Appendix: deriving the conditional posteriors for linear regression

Jasper Ginn

February 18, 2019

- https://stattrek.com/ for matrix algebra
- set up thinning in MCMC
- set up multiple chains in MCMC

1 Preliminaries

1.1 Recognizing a normal distribution

A random variable x is normally distributed when $X \sim N(\mu, \sigma^2)$, or:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]} \tag{1.1.1}$$

Many terms in the equation have the function of a normalizing constant. That is, they make the distribution a proper distribution (adhering to the laws of probability). But in the Bayesian framework, we are not necessarily concerned with these normalizing constants.

In the case of the normal, we will drop any terms that do not contain the random variable x

$$\begin{aligned} f(x) &\propto e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]} \\ &\propto e^{-\left[\frac{(x^2-2x\mu+\mu^2)}{2\sigma^2}\right]} \\ &\propto e^{-\left[\frac{x^2}{2\sigma^2} - \frac{x\mu)}{\sigma^2}\right]} \end{aligned} \tag{1.1.2}$$

Notice that we can drop μ^2 terms because, with respect to x, this term is just a constant. Put another way, by the exponent properties we know that $f(x) = e^{x+Q} = e^x e^Q \propto e^x$. This result leads us to a general form of a normal distribution

$$e^{\left[-Ax^2+Bx\right]} \tag{1.1.3}$$

where

- $A = \frac{1}{2\sigma^2}$
- $B = \frac{\mu}{\sigma^2}$

Hence, if we see $f(x) \sim e^{\left[-Ax^2+Bx\right]}$ with $A > 0$ then we should **recognize a normal distribution**. To derive its parameters μ and σ^2 in this form, we need to manipulate the distribution such that we get (μ, σ^2) from (A, B).

$$\begin{aligned} A &= \frac{1}{2\sigma^2} \longrightarrow \color{blue}{\sigma^2} = \frac{1}{2A} \\ B &= \frac{\mu}{\sigma^2} \longrightarrow \mu = B\color{blue}{\sigma^2} = \frac{B}{2A} \end{aligned} \tag{1.1.4}$$

1.2 Recognizing an inverse gamma distribution

The *inverse gamma* distribution is given by

$$IG(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\frac{\beta}{x}} \tag{1.2.1}$$

Dropping all terms that do not contain x, we get

$$IG(x; \alpha, \beta) \propto x^{-\alpha-1} e^{-\frac{\beta}{x}} \tag{1.2.2}$$

We say a random variable x has an inverse gamma distribution if $f(x) \sim x^A e^{\left[\frac{B}{x}\right]}$ with $B < 0, A < 0$. In this case, we can retrieve the shape parameter α and the scale parameter β by calculating

$$\begin{aligned} \alpha &= -A - 1 \\ \beta &= -B \end{aligned} \tag{1.2.3}$$

1.3 Linear regression equation

The basic linear regression model with two predictors is given by

$$\hat{y}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i \tag{1.3.1}$$

Where:

- $\hat{y}_i$ is the predicted value for the outcome variable for the i^{th} individual
- β_0 is the intercept of the linear model
- β_1, β_2 are the slope coefficients for the linear model
- X_{1i}, X_{2i} are the independent variables values for the i^{th} individual
- e_i is the residual error associated with the i^{th} individual. We assume $e \sim N(0, \sigma^2)$

The parameters of this model are the intercept β_0, the slope coefficients β_1, β_2 and the residual variance σ^2.

As in any bayesian model, we are looking for the posterior

$$f(\text{parameters}|\text{data}) \propto f(\text{data}|\text{parameters}) f(\text{parameters}) \tag{1.3.2}$$

where:

- $f(\text{parameters}|\text{data})$ is the *joint posterior distribution* of the parameters
- $f(\text{data}|\text{parameters})$ is the *likelihood of the data* given the parameters
- $f(\text{parameters})$ is the *joint prior distribution* of the parameters

For the linear regression case, this yields

$$f(\beta_0, \beta_1, \beta_2, \sigma^2|y, X) \propto f(y|X, \beta_0, \beta_1, \beta_2, \sigma^2) f(\beta_0, \beta_1, \beta_2, \sigma^2) \tag{1.3.3}$$

We assume that the priors are independent, and as such, we can state that

$$f(\beta_0, \beta_1, \beta_2, \sigma^2) = f(\beta_0) f(\beta_1) f(\beta_2) f(\sigma^2) \tag{1.3.4}$$

