COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

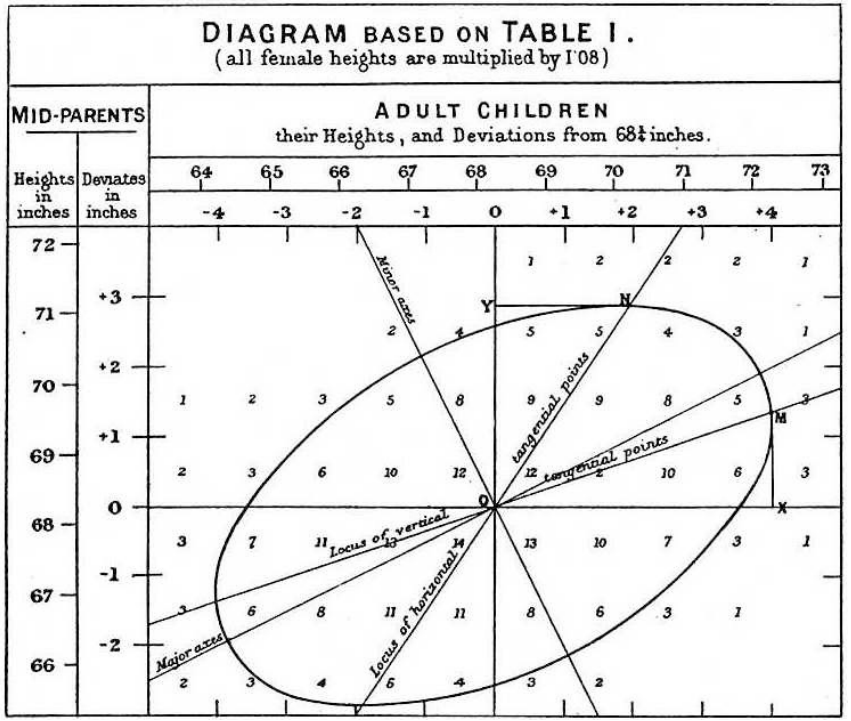
STAT 4224/5224

Bayesian Statistics

Dobrin Marchev

What is Regression Analysis?

A statistical method for estimating the relationship between an outcome or response variable, denoted Y, and one or more input, or predictor variables, denoted $X_1, \ldots, X_p$

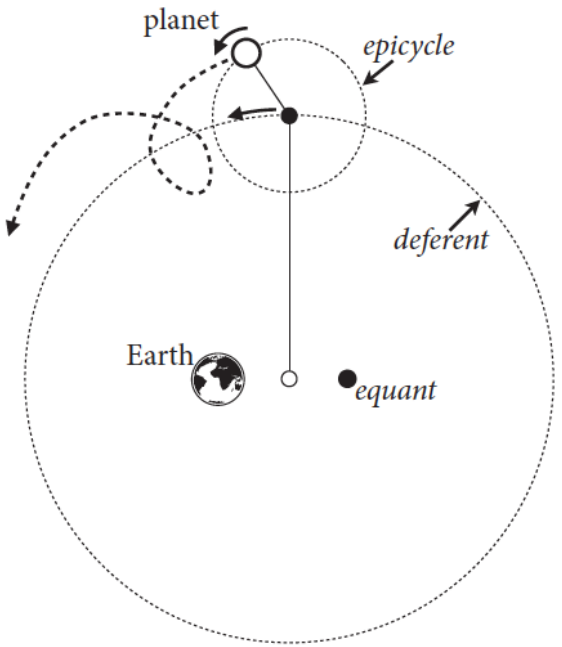


Galton Example – heights of adult children and parents

- Data set from tabulated data set used by Galton in 1885 to study the relationship between a parent's height and their children's height.
- Child = The child's height
- Parent = The "midparent" height
- The midparent's height is an average of the father's height and 1.08 times the mother's height. In the data there are 205 different parents and 928 children. The data here is truncated at the ends for both parents and children so that it can be treated as numeric data. The data were tabulated and consequently made discrete.

