Residualvektorn, $r = -b_\perp$, är ju ortogonal mot bildrummet. Bildrummet utgörs av alla linjärkombinationer av a_1 och a_2 (i vårt specialfall) vilket medför att $a_1^T r = a_2^T r = 0$.
Vi kan skriva dessa likheter på följande form:

$$0 = \begin{bmatrix} a_1^T r \\ a_2^T r \end{bmatrix} = \begin{bmatrix} a_1^T \\ a_2^T \end{bmatrix} r = \begin{bmatrix} a_1 & a_2 \end{bmatrix}^T r = A^T r = A^T(Ax - b)$$

vilket ger oss <u>normalekvationerna</u>:

$$A^T A x = A^T b$$

$\text{rang}(A) = n \Rightarrow A^T A$ symmetrisk och positivt definit. Kan lösa normalekvationerna med hjälp av Choleskyfaktorisering.

Entydighet?

- om A har linjärt oberoende kolonner så har minstakvadratproblemet en entydig lösning. Matrisen har full rang.
- om A har linjärt beroende kolonner (är rangdefekt) så finns det oändligt många lösningar som ger samma residualvektor, ty tag $z \in \mathcal{N}(A)$ då gäller att $A(x + z) = Ax$.

Om A har nästan linjärt beroende kolonner, så är problemet illa konditionerat. Normalekvationerna förvärrar konditionen på problemet, ett elakt problem kan bli omöjligt att lösa. Det gäller att $\kappa(A^T A) = \kappa(A)^2$. Vi ska därför se på en bättre metod baserad på QR-faktorisering.

`x = A \ b` i Matlab använder QR-faktorisering.
Observera att operatorn \ är överlagrad. Om A är kvadratisk så används LU-faktorisering, annars används QR-faktorisering. Matlabkoderna för de två fallen har ingen gemensam del.

Varför vill man ha många mätvärden? Något om statistik.

Antag att vi har modellen $b = p(t)$, där p är ett polynom av grad noll, en konstant, så $p(t) = x$, där x är en skalär. Minstakvadratproblemet kan skrivas:

$$\min_x \left|\left| \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} x - \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \right|\right|_2$$

Om vi låter e beteckna kolonnvektor av n ettor har vi problemet $\min_x ||e\,x - b||_2$ och normalekvationerna ger oss lösningen:

$$e^T e\, x = e^T b \quad \Rightarrow \quad x = \frac{\sum_{k=1}^n b_k}{n}$$

så x är medelvärdet av b_k-värdena.

Antag att $b_k = \beta + \delta_k$, $k = 1, \ldots, n$, där β är det exakta (men okända) b-värdet och δ_k är mätfel. Det gäller

$$x = \frac{1}{n}\sum_{k=1}^n b_k = \frac{1}{n}\sum_{k=1}^n (\beta + \delta_k) = \beta + \frac{1}{n}\sum_{k=1}^n \delta_k$$

vilket ger oss ett uttryck för det absoluta felet:

$$|x - \beta| = \frac{1}{n}\left|\sum_{k=1}^n \delta_k\right|$$

Antag att det finns $\delta > 0$ så att felen har egenskapen:

$$-\delta \le \delta_k \le \delta, \;\; k = 1, \ldots, n$$

Vi kan då begränsa det maximala felet i x:

$$|x - \beta| = \frac{1}{n}\left|\sum_{k=1}^n \delta_k\right| \;\le\; \frac{1}{n}\sum_{k=1}^n |\delta_k| \le \delta$$

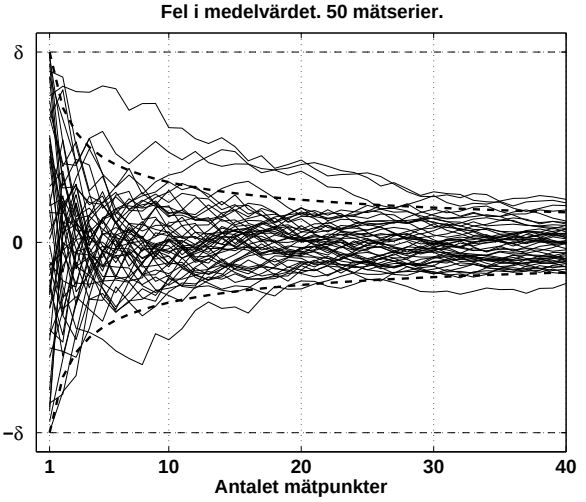
Detta är en pessimistisk uppskattning, eftersom mätfel normalt varierar i storlek och tecken (om vi inte har systematiska fel).