2 Defining the likelihood of the data

Let $k_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$. We assume that the outcome variable y is normally distributed, or $y_i \sim N(k_i, \sigma^2)$. Hence, for the i^{th} example in the data, we can represent the likelihood for this example as

$$f(y_i|x_{1i}, x_{2i}, \beta_0, \beta_1, \beta_2, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left[\frac{(y_i-k_i)^2}{2\sigma^2}\right]} \tag{2.1.1}$$

By virtue of independence, we construct the likelihood of all N examples in the data as

$$f(\text{data}|\text{parameters}) = f(y|x_1, x_2, \beta_0, \beta_1, \beta_2, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left[\frac{(y_i-k_i)^2}{2\sigma^2}\right]} \tag{2.1.2}$$

3 Specifying the prior distribution

For each of the parameters $\beta_0, \beta_1, \beta_2$, we assume a normal prior distribution. For example, the slope coefficient is assumed to be distributed as $f(\beta_0) \sim N(\mu_{0,0}, \tau_{0,0}^2)$, or:

$$f(\beta_0) = \frac{1}{\sqrt{2\pi\tau_{0,0}^2}} e^{-\left[\frac{(\beta_0-\mu_{0,0})^2}{2\tau_{0,0}^2}\right]} \tag{3.1.1}$$

Where the hyperparameters $\tau_{0,0}^2, \mu_{0,0}$ are defined as:

- $\tau_{0,0}^2$ is the prior variance for the intercept
- $\mu_{0,0}$ is the prior mean for the intercept.

For the parameter σ^2, we assume an inverse gamma prior distribution.

$$IG(x; \alpha_0, \beta_0) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} x^{-\alpha_0-1} e^{-\frac{\beta_0}{x}} \tag{3.1.2}$$

Where the hyperparameters α_0 and β_0 are defined as:

- α_0 is the prior shape of the distribution.
- β_0 is the prior scale of the distribution.

4 Deriving the conditional posteriors for each parameter

We can obtain posterior distributions for each of the parameters by constructing their *conditional posterior distributions*. Hence, we want to find

$$
\begin{aligned}
&f(\beta_0|y, X, \beta_1, \ldots, \beta_j, \sigma^2) \propto f(y|X, \beta_0, \beta_1, \ldots, \beta_j, \sigma^2) \times f(\beta_0) \\
&f(\beta_1|y, X, \beta_0, \ldots, \beta_j, \sigma^2) \propto f(y|X, \beta_0, \beta_1, \ldots, \beta_j, \sigma^2) \times f(\beta_1) \\
&\ldots \\
&f(\beta_j|y, X, \beta_0, \beta_1, \ldots, \beta_{j-1}, \sigma^2) \propto f(y|X, \beta_0, \beta_1, \ldots, \beta_j, \sigma^2) \times f(\beta_j) \\
&f(\sigma^2|y, X, \beta_0, \beta_1, \ldots, \beta_j) \propto f(y|X, \beta_0, \beta_1, \ldots, \beta_j, \sigma^2) \times f(\sigma^2)
\end{aligned}
\tag{4.1}
$$

4.1 The conditional distribution for the intercept β_0

Here, we are making the posterior conditional on all parameters other than β_0, and hence we are effectively turning the joint prior distribution into a single prior distribution.

Plugging in the likelihood and prior distribution into (6), we get

$$
f(\beta_0|\ldots) \propto \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left[\frac{(y_i-k_i)^2}{2\sigma^2}\right]}\right] \times \frac{1}{\sqrt{2\pi\tau_{0,0}^2}} e^{-\left[\frac{(\beta_0-\mu_{0,0})^2}{2\tau_{0,0}^2}\right]}
\tag{4.1.1}
$$

To arrive at the conditional posterior for β_0, we simply drop all elements that do not contain β_0. We can do this because, in the conditional posterior distribution, such terms are nothing but normalization constants.

