Some Linear Algebra Basics, Orthogonal Projections and the QR Decomposition

Rob McCulloch

1. Goals

In this set of notes and the next we want to become familiar with some of the basic vector/matrix (Linear Algebra) ideas that are pervasive in statistics.

For example in numpy.linalg (https://numpy.org/doc/stable/reference/routines.linalg.html) we have:

Matrix and vector products

dot(a, b[, out])	Dot product of two arrays.
linalg.multi_dot(arrays, *[, out])	Compute the dot product of two or more arrays in a single function call, while automatically selecting the fastest evaluation order.
vdot(a, b)	Return the dot product of two vectors.
inner(a, b)	Inner product of two arrays.
outer(a, b[, out])	Compute the outer product of two vectors.
matmul(x1, x2, /[, out, casting, order, ...])	Matrix product of two arrays.
tensordot(a, b[, axes])	Compute tensor dot product along specified axes.
einsum(subscripts, *operands[, out, dtype, ...])	Evaluates the Einstein summation convention on the operands.
einsum_path(subscripts, *operands[, optimize])	Evaluates the lowest cost contraction order for an einsum expression by considering the creation of intermediate arrays.
linalg.matrix_power(a, n)	Raise a square matrix to the (integer) power *n*.
kron(a, b)	Kronecker product of two arrays.

Decompositions

linalg.cholesky(a)	Cholesky decomposition.
linalg.qr(a[, mode])	Compute the qr factorization of a matrix.
linalg.svd(a[, full_matrices, compute_uv, ...])	Singular Value Decomposition.

Matrix eigenvalues

linalg.eig(a)	Compute the eigenvalues and right eigenvectors of a square array.
linalg.eigh(a[, UPLO])	Return the eigenvalues and eigenvectors of a complex Hermitian (conjugate symmetric) or a real symmetric matrix.
linalg.eigvals(a)	Compute the eigenvalues of a general matrix.
linalg.eigvalsh(a[, UPLO])	Compute the eigenvalues of a complex Hermitian or real symmetric matrix.

Norms and other numbers

linalg.norm(x[, ord, axis, keepdims])	Matrix or vector norm.
linalg.cond(x[, p])	Compute the condition number of a matrix.
linalg.det(a)	Compute the determinant of an array.

We need to know what some of these are!

Note:

This will not be a formal "intro to Linear Algebra".

Just an informal, hopefully intuitive reminder of basic ideas that are fundamental for us.

That is, I'm not dotting all the i's and crossing all the t's, but I need to be able to say things like "so these vectors are an orthonormal basis" and you know what I mean.

In particular the following matrix decompostions are important:

- QR
- spectral, (eigen values and vectors)
- Cholesky
- Singular value

So, we will review these and get a look at how the play a role in statistics.

2. Vectors, Matrices, and Linear Combinations

A vector x in R^n is

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix}$$

column vector

$$x' = x^T = [x_1, x_2, \ldots x_n]$$

row vector

I tend to use both notations for the transpose.

The default is that a vector is column vector.

We multiply a scalar times a vector and we add vectors:

$$\alpha \in \mathbb{R}, \quad x \in \mathbb{R}^n = [x_i]$$

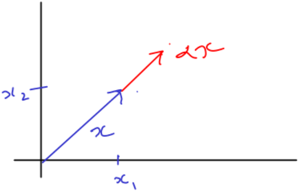
Then $\quad \alpha x = [\alpha x_i]$

$y \in \mathbb{R}^n$

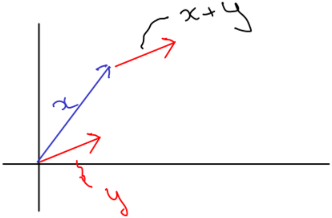
$$x + y = [x_i + y_i]$$

So

$$\alpha x + \beta y = [\alpha x_i + \beta y_i]$$



$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$



$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

Linear Combinations

Let $\{x_1, x_2, \ldots, x_m\}$ be vectors in R^n.

Note that now x_i is the i^{th} vector, not the i^{th} component of the vector x.

A linear combination of the $\{x_i\}$ is $\sum_{i=1}^{m} \alpha_i \, x_i$.

Linearly Independent

$\{x_1, x_2, \ldots, x_m\}$ are linearly independent if,

$$\sum_{i=1}^{m} \alpha_i \, x_i = 0 \quad \Longleftrightarrow \quad \alpha_i = 0, i = 1, 2, \ldots, m.$$

Span

Let $S = \{x_i\}$. The span of S is the $\{\sum \alpha_i x_i, x_i \in S, \alpha_i \in R\}$.

That is, all the linear combinations of vectors in S.

Subspace

A subset S of R^n is a (linear) subspace if
$x, y \in S \implies \alpha x + \beta y \in S$.

Basis

The set of vectors B is a basis for the subspace M if the vectors in B are linearly independent and M is the span of S.

Dimension of a Subspace

The dimension of a subspace is the number of vectors in a basis. You can show this is well defined.

Note Suppose $\{x_i\}_{i=1}^{p}$ is a basis.

Suppose $y = \sum \alpha_i x_i = \sum \beta_i x_i$

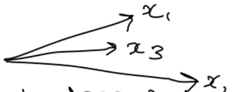
then $\sum (\alpha_i - \beta_i) x_i = 0 \implies \alpha_i = \beta_i$

— coefficients are unique

Note suppose $\{x_i\}$ are linearly dependent.

$\exists\ \alpha_i$ s.t. $\sum \alpha_i x_i = 0 \qquad \alpha_j \neq 0$

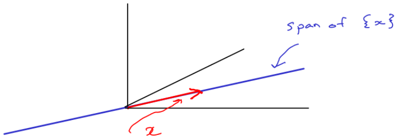
$$x_j = \sum_{i \neq j} \left\{ \frac{-\alpha_i}{\alpha_j} \right\} x_i$$



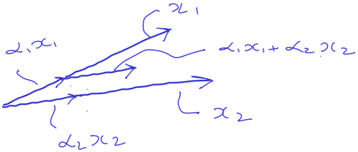
— $\{x_1, x_2, x_3\}$ are linearly dependent

— $x_3 = a x_1 + b x_2$

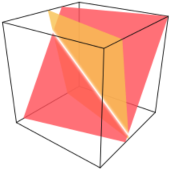
The span of $\{x\}$ is a one dimensional subspace.