Medelvärdet av felen är normalt mindre än det maximala felet. Antag att felen, δ_k, är likformigt fördelade i intervallet $[-\delta, \delta]$, $\delta > 0$. Följande bild (simulering med Matlabs `rand`) visar hur medelvärdena

$$\frac{1}{n}\sum_{k=1}^n \delta_k, \;\; n = 1, 2, \ldots, 40$$

ser ut för 50 mätserier. En heldragen linje visar, för en mätserie:

$$\delta_1, \;\; \frac{\delta_1 + \delta_2}{2}, \;\; \frac{\delta_1 + \delta_2 + \delta_3}{3}, \;\; \ldots, \;\; \frac{\delta_1 + \delta_2 + \delta_3 + \ldots + \delta_{40}}{40}$$



De två horisontella steckade linjerna markerar maxfelet $\pm\delta$.
Vi ser att spridningen minskar med ökande antal mätpunkter. Man kan bevisa att den sk standardavvikelsen minskar som $1/\sqrt{n}$ (de andra streckade linjerna visar $\pm\delta/\sqrt{n}$).

I matematisk statistik (och fysik) formulerar man sig ungefär så här:

med 95% sannolikhet ligger det riktiga värdet i intervallet $(x - fel,\; x + fel)$ där fel kan räknas ut. Detta intervall är ett sk (observerat) konfidensintervall och sannolikheten, 95%, kallas konfidensgrad.

Detta säger att i 19 fall av 20 så kommer det riktiga värdet, β, att ligga i intervallet.

Mer allmänt kan man bestämma en sannolikhet, $0 < p < 1$, där felet är begränsat av $fel(p, n, \delta)$ (en funktion av p, n och δ). För fixt n, så kommer ett större p (säkrare) att ge ett större värde på $fel(p, n, \delta)$. Man skall tänka på att felet kan överstiga $fel(p, n, \delta)$, eftersom vi arbetar med sannolikheter. Om man kräver visshet, $p = 1$, får vi maximalfelet.

Vi kan alltså väljer mellan en säker, men pessimistisk gräns, och en realistisk men osäker.

Bilden ovan är lite missvisande. Man har inte alltid så långa mätserier (40 värden). 10-20 värden är inte ovanligt och ibland har man kanske bara fem värden. Felet behöver inte minska så mycket, med andra ord.

Kort om konditionstal för LS-problemet (LS = Least Squares)

Antag att x resp. y löser följande problem:

$$\min_x ||Ax-b||_2 \quad \text{resp.} \quad \min_y ||(A+F)y-(b+f)||_2$$

y är alltså lösningen till ett stört problem.

Vi vill begränsa $||y-x||_2/||x||_2$ i termer av $||F||_2/||A||_2$ och $||f||_2/||b||_2$.

Att göra detta allmänt är svårt. En första förenkling är att anta att A har full rang och att $||F||_2$ är tillräckligt liten så att $A+F$ har samma rang som A. Härledningen är nu avsevärt enklare, men ändå lite småbesvärlig, så på dessa sidor antar vi att $F=0$, precis som vi gjorde när vi analyserade $Ax=b$-problemet.

Eftersom A har full rang kan vi använda normalekvationerna och får $x=(A^TA)^{-1}A^Tb$ resp. $y=(A^TA)^{-1}A^T(b+f)$.

Lösningen till ett vanligt linjärt ekvationssystem, $Cx=b$, kan skrivas, $x=C^{-1}b$, så det verkar rimligt att betrakta $(A^TA)^{-1}A^T$ som en generaliserad invers. Detta gör man, och denna invers kallas pseudoinversen, beteckna A^+ och kan beräknas med Matlabkommandot `pinv`.

A^+ är ett matematiskt hjälpmedel och den brukar inte användas för att lösa minstakvadratproblem i praktiken. Vi ser att A^+ är en vänsterinvers, $A^+A=(A^TA)^{-1}A^TA=I$. Däremot är inte A^+ en högerinvers, så $AA^+\neq I$. Man kan definiera A^+ även om A är rangdefekt (men då gäller inte att $A^+=(A^TA)^{-1}A^T$).

