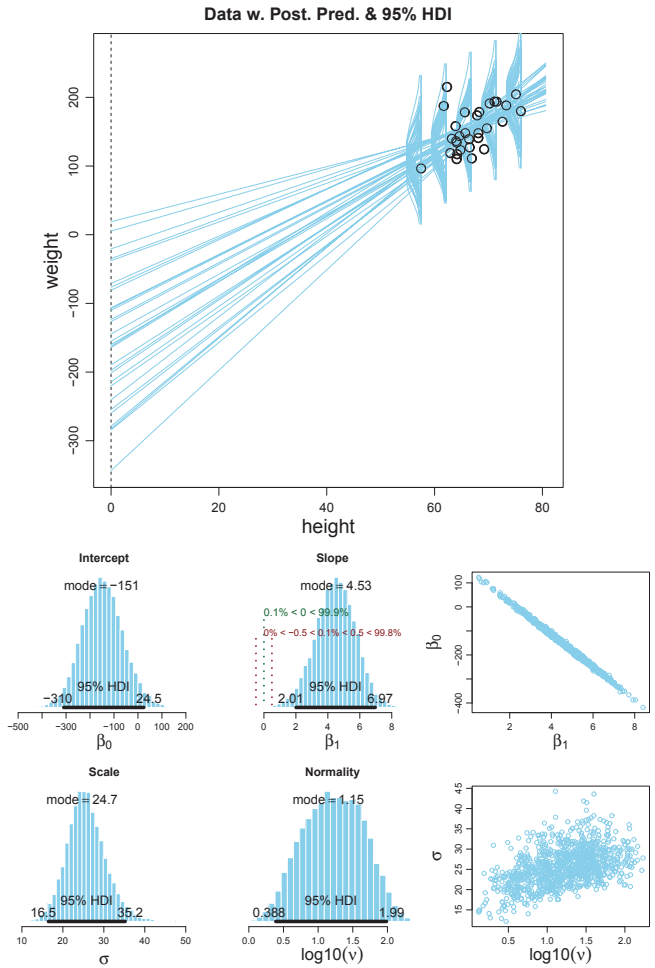


Figure 17.3: Upper panel: Data ($N = 30$) with a smattering of credible regression lines and t noise distributions superimposed. Lower panels: Marginal posterior distribution on parameters. Compare with Figure 17.4. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Multiplicative interaction of metric predictors

- In reality interactions abound
- Interactions can have many functional forms
- Here, we only consider the simplest form, namely multiplicative interaction
- For two predictors, regression with multiplicative interaction can be written as

$$\begin{aligned}\mu &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1\times 2} x_1 x_2 \\ &= \beta_0 + \underbrace{(\beta_1 + \beta_{1\times 2} x_2)}_{\text{slope of } x_1} x_1 + \beta_2 x_2 \\ &= \beta_0 + \beta_1 x_1 + \underbrace{(\beta_2 + \beta_{1\times 2} x_1)}_{\text{slope of } x_2} x_2\end{aligned}$$

- Figure 18.8 (see next page) shows the visual explanation of the three equations