The span of $\{x_1, x_2\}$ is a two dimensional subspace.



The intersection of two subspaces is a subspace.



Note the magic !!!: we imagine these vectors to be in R^n!!.

Standard Basis of R^n:

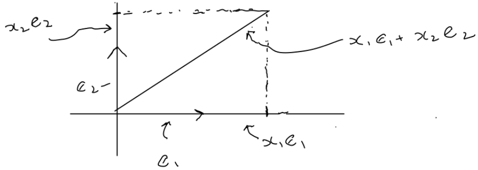
Let $e_i = [0, 0, \ldots, 1, 0, 0, \ldots, 0]'$ where the 1 is in the i^{th} position.

Then, for $x = [x_i]$, $x = \sum_{i=1}^{n} x_i e_i \implies$ span of $\{e_i\}$ is R^n.

$\sum \alpha_i e_i = 0 \implies \alpha_i = 0, i = 1, 2, \ldots, n$, so the set $\{e_i\}$ is linearly independent.

So, dimension of R^n is n.

The set $\{e_i\}$ is called the *standard basis*.



3. Inner Products

$x = [x_i]$, $y = [y_i]$, $x, y \in R^b$.

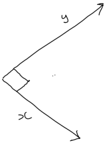
The *inner product* between x and y is:

$$< x, y >= \sum x_i \, y_i .$$

The geometric intuition is the $< x, y >$ tells us about the *angle* between x and y.

Orthogonal vectors:

x is *orthogonal* to y if $< x, y >= 0$.



We write $x \perp y$.

L2 (euclidean) norm:

$$||x|| = \sqrt{< x, x >}.$$

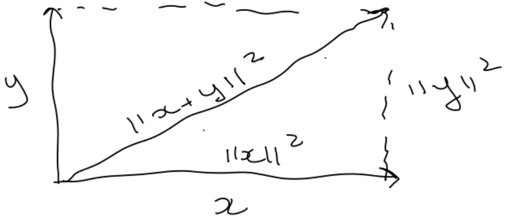
Euclidean distance:

$$x_1, x_2 \in R^n, \quad d(x_1, x_2) = ||x_1 - x_2||.$$

Note:

Suppose $x \perp y$, $z = x + y$, then,

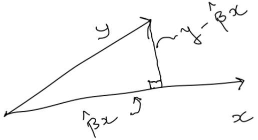
$$
\begin{aligned}
||z||^2 = & \\
& = < x+y, x+y > \\
& = < x, x > +2 < x, y > + < y, y > \\
& = ||x||^2 + ||y||^2
\end{aligned}
$$



Orthogonal Projection of y on x:

We want to "project" y onto x.

The projection is a vector in the span of $\{x\}$ so it equals $\hat{\beta}x$ for some $\hat{\beta}$.



We want the residual, $y - \hat{\beta}x$ to be orthogonal to x:

$$0 = < y - \hat{\beta}x, x > = < y, x > - \hat{\beta} < x, x > \implies \hat{\beta} = \frac{< y, x >}{< x, x >}.$$

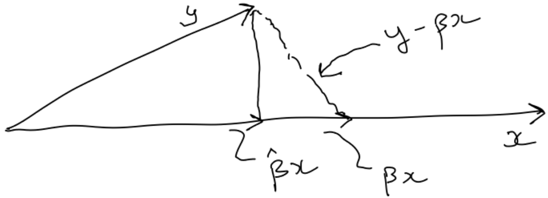
projection gives the minimum distance:

Suppose we want:

$$\underset{\hat{y} \in Span(\{x\})}{\text{minimize}} \; ||y - \hat{y}||^2$$

Which is the same as:

$$\underset{\beta \in R}{\text{minimize}} \; ||y - \beta\, x||^2$$

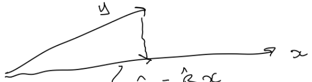


$$y - \beta x = (y - \hat{\beta} x) + (\hat{\beta} x - \beta x)$$

$$\begin{aligned} \|y - \beta x\|^2 &= \|\hat{\beta} x - \beta x\|^2 + \|y - \hat{\beta} x\|^2 \\ &= (\hat{\beta} - \beta)^2 \|x\|^2 + \|y - \hat{\beta} x\|^2 \end{aligned}$$

So clearly the minimum is at $\beta^* = \hat{\beta}$.

Cauchy Swartz Inequality



$$\hat{y} = \hat{\beta} x$$

$$\begin{aligned}
\|y\|^2 &\geq \|\hat{\beta} x\|^2 \\
&= \hat{\beta}^2 \cdot \|x\|^2 \\
&= \left(\frac{<x,y>}{\|x\|^2}\right)^2 \|x\|^2 \\
&= \frac{<y,x>^2}{\|x\|^2}
\end{aligned}$$

So $\left(\text{divide by } \|y\|^2\right) \Rightarrow 1 \geq \dfrac{<y,x>^2}{\|x\|^2 \, \|y\|^2}$

So,

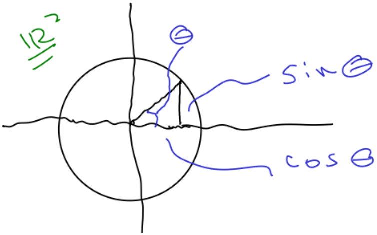
$$-1 \leq \frac{<x,y>}{||x||\,||y||} \leq 1$$

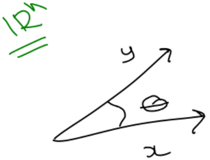
The angle between two vectors:

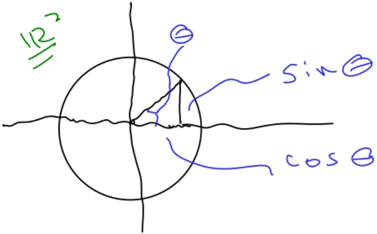
$x, y \in R^n$.

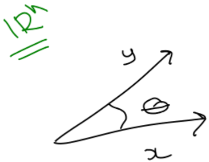
Given the CS inequality, we can let the angle between x and y be given by

$$\cos(\theta) = \frac{< x, y >}{||x||\,||y||}, \quad \theta \in [0, \pi].$$









example: $x \perp y \Rightarrow \cos(\theta) = 0, \theta = \pi/2 = 90$ degrees.

example: $\cos(\theta) = 1, \theta = 0$, x and y are colinear.

4. Matrices

A matrix is a two-way array.

