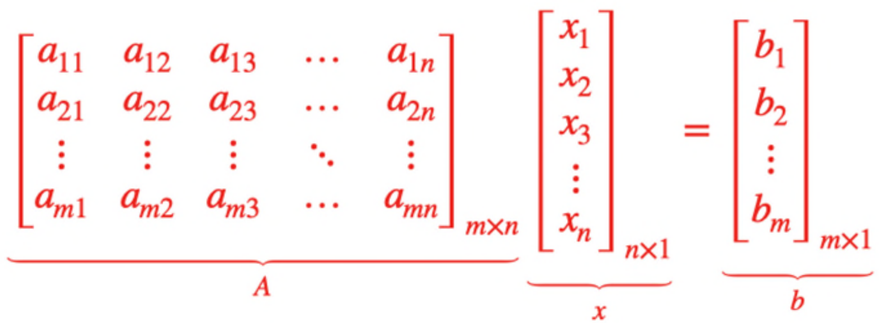


CX1104: **Linear Algebra for Computing**

Chap. No : **7.1.3A**

Lecture : **Least Squares**

Topic : **Least Squares**

Concept : **Supplementary –** Proving Solutions of Normal Equation are Least Square Solutions

Instructor: **A/P Chng Eng Siong**

TAs: **Zhang Su**, **Vishal Choudhari**

Revised: 18 Dec 2024

Rev: 2nd July 2020

Proving Solutions of Normal Equation are Least Square Solutions

The set of least-squares solutions of $A\mathbf{x} = \mathbf{b}$ coincides with the nonempty set of solutions of the normal equations $A^TA\mathbf{x} = A^T\mathbf{b}$.

Proofs by:

1. Gilbert Strang, MIT (2 Proofs)
2. Jeffrey Chasnov, HKUST
3. Alex Townsend, MIT
4. Alexey Grigorev
5. Quora

Gilbert Strang (Proof 1)

In Strang's Lecture 15 video, Strang: 18.06

Projection onto Subspace

https://youtu.be/Y_Ac6KiQ1t0?t=951

1. 3-D, projecting a point b onto a plane span by A's columns (2-D) (18:50)
   - Vector b is not in the plane (span by A's columns). (Fig a)
   - What is that plane? (19:40)
     - The plane is span by columns of A: a_1 and a_2 and hence, form the basis for the subspace.
     - The columns of A are independent, and need NOT be perpendicular.
   - (22:20) $e = b - p$ (is orthogonal to the plane of A) (crucial fact)

2. What is small p (the projected vector of b on subspace spanned by A's columns)?
   - 23:45: p is some linear combination of columns of A. (Fig b)
   - $p = A\hat{x}$, we are looking for $\hat{x}$ such that error vector e is orthogonal to plane.

3. Now e is orthogonal to the column space of A. (26:00)
   - We have 2 equations to solve for 2 unknowns (fig c)
   - Reducing it to $A^T(b - A\hat{x}) = 0$ (fig d) -> The normal equation

Fig (a)

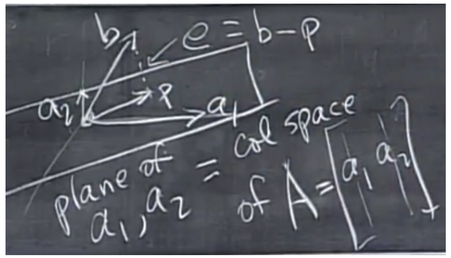


Fig (b)

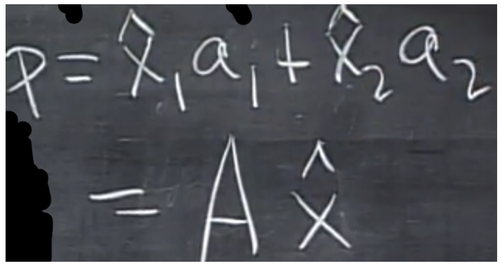


Fig (c)