Vi ser att

$$y-x=A^+(b+f)-A^+b=A^+f \Rightarrow ||y-x||_2 \le ||A^+||_2\,||f||_2$$

Vi måste få en undre begränsning av $||x||_2$ och använder sambandet $Ax=b_A$, där b_A är den ortogonala projektionen av b på A:s bildrum. Antag vidare att $b_A\neq 0$ vilket medför att $x\neq 0$. Vi får

$$||b_A||_2=||Ax||_2 \le ||A||_2\,||x||_2 \Rightarrow 1/||x||_2 \le ||A||_2/||b_A||_2$$

Slutligen:

$$\frac{||y-x||_2}{||x||_2} \le \underbrace{||A||_2\,||A^+||_2}_{\kappa_2(A)}\,\frac{||f||_2}{||b_A||_2}$$

Denna gräns liknar den för linjära ekvationssystem. En viktig skillnad är att det inte står $||f||_2/||b||_2$.
Låt oss skriva om uppskattningen:

$$\frac{||y-x||_2}{||x||_2} \le ||A||_2\,||A^+||_2\,\frac{||b||_2}{||b_A||_2}\,\frac{||f||_2}{||b||_2}$$

Om modell och mätdata stämmer väl överens så kommer $||b||_2/||b_A||_2$ att vara nära ett (kvoten är alltid ≥ 1), men om modell och data inte passar ihop så kan kvoten bli stor. Extremfallet är att b är ortogonal mot A:s bildrum i vilket fall $b_A=0$ och kvoten är oändlig.

Skulle kvoten vara väldigt stor är det kanske inte så meningsfullt att lösa minstakvadratproblemet. Stör vi nu även A med F så tillkommer ytterligare en term i feluppskattningen och det visar sig att man även får en faktor $\kappa_2^2(A)\;||b_\perp||_2/||b_A||_2$ gånger de relativa störningarna.

När vi studerade $Ax=b$-problemet sa vi att $||A||/\kappa(A)$ är normen på den minsta störning, E, som gör $A+E$ singulär. Analogt gäller för minstakvadratproblemet att $||A||_2/\kappa_2(A)$ är tvånormen på den minsta E som gör att $A+E$ är rangdefekt ($A+E$ har linjärt beroende kolonner).

Exempel: låt $\epsilon,\mu>0$ vara små och antag att $0\le\psi<\pi/2$. Sätt;

$$A=\begin{bmatrix}1 & 1\\ 0 & \epsilon\\ 0 & 0\end{bmatrix},\quad b=\begin{bmatrix}\cos\psi\\ 0\\ \sin\psi\end{bmatrix},\quad f=\begin{bmatrix}0\\ \mu\\ 0\end{bmatrix}$$

Det gäller att $\kappa_2(A)\approx 2/\epsilon$. Pseudoinversen blir:

$$A^+=(A^TA)^{-1}A^T=\frac{1}{\epsilon}\begin{bmatrix}\epsilon & -1 & 0\\ 0 & 1 & 0\end{bmatrix}$$

varför lösningarna x och y ges av

$$x=\begin{bmatrix}\cos\psi\\ 0\end{bmatrix},\quad y=\begin{bmatrix}\cos\psi-\mu/\epsilon\\ \mu/\epsilon\end{bmatrix}$$

och

$$\frac{||y-x||_2}{||x||_2}=\sqrt{2}\,\frac{\mu}{\epsilon\cos\psi}$$

Det gäller att

$$\frac{||f||_2}{||b||_2}=\frac{\mu}{1}=\mu$$

och att

$$\frac{||b||_2}{||b_A||_2}=\frac{1}{\cos\psi}$$

Vår uppskattning stämmer bra i detta exempel:

$$\underbrace{\frac{||y-x||_2}{||x||_2}}_{\sqrt{2}\;\mu/(\epsilon\cos\psi)} \le \underbrace{\kappa_2(A)}_{2/\epsilon}\;\underbrace{\frac{||b||_2}{||b_A||_2}}_{1/\cos\psi}\;\underbrace{\frac{||f||_2}{||b||_2}}_{\mu}$$

Om $\psi\approx\pi/2$ så är b nästan ortogonal mot A:s bildrum (ψ är i själva verket vinkeln som b bildar mot bildrummet) och felet ökar med (den då stora) faktorn $1/\cos\psi$.