The $n \times m$ matrix X is $[x_{ij}]$, $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$.

$$X = \begin{bmatrix} x_{11} & x_{21}, & \cdots & x_{1j} & \cdots & x_{1m} \\ x_{21} & & & & & \vdots \\ \vdots & & & & & \vdots \\ x_{i1} & \cdots & & x_{ij} & \cdots & x_{im} \\ \vdots & & & & & \\ x_{n1} & \cdots & & x_{nj} & \cdots & x_{n,m} \end{bmatrix}$$

ith row

jth column

It often helps to think of a matrix as a bunch of columns:

$$X = [x_1, x_2, \ldots x_j, \ldots x_m]$$

$$x_j \in \mathbb{R}^n, \; j = 1, 2, \ldots m$$

It often helps to think of a matrix as a bunch of rows:

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_i' \\ \vdots \\ x_n' \end{bmatrix} \qquad x_i \in \mathbb{R}^m, \; i = 1, 2 \ldots n$$

The Transpose:

To transpose a matrix we flip the rows and columns.

$$X^T = \{x_1, x_2, \ldots x_m\}^T$$

$$= \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_m^T \end{bmatrix}$$

So if X is $n \times m$ then X' is $m \times n$.

Symmetric Matrices

A square matrix $(n \times n)$ is symmetric if $A = A'$.

Matrix Multiplication:

$$A_{n \times m} = \begin{bmatrix} a_1' \\ a_2' \\ \vdots \\ a_n' \end{bmatrix}$$

$$B_{m \times p} = [b_1, b_2 \dots b_p]$$

$$a_i \in \mathbb{R}^m \quad b_j \in \mathbb{R}^m$$

$$A_{n \times m} B_{m \times p} = [\langle a_i, b_j \rangle] = [a_i^T b_j]$$

$$i = 1, 2 \dots n$$
$$j = 1, 2 \dots p$$

So AB is $n \times p$

Several ways to think about matrix multiplication.

A is $n \times p$. $b \in R^p$.

Ab is a linear combination of the the columns of A.

$$X_{n \times p}, \quad b_{p \times 1}$$

$$Xb = [x_1, x_2 \cdots x_p] \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} = \sum x_i b_i$$

$$B = [b_1, b_2, \cdots b_m]$$

$$XB = [Xb_1, Xb_2, \cdots Xb_m]$$

Similarly $b'A$ is a linear combination of the rows for A.

Note

$$(AB)' = B'A'.$$

$$A = [a_{ij}], B = [b_{ij}], \text{same dimensions}, \quad aA + bB = [a\, a_{ij} + b\, b_{ij}].$$

$$C(A + B) = CA + CB$$

and so on...

Note

$$x, y \in R^n. \quad < x, y >= x'y = y'x.$$

$$x \in R^n,\ y \in R^m. \ \ xy' = [x_i y_j].$$

Linear Transformation

A fundamental way to think about a matrix is as a linear transformation.

For A, $n \times p$:

$$A(x) = A\,x.$$

$$A : R^p \Rightarrow R^n$$

Linear:

$$A(\alpha\,x + \beta\,z) = \alpha\,A(x) + \beta\,A(z)$$

Diagonal Matrices

A square matrix A is diagonal if $A = [a_{ij}]$ has $a_{ij} = 0, \forall i \neq j$.

We write $A = diag(a) = diag(a_1, a_2, \ldots, a_n)$ means:

$A =$

$$\begin{bmatrix} a_1 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & a_2 & 0 & \cdot & \cdots & 0 \\ 0 & & & & & \\ \vdots & & & & & \\ 0 & - & - & - & 0 & a_n \end{bmatrix}$$

The Identity

$I = diag(1, 1, \ldots, 1)$.

$Ix = x$.

Rank of a matrix

Suppose

$$X = [x_1, x_2, \ldots, x_p].$$

Let $sp(X)$ be be span of the columns of X:

$$sp(X) = \{Xb, b \in R^p\}.$$

The *column rank* is the dimension of $sp(X)$.

Similarly, the *row rank* is the dimension of $sp(X')$, the rows of X.

It runs out that the row rank is the same as the column rank so we can define the rank of a matrix to be the column rank.

Inverse of a Matrix

Suppose A is an $n \times n$ square matrix.

Suppose the rank of A is n.

Then the columns of A form a basis for R^n.

Hence, for any $y \in R^n$ there is a unique $b \in R^n$ such that $y = Ab$.

Hence, $\exists$ a matrix A^{-1} which is the inverse of A.
That is,

$$y = Ab \Rightarrow b = A^{-1}y.$$

Note

$$AA^{-1} = A^{-1}A = I.$$

$$(AB)^{-1} = B^{-1}A^{-1}.$$

$$I = I' = (AA^{-1})' = (A^{-1})'A'.$$

$$(A')^{-1} = (A^{-1})'$$

Trace of a Matrix

$A = [a_{ij}], \quad n \times n.$

Trace of A:

$$tr(A) = \sum_{i=1}^{n} a_{ii}$$

$A, n \times k, \quad B, k \times n,$

$$tr(AB) = tr(BA)$$

example:

$y, x \in R^n.$
$y'x = tr(y'x) = tr(xy').$

5. Orthogonal Projections and Orthogonal Matrices

Suppose V is a subspace of R^n with $dim(V) = p < n$.

For any $y \in R^n$, we want to *orthogonally project* y onto V.

Let P_V denote the map such that $P_V y$ is the projection.

That is,

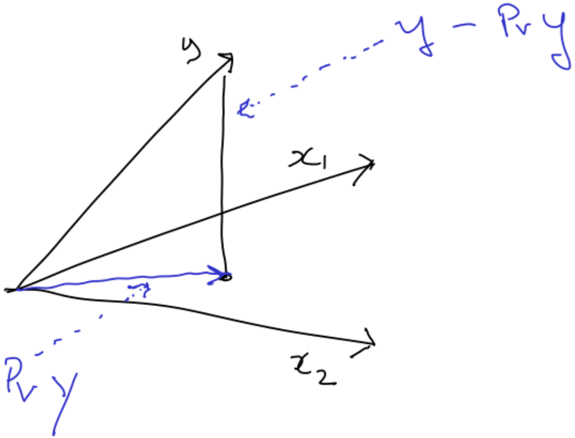
$$P_V y \in V, \;\; y - P_V y \perp V.$$

That is,

$$< y - P_V y, v >= 0, \forall\, v \in V.$$

We can always find a basis for V.

