Beta Derivation with Matrix Computation

Viveka Kulharia

July 2015

1 Introduction

Let $[X_n^T X^T]^T$ be an $N \times m$ matrix and $[Y_n^T Y^T]^T = [X_n^T X^T]^T \beta + \epsilon$ whose first n elements correspond to missing values and β is a vector of dimension m. Now, we need to solve the QP:
minimize

$$\log \sigma + \frac{(Y - X\beta)^T (Y - X\beta)}{2\sigma^2} + \frac{(Y_n - X_n\beta)^T (Y_n - X_n\beta)}{2\sigma^2} \tag{1}$$

subject to:

$$Y_n^T \mathbf{1} = C \tag{2}$$

where $\mathbf{1}$ is a column vector of ones with length n. Its solution can be obtained by solving $\nabla L = 0$ where

$$L(Y, \beta, \lambda) = \log \sigma + \frac{(Y - X\beta)^T (Y - X\beta)}{2\sigma^2} + \frac{(Y_n - X_n\beta)^T (Y_n - X_n\beta)}{2\sigma^2} + \lambda (Y_n^T \mathbf{1} - C) \tag{3}$$

where λ is a lagrangian multiplier.
On solving it,

$$\beta = (X^T X + \frac{X_n^T \mathbf{1}\mathbf{1}^T X_n}{n})^{-1} (X^T Y + \frac{X_n^T \mathbf{1} C}{n}) \tag{4}$$

And,

$$Y_n = (I - \frac{\mathbf{1}\mathbf{1}^T}{n}) X_n (X^T X + \frac{X_n^T \mathbf{1}\mathbf{1}^T X_n}{n})^{-1} (X^T Y + \frac{X_n^T \mathbf{1} C}{n}) + \frac{C\mathbf{1}}{n} \tag{5}$$

where I is an $n \times n$ identity matrix.

2 Method

Differentiating L w.r.t σ :

$$\begin{aligned} &\frac{\partial L}{\partial \sigma} = 0 \\ \implies &\frac{1}{\sigma} - \frac{(Y - X\beta)^T (Y - X\beta)}{\sigma^3} - \frac{(Y_n - X_n\beta)^T (Y_n - X_n\beta)}{\sigma^3} = 0 \\ \implies &\sigma^2 = (Y - X\beta)^T (Y - X\beta) + (Y_n - X_n\beta)^T (Y_n - X_n\beta) \end{aligned} \tag{6}$$

Differentiating L w.r.t Y_n :

$$\begin{aligned} \frac{\partial L}{\partial Y_n} &= \frac{Y_n - X_n\beta}{\sigma^2} + \lambda\mathbf{1} = 0 \\ &\implies Y_n = -\sigma^2\lambda\mathbf{1} + X_n\beta \end{aligned} \tag{7}$$

Differentiating L w.r.t λ :

$$\begin{aligned} &\frac{\partial L}{\partial \lambda} = 0 \\ &\implies Y_n^T\mathbf{1} = C \end{aligned} \tag{8}$$

Differentiating L w.r.t β :

$$\begin{aligned} &\frac{\partial L}{\partial \beta} = 0 \\ &\implies -\frac{X^T(Y - X\beta)}{\sigma^2} - \frac{X_n^T(Y_n - X_n\beta)}{\sigma^2} = 0 \end{aligned} \tag{9}$$

Substituting 7 to 9:

$$\begin{aligned} &-\frac{X^T(Y - X\beta)}{\sigma^2} + X_n^T\lambda\mathbf{1} = 0 \\ &\implies \frac{X^T(Y - X\beta)}{\sigma^2} - \lambda X_n^T\mathbf{1} = 0 \end{aligned} \tag{10}$$

Substituting Y_n from 7 to 8:

$$\begin{aligned} &-\sigma^2\lambda\mathbf{1}^T\mathbf{1} + \beta^T X_n^T\mathbf{1} = C \\ &\implies \beta^T X_n^T\mathbf{1} = C + \sigma^2\lambda n \\ &\implies \lambda = \frac{\mathbf{1}^T X_n\beta - C}{\sigma^2 n} \end{aligned} \tag{11}$$

