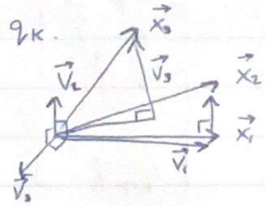
M342W 3/8

$X = QR$ decomposition has two steps 1) Gram-Shmidt algorithm which converts X into Q column-by-column and 2) reconstruction of the upper triangular change-of-basis matrix R. X has dimension $n \times K$ and columns $x_1, \dots, x_K$

In 1), we first a) create a orthogonal basis $v_1, \dots, v_K$ and then b) normalize its component vectors into $q_1, \dots, q_K$.



1a) $\vec{v}_1 = \vec{x}_1$

$\vec{v}_2 = \vec{x}_2 - \text{proj}_{\vec{v}_1}(\vec{x}_2)$

$\text{Span}\{\vec{x}_1, \vec{x}_2\} = \text{Span}\{\vec{v}_1, \vec{v}_2\}$ but $\vec{v}_1 \perp \vec{v}_2$

$\vec{v}_3 = \vec{x}_3 - \text{proj}_{[\vec{v}_1|\vec{v}_2]}(\vec{x}_3)$

$\vdots$

$\vec{v}_K = \vec{x}_K - \text{proj}_{[\vec{v}_1|\cdots|\vec{v}_{K-1}]}(\vec{x}_K)$

1b) $\vec{q}_1 := \frac{\vec{v}_1}{\|\vec{v}_1\|}$, $\vec{q}_2 := \frac{\vec{v}_2}{\|\vec{v}_2\|}$, $\dots$, $\vec{q}_K := \frac{\vec{v}_K}{\|\vec{v}_K\|}$

$\Rightarrow Q = [\vec{q}_1|\cdots|\vec{q}_K]$

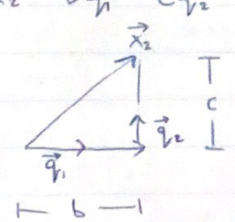
2) $X = QR$

$$[\vec{x}_1|\vec{x}_2|\cdots|\vec{x}_K] = [\vec{q}_1|\vec{q}_2|\cdots|\vec{q}_K] \overbrace{\begin{bmatrix} a & b & d & & - \\ 0 & c & e & & - \\ 0 & 0 & f & \cdots & - \\ \vdots & \vdots & 0 & & - \\ & & \vdots & & - \\ 0 & 0 & 0 & & - \end{bmatrix}}^{R}$$

$\vec{x}_1 = \overbrace{\|\vec{x}_1\|}^{a}\vec{q}_1$

$\vec{x}_2 = b\vec{q}_1 + c\vec{q}_2 = H_1\vec{x}_2 + H_2\vec{x}_2$

$= \vec{q}_1\underbrace{\vec{q}_1^T\vec{x}_2}_{b} + \vec{q}_2\underbrace{\vec{q}_2^T\vec{x}_2}_{c}$



$\vec{x}_3 = d\vec{q}_1 + e\vec{q}_2 + f\vec{q}_3$

$d = \vec{q}_1 \cdot \vec{x}_3$, $e = \vec{q}_2 \cdot \vec{x}_3$, $f = \vec{q}_3 \cdot \vec{x}_3$

Sidebar: QR decomposition helps to speedup the OLS estimate computation in the following way:

$$\vec{b} = (X^T X)^{-1} X^T \vec{y}$$

$$\Downarrow$$

$$X^T X \vec{b} = X^T \vec{y} \Rightarrow (QR)^T QR\vec{b} = (QR)^T \vec{y} \Rightarrow R^T \overbrace{Q^T Q}^{I} R\vec{b} = R^T \overbrace{Q^T \vec{y}}^{\vec{z}}$$

$$\Rightarrow R^T R b = R^T \vec{z} \Rightarrow R^{T^{-1}} R^T R \vec{b} = R^{T^{-1}} R^T \vec{z} \Rightarrow R\vec{b} = \vec{z}$$

e.g. $p+1=3$

by back-substitution

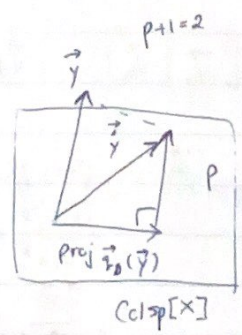
$$\underbrace{\begin{bmatrix} a & b & d \\ 0 & c & e \\ 0 & 0 & f \end{bmatrix}}_{R} \underbrace{\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}}_{\vec{b}} = \begin{bmatrix} z_0 \\ z_1 \\ z_2 \end{bmatrix} \quad \begin{aligned} &\Rightarrow f b_2 = z_2 \Rightarrow b_2 = \frac{z_2}{f} \\ &\Rightarrow c b_1 + e b_2 = z_1 \Rightarrow b_1 = \frac{1}{c}\left(z_1 - e\frac{z_2}{f}\right), \ldots \end{aligned}$$

$$SST = SSR + SSE \Rightarrow SSR\uparrow \Leftrightarrow SSE\downarrow$$

$$\Updownarrow \qquad \Updownarrow$$

$$R^2 \uparrow \Leftrightarrow RMSE\downarrow$$

$\uparrow$ fixed quantity only a function of $\vec{y}$



$$\hat{\vec{y}} = H\vec{y} = QQ^T\vec{y} = \sum_{j=0}^{p} proj_{\vec{q}_j}(\vec{y})$$