First, we drop the leading coefficients from (7)

$$
f(\beta_0|\ldots) \propto \left[\prod_{i=1}^{N} e^{-\left[\frac{(y_i-k_i)^2}{2\sigma^2}\right]}\right] \times e^{-\left[\frac{(\beta_0-\mu_{0,0})^2}{2\tau_{0,0}^2}\right]}
\tag{4.1.2}
$$

Next, we examine each of the exponents separately and expand the factored quadratic in the numerator. If we forget about the product for a moment and focus on a single example of the likelihood, we see

$$
\begin{aligned}
&e^{-\left[\frac{(y_i-k_i)^2}{2\sigma^2}\right]} \longrightarrow \\
&\frac{(y_i-k_i)^2}{2\sigma^2} = \\
&\frac{1}{2\sigma^2}\left[y_i^2 - 2y_ik_i + k_i^2\right] = \\
&\frac{1}{2\sigma^2}\left[y_i^2 - 2y_i[\beta_0 + \beta_1X_{1i} + \beta_2X_{2i}] + (\beta_0 + \beta_1X_{1i} + \beta_2X_{2i})^2\right]
\end{aligned}
\tag{4.1.3}
$$

Expanding the terms and dropping all terms without β_0 yields

$$f(y|\dots) \propto \prod_{i=1}^{N} e^{\left[-\frac{\beta_0^2}{2\sigma^2}+\beta_0\frac{y_i-\beta_1 x_{1i}-\beta_2 x_{2i}}{\sigma^2}\right]} \tag{4.1.4}$$

If we repeat the above procedure for the prior distribution $f(\beta_0)$, we get

$$f(\beta_0) \propto e^{\left[-\frac{\beta_0^2}{2\tau_{0,0}^2}+\frac{\beta_0\mu_{0,0}}{\tau_{0,0}^2}\right]} \tag{4.1.5}$$

Recall the following exponent rules

- $e^a e^b = e^{[a+b]}$
- $\prod_{i=1}^{N} e^{[i+j]} = e^{\left[\left(\sum_{i=1}^{N} i\right)+\left(\sum_{i=1}^{N} j\right)\right]} = e^{\left[\left(\sum_{i=1}^{N} i\right)+Nj\right]}$

Applying these rules to (9) and (10) and factoring the exponent yields

$$\begin{aligned} f(\beta_0|\dots) &\propto e^{\left[-\frac{\beta_0^2 N}{2\sigma^2}+\beta_0\frac{\sum_{i=1}^{N} y_i-\beta_1 x_{1i}-\beta_2 x_{2i}}{\sigma^2}\right]} \times e^{\left[-\frac{\beta_0^2}{2\tau_{0,0}^2}+\frac{\beta_0\mu_{0,0}}{\tau_{0,0}^2}\right]} \\ &\propto e^{\left[-\frac{\beta_0^2 N}{2\sigma^2}+\beta_0\frac{\sum_{i=1}^{N} y_i-\beta_1 x_{1i}-\beta_2 x_{2i}}{\sigma^2}-\frac{\beta_0^2}{2\tau_{0,0}^2}+\frac{\beta_0\mu_{0,0}}{\tau_{0,0}^2}\right]} \\ &\propto e^{\left[-\beta_0^2\left(\frac{N}{2\sigma^2}+\frac{1}{2\tau_{0,0}^2}\right)+\beta_0\left(\frac{\sum_{i=1}^{N} y_i-\beta_1 x_{1i}-\beta_2 x_{2i}}{\sigma^2}+\frac{\mu_{0,0}}{\tau_{0,0}^2}\right)\right]} \end{aligned} \tag{4.1.6}$$

This we should recognize as the form

$$e^{\left[-Ax^2+Bx\right]} \tag{4.1.7}$$

Recall that:

$$\begin{aligned} A &= \frac{1}{2\sigma^2} \longrightarrow \sigma^2 = \frac{1}{2A} \\ B &= \frac{\mu}{\sigma^2} \longrightarrow \mu = B\sigma^2 = \frac{B}{2A} \end{aligned} \tag{4.1.8}$$

And so we get

$$
\begin{aligned}
&\tau^2_{0,1}\frac{1}{\left(\frac{N}{2\sigma^2}+\frac{1}{2\tau^2_{0,0}}\right)}\\
&\mu_{0,1}\frac{\left(\frac{\sum_{i=1}^{N} y_i-\beta_1 x_{1i}-\beta_2 x_{2i}}{\sigma^2}+\frac{\mu_{0,0}}{\tau^2_{0,0}}\right)}{\left(\frac{N}{\sigma^2}+\frac{1}{\tau^2_{0,0}}\right)}
\end{aligned}
\tag{4.1.9}
$$

From (12) we can see the effect of the prior distribution. If the prior variance $\tau^2_{0,0}$ is large relative to the prior mean $\mu_{0,0}$, then the posterior distribution for $\mu_{0,1}$ will depend mainly on the likelihood of the data. In the case of the posterior $\tau^2_{0,1}$, we see that it reduces to

$$
\tau^2_{0,1}=\frac{1}{\frac{N}{2\sigma^2}}=\frac{2\sigma^2}{N}=\sqrt{2}\frac{\sigma}{\sqrt{N}} \tag{4.1.10}
$$

which is simply the variance scaled by the number of examples in the data and similar to the standard error obtained by the central limit theorem.