Alternativ till normalekvationerna

Först ett exempel som visar en nackdel med normalekvationerna.

$$A=\begin{bmatrix}1 & 1\\ 0 & \epsilon\\ 0 & 0\end{bmatrix},\ \text{med } \epsilon>0.\ A^TA=\begin{bmatrix}1 & 0 & 0\\ 1 & \epsilon & 0\end{bmatrix}\begin{bmatrix}1 & 1\\ 0 & \epsilon\\ 0 & 0\end{bmatrix}=\begin{bmatrix}1 & 1\\ 1 & 1+\epsilon^2\end{bmatrix}$$

Om $0<\epsilon\le\sqrt{\epsilon_{\text{mach}}}$ så är $fl(1+\epsilon^2)=1$ varför A^TA blir singulär och $A^TAx=A^Tb$ har inte entydig lösning. Minstakvadratproblemet, $\min_x||Ax-b||_2$ har dock entydig lösning så länge som $\epsilon\neq 0$.

Idé: vi utnyttjar att tvånormen är unitärt invariant, dvs.

$$||QAP||_2=||A||_2,\ \text{om } Q^TQ=I,\ P^TP=I$$

förutsatt att P är kvadratisk (Q behöver dock inte vara kvadratisk). Speciellt kan A vara en vektor, v säg, så:

$$||Qv||_2=||v||_2$$

En komplex matris, Q, är unitär då $Q^HQ=I$. Så unitär är motsvarigheten till ortogonal för reella matriser.

Bevis av $||Qv||_2=||v||_2$. Utnyttja att $||\cdot||_2\ge 0$ och att

$$||Qv||_2^2=(Qv)^TQv=v^TQ^TQv=v^TIv=v^Tv=||v||_2^2$$

Sats: Antag att A har linjärt oberoende kolonner. A har då en QR-faktorisering: $A=QR$ där $Q^TQ=I$ och R är övertriangulär med positiva diagonalelement.

Bevis: Beviset är konstruktivt (men ger inte en lämplig algoritm). Låt R definieras av $A^TA=R^TR$, Choleskyfaktoriseringen av A^TA. Denna existerar eftersom A har full kolonnrang, och R är övertriangulär med positiva diagonalelement. Sätt nu $Q=AR^{-1}$. Inversen existerar och Q blir ortogonal, ty: $(AR^{-1})^TAR^{-1}=R^{-T}A^TAR^{-1}=R^{-T}R^TRR^{-1}=I$.

QR-faktoriseringen ovan är av ”economy size”(som det står i `help qr` i Matlab).

$$\underbrace{\begin{bmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \end{bmatrix}}_{R}$$

Följande bild visar den fullständiga varianten:

$$\underbrace{\begin{bmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \end{bmatrix}}_{Z} \underbrace{\begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{\begin{bmatrix} R \\ 0 \end{bmatrix}}$$

eller mer kortfattat

$$A = [Q, Z] \begin{bmatrix} R \\ 0 \end{bmatrix}$$

Varför går detta? Givet en bas, Q, för ett underrum, kan man utvidga den (med Z) så att man får en bas för hela rummet.

Vi använder nu den fullständiga faktoriseringen för att lösa $\min_x ||Ax - b||_2$.

$$||Ax - b||_2 = || \, [Q, Z]^T (Ax - b)||_2 = \left|\left| \begin{bmatrix} Q^T(Ax - b) \\ Z^T(Ax - b) \end{bmatrix} \right|\right|_2 =$$

$$\left|\left| \begin{bmatrix} Rx - Q^T b \\ 0 - Z^T b \end{bmatrix} \right|\right|_2 = \left[\, ||Rx - Q^T b||_2^2 + ||Z^T b||_2^2 \, \right]^{1/2} \geq ||Z^T b||_2$$

där minimum antas när $Rx = Q^T b$.

Vi behöver inte Z för att lösa $Rx = Q^T b$. Vi behöver inte ens Q utan endast $Q^T b$. Q är en matris men $Q^T b$ är en vektor.

En av nackdelarna med normalekvationerna är ju att $\kappa(A^T A) = \kappa(A)^2$. Man kan visa att $\kappa(R) = \kappa(A)$.

Hur ska vi beräkna QR-faktoriseringen på ett bra sätt?

