The Gram-Schmidt Process and the $\boldsymbol{QR}$ Decomposition of $\boldsymbol{X}$

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Linear models notation

We continue to use the notation

$$\underset{N\times p}{\mathbf{X}} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_N' \end{pmatrix} \text{ and potentially } \underset{N\times 1}{\mathbf{Y}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

and recall that it is standard linear models fare that ordinary least squares projects $\mathbf{Y}$ onto $C(\mathbf{X})$, the column space of $\mathbf{X}$, in order to produce the vector of fitted values

$$\underset{N\times 1}{\widehat{\mathbf{Y}}} = \begin{pmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_N \end{pmatrix}$$

Orthogonal bases for inner product spaces

For many purposes it would be convenient if the columns of a full rank $(rank = p)$ matrix $\mathbf{X}$ were orthogonal. In fact, it would be useful to replace the $N \times p$ matrix $\mathbf{X}$ with an $N \times p$ matrix $\mathbf{Z}$ with orthogonal columns and having the property that for each l if $\mathbf{X}_l$ and $\mathbf{Z}_l$ are $N \times l$ consisting of the first l columns of respectively $\mathbf{X}$ and $\mathbf{Z}$, then $C(\mathbf{Z}_l) = C(\mathbf{X}_l)$. Such a matrix can be constructed using the so-called Gram-Schmidt process.

This process generalizes beyond application to $\Re^N$ to general inner product spaces, and in recognition of that important fact we'll first describe it in general terms and then consider its implications for a $(rank = p)$ matrix $\mathbf{X}$.

Gram-Schmidt process

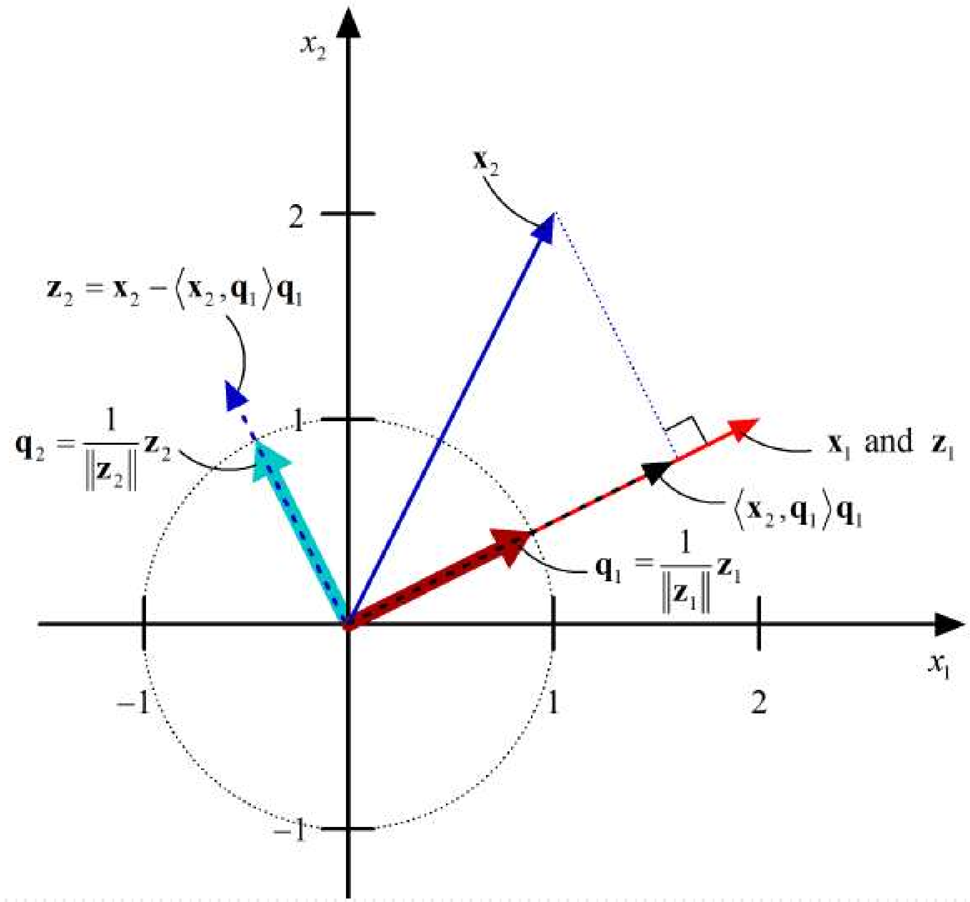
Consider p vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$. (These could be N-vectors where $\mathbf{x}_j$ is the jth column of $\mathbf{X}$, in notational conflict with the convention that made $\mathbf{x}_i$ the p-vector of inputs for the ith case in the training data. In that case one must read the following keeping in mind the local convention.) The **Gram-Schmidt process** proceeds as follows:

1. Set
$$\mathbf{z}_1 = \mathbf{x}_1 \quad \text{and} \quad \mathbf{q}_1 = \langle \mathbf{z}_1, \mathbf{z}_1 \rangle^{-1/2} \mathbf{z}_1 = \frac{1}{\|\mathbf{z}_1\|} \mathbf{z}_1$$

2. Having constructed $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{l-1}\}$, let
$$\mathbf{z}_l = \mathbf{x}_l - \sum_{j=1}^{l-1} \frac{\langle \mathbf{x}_l, \mathbf{z}_j \rangle}{\langle \mathbf{z}_j, \mathbf{z}_j \rangle} \mathbf{z}_j = \mathbf{x}_l - \sum_{j=1}^{l-1} \langle \mathbf{x}_l, \mathbf{q}_j \rangle \mathbf{q}_j \quad \text{and} \quad \mathbf{q}_l = \frac{1}{\|\mathbf{z}_l\|} \mathbf{z}_l$$

Geometry

The figure below illustrates this construction for a simple case of $p = 2$.



Orthonormal basis for the span

It is easy to see that $\langle \mathbf{z}_l, \mathbf{z}_j \rangle = 0$ for all $j < l$ (building the orthogonality of $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{l-1}$ by induction), since

$$\langle \mathbf{z}_l, \mathbf{z}_j \rangle = \langle \mathbf{x}_l, \mathbf{z}_j \rangle - \langle \mathbf{x}_l, \mathbf{z}_j \rangle$$

as at most one term of the sum in step 2. is non-zero. Further, assume that the span of $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{l-1}\}$ is the same as the span of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{l-1}\}$. $\mathbf{z}_l$ is in the span of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l\}$ so that the span of $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_l\}$ is a subset of the span of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l\}$. And since any element of the span of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l\}$ can be written as a linear combination of an element of the span of $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{l-1}\}$ (span of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{l-1}\}$) and $\mathbf{x}_l$ the span of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l\}$ is a subset of the span of $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_l\}$. That is that $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_l\}$ and $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l\}$ have the same span, and vectors $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_l$ comprise an **orthonormal basis** for the span of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l\}$.

Projections

Since the $\mathbf{z}_j$ are perpendicular, for any vector $\mathbf{w}$,

$$\sum_{j=1}^{l} \frac{\langle \mathbf{w}, \mathbf{z}_j \rangle}{\langle \mathbf{z}_j, \mathbf{z}_j \rangle} \mathbf{z}_j = \sum_{j=1}^{l} \langle \mathbf{w}, \mathbf{q}_j \rangle \mathbf{q}_j$$

is the projection of $\mathbf{w}$ onto the span of $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_l\}$ (of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l\}$). (To see this, consider minimization of the quantity $\left\langle \mathbf{w} - \sum_{j=1}^{l} c_j \mathbf{z}_j, \mathbf{w} - \sum_{j=1}^{l} c_j \mathbf{z}_j \right\rangle = \left\| \mathbf{w} - \sum_{j=1}^{l} c_j \mathbf{z}_j \right\|^2$ by choice of the constants c_j.) In particular, $\sum_{j=1}^{l-1} \langle \mathbf{x}_l, \mathbf{q}_j \rangle \mathbf{q}_j$ in step 2. of the Gram-Schmidt process is the projection of $\mathbf{x}_l$ onto the span of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{l-1}\}$.