4.2 The conditional distribution for the slope coefficients

The slope coefficients $\beta_1, \ldots, \beta_j$ are derived similarly to the intercept coefficient β_0. For the moment, assume that β_j is the *last slope coefficient*. That is, assume that $j = \max(j)$

First, we define the posterior for β_j as the product of the likelihood of the data and the prior distribution for β_j and drop all terms in the prior distribution that do not contain this parameter

$$
f(\beta_j|\beta_0,\beta_1,\ldots,\beta_j,\sigma^2)\propto\left[\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\left[\frac{(y_i-k_i)^2}{2\sigma^2}\right]}\right]\times\frac{1}{\sqrt{2\pi\tau^2_{j,0}}}e^{-\left[\frac{(\beta_j-\mu_{j,0})^2}{2\tau^2_{j,0}}\right]} \tag{1}
$$

Next, we drop all terms that do not contain β_j from the prior distribution

$$
f(\beta_j|\ldots)\propto\left[\prod_{i=1}^{N}e^{-\left[\frac{(y_i-k_i)^2}{2\sigma^2}\right]}\right]\times e^{-\left[\frac{(\beta_j-\mu_{j,0})^2}{2\tau^2_{j,0}}\right]} \tag{1}
$$

Just like in equation (4.1.3) we expand the part in green and keep all elements that contain β_j

$$
\begin{aligned}
(y_i - k_i)^2 &= y_i^2 - 2y_ik_i + k_i^2 \\
&= y_i^2 - 2y_i[\beta_0 + \beta_1 x_{1i} + \cdots + \beta_j x_{ji}] + (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_j x_{ji})^2 \\
&\propto -2y_i\beta_j x_{ji} + \left[(\beta_0^2 + \beta_0\beta_1 x_{1i} + \cdots + \beta_0\beta_j x_{ji}) + \cdots + ([\beta_j x_{ji}]^2 + \beta_0\beta_j x_{ji} + \cdots + \beta_{j-1}x_{(j-1)i}\beta_j x_{ji})\right] \\
&\propto -2y_i\beta_j x_{ji} + (\beta_j x_{ji})^2 + 2\beta_0\beta_j x_{ji} + 2\beta_1 x_{1i}\beta_j x_{2i} + \cdots + 2\beta_{j-1}x_{(j-1)i}\beta_j x_{ji} \\
&\propto (\beta_j x_{ji})^2 - 2\beta_j x_{ji}(y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_{j-1}x_{(j-1)i})
\end{aligned}
$$

Plugging this back into equation XX and expanding the exponential containing the priors yields

$$
\begin{aligned}
f(\beta_j|\dots) &\propto \left[\prod_{i=1}^{N} e^{-\left[\frac{(\beta_j x_{ji})^2 - 2\beta_j x_{ji}(y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_{j-1}x_{(j-1)i})}{2\sigma^2}\right]}\right] \times e^{\left[-\frac{\beta_j^2}{2\tau_{j,0}^2} + \frac{\beta_j \mu_{j,0}}{\tau_{j,0}^2}\right]} \\
&\propto \left[\prod_{i=1}^{N} e^{\left[-\frac{(\beta_j x_{ji})^2}{2\sigma^2} + \beta_j x_{ji}\frac{y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_{j-1}x_{(j-1)i}}{\sigma^2}\right]}\right] \times e^{\left[-\frac{\beta_j^2}{2\tau_{j,0}^2} + \frac{\beta_j \mu_{j,0}}{\tau_{j,0}^2}\right]}
\end{aligned}
$$

