Applied Statistical Methods - Exercise 6

Peter von Rohr

2020-03-30

Problem 1: Bayesian Regression Analysis

Given is the earlier used dataset of `breast circumference` and `body weight`.

Table 1: Dataset for Regression of Body Weight on Breast Circumference for ten Animals

Animal	Breast Circumference	Body Weight
1	176	471
2	177	463
3	178	481
4	179	470
5	179	496
6	180	491
7	181	518
8	182	511
9	183	510
10	184	541

The model that is used is a simple linear regression model given by

$$y_i = \beta_0 + \beta_1 * x_i + \epsilon_i$$

.

where y_i corresponds to the body weight of animal i, x_i is the breast circumference of animal i, β_0 is the unknown intercept and β_1 is the unknown regression coefficient. For reasons of simplicity, we assume the residual variance σ^2 to be known. For the later computations, we insert the estimate that is obtained from the `lm()` function. This value corresponds to $\sigma^2 = 122.8$.

Bayesian Estimation Of Unknowns

As already mentioned during the lecture, Bayesian estimates of unknowns are based on the posterior distribution of the unknowns given the knowns. For our regression model the unknowns correspond to

$$\beta = \left[\begin{array}{c} \beta_0 \\ \beta_1 \end{array} \right]$$

The posterior distribution of the unknowns given the knowns is $f(\beta|y)$. Using Bayes' Theorem we can write $f(\beta|y)$ as

$$
\begin{aligned}
f(\beta|y) &= \frac{f(\beta, y)}{f(y)} \\
&= \frac{f(y|\beta)f(\beta)}{f(y)} \\
&\propto f(y|\beta)f(\beta)
\end{aligned}
$$

When we do not have any specific prior knowledge about β, the prior distribution $f(\beta)$ for the unknown β is set to a constant. Therefore we can write

$$
\begin{aligned}
f(\beta|y) &\propto f(y|\beta)f(\beta) \\
&\propto f(y|\beta)
\end{aligned}
$$

Assuming a normal distribution for the data causes the likelihood $f(y|\beta)$ to be a multivariate normal distribution.

$$
\begin{aligned}
f(\beta|y) &\propto f(y|\beta) \\
&= (2\pi\sigma^2)^{-n/2} exp\left\{-\frac{1}{2}\frac{(y - X\beta)^T(y - X\beta)}{\sigma^2}\right\}
\end{aligned} \tag{1}
$$

The above expression (1) is an $n-$ dimensional normal distribution with expected value $X\beta$ and variance-covariance matrix corresponding to $I\sigma^2$. But because we have just two unknowns β_0 and β_1 the posterior distribution $f(\beta|y)$ must have two dimensions and not n. The following re-arrangement can solve this problem. Let us set the variable Q to

$$
Q = (y - X\beta)^T(y - X\beta) = y^Ty - 2y^TX\beta + \beta^T(X^TX)\beta
$$

Introducing the least squares estimate $\hat{\beta} = (X^TX)^{-1}X^Ty$ into the above equation by replacing y^TX with $\hat{\beta}^T(X^TX)$ results in

$$
Q = y^Ty - 2\hat{\beta}^T(X^TX)\beta + \beta^T(X^TX)\beta = y^Ty + (\beta - \hat{\beta})^T(X^TX)(\beta - \hat{\beta}) - \hat{\beta}^T(X^TX)\hat{\beta}
$$

Inserting this last result back into (1) gives

$$
\begin{aligned}
f(\beta|y) &\propto f(y|\beta) \\
&= (2\pi\sigma^2)^{-n/2} exp\left\{-\frac{1}{2}\frac{(y - X\beta)^T(y - X\beta)}{\sigma^2}\right\} \\
&= (2\pi\sigma^2)^{-n/2} exp\left\{-\frac{1}{2}\frac{y^Ty + (\beta - \hat{\beta})^T(X^TX)(\beta - \hat{\beta}) - \hat{\beta}^T(X^TX)\hat{\beta}}{\sigma^2}\right\} \\
&= (2\pi\sigma^2)^{-n/2}\left[exp\left\{-\frac{1}{2}\frac{y^Ty}{\sigma^2}\right\} * exp\left\{-\frac{1}{2}\frac{(\beta - \hat{\beta})^T(X^TX)(\beta - \hat{\beta})}{\sigma^2}\right\} * exp\left\{-\frac{1}{2}\frac{-\hat{\beta}^T(X^TX)\hat{\beta}}{\sigma^2}\right\}\right] \\
&\propto exp\left\{-\frac{1}{2}\frac{(\beta - \hat{\beta})^T(X^TX)(\beta - \hat{\beta})}{\sigma^2}\right\}
\end{aligned} \tag{2}
$$

The last proportionality results from the fact that only the term depending on β is retained. All other terms not depending on β are constant factors with respect to β and can therefore be dropped. Thus $f(\beta|y)$ can be written as

$$f(\beta|y) \propto exp\left\{-\frac{1}{2}\frac{(\beta-\hat{\beta})^T(X^TX)(\beta-\hat{\beta})}{\sigma^2}\right\}$$

which is recognized as proportional to a two dimensional normal density with mean $\hat{\beta}$ and variance $(X^TX)^{-1}\sigma^2$. Thus in the simple setting the mean of the posterior mean can already be seen from the above formula. But in a more complex setting, the posterior distribution does not have a standard form and we need to setup a sampling scheme which allows us to draw random numbers from the posterior distribution. The sampling scheme that we are introducing here is called the **Gibbs Sampler**.

Gibbs Sampler for β

The simple regression model that we are using for the breast circumference and the body weight data can be written in matrix-vector notation as

$$y = 1\beta_0 + x\beta_1 + \epsilon$$

In the Gibbs sampling scheme both unknowns β_0 and β_1 are sampled from their full conditional distributions. For β_0 the full conditional posterior distribution is $f(\beta_0|\beta_1, y)$ which is computed for the current value of β_1. Separating β_0 from the other unknowns yields the linear model

$$w_0 = 1\beta_0 + \epsilon$$

where $w_0 = y - x\beta_1$. The least squares estimator of β_0 is

$$\hat{\beta}_0 = (1^T1)^{-1}1^Tw_0$$

with variance

$$var(\hat{\beta}_0) = (1^T1)^{-1}\sigma^2$$

Applying the same strategy as for $f(\beta|y)$, it can be shown that $f(\beta_0|\beta_1, y)$ is a normal distribution with mean $\hat{\beta}_0$ as mean and $(1^T1)^{-1}\sigma^2$ as variance. The full-conditional posterior of β_1 can be derived the same way, leading to

$$\hat{\beta}_1 = (x^Tx)^{-1}x^Tw_1$$

with variance $var(\hat{\beta}_1) = (x^Tx)^{-1}\sigma^2$ where $w_1 = y - 1\beta_0$.

Your Task

- Create a Gibbs Sampling scheme for the dataset shown in Table 1.
- Use the mean of the generated samples as an estimate for the unknowns β_0 and β_1.