Models

- Representation of some phenomenon
- Mathematical model is a mathematical expression of the phenomenon
- Often describes relationships between variables
- Types
  - Deterministic models (Functional relation)
  - Probabilistic models (Statistical relation)

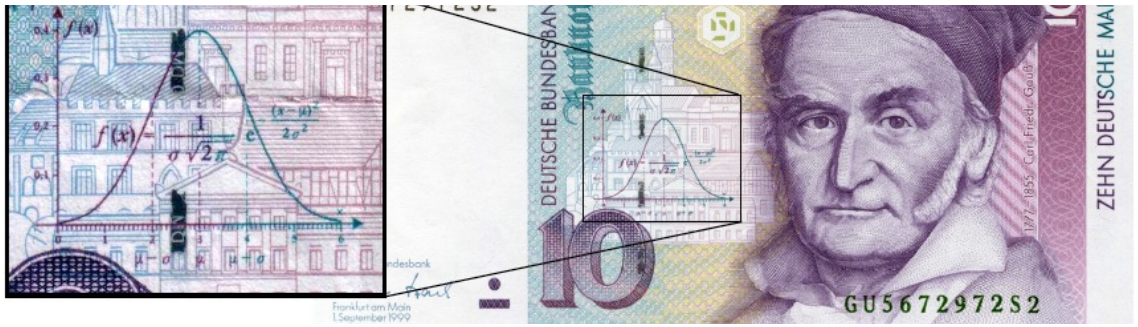


Probabilistic Models

- Probabilistic modeling is a statistical approach that uses the effect of random occurrences or actions to forecast the possibility of future results.
- Hypothesize two components: Deterministic part & random error
- Example: sales volume (y in \$1000) is 0.7 times advertising spending (x \$100) + random error:

$$y = 0.7x + \varepsilon$$

– Random error ε may be due to factors other than advertising, or even unforeseen events, and is assumed Gaussian or *normally* distributed.



General Linear Form of Probabilistic Models

Y = Deterministic component + Random error

where:

- Y is the output variable of interest;
- Deterministic component is aka structural assumption
- We assume that the mean value of the random error is 0
- This is equivalent to assuming that

$$E(Y \mid \boldsymbol{x}) = \text{Deterministic component}$$

A First-Order (Straight Line) Probabilistic Model or Simple Linear Regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where

y = **Dependent** *or* **response variable** (variable to be modeled)

x = **Independent** *or* **predictor variable** (variable used as a predictor of y)

$i = 1, \ldots, n$ is the i^{th} observation in the sample

Simple Linear Regression

$$y = \beta_0 + \beta_1 x + \varepsilon$$

The unknown parameters are:

β_0 (beta zero) = ***y*-intercept of the line**, that is, the point at which the line *intercepts or cuts through the y-axis*

β_1 (beta one) = **slope of the line**, that is, the change (amount of increase or decrease) in the deterministic component of y for every 1-unit increase in x

Simple Linear Regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$\beta_0 + \beta_1 x$ = Deterministic component

ε = Random error component, assumed normally distributed

Probabilistic model assumes $(X, Y) \sim N_2$. Then:

$$Y_i | X = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), i = 1, \ldots, n$$

Then the deterministic component is:

$$\mathrm{E}(Y_i | X = x_i) = \beta_0 + \beta_1 x_i, i = 1, \ldots, n$$

Models with Multiple Predictors

- All practical problems have *more than one* potential predictor variable
- The goal is to determine effects (if any) of each predictor, *controlling for the others*
- Can include polynomial terms to allow for *nonlinear* relations
- Can include product terms to allow for *interactions* when effect of one variable depends on level of another variable
- Can include "dummy" variables for *categorical* predictors

Linear Model with Two Numeric Predictors

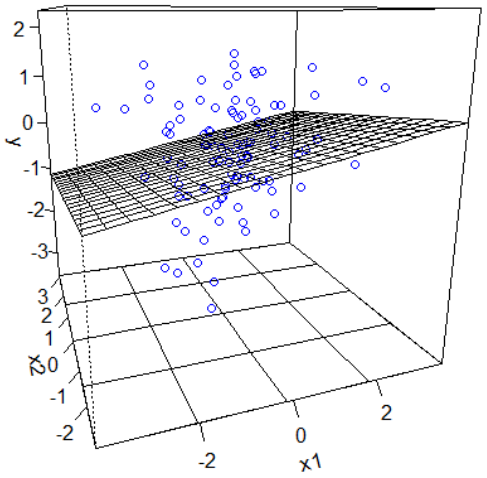
In the case of two predictors, the regression model is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_1 X_{i1} + \varepsilon_i$$

Assuming $E(\varepsilon_i) = 0$, the regression equation becomes

$$E(Y_i|x) = \beta_0 + \beta_1 X_{i1} + \beta_1 X_{i1}$$

Geometrically, this is a plane in 3-D.