Let X be the matrix whose columns are the basis vectors, $sp(X) = V$.



Let $X = [x_1, x_2, \ldots, x_p]$, $rank(X) = p$.

Given y, there is some b such that $P_V y = Xb$.

We need:

$$< y - Xb, x_j >= 0, \forall j, \quad \Longleftrightarrow X'(y - Xb) = 0.$$

$$X = \{x_1, x_2, \ldots x_p\}$$

$$X^T (y - Xb)$$

$$= \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_p^T \end{bmatrix} (y - Xb) = \begin{bmatrix} x_1^T (y - Xb) \\ x_2^T (y - Xb) \\ \vdots \\ x_p^T (y - Xb) \end{bmatrix}$$

$$= \begin{bmatrix} \langle x_1, y - Xb \rangle \\ \langle x_2, y - Xb \rangle \\ \vdots \\ \langle x_p, y - Xb \rangle \end{bmatrix}$$

$$X^T(y - Xb) = 0$$

$$X^T y = X^T X b$$

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

$$P_V = X(X^T X)^{-1} X^T$$

Very cool.

Incredibly important.

V perp

Let V be a subspace.

$$V^{\perp} = \{x \text{ such that } x \perp v, \forall\, v \in V\}.$$

$V^{\perp}$ is a subspace.

$P_{V^{\perp}} = I - P_V$

$$y = P_V y + P_{V^{\perp}} y, \;\; ||y||^2 = ||P_V y||^2 + ||P_{V^{\perp}} y||^2.$$

Minimum Distance to a Linear Subspace

Let y be a vector in R^n.

Let V be a p dimensional subspace.

Let X be a $n \times p$ matrix whose columns are a basis for V.

$$\underset{v \in V}{\text{minimize}} \; ||y - v||^2$$

Which is the same as

$$\underset{b \in R^p}{\text{minimize}} \; ||y - Xb||^2$$

Let $\hat{b} = (X'X)^{-1}X'y$.

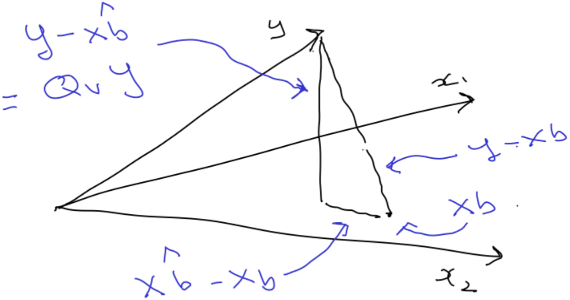
So, $P_V y = X\hat{b}$.

Let $Q_V = P_{V^\perp}$.

$$
\begin{aligned}
||y - Xb||^2 = & \\
& = ||P_V y + Q_V y - Xb||^2 \\
& = ||X(\hat{b} - b) + Q_V y||^2 \\
& = ||Q_V y||^2 + ||X(\hat{b} - b)||^2 \\
& = ||Q_V y||^2 + (\hat{b} - b)'X'X(\hat{b} - b)
\end{aligned}
$$

So, the min is at $b^* = \hat{b}$.

$V = \mathrm{span}(\{x_1, x_2\})$



$y - xb = Q_V y + x\hat{b} - xb$

Sum of Subspaces

V, W subspaces.

$$V + W = \{v + w, v \in V, w \in W\}.$$

$V + W$ is a subspace.

Orthogonal Subspaces

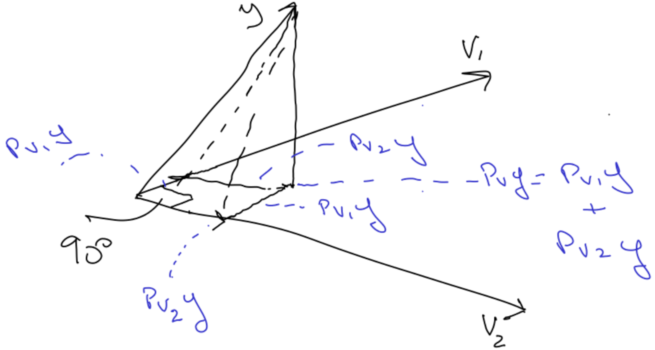
V, W subspaces.

$$V \perp W \iff v \perp w, \ \forall\, v \in V, \ w \in W.$$

Key result

V_1, V_2 orthogonal subspaces.

$$P_{V_1+V_2} = P_{V_1} + P_{V_2}$$



Let $V_i = span(X_i)$.

$V_1 \perp V_2 \Rightarrow X_1' X_2 = 0$.

$$X = [X_1, X_2] \quad X_1' X_2 = 0$$

$$X'X = \begin{bmatrix} X_1' \\ X_2' \end{bmatrix} [X_1, X_2] = \begin{bmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{bmatrix}$$

$$X(X'X)^{-1}X' = [X_1, X_2] \begin{bmatrix} (X_1'X_1)^{-1} & 0 \\ 0 & (X_2'X_2)^{-1} \end{bmatrix} \begin{bmatrix} X_1' \\ X_2' \end{bmatrix}$$

$$= [X_1, X_2] \begin{bmatrix} (X_1'X_1)^{-1} X_1' \\ (X_2'X_2)^{-1} X_2' \end{bmatrix}$$

$$= X_1 (X_1'X_1)^{-1} X_1' + X_2 (X_2'X_2)^{-1} X_2'$$

$$\equiv P_{V_1} + P_{V_2}$$

Projecting onto the sum of orthogonal subspaces

V_i is a subspace, $i = 1, 2, \ldots, m$.

$V_i \perp V_j, \ i \neq j$.

$$P_{\sum_{i=1}^m V_i} = \sum_{i=1}^m P_{V_i}.$$

To project onto the sum of orthogonal subspaces, you can project onto each subspace one at a time and then add up the projections.

This underlies a ton of stuff in statistics (e.g. ANOVA).

$$||P_{\sum_{i=1}^m V_i} y||^2 = \sum_{i=1}^m ||P_{V_i} y||^2.$$

Orthonormal vectors

A set of vectors $\{o_1, o_2, \ldots, o_p\}$ is *orthonormal* if

$$||o_i|| = 1, \forall i, \;\; < o_i, o_j >= 0, \; i \neq j.$$

A set of orthonormal vectors is always linearly independent.

