ESL- multiple linear regression notes

James Chuang

December 21, 2016

The linear model $f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$ with $p > 1$ inputs is called the ***multiple linear regression model***. The least squares estimates $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ for this model are best understood in terms of the estimates for the ***univariate*** linear model.

Suppose first a univariate model with no intercept, i.e.,

$$Y = X\beta + \epsilon$$

. In the univariate case, the least squares estimate and residuals can be written as inner products:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \quad \text{in the univariate case is}$$
$$\hat{\beta} = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle}$$

$$\mathbf{r} = \mathbf{y} - \mathbf{x}\hat{\beta}$$

. This simple univariate regression provides the building block for multiple linear regression. Suppose next that the inputs $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ (the columns of the data matrix X) are orthogonal, i.e. $\langle \mathbf{x}_j, \mathbf{x}_k \rangle = 0$ for all $j \neq k$. Then, the multiple least squares estimates $\hat{\beta}_j$ are equal to the univariate estimates $\langle \mathbf{x}_j, \mathbf{y} \rangle / \langle \mathbf{x}_j, \mathbf{x}_j \rangle$:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\hat{\beta} = \left(\begin{bmatrix} - & \mathbf{x}_1 & - \\ - & \mathbf{x}_2 & - \\ & \vdots & \\ - & \mathbf{x}_p & - \end{bmatrix} \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \\ | & | & & | \end{bmatrix} \right)^{-1} \begin{bmatrix} - & \mathbf{x}_1 & - \\ - & \mathbf{x}_2 & - \\ & \vdots & \\ - & \mathbf{x}_p & - \end{bmatrix} \mathbf{y}$$

$$= \begin{bmatrix} \langle \mathbf{x}_1, \mathbf{x}_1 \rangle & \langle \mathbf{x}_1, \mathbf{x}_2 \rangle & \cdots & \langle \mathbf{x}_1, \mathbf{x}_p \rangle \\ \langle \mathbf{x}_2, \mathbf{x}_1 \rangle & \langle \mathbf{x}_2, \mathbf{x}_2 \rangle & \cdots & \langle \mathbf{x}_2, \mathbf{x}_p \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{x}_p, \mathbf{x}_1 \rangle & \langle \mathbf{x}_p, \mathbf{x}_2 \rangle & \cdots & \langle \mathbf{x}_p, \mathbf{x}_p \rangle \end{bmatrix}^{-1} \begin{bmatrix} \langle \mathbf{x}_1, \mathbf{y} \rangle \\ \langle \mathbf{x}_2, \mathbf{y} \rangle \\ \vdots \\ \langle \mathbf{x}_p, \mathbf{y} \rangle \end{bmatrix}$$

$$= \begin{bmatrix} \langle \mathbf{x}_1, \mathbf{x}_1 \rangle & 0 & \cdots & 0 \\ 0 & \langle \mathbf{x}_2, \mathbf{x}_2 \rangle & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \langle \mathbf{x}_p, \mathbf{x}_p \rangle \end{bmatrix}^{-1} \begin{bmatrix} \langle \mathbf{x}_1, \mathbf{y} \rangle \\ \langle \mathbf{x}_2, \mathbf{y} \rangle \\ \vdots \\ \langle \mathbf{x}_p, \mathbf{y} \rangle \end{bmatrix}$$

$$= \begin{bmatrix} \langle \mathbf{x}_1, \mathbf{x}_1 \rangle^{-1} & 0 & \cdots & 0 \\ 0 & \langle \mathbf{x}_2, \mathbf{x}_2 \rangle^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \langle \mathbf{x}_p, \mathbf{x}_p \rangle^{-1} \end{bmatrix} \begin{bmatrix} \langle \mathbf{x}_1, \mathbf{y} \rangle \\ \langle \mathbf{x}_2, \mathbf{y} \rangle \\ \vdots \\ \langle \mathbf{x}_p, \mathbf{y} \rangle \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} \frac{\langle \mathbf{x}_1, \mathbf{y} \rangle}{\langle \mathbf{x}_1, \mathbf{x}_1 \rangle} \\ \frac{\langle \mathbf{x}_2, \mathbf{y} \rangle}{\langle \mathbf{x}_2, \mathbf{x}_2 \rangle} \\ \vdots \\ \frac{\langle \mathbf{x}_p, \mathbf{y} \rangle}{\langle \mathbf{x}_p, \mathbf{x}_p \rangle} \end{bmatrix}$$

I.e., when the inputs are orthogonal, they have no effect on each other's parameter estimates in the model.

Orthogonal inputs occur most often with balanced, designed experiments where orthogonality is enforced, but almost never with observational data. Hence we will have to orthogonalize them to carry this idea further. Suppose next that we have an intercept and a single input $\mathbf{x}$, i.e. $\mathbf{y} = \beta_o + \beta_1 \mathbf{x}$. Then the least squares coefficient of x has the form

$$\hat{\beta}_1 = \frac{\langle \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{y} \rangle}{\langle \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{x} - \bar{x}\mathbf{1} \rangle}$$

, where $\bar{x} = \sum_i x_i / N$, and $\mathbf{1} = \mathbf{x}_0$, the vector of N ones.