Substituting λ from 11 to 10:

$$\begin{aligned} &\frac{X^T(Y - X\beta)}{\sigma^2} - \frac{\mathbf{1}^T X_n\beta - C}{\sigma^2 n} X_n^T\mathbf{1} = 0 \\ &\implies (X^TY + \frac{CX_n^T\mathbf{1}}{n}) - (X^TX\beta + \frac{\mathbf{1}^T X_n\beta X_n^T\mathbf{1}}{n}) = 0 \end{aligned} \tag{12}$$

Now, as $\mathbf{1}^T X_n\beta$ is a scalar, we can write 12 as,

$$\begin{aligned} &\implies (X^TY + \frac{CX_n^T\mathbf{1}}{n}) - (X^TX\beta + \frac{X_n^T\mathbf{1}\mathbf{1}^T X_n\beta}{n}) = 0 \\ &\implies \beta = (X^TX + \frac{X_n^T\mathbf{1}\mathbf{1}^T X_n}{n})^{-1}(X^TY + \frac{X_n^T\mathbf{1}C}{n}) \end{aligned} \tag{13}$$

which is quite different from the normal regression estimate:

$$\hat{\beta} = (X^TX)^{-1}X^TY \tag{14}$$

For estimating the missing values Y_n, let's use 7,

$$\begin{aligned} Y_n &= X_n\beta - \sigma^2\lambda\mathbf{1} \\ &= X_n\beta - \mathbf{1}\sigma^2\lambda \end{aligned} \tag{15}$$

Substituting 11 to 15,

$$\begin{aligned} Y_n &= X_n\beta - \frac{\mathbf{1}\mathbf{1}^T X_n\beta - \mathbf{1}C}{n} \\ \implies Y_n &= \frac{nI - \mathbf{1}\mathbf{1}^T}{n} X_n\beta + \frac{C\mathbf{1}}{n} \end{aligned} \tag{16}$$

Substituting 13 to 16,

$$\implies Y_n = (I - \frac{\mathbf{1}\mathbf{1}^T}{n})X_n(X^TX + \frac{X_n^T\mathbf{1}\mathbf{1}^TX_n}{n})^{-1}(X^TY + \frac{X_n^T\mathbf{1}C}{n}) + \frac{C\mathbf{1}}{n} \tag{17}$$

3 Observation

Using Sherman–Morrison formula:

$$(A + bc^T)^{-1} = A^{-1} - \frac{A^{-1}bc^TA^{-1}}{1 + c^TA^{-1}b} \tag{18}$$

where A is an invertible matrix while b and c are column vectors.
So, if X^TX is invertible, then:

$$\begin{aligned} &(X^TX + \frac{X_n^T\mathbf{1}\mathbf{1}^TX_n}{n})^{-1} = \\ &\qquad \frac{(X^TX)^{-1} - (X^TX)^{-1}\frac{X_n^T\mathbf{1}\mathbf{1}^TX_n}{n}(X^TX)^{-1}}{1 + \frac{\mathbf{1}^TX_n}{\sqrt{n}}(X^TX)^{-1}\frac{X_n^T\mathbf{1}}{\sqrt{n}}} \end{aligned} \tag{19}$$

which can be used to compute β. It is clear that if X^TX is invertible, β is almost always computable as denominator of RHS is very rarely 0. Hence, if $\hat{\beta}$ is computable, β is also computable.
It should be noted that if all columns of X are linearly independent then X^TX is always invertible.

Also, the sum of estimated Y_n comes out to be C as

$$\begin{aligned} \mathbf{1}^TY_n &= (\mathbf{1}^T - \frac{n\mathbf{1}^T}{n})\ldots + C \\ &= 0 + C \\ &= C \end{aligned} \tag{20}$$

Difference between estimate using normal regression and QP seem to be unaffected as missing values increase. It was checked with β of dimension 10, total fully observed samples at 50 and samples with missing values increased from 50 to 1049 and then from 10000 to 10049 and test data containing 2000 missing values. In the first set of missing values, there was a slight decrease for both normal regression and QP initially but remained almost same later. In fact, it was almost same to the error obtained in second set of missing values as well. And every time, estimate using QP performed better than that using normal regression.