If $V = span(\{o_i\})$, then

$$P_V y = \sum_{i=1}^{p} < o_i, y > o_i.$$

$$||P_V y||^2 = \sum_{i=1}^{p} (< o_i, y >)^2.$$

We can see this with the matrix formula for the projection.

Again let $\{o_1, o_2, \ldots, o_p\}$ be orthonormal.
Then,

$$O = [o_1, o_2, \cdots o_p]$$

$$O^T O = I$$

$$O(O^T O)^{-1} O^T = O O^T$$

$$O O^T y = [o_1, o_2, \cdots o_p] \begin{bmatrix} o_1^T \\ o_2^T \\ \vdots \\ o_p^T \end{bmatrix} y$$

$$= \sum o_i \langle o_i, y \rangle$$

Orthogonal matrices

If $p = n$ we have $O = [o_1, o_2, \ldots, o_n]$ with

$$O'O = OO' = I$$

O is an *orthogonal matrix*.

Orthogonal matrices play a key role in 3 out of 4 of our important matrix decompositions !!!!

Two ways to look at orthogonal matrices.

$O'O = I \implies$ ON matrix is a rotation

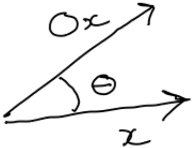
Thinking of $x \Rightarrow Ox$ as a map from R^n to R^n, O is a rotation.

Because $O'O = I$,

$$< Ox, Oy >= x'O'Oy = x'y =< x, y > .$$

and,

$$||Ox|| = ||x||.$$



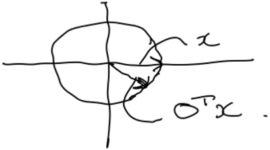
In R^2

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$O = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} = \text{counter clockwise rotation by } \theta$$

$$O^T = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} = \text{clockwise rotation by } \theta$$

$$x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$



$OO' = I \implies$ ON matrix is change of basis

If $\{v_i\}_{i=1}^n$ is a basis for R^n then any $x \in R^n$ can be written as $\sum c_i v_i$.

By a *change of basis* we mean writing vectors in terms of an alternative basis.

If $\{u_i\}_{i=1}^n$ is also a basis for R^n, then $x = \sum d_i u_i$, for some d_i.

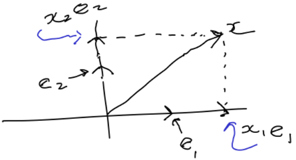
Let $O = [o_1, o_2, \ldots, o_n]$.

$$x = Ix = OO'x = \sum o_i < o_i, x >$$

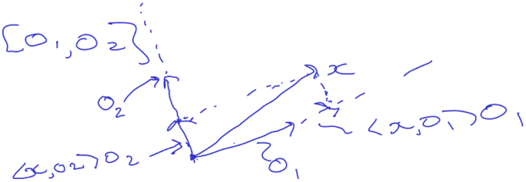
In R^2

$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ $e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ $I = [e_1, e_2]$

$x = x_1 e_1 + x_2 e_2$



$O = [O_1, O_2]$



$$x = < o_1, x > o_1 + < o_2, x > o_2$$

6. Gram Schmidt and QR

Let $X = [x_1, x_2, \ldots, x_p]$. $V_j = span(\{x_1, x_2, \ldots, x_j\})$.

Let $e_1 = x_1$ $\quad o_1 = \frac{x_1}{\|x_1\|}$

for j in 2:p:

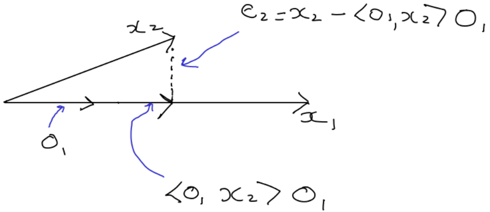
$$e_j = x_j - \sum_{k=1}^{j-1} \langle o_k, x_j \rangle o_k$$

$$= x_j - P_{V_{j-1}} x_j = Q_{V_{j-1}} x_j$$

$$o_j = \frac{e_j}{\|e_j\|}$$

- $span(\{x_1, x_2, \ldots, x_j\}) = span(\{o_1, o_2, \ldots, o_j\})$.
- for $O = [o_1, o_2, \ldots, o_p]$, $O'O = I_p$.

$p = 2$



$$e_1 = x_1 \qquad o_1 = \frac{e_1}{\|e_1\|}$$

$$e_2 = x_2 - \langle o_1, x_2 \rangle o_1$$

$$o_2 = \frac{e_2}{\|e_2\|}$$

$$x_1 = \|e_1\| o_1 \equiv r_{11} o_1$$

$$x_2 = o_2 \|e_2\| + \langle o_1, x_2 \rangle o_1 \equiv r_{22} o_2 + r_{12} o_1$$

$$X = [x_1, x_2] = [o_1, o_2] \begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix} \equiv QR$$

The QR Decomposition

In general, since x_j always is a linear combination of $\{o_1, o_2, \ldots, o_j\}$, we can always write $X = [x_1, x_2, \ldots, x_p]$ as

$$X = Q\,R$$

where

- $Q'Q = I$, if $p = n$, Q is orthogonal.
- R is upper triangular, $p \times p$.

Upper Triangular: $R = [r_{ij}], r_{ij} = 0$ for $i < j$.

r_{ij}: $x_j \in span(\{o_1, o_2, \ldots, o_j\}) \Rightarrow$
$x_j = \sum_{i=1}^{j} r_{ij} o_i = \sum_{i=1}^{j} < x_j, o_i > o_i$.

Note

- Given a basis for a subspace, you can always construct an orthonormal basis.
- The inverse of an upper triangular matrix is upper triangular.
- For X, $n \times n$, $\sim O(n^3)$ operations.