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= \left(\begin{bmatrix} - & \mathbf{1} & - \\ - & \mathbf{x}_1 & - \end{bmatrix} \begin{bmatrix} | & | \\ \mathbf{1} & \mathbf{x}_1 \\ | & | \end{bmatrix} \right)^{-1} \begin{bmatrix} - & \mathbf{1} & - \\ - & \mathbf{x}_1 & - \end{bmatrix} \mathbf{y} \\
&= \begin{bmatrix} \langle \mathbf{1}, \mathbf{1} \rangle & \langle \mathbf{1}, \mathbf{x}_1 \rangle \\ \langle \mathbf{x}_1, \mathbf{1} \rangle & \langle \mathbf{x}_1, \mathbf{x}_1 \rangle \end{bmatrix}^{-1} \begin{bmatrix} \langle \mathbf{1}, \mathbf{y} \rangle \\ \langle \mathbf{x}, \mathbf{y} \rangle \end{bmatrix} \\
&= \\
&= \\
&= \langle \quad \rangle \langle \mathbf{1}, \mathbf{y} \rangle + \langle \quad \rangle \langle \mathbf{x}, \mathbf{y} \rangle \\
&= \frac{-\langle \mathbf{x}_1, \mathbf{1} \rangle \langle \mathbf{1}, \mathbf{y} \rangle + \langle \mathbf{1}, \mathbf{1} \rangle \langle \mathbf{x}_1, \mathbf{y} \rangle}{\langle \mathbf{1}, \mathbf{1} \rangle \langle \mathbf{x}_1, \mathbf{x}_1 \rangle - \langle \mathbf{x}_1, \mathbf{1} \rangle \langle \mathbf{1}, \mathbf{x}_1 \rangle} \\
&= \frac{-\sum_i x_i \sum_i y_i + N \sum_i x_i y_i}{N \sum_i x_i^2 - \sum_i x_i \sum_i x_i} \\
&= \frac{\langle \mathbf{x} - \left(\frac{\sum_i x_i}{N} \right) \mathbf{1}, \mathbf{y} \rangle}{\langle \mathbf{x} - \left(\frac{\sum_i x_i}{N} \right) \mathbf{1}, \mathbf{x} - \left(\frac{\sum_i x_i}{N} \right) \mathbf{1} \rangle} \\
&= \frac{\langle \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{y} \rangle}{\langle \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{x} - \bar{x}\mathbf{1} \rangle}
\end{aligned}$$

We can view this estimate as the result of two applications of the simple regression. The steps are:

1. regress $\mathbf{x}$ on $\mathbf{1}$ to produce the residual $\mathbf{z} = \mathbf{x} = \bar{x}\mathbf{1}$;
2. regress $\mathbf{y}$ on the residual $\mathbf{z}$ to give the coefficient $\hat{\beta}_1$.

In this procedure, "regress $\mathbf{b}$ on $\mathbf{a}$" means a simple univariate regression of $\mathbf{b}$ on $\mathbf{a}$ with no intercept, producing coefficient $\hat{\gamma} = \langle \mathbf{a}, \mathbf{b} \rangle / \langle \mathbf{a}, \mathbf{a} \rangle$ and residual vector $\mathbf{b} - \hat{\gamma}\mathbf{a}$. We say that $\mathbf{b}$ is adjusted for $\mathbf{a}$, or is "orthogonalized" with respect to $\mathbf{a}$.

Gram-Schmidt procedure for multiple regression

1. Initialize $\mathbf{z}_0 = \mathbf{x}_0 = \mathbf{1}$
2. For $j = 1, 2, \ldots, p$
   - Regress $\mathbf{x}_j$ on $\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_{j-1}$ to produce coefficients $\hat{\gamma}_{\ell j}$:

$$\hat{\gamma}_{\ell j} = \frac{\langle \mathbf{z}_\ell, \mathbf{x}_j \rangle}{\langle \mathbf{z}_\ell, \mathbf{z}_\ell \rangle}, \quad \ell = 0, \ldots, j-1$$

   and residual vector $\mathbf{z}_j$:

$$\mathbf{z}_j = \mathbf{x}_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} \mathbf{z}_k$$

3. Regress $\mathbf{y}$ on the residual $\mathbf{z}_p$ to give the estimate $\hat{\beta}_p$, i.e.:

$$\hat{\beta}_p = \frac{\langle \mathbf{z}_p, \mathbf{y} \rangle}{\langle \mathbf{z}_p, \mathbf{z}_p \rangle}$$