Interpretation of Regression Coefficients

- Regression coefficients are more complicated to interpret with multiple predictors because the interpretation for any given coefficient is, in part, contingent on the other variables in the model.
- Additive: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \equiv$ Mean of Y @ X_1, X_2
- $\beta_0 \equiv$ Intercept, Mean of Y when $X_1 = X_2 = 0$
- $\beta_1 \equiv$ Slope with respect to X_1 (effect of increasing X_1 by 1 unit, while *holding X_2 constant*)
- $\beta_2 \equiv$ Slope with respect to X_2 (effect of increasing X_2 by 1 unit, while *holding X_1 constant*)

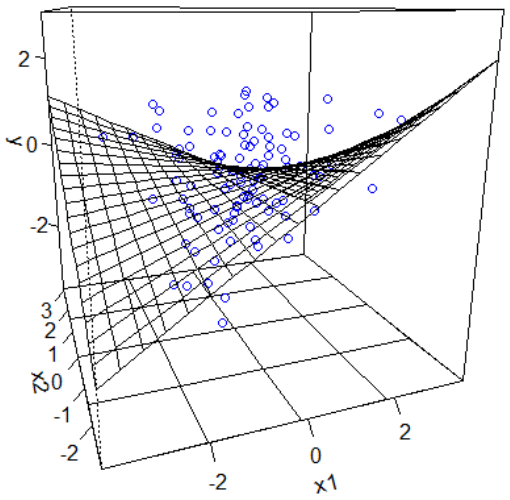
Interaction Model with Two Numeric Predictors

Interaction model:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2)$$

When $X_2 = 0$: Effect of increasing X_1 by 1: $\beta_1(1) + \beta_3(1)(0) = \beta_1$
When $X_2 = 1$: Effect of increasing X_1 by 1: $\beta_1(1) + \beta_3(1)(1) = \beta_1 + \beta_3$

That is, the effect of increasing X_1 *depends on* level of X_2, and vice versa
The geometric surface is called a "saddle".



General Linear Regression Model

Assume that the conditional distribution $f(y|\boldsymbol{x})$ changes smoothly as a function of $\boldsymbol{x}$, so that data we have at one value of $\boldsymbol{x}$ can inform us about what might be going on at a different value:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

$$\Rightarrow Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i$$

$$\Rightarrow Y_i = \sum_{k=0}^{p} \beta_k x_{ik} + \varepsilon_i, \text{where } x_{io} = 1$$

Assume E(ε_i) = 0

$$\Rightarrow E(Y|x) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \boldsymbol{\beta}'\boldsymbol{x}$$

This is an equation of a hyperplane in in p+1 dimensional space.

Note: When p = 1 we obtain the SLR case.

In addition, classic regression assumes normality, independence and constant variance:

$$\varepsilon_i \sim iid\ N(0, \sigma^2)$$

Likelihood

Under the assumption

$$Y_i = \sum_{k=0}^{p} \beta_k x_{ik} + \varepsilon_i, \varepsilon_i \sim iid\ N(0, \sigma^2)$$

This means that:

$$f(y_1, \dots, y_n | \boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f(y_i | \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2)$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{\beta}' \boldsymbol{x}_i)^2}$$

We need to develop matrix notation and express this model in terms of multivariate normal distribution.

Matrix Form of Multiple Regression

Model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ ... + \beta_p X_{ip} + \varepsilon_i, \qquad i = 1, ..., n$$

Matrix form: $\boldsymbol{Y}$ is $n\times 1$ vector and $\boldsymbol{X}$ is $n\times(p+1)$ matrix

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & ... & X_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & ... & X_{np} \end{bmatrix}$$

Then with $\boldsymbol{\beta}$ being $p\times 1$ vector and $\boldsymbol{\varepsilon}$ $n\times 1$ vector, we have:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \mathrm{E}(\boldsymbol{\varepsilon}) = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \boldsymbol{\sigma^2}(\boldsymbol{\varepsilon}) = \begin{bmatrix} \sigma^2 & 0 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \sigma^2 \end{bmatrix} = \sigma^2 \boldsymbol{I}$$

Model in matrix form:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \Rightarrow \mathrm{E}(\boldsymbol{Y}|\boldsymbol{X}) = \mathrm{E}(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{X\beta}, \sigma^2(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I}$$