- Klassisk Gram-Schmidt (enkelt men inte så bra). GS.
- Modifierad Gram-Schmidt (mindre dåligt). MGS. Läs själv.
- Householderspeglingar. Bra!
- Householderspeglingar med pivotering (ännu bättre; tar jag inte upp). Standardmetoden, används i Matlab till exempel.

Klassisk Gram-Schmidt via exempel

Låt oss se på ett exempel där A har två kolonner: Gör följande ansats:

$$\underbrace{\begin{bmatrix} a_1 & a_2 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} q_1 & q_2 \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} r_{1,1} & r_{1,2} \\ 0 & r_{2,2} \end{bmatrix}}_{R}$$

Vi ser att $a_1 = q_1 r_{1,1}$ så att $||a_1||_2 = ||q_1||_2 \, |r_{1,1}|$. Men $||q_1||_2 = 1$ och vi kan välja $r_{1,1} > 0$ så att $r_{1,1} = ||a_1||_2$ och $q_1 = a_1 / r_{1,1}$. Med andra ord: $r_{1,1}$ är längden av a_1 och q_1 är den vektor man får när a_1 normeras.

Nu till nästa kolonn. $a_2 = q_1 r_{1,2} + q_2 r_{2,2}$. Vi får:

$$q_1^T a_2 = \underbrace{q_1^T q_1}_{1} r_{1,2} + \underbrace{q_1^T q_2}_{0} r_{2,2}$$

Alltså är $r_{1,2} = q_1^T a_2$. Notera att $a_2 - q_1 r_{1,2} = q_2 r_{2,2}$ och q_1 är ortogonal mot q_2. Vi kan betrakta $r_{1,2}$ som det värde som gör $a_2 - q_1 r_{1,2}$ ortogonal mot q_1. $r_{2,2}$ är längden av $a_2 - q_1 r_{1,2}$.

Här ett numeriskt exempel:

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}, \; r_{1,1} = \sqrt{2}, \quad q_1 = a_1 / r_{1,1} = \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix}$$

$$r_{1,2} = q_1^T a_2 = \begin{bmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = 1/\sqrt{2}$$

$$a_2 - q_1 r_{1,2} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} - \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix} 1/\sqrt{2} = \begin{bmatrix} -1/2 \\ 0 \\ 1/2 \end{bmatrix}$$

Notera att denna vektor är ortogonal mot q_1. Slutligen:

$$r_{2,2} = 1/\sqrt{2}, \quad q_2 = \begin{bmatrix} -1/2 \\ 0 \\ 1/2 \end{bmatrix} / (1/\sqrt{2}) = \begin{bmatrix} -1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix}$$

QR-faktoriseringen kan alltså skrivas:

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} \sqrt{2} & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} \end{bmatrix}$$

Säg att vi vill lösa $\min_x ||Ax - b||_2$ där $b^T = [1, \; 1, \; 1]$.

$$\underbrace{\begin{bmatrix} \sqrt{2} & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} \end{bmatrix}}_{R} x = \underbrace{\begin{bmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix}}_{Q^T} \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}_{b} \quad \Rightarrow \quad x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Z, som vi inte behöver, ges av

$$Z = \begin{bmatrix} 0 \\ \pm 1 \\ 0 \end{bmatrix}$$

Householdermatriser (speglingar)

Låt u vara en kolonnvektor och definiera:

$$H = I - 2 \frac{u u^T}{u^T u}$$

Notera att uu^T är en ytterprodukt (matris) och $u^T u$ är en innerprodukt (skalär).

Övning: visa att $H = H^{-1} = H^T$ så H är symmetrisk och ortogonal och lika med sin egen invers.

H kallas spegling ty H speglar varje vektor i planet $\text{span}(u)^\perp$, ortogonala komplementet till $\text{span}(u)$ (mängden av vektorer ortogonala mot u). Övning: visa detta. Låt a vara godtycklig och sätt $a = \alpha u + v$ där $u^T v = 0$. Studera Ha.

Påstående: låt a vara en godtycklig vektor med n element.
Vi kan hitta H så att element $2, \ldots, n$ i Ha är noll.

Använder vi standardbeteckningen e_j för j-te kolonnen i enhetsmatrisens gäller alltså att $Ha = \alpha \, e_1$ för något α.