QR and Regression

http://madrury.github.io/jekyll/update/statistics/2016/07/20/lm-in-R.html

```
c     dqrdc2 uses householder transformations to compute the qr
c     factorization of an n by p matrix x.
```

This is where the actual work is done. We are going to decompose X into its `QR` factorization.

$$X = QR, \; Q \text{ orthogonal}, \; R \text{ upper triangular}$$

This is a smart thing to do, because once you have Q and R you can solve the linear equations for regression

$$X^t X \beta = X^t y$$

very easily. Indeed

$$X^t X = R^t Q^t Q R = R^t R$$

so the whole system becomes

$$R^t R \beta = R^t Q^t y$$

R is upper triangular, so it has the same rank as $X^t X$, and if our problem is well posed then $X^t X$ has full rank. So, as R is a full rank matrix, we can ignore the R^t factor in the equations above, and simply seek solutions to the equation

$$R\beta = Q^t y$$

But here's the awesome thing. Again, R is upper triangular, so the last linear equation here is just `constant * beta_n = constant`, so solving for β_n is trivial. We can then go up the rows, one by one, and substitute in the βs we already know, each time getting a simple one variable linear equation to solve. So, once we have Q and R, the whole thing collapses to what is called *backwards substitution*, which is easy.

The simplest and most intuitive way to compute the QR factorization of a matrix is with the Ghram-Schmidt procedure, which unfortunately is not suitable for serious numeric work due to it's instability. Linpack instead uses Householder reflections, which have better computational properties.

Backsolve

We often want to solve $Ax = y$ for x given y and A.
If A is triangular, this is easy.

This is often called "backsolve".

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$y_2 = r_{22} x_2 \Rightarrow x_2 = \frac{y_2}{r_{22}}$$

$$y_1 = r_{11} x_1 + r_{12} x_2$$

$$\Rightarrow x_1 = \frac{y_1 - r_{12} x_2}{r_{11}}$$

all of our matrix decompositions involve

- Orthogonal matrices
- diagonal matrices
- upper/lower triangular matrices

7. Determinant

Let A be a square matrix.

The determinant of a square matrix will play a key role in some statistical compuations.

For example, the densities of the multivariate normal and multivariate t involve determinants.

The determinant of a $n \times n$ matrix is a number.

$$det : R^{n \times n} \Rightarrow R.$$

Here is an intuitive definition of the determinant.

Let C^n be the unit cube in R^n. That is, $C^n = [0, 1]^n$.

$$det(A) \equiv |A| = \text{Volume}(\{Ax, x \in C^n\}) \times (-1)^k$$

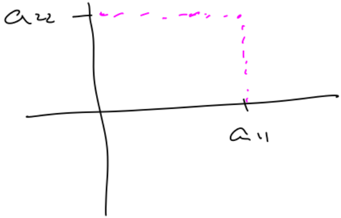
where k is the number of orientation flips.

I can get away with being vague about "orientation flips" because most of the time either it will be zero, or we just need the absolute value of the determinant.

Example

$$A = \begin{bmatrix} a_{11} & 0 \\ 0 & a_{12} \end{bmatrix}$$

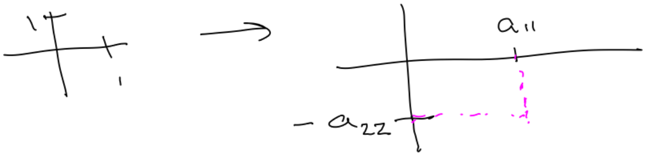
$\Rightarrow$



$\Rightarrow \quad |A| = a_{11} a_{22}$

Example

$$A = \begin{bmatrix} a_{11} & 0 \\ 0 & -a_{22} \end{bmatrix} \qquad a_{11}, a_{22} > 0$$



$$|A| = -a_{22}a_{11}$$

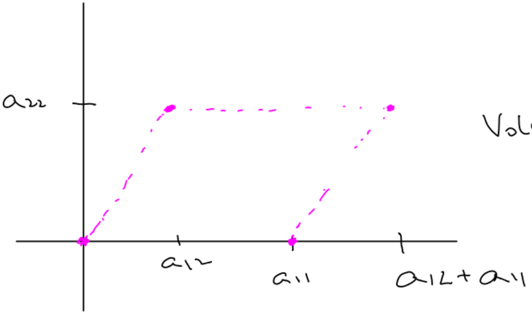
$$= (-1)\text{ Volume.}$$

One orientation flip.

Example

$$A = \begin{bmatrix} a_{11} & a_{12} \\ 0 & a_{22} \end{bmatrix}$$

$$[0,1] \times [0,1] \Rightarrow$$



Volume =

$a_{11}a_{22}$

Key Properties of the Determinant

A and B are square, U is upper trianguler, L is lower triangular, and O is orthogonal.

Diagonals of U and L are positive. $|A|_+$ is the absolute value of the determinant of A.

$$|\mathrm{diag}(a_i)| = \prod a_i$$

$$|U| = \prod U_{ii}$$

$$|L| = \prod L_{ii}$$

$$|AB| = |A||B|$$

$$A^{-1} \text{ exists} \iff |A| \neq 0$$

$$|A^{-1}| = \frac{1}{|A|}$$

$$|A^T| = |A|$$

$$|\lambda A| = \lambda^n |A|$$

$$\boxed{\begin{aligned} \begin{vmatrix} A & 0 \\ 0 & B \end{vmatrix} &= |A||B| \\ |O|_+ &= 1 \end{aligned}}$$

Key

All of our matrix decompostions involve upper and lower triangular, diagonal, and orthogonal matrics, and products of matrices.

For of these cases, the determinant is simple and intuitive.

Example

X is $n \times n$. $X = QR$.

$det(X) = \prod R_{ii}$.

8. Random Variables and Vectors

Recall that for a discrete random variable X we have:

$$P(X = x_k) = p_k, k = 1, \ldots, m, \;\; E(X) = \sum p_k x_k.$$

Recall that for a continuous random variable X we have:

$$P(X \in A) = \int_A f(x)dx, \;\; E(X) = \int f(x)\, x\, dx.$$

$$Var(X) = E((X - E(X))^2), \;\; Cov(X, Y) = E((X - E(X))(Y - E(Y))).$$

We need to work the vectors of random variables.

Expectation of a Random Vector

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$$

$$E\{\sum a_i X_i\} = \sum a_i E[X_i]$$

$$E\{X\} = \begin{bmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_p] \end{bmatrix} \overset{\text{often}}{=} \mu$$

Expectation of a Random Matrix

$$X = [x_{ij}]$$

$$E[X] = \left[E[x_{ij}] \right]$$

Variance (or Covariance) of a Random Vector

$$X = [X_i]$$

$$cov(X) = E((X-\mu)(X-\mu)')$$

$$= [E((X_i-\mu_i)(X_j-\mu_j))]$$

$$\overset{\text{often}}{=} \Sigma = [\sigma_{ij}]$$

Σ is symmetric: $\Sigma' = \Sigma$.