That is,

$$\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$$

Least Squares Estimation of Regression Coefficients

The likelihood depends on $\boldsymbol{\beta}$ only through the residuals $r_i = y_i - \boldsymbol{\beta}'\boldsymbol{x}_i$. Therefore, it is maximized when when the sum of the squared residuals in the exponent is minimized. That is, the goal is to minimize

$$SSE(\boldsymbol{\beta}) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}'\boldsymbol{x}_i)^2 = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$$
$$= \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta}$$

Using matrix calculus, we obtain:

$$\frac{\partial\, SSE(\boldsymbol{\beta})}{\partial\, \boldsymbol{\beta}} = -2\boldsymbol{X}'\boldsymbol{y} + \mathbf{2}\boldsymbol{X}'\boldsymbol{X\beta} = \mathbf{0}$$
$$\Leftrightarrow \boldsymbol{X}'\boldsymbol{X\beta} = \boldsymbol{X}'\boldsymbol{y} \Leftrightarrow \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}'\boldsymbol{y}$$

The value $\widehat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}'\boldsymbol{y}$ is called the *ordinary least squares estimator* of $\boldsymbol{\beta}$. It is unique as long as $(\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}$ exists.

Fitted Values, Residuals and Covariances

Fitted values $\widehat{\boldsymbol{Y}}$ and residuals $\boldsymbol{e}$ are now $n\times 1$ vectors:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{Xb} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{HY}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the so-called "hat matrix", such that $\boldsymbol{H} = \boldsymbol{H}' = \boldsymbol{HH}$

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{Xb} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

The covariance matrices are:

$$\mathrm{cov}\left(\widehat{\boldsymbol{Y}}\right) = \mathrm{cov}(\boldsymbol{HY}) = \boldsymbol{H}\mathrm{cov}(\boldsymbol{Y})\boldsymbol{H}' = \sigma^2\boldsymbol{H}$$

where $\sigma^2 = \mathrm{Var}(\varepsilon_i)$, so the covariance of $\widehat{\boldsymbol{Y}}$ is estimated with

$$\widehat{\mathrm{cov}}\left(\widehat{\boldsymbol{Y}}\right) = s^2\left(\widehat{\boldsymbol{Y}}\right) = \mathrm{MSE} * \boldsymbol{H}$$

Similarly, for the residuals:

$$\mathrm{cov}(\boldsymbol{e}) = \mathrm{cov}((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}) = (\boldsymbol{I} - \boldsymbol{H})\mathrm{cov}(\boldsymbol{Y})(\boldsymbol{I} - \boldsymbol{H})' = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$

so, the covariance of $\boldsymbol{e}$ is estimated with

$$\widehat{\mathrm{cov}}(\boldsymbol{e}) = s^2(\boldsymbol{e}) = \mathrm{MSE} * (\boldsymbol{I} - \boldsymbol{H})$$

Bayesian Estimation

The likelihood function is:

$$f(\boldsymbol{y}\,|\boldsymbol{X},\boldsymbol{\beta},\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}SSE(\boldsymbol{\beta})}$$
$$= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\boldsymbol{y'y}-2\boldsymbol{\beta' X' y}+\boldsymbol{\beta' X' X\beta})}$$

We will use a multivariate conjugate prior on $\boldsymbol{\beta} \sim \mathrm{N}_{p+1}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$

Then

$$f(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},\sigma^2) \propto f(\boldsymbol{y}\,|\boldsymbol{X},\boldsymbol{\beta},\sigma^2) \times \pi(\boldsymbol{\beta})$$
$$\propto e^{-\frac{1}{2}\left(-\frac{2\boldsymbol{\beta' X' y}}{\sigma^2}+\frac{\boldsymbol{\beta' X' X\beta}}{\sigma^2}\right)-\frac{1}{2}\left(-\frac{2\boldsymbol{\beta'\Sigma_0^{-1}\beta_0}}{\sigma^2}+\frac{\boldsymbol{\beta'\Sigma_0^{-1}\beta}}{\sigma^2}\right)}$$
$$= e^{\boldsymbol{\beta'}\left(\boldsymbol{\Sigma_0^{-1}\beta_0}+\frac{\boldsymbol{X'y}}{\sigma^2}\right)-\frac{1}{2}\boldsymbol{\beta'}\left(\boldsymbol{\Sigma_0^{-1}}+\frac{\boldsymbol{X'X}}{\sigma^2}\right)\boldsymbol{\beta}}$$

You can recognize this a multivariate normal distribution.