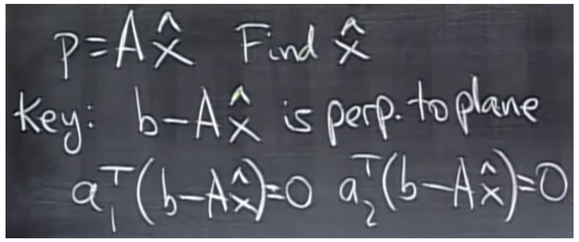
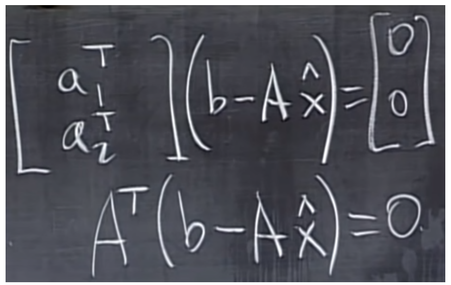
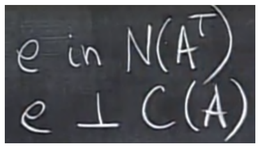


Fig (d)





Gilbert Strang (Proof 2 - Using Dot Product)

Projections onto subspaces

Projections

If we have a vector $\mathbf{b}$ and a line determined by a vector $\mathbf{a}$, how do we find t
point on the line that is closest to $\mathbf{b}$?

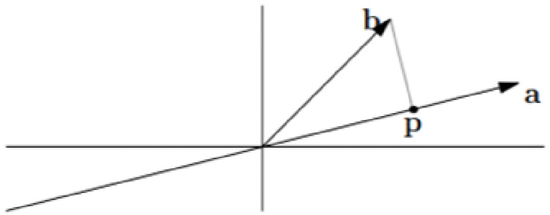


Figure 1: The point closest to $\mathbf{b}$ on the line determined by $\mathbf{a}$.

We can see from Figure 1 that this closest point $\mathbf{p}$ is at the intersecti
formed by a line through $\mathbf{b}$ that is orthogonal to $\mathbf{a}$. If we think of $\mathbf{p}$ as
approximation of $\mathbf{b}$, then the length of $\mathbf{e} = \mathbf{b} - \mathbf{p}$ is the error in that appro
mation.

We could try to find $\mathbf{p}$ using trigonometry or calculus, but it's easier to u
linear algebra. Since $\mathbf{p}$ lies on the line through $\mathbf{a}$, we know $\mathbf{p} = x\mathbf{a}$ for so
number x. We also know that $\mathbf{a}$ is perpendicular to $\mathbf{e} = \mathbf{b} - x\mathbf{a}$:

$$\begin{aligned} \mathbf{a}^T(\mathbf{b} - x\mathbf{a}) &= 0 \\ x\mathbf{a}^T\mathbf{a} &= \mathbf{a}^T\mathbf{b} \\ x &= \frac{\mathbf{a}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a}}, \end{aligned}$$

and $\mathbf{p} = \mathbf{a}x = \mathbf{a}\frac{\mathbf{a}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a}}$. Doubling $\mathbf{b}$ doubles $\mathbf{p}$. Doubling $\mathbf{a}$ does not affect $\mathbf{p}$.

Why project?

As we know, the equation $A\mathbf{x} = \mathbf{b}$ may have no solution. The vector $A\mathbf{x}$ is always in the column space of A, and $\mathbf{b}$ is unlikely to be in the column space. So, we project $\mathbf{b}$ onto a vector $\mathbf{p}$ in the column space of A and solve $A\hat{\mathbf{x}} = \mathbf{p}$.

Projection in higher dimensions

In $\mathbb{R}^3$, how do we project a vector $\mathbf{b}$ onto the closest point $\mathbf{p}$ in a plane?

If $\mathbf{a}_1$ and $\mathbf{a}_2$ form a basis for the plane, then that plane is the column space of the matrix $A = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 \end{bmatrix}$.