Applying exponent properties to YY and combining the terms yields

$$
\begin{aligned}
f(\beta_j|\dots) &\propto e^{\left[-\frac{\sum_i (\beta_j x_{ji})^2}{2\sigma^2} + \beta_j x_{ji}\frac{\sum_i (y_i - \beta_0 - \left[\sum_{k=1}^{j-1} \beta_k x_{ki}\right])}{\sigma^2}\right]} \times e^{\left[-\frac{\beta_j^2}{2\tau_{j,0}^2} + \frac{\beta_j \mu_{j,0}}{\tau_{j,0}^2}\right]} \\
&\propto e^{\left[\left\{-\frac{\sum_i (\beta_j)^2 (x_{ji})^2}{2\sigma^2} - \frac{\beta_j^2}{2\tau_{j,0}^2}\right\} + \left\{\beta_j x_{ji}\frac{\sum_i (y_i - \beta_0 - \left[\sum_{k=1}^{j-1} \beta_k x_{ki}\right])}{\sigma^2} + \frac{\beta_j \mu_{j,0}}{\tau_{j,0}^2}\right\}\right]} \\
&\propto e^{\left[-\beta_j^2\left\{\frac{\sum_i x_{ji}^2}{2\sigma^2} + \frac{1}{2\tau_{j,0}^2}\right\} + \beta_j\left\{\frac{x_{ij}\sum_i (y_i - \beta_0 - \left[\sum_{k=1}^{j-1} \beta_k x_{ki}\right])}{\sigma^2} + \frac{u_{j,0}}{\tau_{j,0}^2}\right\}\right]}
\end{aligned}
$$

Again, we should recognize this as the form

$$
e^{\left[-Ax^2 + Bx\right]} \tag{4.1.7}
$$

Recall that:

$$
\begin{aligned}
A &= \frac{1}{2\sigma^2} \longrightarrow \sigma^2 = \frac{1}{2A} \\
B &= \frac{\mu}{\sigma^2} \longrightarrow \mu = B\sigma^2 = \frac{B}{2A}
\end{aligned} \tag{4.1.8}
$$

And so we get

$$
\begin{aligned}
\tau_{j,1}^2 &= \frac{1}{\left(\frac{\sum_i x_{ji}^2}{2\sigma^2} + \frac{1}{2\tau_{j,0}^2}\right)} \\
\mu_{j,1} &= \frac{\left(\frac{x_{ji}\sum_i (y_i - \beta_0 - \left[\sum_{k=1}^{j-1} \beta_k x_{ki}\right])}{\sigma^2} + \frac{u_{j,0}}{\tau_{j,0}^2}\right)}{\left(\frac{N x_{ji}^2}{\sigma^2} + \frac{1}{\tau_{0,0}^2}\right)}
\end{aligned} \tag{4.1.9}
$$

Finally, we need to generalize this result to cases where $j \neq \max(j)$. Assume that we have $j = 8$ parameters, then the above equation works fine if $j = 8$. For $j < 8$ it does not work. For example, if we have $j = 4$, then the posterior mean $\mu_{4,1}$ does not depend on all coefficients except β_4 (which is what we desire). Rather, it depends on $\beta_1, \beta_2, \beta_3$ only. The simplest way to get around this problem is to rewrite B such that we exclude the j^{th} coefficient in the summation

$$
\mu_{j,1} \frac{\left(\frac{\sum_i x_{ji}(y_i - \beta_0 - \left[\sum_{k\neq j} \beta_k x_{ki}\right])}{\sigma^2} + \frac{u_{j,0}}{\tau_{j,0}^2}\right)}{\left(\frac{N x_{ji}^2}{\sigma^2} + \frac{1}{\tau_{j,0}^2}\right)}
$$

4.3 The conditional distribution for $sigma^2$

The process to find the posterior for σ^2 is the same as for the previous parameters, with the difference being the prior $f(\sigma^2)$, which is inverse-gamma distributed. Again, we drop all terms that do not contain the parameter σ^2, which yields

$$
\begin{aligned}
f(\sigma^2|\dots) &\propto \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left[\frac{(y_i-k_i)^2}{2\sigma^2}\right]}\right] \times \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} x^{-\alpha_0-1} e^{-\frac{\beta_0}{\sigma^2}} \\
&\propto \frac{N}{\sqrt{\sigma^2}} e^{-\left[\frac{\sum_i (y_i-k_i)^2}{2\sigma^2}\right]} \times x^{-\alpha_0-1} e^{-\frac{\beta_0}{\sigma^2}}
\end{aligned} \tag{13}
$$