Bayesian Estimation

Since

$$f(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},\sigma^2) \propto e^{\boldsymbol{\beta}'\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0+\frac{\boldsymbol{X}'\boldsymbol{y}}{\sigma^2}\right)-\frac{1}{2}\boldsymbol{\beta}'\left(\boldsymbol{\Sigma}_0^{-1}+\frac{\boldsymbol{X}'\boldsymbol{X}}{\sigma^2}\right)\boldsymbol{\beta}}$$

we can conclude that:

$$Var(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},\sigma^2) = \left(\boldsymbol{\Sigma}_0^{-1} + \frac{\boldsymbol{X}'\boldsymbol{X}}{\sigma^2}\right)^{-1}$$

$$E(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},\sigma^2) = \left(\boldsymbol{\Sigma}_0^{-1} + \frac{X'X}{\sigma^2}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\beta_0 + \frac{X'y}{\sigma^2}\right)$$

Note that if the prior precision matrix $\boldsymbol{\Sigma}_0^{-1}$ is "small", then the the conditional expectation $E(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},\sigma^2)$ is approximately $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, which is the OLS estimator.

Bayesian Estimation

For σ^2 we will again use a semiconjugate prior:

$$\sigma^2 \sim IG\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$$

Then:

$$f(\sigma^2|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}) \propto \pi(\sigma^2) f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0}{2}-1} e^{-\frac{1}{\sigma^2}\frac{\nu_0\sigma_0^2}{2}} \times \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{1}{\sigma^2}\frac{SSE(\beta)}{2}}$$

$$= \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0+n}{2}-1} e^{-\frac{1}{\sigma^2}\left(\frac{\nu_0\sigma_0^2}{2}+\frac{SSE(\beta)}{2}\right)}$$

$$\Rightarrow \sigma^2|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta} \sim \text{IG}\left(\frac{\nu_0 + n}{2}, \frac{\nu_0\sigma_0^2 + SSE(\boldsymbol{\beta})}{2}\right)$$

Gibbs Sampler

Constructing a Gibbs sampler to approximate the joint posterior distribution $f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X})$ is then straightforward: Given current values $\{\boldsymbol{\beta}^{(s)}, \sigma^{2(s)}\}$, new values can be generated by:

1. Updating $\boldsymbol{\beta}$:

a) Compute $V = Var\left(\boldsymbol{\beta}\middle|\boldsymbol{y}, \boldsymbol{X}, \sigma^{2(s)}\right), m = E\left(\boldsymbol{\beta}\middle|\boldsymbol{y}, \boldsymbol{X}, \sigma^{2(s)}\right)$

b) Sample $\beta^{(s+1)} \sim N_{p+1}(m, V)$

2. Updating σ^2

a) Compute SSE($\boldsymbol{\beta}^{(s+1)}$)

b) Sample $\sigma^{2(s+1)} \sim IG\left(\frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2+SSE\left(\boldsymbol{\beta}^{(s+1)}\right)}{2}\right)$

Weakly Informative Priors

Sometimes an analysis must be done in the absence of precise prior information, or information that is easily converted into the parameters of a conjugate prior distribution. One idea is that, if the prior distribution is not going to represent real prior information about the parameters, then it should be as minimally informative as possible. To some, such an analysis would give a "more objective" result than using an informative prior distribution. One type of weakly informative prior is the *unit information prior*. A unit information prior is one that contains the same amount of information as that would be contained in only a single observation. For example, the precision of $\widehat{\boldsymbol{\beta}}_{OLS}$ is the inverse of variance or $\frac{\boldsymbol{X'X}}{\sigma^2}$. Since this can be viewed as the amount of information in *n* observations, the amount of information in one observation should be $\frac{\boldsymbol{X'X}}{n\sigma^2}$. The unit information prior thus sets $\Sigma_0^{-1} = \frac{\boldsymbol{X'X}}{n\sigma^2}$. Further suggestion is to set $\boldsymbol{\beta_0} = \widehat{\boldsymbol{\beta}}_{OLS}$. For σ^2 we can choose $\nu_0 = 1$ and $\sigma_0^2 = \hat{\sigma}_{OLS}^2$.