Figure 18.8

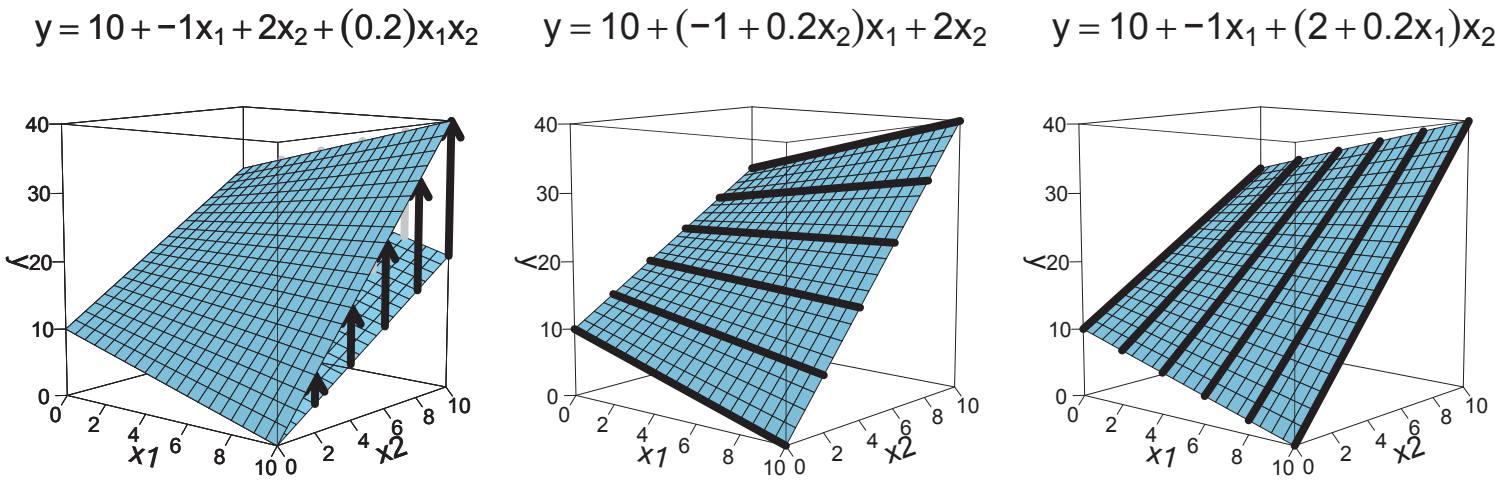


Figure 18.8: A multiplicative interaction of x_1 and x_2, parsed three ways. The left panel emphasizes that the interaction involves a multiplicative component that adds a vertical amount to the planar additive model, as indicated by the arrows that mark $\beta_{1\times2}x_1x_2$. The middle panel shows the same function, but with the terms algebraically re-grouped to emphasize that the slope in the x_1 direction depends on the value of x_2, as shown by the darkened lines that mark $\beta_1 + \beta_{1\times2}x_2$. The right panel again shows the same function, but with the terms algebraically re-grouped to emphasize that the slope in the x_2 direction depends on the value of x_1, as shown by the darkened lines that mark $\beta_2 + \beta_{1\times2}x_1$. Compare with Figure 15.3 (p. 400).

Multiplicative interaction of metric predictors

- Figure 18.9 (see next page) shows the posteriors with the multiplicative interaction
- See `Jags-Ymet-XmetMulti-Mrobust-Example.R` for details

Figure 18.9

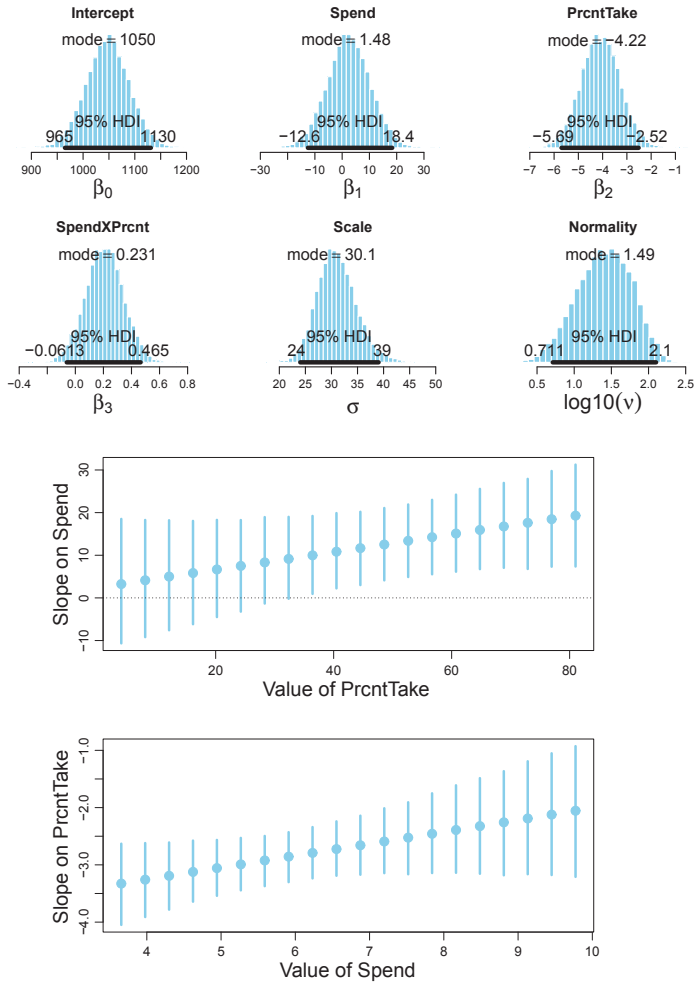


Figure 18.9: Posterior distribution when including a multiplicative interaction of Spend and PrcntTake. The marginal distribution of β_1 is the slope on Spend when PrcntTake=0, and the marginal distribution of β_2 is slope on PrcntTake when Spend=0. Lower panels show 95% HDIs and median values of slopes for other values of predictors. Slope on Spend is $\beta_1 + \beta_3 \cdot$ PrcntTake and slope on PrcntTake is $\beta_2 + \beta_3 \cdot$ Spend.