Vad är α? Jo, $||Ha||_2 = ||\alpha e_1||_2$ så att $\alpha = \pm ||a||_2$.
Hur ser H ut, dvs. hur skall u väljas?

$$\alpha e_1 = \underbrace{\left[I - 2 \frac{u u^T}{u^T u} \right]}_{H} a = a - u \underbrace{\left[\frac{2 u^T a}{u^T u} \right]}_{\text{skalär}}$$

så u måste vara en multipel av $a - \alpha e_1$ (observera att $uu^T / (u^T u)$ är skalningsoberoende).

Övning: visa att $u = a - \alpha e_1$ faktiskt fungerar.
Vi tar $\alpha = -\text{sign}(a_1) ||a||_2$ för att slippa kancellation.

Vi bildar en H utifrån a. Det kräver lite minne och få operationer för att bilda Hb, för en godtycklig vektor b (Ha bildar vi inte eftersom vi vet att $Ha = \alpha e_1$). Så här kan man göra:

1. $\alpha = -\text{sign}(a_1)||a||_2$.
2. $u_1 = a_1 - \alpha$, $u_k = a_k, k = 2, \ldots, n$.
3. $\beta = 2(u^T b)/(u^T u)$ en skalär.
4. $Hb = b - \beta u$, en linjärkombination av två vektorer.

Det enda extraminne vi behöver är u, en vektor.

Punkt 3 kan alternativt skrivas: Skala om u, $u = u/||u||_2$. Notera att $H = I - 2uu^T$ med detta u. $Hb = b - (2u^T b)u$. Notera att vi aldrig bildar H.

Att applicera H på en matris A är heller inte svårt. Antag att A har tre kolonner. $HA = H[a_1,\ a_2,\ a_3] = [Ha_1,\ Ha_2,\ Ha_3]$.

Vi är nu redo att att beräkna den fullständiga QR-faktoriseringen. När vi räknade ut LU-faktoriseringen multiplicerade vi med L_k-matriser så att A överfördes till övertriangulär form, U. Så, $L_{n-1} \cdots L_2 L_1 A = U$ så att $A = (L_{n-1} \cdots L_2 L_1)^{-1} U$ som ger faktoriseringen. Nu gör vi ungefär på samma sätt med L_k ersatt av H_k. Låt oss anta att A har tre kolonner. Vi kommer att välja H_1, H_2, H_3 så att

$$H_3 H_2 H_1 A = \begin{bmatrix} R \\ 0 \end{bmatrix}, \quad A = \underbrace{(H_3 H_2 H_1)^{-1}}_{[Q,Z]} \begin{bmatrix} R \\ 0 \end{bmatrix}$$

Produkter av ortogonal matriser är ortogonal så att $(H_3 H_2 H_1)^{-1}$ existerar och är ortogonal.

Vi antar att A är en 5×3-matris. Vi börjar med att välja H_1 så att $H_1 a_1 = \alpha_1 e_1$, dvs. så att H_1 applicerat på $\square$-vektorn blir en multipel av e_1. $\times$, $\square$ etc. markerar godtyckliga element (som ej behöver vara noll).

$$H_1 \underbrace{\begin{bmatrix} \square & \times & \times \\ \square & \times & \times \\ \square & \times & \times \\ \square & \times & \times \\ \square & \times & \times \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} \alpha_1 & \times & \times \\ 0 & \triangle & \times \\ 0 & \triangle & \times \\ 0 & \triangle & \times \\ 0 & \triangle & \times \end{bmatrix}}_{A^{(1)}}$$

Tag nu H_2 så att H_2 applicerat på $\triangle$-vektorn blir en multipel av e_1. Observera att denna vektor har 4 och inte 5 element och att H_2 är en 4×4-matris. Vi multiplicerar givetvis inte med ettor eller nollor utan arbetar endast med element under linjen och i andra och tredje kolonnen. Jag har bytt beteckningar så H_2 och H_3 är inte samma matriser som på föregående sida.

$$\left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & H_2 \end{array}\right] \underbrace{\left[\begin{array}{ccc} \alpha_1 & \times & \times \\ \hline 0 & \triangle & \times \\ 0 & \triangle & \times \\ 0 & \triangle & \times \\ 0 & \triangle & \times \end{array}\right]}_{A^{(1)}} = \underbrace{\left[\begin{array}{ccc} \alpha_1 & \times & \times \\ \hline 0 & \alpha_2 & \times \\ 0 & 0 & \diamond \\ 0 & 0 & \diamond \\ 0 & 0 & \diamond \end{array}\right]}_{A^{(2)}}$$

