Chapter 12.2a Bayesian Multiple Regression

Jim Albert and Monika Hu

Chapter 13 Bayesian Multiple Regression and Logistic Models

A Multiple Regression Model

- Similar to a simple linear regression model, a multiple linear regression model assumes a observation specific mean μ_i for the i-th response variable Y_i.

$$Y_i \mid \mu_i, \sigma \overset{ind}{\sim} \mathrm{Normal}(\mu_i, \sigma), \; i = 1, \cdots, n.$$

- Assume that the mean of Y_i, μ_i, is a linear function of all predictors. One writes

$$\mu_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_r x_{i,r},$$

  where $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,r})$ is a vector of r known predictors for observation i, and $\beta = (\beta_0, \beta_1, \cdots, \beta_r)$ is a vector of unknown regression coefficients shared among all observations.

Interpretation for continuous predictors

- For studies where all r predictors are continuous, one interprets the intercept parameter β_0 as the expected response μ_i for observation i, where all of its predictors take values of 0
- One can also interpret the slope parameter β_i as the change in the expected response μ_i, when the j-th predictor, $x_{i,j}$, of observation i increases by a single unit while all remaining $(r-1)$ predictors are constant.

Categorical Predictors

- In the household expenditures example from the CE data sample, the urban/rural status variable is a binary categorical variable coded as 1 (urban) or 2 (rural).
- It is much more common to consider this variable as a binary (0 or 1) categorical variable that classifies the observations into two distinct groups: the urban group and the rural group.
- Define a new indicator variable that takes a value of 0 if the CU is in an urban area, and a value of 1 if the CU is in a rural area.

Regression with an indicator variable

- Consider a simplified regression model with a single predictor, the binary indicator for rural area x_i.
- This simple linear regression model is given by

$$\mu_i = \beta_0 + \beta_1 x_i = \begin{cases} \beta_0, & \text{the urban group;} \\ \beta_0 + \beta_1, & \text{the rural group.} \end{cases}$$

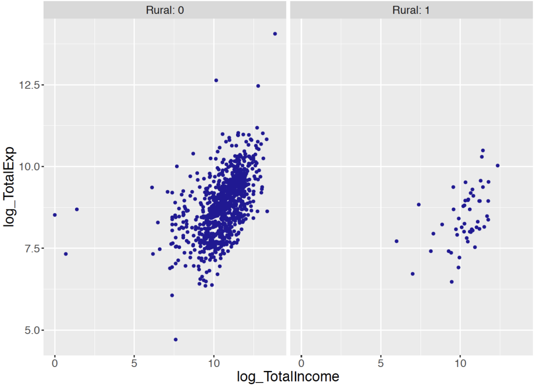
- Here β_1 represents the change in the expected response μ_i from the urban group to the rural group. That is, β_1 represents the effect of being a member of the rural group.

Some Data Transformation

- Both the expenditure and income variables are highly skewed.
- Both variables have more even distributions if we apply logarithm transformations.
- So the response variable will be the logarithm of the CU's total expenditure and the continuous predictor will be the logarithm of the CU 12-month income.

Graph

- Figure displays scatterplots of log income and log expenditure where the two panels correspond to urban and rural residents.
- In each panel there appears to be a positive association between log income and log expenditure.



Multiple Regression Model

- Set up a multiple linear regression model for the log expenditure response including one continuous predictor and one binary categorical predictor.
- Assume response

$$Y_i \sim N(\mu_i, \sigma)$$

- Expected response μ_i is expressed as a linear combination of the log income variable and the rural indicator variable.

$$\mu_i = \beta_0 + \beta_1 x_{i,income} + \beta_2 x_{i,rural}.$$

Interpret regression coefficients

- Intercept parameter β_0 is the expected log expenditure when $x_{i,income} = x_{i,rural} = 0$.
- This intercept β_0 represents the mean log expenditure for an urban CU with a log income of 0.
- The slope β_1 can be interpreted as the change in the expected log expenditure when the predictor log income of record i increases by one unit, while x_2 stays unchanged.
- The coefficient β_2 is the change in the expected log expenditure of a rural CU comparing to an urban CU, when the two CUs have the same log income.