Recall that $k_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$. It is helpful to redefine the part in orange in the equation above such that

$$
S = \sum_{i=1}^{N} (y_i - k_i)^2 \tag{14}
$$

Furthermore, we take the 1/2 from the denominator in the exponent and divide S by this constant. (look at the preliminary section on the inverse gamma distribution to see why). Hence, we get

$$f(\sigma^2|\dots) \propto \frac{N}{\sqrt{\sigma^2}} e^{-\left[\frac{S/2}{\sigma^2}\right]} \times (\sigma^2)^{-\alpha_0 - 1} e^{-\frac{\beta_0}{\sigma^2}} \tag{14}$$

$$k_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_j x_{ji}$$

Now, collecting the like terms yields

$$\begin{aligned} f(\sigma^2|\dots) &\propto \left[(\sigma^2)^{-\frac{N}{2}} (\sigma^2)^{-\alpha_0 - 1}\right] \times \left[e^{-\left[\frac{S/2}{\sigma^2}\right]} e^{-\frac{\beta_0}{x}}\right] \\ &\propto (\sigma^2)^{\left[-\frac{N}{2} - \alpha_0 - 1\right]} e^{\left[-\frac{(S/2 + \beta_0)}{\sigma^2}\right]} \end{aligned} \tag{14}$$

Which you should recognize as an inverse gamma distribution with $A = -\frac{N}{2} - \alpha_0 - 1$ and $B = -(S/2 + \beta_0)$. We can retrieve the posterior shape α_1 and scale β_1 by using

$$\begin{aligned} \alpha_1 &= -A - 1 = -(-\frac{N}{2} - \alpha_0 - 1) - 1 = \frac{N}{2} + \alpha_0 \\ \beta_1 &= -B = -(-(S/2 + \beta_0)) = \frac{S}{2} + \beta_0 \end{aligned}$$

5 Posterior distributions in vector notation

The notation we used for the posteriors can become a little unwieldy when we implement it in $\mathbb{R}$. This has to do with the summations in the calculations for the posterior means of the coefficients and intercepts. If we would implement it as is, we would have to create a for loop in which we loop over the j parameters.

Alternatively, we can use some linear algebra to simplify the equations. Recall that $k_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_j x_{ji}$. This is a system of linear equations for n examples. Accordingly, we can rewrite k_i as

$$\vec{k} = \mathbf{X}\vec{w}$$

Where

- $\vec{k}$ is a column vector of length n
- $\mathbf{X}$ is a matrix with n rows and m columns
- $\vec{w}$ is a column vector of length m

$$\begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_j x_{1m} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_j x_{2m} \\ \dots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_j x_{nm} \end{bmatrix}$$

Note that the subscripts in x have switched to accomodate the rows $\times$ columns notation that is usual in linear algebra. Instead of $x_{\text{coefficient}\times\text{observation}} = x_{ji}$ we now have $x_{\text{observation}\times\text{coefficient}} = x_{ij} = x_{nm}$.

We will use the following notation to denote that the j^{th} *column* of the design matrix $\mathbf{X}$ and the j^{th} row of the coefficient matrix $\vec{w}$ have been removed

$$\vec{k}_{*j} = \mathbf{X}^{-j} \vec{w}_{-j}$$

We will use x_j to denote the j^{th} column vector from the design matrix $\mathbf{X}$

5.0.1 Intercept and slope posteriors using linear algebra notation

When we examine the posterior means for the intercept and slope coefficients, we should recognize that we can rewrite both equations as one using vector notation.

$$\mu_{j,1} = \frac{1}{\frac{\vec{x}_j^T \cdot \vec{x}_j}{\sigma^2} + \frac{1}{\tau_{j,0}^2}} \times \left(\frac{\sum \vec{x}_j \circ [\vec{y} - \vec{k}_{*j}]}{\sigma^2} + \frac{u_{j,0}}{\tau_{j,0}^2} \right)$$

Where:

- $\vec{x}_j^T \vec{x}_j$ is the sum of the squared elements of $\vec{x}_j$
- X^{-j} is the design matrix without the $j^t h$ column corresponding to β_j
- w_{-j} is the weight matrix without the $j_t h$ coefficient

For the variance, we get

$$\tau_{j,1}^2 = \frac{1}{\left(\frac{\vec{x}_j^T \cdot \vec{x}_j}{2\sigma^2} + \frac{1}{2\tau_{j,0}^2} \right)}$$