Re-arranging the residual in step 2 ($\mathbf{x}_j = \mathbf{z}_j + \sum_{k=0}^{j-1} \hat{\gamma}_{kj} \mathbf{z}_k$), we can see that each of the $\mathbf{x}_j$ is a linear combination of the $\mathbf{z}_k, k \leq j$. Since the $\mathbf{z}_j$ are all orthogonal, they form a basis for the column space of $\mathbf{X}$, and hence the least squares projection onto this subspace is $\hat{\mathbf{y}}$. Since $\mathbf{z}_p$ alone involves $\mathbf{x}_p$ (with coefficient 1), we see that the coefficient $\hat{\beta}_p = \frac{\langle \mathbf{z}_p, \mathbf{y} \rangle}{\langle \mathbf{z}_p, \mathbf{z}_p \rangle}$ is indeed the multiple regression coefficient of $\mathbf{y}$ on $\mathbf{x}_p$. This key result exposes the effect of correlated inputs in multiple regression. Note also that by rearranging the $\mathbf{x}_j$, any one of them could be in the last position, and a similar result holds. Hence, stated more generally, we have shown that the jth multiple regression coefficient is the univariate regression coefficient of $\mathbf{y}$ on $\mathbf{x}_{0,1,\ldots(j-1)(j+1)\ldots,p}$, the residual after regressing $\mathbf{x}_j$ on $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_p$:

The multiple regression coefficient $\hat{\beta}_j$ represents the **additional** contribution of $\mathbf{x}_j$ on $\mathbf{y}$, **after $\mathbf{x}_j$ has been adjusted for $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_p$.**

If $\mathbf{x}_p$ is highly corelated with some of the other $\mathbf{x}_k$'s, the residual vector $\mathbf{z}_p$ will be close to zero, and hcne the coefficient $\hat{\beta}_p = \frac{\langle \mathbf{z}_p, \mathbf{y} \rangle}{\langle \mathbf{z}_p, \mathbf{z}_p \rangle}$ will be very unstable. This will be true for all the variables in the correlated set. In such situations, we might have all the Z-scores be small – any one of the set can be deleted – yet we cannot delete them all.

We can also obtain an alternate formula for the variance estimates of the coefficients:

$$\text{Var}(\hat{\beta}_p) = \frac{\sigma^2}{\langle \mathbf{z}_p, \mathbf{z}_p \rangle} = \frac{\sigma^2}{\|\mathbf{z}_p\|^2}$$

. In other words, the precision with which we can estimate $\hat{\beta}_p$ depends on the length of the residual vector $\mathbf{z}_p$; this represents how much of $\mathbf{x}_p$ is unexplained by the other $\mathbf{x}_k$'s.

Step 2 of the algorithm can be represented in matrix form:

$$\mathbf{X} = \mathbf{Z}\mathbf{\Gamma}$$

, where $\mathbf{z}_j$ has as columns the $\mathbf{z}_j$ (in order), and $\mathbf{\Gamma}$ is the upper triangular matrix with entries $\hat{\gamma}_{kj}$. I.e.:

$$\mathbf{X} = \mathbf{Z}\mathbf{\Gamma} = \begin{bmatrix} | & | & & | \\ \mathbf{z}_0 & \mathbf{z}_1 & \cdots & \mathbf{z}_p \\ | & | & & | \end{bmatrix} \begin{bmatrix} \gamma_{0,0} & \gamma_{0,1} & \cdots & \gamma_{0,p} \\ 0 & \gamma_{1,1} & \cdots & \gamma_{1,p} \\ \vdots & \vdots & \ddots & \cdots \\ 0 & 0 & \cdots & \gamma_{p,p} \end{bmatrix}$$

Introducing the diagonal matrix $\mathbf{D}$ with jth diagonal entry $D_{j,j} = \|\mathbf{z}_j\|$, we get:

$$\begin{aligned} \mathbf{X} &= \mathbf{Z}\mathbf{D}^{-1}\mathbf{D}\mathbf{\Gamma} \\ &= \mathbf{Q}\mathbf{R} \end{aligned}$$

, the so-called QR decomposition of $\mathbf{X}$. Here $\mathbf{Q}$ is an $N \times (p+1)$ orthogonal matrix, $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, and $\mathbf{R}$ is a $(p+1) \times (p+1)$ upper triangular matrix.

The $\mathbf{QR}$ decomposition represents a convenient orthogonal basis for the column space of $\mathbf{X}$. For example, the least squares solution

is given by:

$$
\begin{aligned}
\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= \left[(\mathbf{QR})^T(\mathbf{QR})\right]^{-1}(\mathbf{QR})^T\mathbf{y} \\
&= \left[\mathbf{R}^T\mathbf{Q}^T\mathbf{QR}\right]^{-1}\mathbf{R}^T\mathbf{Q}^T\mathbf{y} \\
&= \mathbf{R}^{-1}\left(\mathbf{R}^T\right)^{-1}\mathbf{R}^T\mathbf{Q}^T\mathbf{y} \\
\hat{\beta} &= \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y},
\end{aligned}
$$

$$
\begin{aligned}
\hat{\mathbf{y}} &= \mathbf{X}\hat{\beta} \\
&= (\mathbf{QR})\left(\mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y}\right) \\
\hat{\mathbf{y}} &= \mathbf{QQ}^T\mathbf{y}
\end{aligned}
$$

$\hat{\beta} = \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y}$ is easy to solve because $\mathbf{R}$ is upper triangular.