Gram-Schmidt in Euclidean spaces

Since the $\mathbf{z}_j$ are perpendicular, for any vector $\mathbf{w}$,

$$\sum_{j=1}^{l} \frac{\langle \mathbf{w}, \mathbf{z}_j \rangle}{\langle \mathbf{z}_j, \mathbf{z}_j \rangle} \mathbf{z}_j = \sum_{j=1}^{l} \langle \mathbf{w}, \mathbf{q}_j \rangle \mathbf{q}_j$$

is the projection of $\mathbf{w}$ onto the span of $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_l\}$ (of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l\}$). (To see this, consider minimization of the quantity $\left\langle \mathbf{w} - \sum_{j=1}^{l} c_j \mathbf{z}_j, \mathbf{w} - \sum_{j=1}^{l} c_j \mathbf{z}_j \right\rangle = \left\| \mathbf{w} - \sum_{j=1}^{l} c_j \mathbf{z}_j \right\|^2$ by choice of the constants c_j.) In particular, $\sum_{j=1}^{l-1} \langle \mathbf{x}_l, \mathbf{q}_j \rangle \mathbf{q}_j$ in step 2. of the Gram-Schmidt process is the projection of $\mathbf{x}_l$ onto the span of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{l-1}\}$.

QR decomposition of X

The construction of the orthogonal variables $\mathbf{z}_j$ can be represented in matrix form as

$$\underset{N\times p}{\mathbf{X}} = \underset{N\times p}{\mathbf{Z}}\;\underset{p\times p}{\boldsymbol{\Gamma}}$$

where $\boldsymbol{\Gamma}$ is upper triangular with

$$\begin{aligned}\gamma_{kj} &= \text{the value in the } k\text{th row and } j\text{th column of } \boldsymbol{\Gamma} \\ &= \begin{cases} 1 & \text{if } j = k \\ \dfrac{\langle \mathbf{z}_k, \mathbf{x}_j \rangle}{\langle \mathbf{z}_k, \mathbf{z}_k \rangle} & \text{if } j > k \end{cases}\end{aligned}$$

$\boldsymbol{QR}$ decomposition of $\boldsymbol{X}$ cont.

Defining

$$\mathbf{D} = \mathbf{diag}\left(\langle \mathbf{z}_1, \mathbf{z}_1\rangle^{1/2}, \ldots, \langle \mathbf{z}_p, \mathbf{z}_p\rangle^{1/2}\right) = \mathbf{diag}\left(\|\mathbf{z}_1\|, \ldots, \|\mathbf{z}_p\|\right)$$

and letting

$$\mathbf{Q} = \mathbf{Z}\mathbf{D}^{-1} \quad \text{and} \quad \mathbf{R} = \mathbf{D}\boldsymbol{\Gamma}$$

one may write

$$\mathbf{X} = \mathbf{QR}$$

that is the so-called **QR** decomposition of **X**.

Note that the notation used here is consistent, in that for $\mathbf{q}_j$ the jth column of **Q**, $\mathbf{q}_j = \left(\langle \mathbf{z}_j, \mathbf{z}_j\rangle\right)^{-1/2}\mathbf{z}_j$ as was used in defining the Gram-Schmidt process.

$\boldsymbol{QR}$ and OLS

$\mathbf{Q}$ is $N \times p$ with

$$\mathbf{Q}'\mathbf{Q} = \mathbf{D}^{-1}\mathbf{Z}'\mathbf{Z}\mathbf{D}^{-1} = \mathbf{D}^{-1}\mathbf{diag}\left(\langle \mathbf{z}_1, \mathbf{z}_1 \rangle, \ldots, \langle \mathbf{z}_p, \mathbf{z}_p \rangle\right)\mathbf{D}^{-1} = \mathbf{I}$$

consistent with the fact that $\mathbf{Q}$ has for columns perpendicular unit vectors that form a basis for $C(\mathbf{X})$. $\mathbf{R}$ is upper triangular so that only the first l of these unit vectors are needed to create $\mathbf{x}_l$ and there are efficient ways to compute $\mathbf{R}^{-1}$.

It is computationally useful that the projection of response vector $\mathbf{Y}$ onto $C(\mathbf{X})$ is

$$\widehat{\mathbf{Y}} = \sum_{j=1}^{p} \langle \mathbf{Y}, \mathbf{q}_j \rangle \mathbf{q}_j = \mathbf{Q}\mathbf{Q}'\mathbf{Y}$$

and

$$\widehat{\beta}^{\mathrm{OLS}} = \mathbf{R}^{-1}\mathbf{Q}'\mathbf{Y}$$