I will probably use both $Var(X)$ and $cov(X)$ for the same thing.

Expectation of a Matrix (Matrices) time a Random Vector (Matrix)

$X = \{x_i\}$ $\quad E\{AX\} = A\,E\{X\}$

$X = \{x_{ij}\}$ $\quad E\{AXB\}$

$= A\,E\{X\}\,B$

Variance (Covariance) of a Matrix times a Random Vector

$$\mathrm{Var}(AX) = \;?$$

note $\quad AX - E[AX] = AX - A\mu$

$$= A(X-\mu)$$

$$\mathrm{Var}(AX) = E\{(A(X-\mu))(A(X-\mu))'\}$$

$$= E\{A(X-\mu)(X-\mu)'A'\}$$

$$= A\Sigma A'$$

A Single Linear Combination

$X = [X_i], i = 1, 2, \ldots, p.\ a \in R^p.$

$$E(a'X) = a'\mu.$$

$$Var(a'X) = a'\Sigma a = \sum a_i a_j \sigma_{ij} = \sum a_i^2 \sigma_{ii} + \sum_{i<j} 2a_i a_j \sigma_{ij}.$$

$$Var(a'X) = a'\Sigma a$$

Since $Var(a'X) \geq 0$ we have

$$a'\Sigma a \geq 0, \forall a.$$

Σ is *positive semi-definite*.

If $a'\Sigma a > 0, \forall a$, Σ is *positive definite*.

9. Statistical Connections

Let's go back through the linear algebra and explore some of the basic statistical connections.

We have already seen how the QR decomposition is used in linear regression.

Sample variance and standard deviation

Suppose we have observation on a single numeric x.

$$x = (x_1, x_2, \ldots x_n)$$

$$\bar{x} = \frac{1}{n} \sum x_i \,, \quad \tilde{x}_i = x_i - \bar{x}$$

$$Var(x) = s_x^2 =$$

$$\frac{1}{n-1} \sum (x_i - \bar{x})^2 = \frac{1}{n-1} \sum \tilde{x}_i^2 = \frac{\|\tilde{x}\|^2}{n-1}$$

$$sd(x) = s_x = \sqrt{s_x^2} = \frac{\|\tilde{x}\|}{\sqrt{n-1}}$$

Here, $Var(x)$ is the sample variance of x, also often denoted by s_x^2.

Covariance

$$x = (x_1, x_2, \cdots x_n)$$

$$y = (y_1, y_2, \cdots y_n)$$

$$\tilde{x}_i = x_i - \bar{x} \quad , \quad \tilde{y}_i = y_i - \bar{y}$$

$$\mathrm{cov}(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\langle \tilde{x}, \tilde{y} \rangle}{n-1}$$

Correlation

$$r_{xy} \equiv$$

$$\operatorname{cor}(x,y) = \frac{\operatorname{cov}(x,y)}{s_x \, s_y}$$

$$= \frac{\langle \tilde{x}, \tilde{y} \rangle / (n-1)}{\frac{\|\tilde{x}\|}{\sqrt{n-1}} \; \frac{\|\tilde{y}\|}{\sqrt{n-1}}} = \frac{\langle \tilde{x}, \tilde{y} \rangle}{\|\tilde{x}\| \, \|\tilde{y}\|}$$

So by the C S.

$$-1 \leq r_{xy} \leq 1$$

Simple Regression Likelihood

$$Y_i = \beta x_i + \epsilon_i, \;\; \epsilon_i \sim N(0, \sigma^2), \; iid.$$

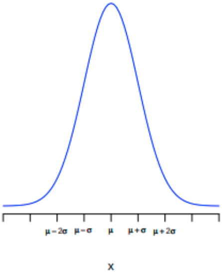
So,

$$Y_i \,|\, x_i \sim N(x_i \beta, \sigma^2).$$

$$f(y|x, \beta, \sigma) = \prod_{i=1}^{n} n(y_i | x_i \beta, \sigma^2).$$

Where $n(y|\mu, \sigma^2)$ is the normal density with mean μ and standard deviation σ.

$$n(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp(-\frac{1}{2\sigma^2} (y - \mu)^2).$$



$$Prob(\mu - \sigma < Y < \mu + \sigma) = .68$$

$$Prob(\mu - 1.96\sigma < Y < \mu + 1.96\sigma) = .95.$$

$$E(Y) = \mu, \quad Var(Y) = \sigma^2.$$

We write our model in vector notation

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$y = x\beta + \varepsilon$$

Mle:

We estimate β and σ by maximizing the likelihood:

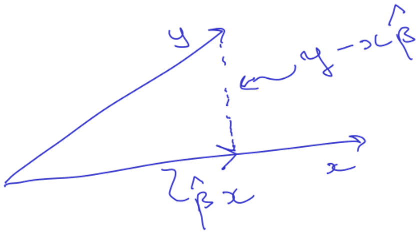
$$\max_{\beta,\sigma} \; L(\beta, \sigma \,|\, y, x) \propto f(y \,|\, x, \beta, \sigma).$$

$$x = (x_1, x_2, \cdots x_n)'$$
$$y = (y_1, y_2, \cdots y_n)'$$

$$f(y|x, \beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{1}{2\sigma^2}(y_i - x_i\beta)^2}$$

$$\propto \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2} \|y - x\beta\|^2}$$

$$\hat{\beta} = \frac{\langle y, x \rangle}{\langle x, x \rangle}$$



$$\hat{\beta} = \frac{\langle y, x \rangle}{\langle x, x \rangle} \qquad s^2 = \|y - x\hat{\beta}\|^2$$

$$S^2 \equiv \|y - \hat{\beta}x\|^2$$

$$f(y|x, \hat{\beta}, \sigma) \propto \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2} S^2}$$

$$\log f = -n\log(\sigma) - \frac{1}{2\sigma^2} S^2$$

$$\frac{d\log f}{d\sigma} = \frac{-n}{\sigma} + \frac{S^2}{\sigma^3}$$

$$\frac{d\log f}{d\sigma} = 0 \Rightarrow -n\sigma^2 + S^2 = 0$$

$$\Rightarrow \hat{\sigma}^2 = \frac{S^2}{n}$$

The Sample Mean and the one vector

A basic model in statistics is

$$Y_i \sim N(\mu, \sigma^2), \;\; \text{iid}.$$

Or,

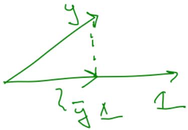
$$Y_i = \mu + \epsilon_i \sim N(0, \sigma^2), \; iid.$$