$$\|\hat{\vec{y}}\|^2 \overset{\text{Pythag thm}}{=} \sum_{j=0}^{p} \|proj_{\vec{q}_j}(\vec{y})\|^2 = \underbrace{\|proj_{\vec{q}_0}(\vec{y})\|^2} + \sum_{j=1}^{p} \|proj_{\vec{q}_j}(\vec{y})\|^2$$

$$= \|proj_{\vec{1}}(\vec{y})\|^2 = (H_0\vec{y})^T(H_0\vec{y}) = \vec{y}^T H_0 H_0 \vec{y} = (\bar{y}\vec{1}_n)^T(\bar{y}\vec{1}_n) = \bar{y}^2 \vec{1}^T\vec{1} = n\bar{y}^2$$

$$SSR := (\hat{\vec{y}} - \bar{y}\vec{1})^T(\hat{\vec{y}} - \bar{y}\vec{1}) = \hat{\vec{y}}^T\hat{\vec{y}} - \bar{y}\vec{1}^T\hat{\vec{y}} - \bar{y}\hat{\vec{y}}^T\vec{1} + \bar{y}^2\vec{1}^T\vec{1}$$

$$= \|\hat{\vec{y}}\|^2 - 2\bar{y}\underbrace{\hat{\vec{y}}^T\vec{1}}_{(H\vec{y})^T\vec{1} = \vec{y}^T H\vec{1} = \vec{y}^T\vec{1} = n\bar{y}} + n\bar{y}^2 = \|\hat{\vec{y}}\|^2 - 2n\bar{y}^2 + n\bar{y}^2 = \|\hat{\vec{y}}\|^2 - n\bar{y}^2 = \sum_{j=1}^{p} \|proj_{\vec{q}_j}(\vec{y})\|^2$$

Pretend your friend gave you a new feature, i.e. a new $\vec{x}_*$. You want to now update your OLS model to use it.

$$X_* = [\, X \mid \vec{x}_{\cdot *} \,]$$

$$SSR_* = SSR + \underbrace{\|proj_{\vec{q}_*}(\vec{y})\|^2}_{\geq 0} \Rightarrow SSR_* \geq SSR \Leftrightarrow SSE_* \leq SSE$$

Now your friend says "btw, I made up that vector... it's just a bunch of random nonsense". Any new column vector in X would have the ostensible effect of improving your model. If that new column is ~~ind~~ independent of the true causal inputs to y (i.e. the z's), we call this "overfitting".

Let's keep going. Your friend keeps supplying you with more and more garbage vectors. What happens when you have the same # of vectors $p+1 = n$?

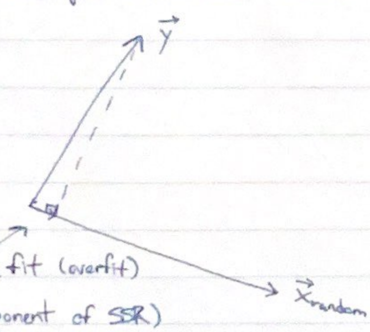
X_* will be $n \times n$ invertible so... $\text{colsp}[X_*] = \mathbb{R}^n$

$H_* = X_*(X_*^T X_*)^{-1} X_*^T = X_* X_*^{-1} X_*^{T^{-1}} X_*^T = I$

$\hat{\vec{y}} = H_* \vec{y} = \vec{y} \Rightarrow \vec{e} = \vec{0}_n \Rightarrow SSE = 0 \Leftrightarrow R^2 = 1 \Leftrightarrow RMSE = 0$

"perfect fit" or "maximal overfitting"

How did we get into this mess? Consider a random vec $\vec{x}_{random}$:



added fake fit (overfit)
(fake component of SSR)

Overfitting becomes a problem with lots of features relative to n. If you have a small number of features, it's not too bad (i.e. it won't reduce your predictive accuracy).

We proved this in the context of OLS regression, but this is true in every other modeling context. Overfitting increases "generalization error" which is error on future predictions.

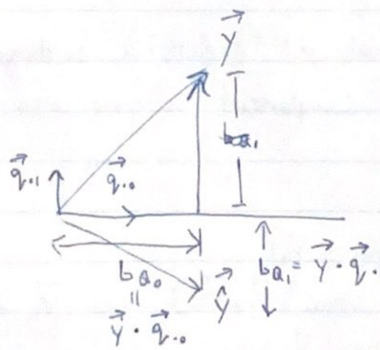
$H = QR^T$

$\vec{b} = (X^T X)^{-1} X^T \vec{y} = ((QR)^T (QR))^{-1} (QR)^T \vec{y}$

$= (R^T \underbrace{Q^T Q}_{I} R)^{-1} R^T Q^T \vec{y} = (R^T R)^{-1} R^T Q^T \vec{y} = R^{-1} \underbrace{R^{T^{-1}} R^T}_{I} Q^T \vec{y}$

$= R^{-1} Q^T \vec{y}$

$$\vec{b}_Q = Q^T \vec{y} = \begin{bmatrix} \vec{q}_{\cdot 0}^{\,T} \vec{y} \\ \vec{q}_{\cdot 1}^{\,T} \vec{y} \\ \vdots \\ \vec{q}_{\cdot p}^{\,T} \vec{y} \end{bmatrix}$$



$\text{Colsp}[\vec{q}_{\cdot 0}, \vec{q}_{\cdot 1}]$

$=$

$\text{Colsp}[\vec{1},$