Variable selection

- Variable selection in essence is model comparsion using hierarchical modeling
- In the model that facilitates variable selection, the predicted mean value of y can be written as

$$\mu = \beta_0 + \sum_j \delta_j \beta_j x_j$$

- Here, $\delta_j \in \{0, 1\}$ is the inclusion indicator, indicating whether the predicator should be included
- **Q:** If there are two predicators, how many models the equation above represents?

Variable selection

- Variable selection in essence is model comparsion using hierarchical modeling
- In the model that facilitates variable selection, the predicted mean value of y can be written as

$$\mu = \beta_0 + \sum_j \delta_j \beta_j x_j$$

- Here, $\delta_j \in \{0, 1\}$ is the inclusion indicator, indicating whether the predicator should be included
- **Q:** If there are two predicators, how many models the equation above represents?
- **A:** 4

Variable selection

- Figure 18.13 (see next page) shows the posteriors with variable selection
- See `Jags-Ymet-XmetMulti-MrobustVarSelect-Example.R` for details

Figure 18.13

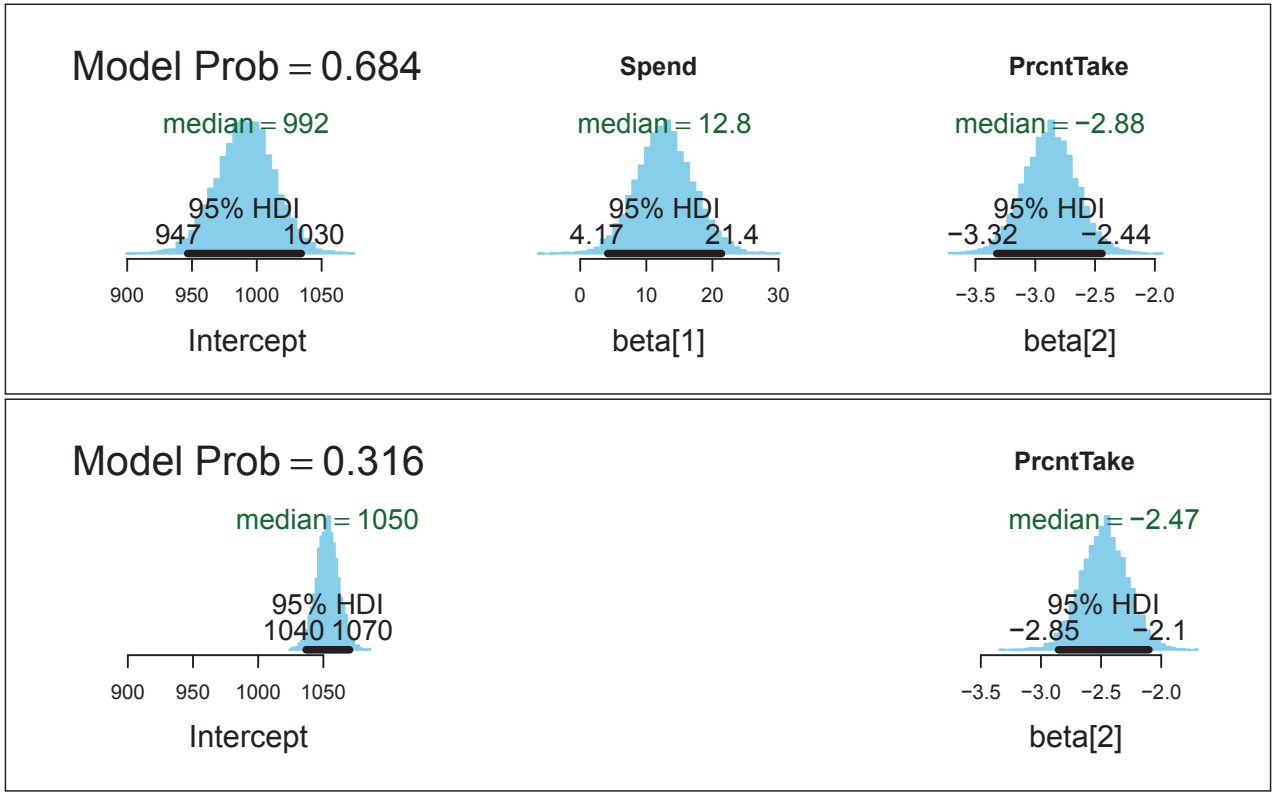


Figure 18.13: Posterior probabilities of different subsets of predictors along with the marginal posterior distributions of the included regression coefficients. The two other possible models, involving only Spend or only the intercept, had essentially zero probability. The prior probability of each model was $0.5^2 = 0.25$. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.