Write the model as a regression:

$$y = (y_1, y_2, \ldots, y_n)'$$

$$\mathbf{1} = (1, 1, \ldots, 1)'$$

$$\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$$



$$y = \mu \mathbf{1} + \varepsilon$$

$$\hat{\mu} = \frac{\langle y, \mathbf{1} \rangle}{\langle \mathbf{1}, \mathbf{1} \rangle} = \frac{\sum y_i}{n} = \bar{y}$$

$$\hat{\sigma}^2 = \frac{\| y - \hat{\mu} \mathbf{1} \|^2}{n} = \frac{\sum (y_i - \bar{y})^2}{n}$$

Multiple Regression Model

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix} \qquad \varepsilon_i \sim N(0, \sigma^2) \text{ iid.}$$

$$y_i = x_i' \beta + \varepsilon_i \,, \quad \beta \in \mathbb{R}^p$$

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix}_{n \times p} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n \times 1}$$

$$Y = X\beta + \varepsilon$$

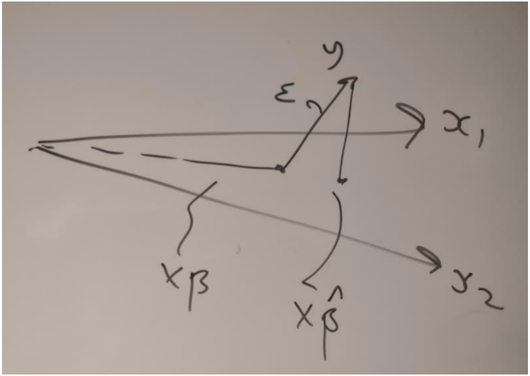
MLE

$$f(y \mid x, \beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{1}{2\sigma^2}(y_i - x_i^T \beta)^2}$$

$$\propto \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2}\|y - X\beta\|^2}$$

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

$$\hat{\sigma}^2 = \frac{\|y - X\hat{\beta}\|^2}{n}$$


y
ε
x₁
x₂
Xβ
Xβ̂

Mean and Variance of $\hat{\beta}$

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

$$E(\hat{\beta}) = (X^T X)^{-1} X^T E[Y|X]$$

$$Var(\hat{\beta}) = (X^T X)^{-1} X^T \, Var(y|x) \, X (X^T X)^{-1}$$

$$E[Y|X] = X\beta \qquad Var(y|x) = \sigma^2 I$$

$$E[\hat{\beta}] = (X^T X)^{-1} X^T (X\beta) = \beta.$$

$$Var(\hat{\beta}) = (X^T X)^{-1} X^T [\sigma^2 I] X (X^T X)^{-1}$$

$$= \sigma^2 (X^T X)^{-1}$$

QR and $Var(\hat{\beta})$

$$X = QR$$

$$X^TX = R^TQ^TQR$$

$$= R^TR$$

$$(X^TX)^{-1} = (R^TR)^{-1}$$

$$= R^{-1}(R^{-1})^T$$

Easy to invert an upper triangular and the inverse is upper triangular.

Note $X = [X_1, X_2]$ $V_1 = \text{span}(X_1)$

$$P_V y = X\hat{b} = (X_1 \hat{b}_1 + X_2 \hat{b}_2)$$

$$= (X_1 \hat{b}_1 + [P_{V_1} X_2 + Q_{V_1} X_2] \hat{b}_2)$$

$$= X_1 (\hat{b}_1 + (X_1^T X_1) X_1^T X_2 \hat{b}_2)$$

$$+ Q_{V_1} X_2 \hat{b}_2$$

You can get $\hat{b}_2$ by regressing y on the residuals from the regression of X_2 on X_1

Example $X = [X_{-p}, x_p]$

e_p = resids from x_p on X_{-p}

$$\hat{b}_p = \frac{\langle e_p, y \rangle}{\langle e_p, e_p \rangle}$$

just relate y to the part of x_p unrelated to the other x's !!

$$\mathrm{Var}(\hat{b}_p) = \frac{1}{\|e_p\|^4} \left(e_p^T [\sigma^2 I] e_p \right)$$

$$= \frac{\sigma^2}{\|e_p\|^2}$$

$$\frac{\mathrm{Var}(\varepsilon)}{\text{variability in } x_p \text{ unrelated to other } x\text{'s}}$$

Note

Back to QR, $R\hat{b} = Q^T y$

e_p = resid x_p on others

$r_{pp} = \|e_p\| \qquad o_p = \dfrac{e_p}{\|e_p\|}$

$$r_{pp}\hat{b}_p = o_p^T y$$

$$\|e_p\|\,\hat{b}_p = \frac{e_p^T y}{\|e_p\|}$$

$$\hat{b}_p = \frac{e_p^T y}{\|e_p\|^2}$$

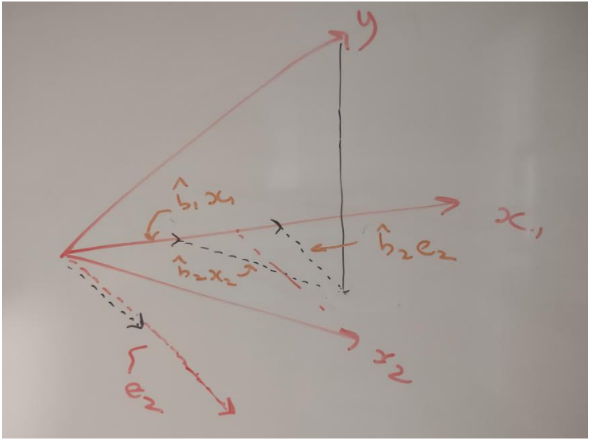
$$r_{jj} = \langle o_j, x_j \rangle$$

$$= \frac{\langle e_j, x_j \rangle}{\|e_j\|}$$

$$= \frac{\langle x_j - P_{j-1}x_j,\ x_j \rangle}{\|e_j\|}$$

$$= \frac{\langle x_j - P_{j-1}x_j,\ x_j - P_{j-1}x_j \rangle}{\|e_j\|}$$

$$= \|e_j\|$$


y
$\hat{b}_1 x_1$
$x_{.1}$
$\hat{b}_2 e_2$
$\hat{b}_2 x_2$
x_2
e_2