Vi väljer nu H_3 som applicerat på $\diamond$-vektorn blir en multipel av e_1. Observera att denna vektor har 3 element och att H_3 är en 3×3-matris.

$$\left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & H_3 \end{array}\right] \underbrace{\left[\begin{array}{ccc} \alpha_1 & \times & \times \\ 0 & \alpha_2 & \times \\ \hline 0 & 0 & \times \\ 0 & 0 & \times \\ 0 & 0 & \times \end{array}\right]}_{A^{(2)}} = \underbrace{\left[\begin{array}{ccc} \alpha_1 & \times & \times \\ 0 & \alpha_2 & \times \\ \hline 0 & 0 & \alpha_3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right]}_{A^{(3)}} = \left[\begin{array}{c} R \\ \hline 0 \end{array}\right]$$

Vi sammanställer och får:

$$\underbrace{\left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & H_3 \end{array}\right] \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & H_2 \end{array}\right] H_1}_{H} \quad A = \begin{bmatrix} R \\ 0 \end{bmatrix}$$

så att

$$A = H^{-1} \begin{bmatrix} R \\ 0 \end{bmatrix} = H^T \begin{bmatrix} R \\ 0 \end{bmatrix} = [Q, Z] \begin{bmatrix} R \\ 0 \end{bmatrix}$$

Q består av de tre första kolonnerna i H^T och de resterande två bildar Z.

Om vi bara vill lösa minstakvadratproblemet behöver vi aldrig bilda Q (eller H) explicit utan det räcker att känna R och $Q^T b$ (vi löser ju $Rx = Q^T b$). Så i vårt exempel ges R av de tre första raderna i $A^{(3)}$ och $A^{(3)}$ har vi ju bildat utan att explicit bilda H.

$Q^T b$ utgörs av de tre första raderna i Hb. Varför? $H^T = [Q, Z]$ så att:

$$Hb = [Q, Z]^T b = \begin{bmatrix} Q^T \\ Z^T \end{bmatrix} b = \begin{bmatrix} Q^T b \\ Z^T b \end{bmatrix}$$

Men vi kan ju bilda Hb genom att applicera H_1, H_2 och H_3 på b och sedan plocka ut de tre första elementen. Vi behöver således inte Q här heller.

Notera att vi inte måste spara alla u-vektorer som bildar H-matriserna heller. I praktiken brukar man skriva över A med R (plus nollor) så vi behöver inget extra matris-minne.

Illakonditionerade problem, regularisering

Tyvärr räcker inte alltid ovanstående metoder till. Antag att vi har beräknat en QR-faktorisering och att R har utseendet (hårt avrundat):

$$R = \begin{bmatrix} 2.4 & -1.7 & 3.2 \\ 0 & 1.2 \cdot 10^{-5} & 3 \\ 0 & 0 & 2.3 \cdot 10^{-10} \end{bmatrix}$$

I detta fall är R (och därmed A) mycket illa konditionerad. $\kappa(A) \approx 7 \cdot 10^{14}$.

När vi löser $Rx = Q^T b$ kommer vi (implicit) att invertera R. De små diagonalelementen kommer då att bestämma hur lösningen ser ut eftersom R^{-1} kommer att innehålla inversen av dessa element (bland annat). T.ex. gäller att $(R^{-1})_{kk} = 1/r_{kk}$.

Frågan är nu hur mycket man litar på de små r_{kk}. Är de säkra värden eller består de bara av mätfel och avrundningsfel?

Vi litar nog mycket på r_{11}-elementet, det är ju relativt stort. r_{22} tror vi kanske rätt mycket på, men $r_{33} = 2.3 \cdot 10^{-10}$ kanske vi tvivlar på. Problemet är att $1/r_{33}$ i stor utsträckning kommer att bestämma utseendet på lösningen x (om inte b är väldigt speciell).

Det säkra värdet, 2.4 bidrar knappast något till lösningen, eftersom $1/2.4$ är så litet. Hur mycket vi litar på värdena beror på mätfel och modell.

Om man tror att ett värde helt eller delvis består av brus är det ingen mening att bara räkna på. Den lösning man får fram är nog tämligen meningslös. GIGO = Garbage In, Garbage Out.

Hur kan ett sådant illa konditionerat problem uppstå? Det finns flera orsaker. På nästa följer några exempel.