We know that $\mathbf{p} = \hat{x}_1\mathbf{a}_1 + \hat{x}_2\mathbf{a}_2 = A\hat{\mathbf{x}}$. We want to find $\hat{\mathbf{x}}$. There are many ways to show that $\mathbf{e} = \mathbf{b} - \mathbf{p} = \mathbf{b} - A\hat{\mathbf{x}}$ is orthogonal to the plane we're projecting onto, after which we can use the fact that $\mathbf{e}$ is perpendicular to $\mathbf{a}_1$ and $\mathbf{a}_2$:

$$\mathbf{a}_1^T(\mathbf{b} - A\hat{\mathbf{x}}) = 0 \quad \text{and} \quad \mathbf{a}_2^T(\mathbf{b} - A\hat{\mathbf{x}}) = 0.$$

In matrix form, $A^T(\mathbf{b} - A\hat{\mathbf{x}}) = \mathbf{0}$. When we were projecting onto a line, A only had one column and so this equation looked like: $a^T(\mathbf{b} - x\mathbf{a}) = \mathbf{0}$.

Note that $\mathbf{e} = \mathbf{b} - A\hat{\mathbf{x}}$ is in the nullspace of A^T and so is in the left nullspace of A. We know that everything in the left nullspace of A is perpendicular to the column space of A, so this is another confirmation that our calculations are correct.

We can rewrite the equation $A^T(\mathbf{b} - A\hat{\mathbf{x}}) = \mathbf{0}$ as:

$$A^TA\hat{\mathbf{x}} = A^T\mathbf{b}.$$

When projecting onto a line, A^TA was just a number; now it is a square matrix. So instead of dividing by $\mathbf{a}^T\mathbf{a}$ we now have to multiply by $(A^TA)^{-1}$

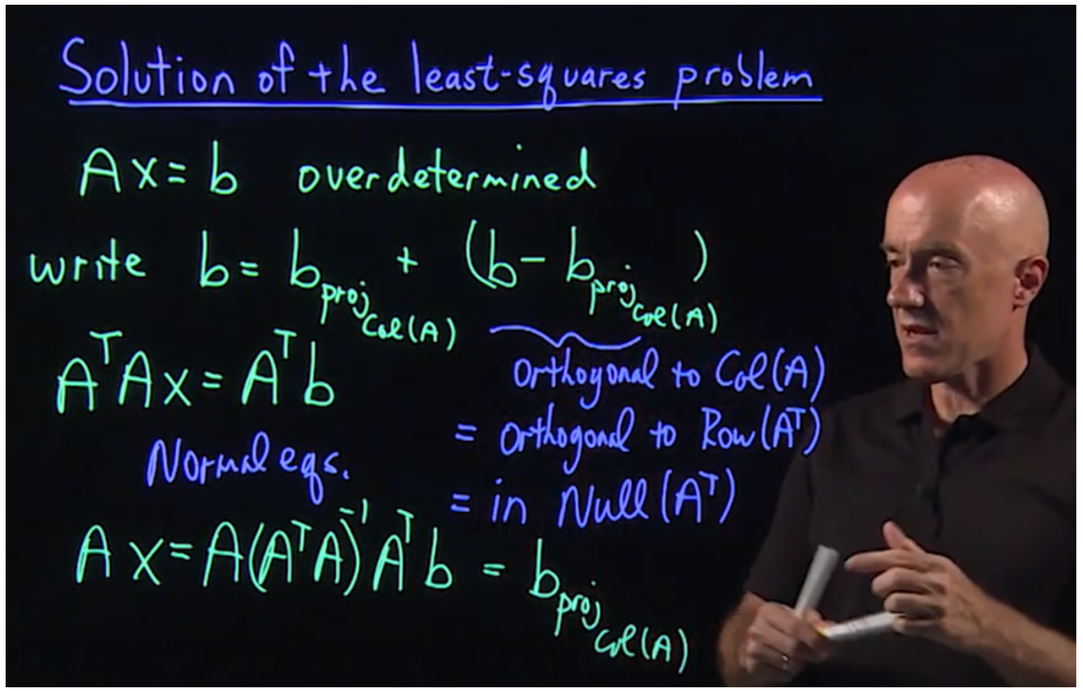
In n dimensions,

$$\begin{aligned} \hat{\mathbf{x}} &= (A^TA)^{-1}A^T\mathbf{b} \\ \mathbf{p} = A\hat{\mathbf{x}} &= A(A^TA)^{-1}A^T\mathbf{b} \\ P &= A(A^TA)^{-1}A^T. \end{aligned}$$

Ref: Lecture Writeup
https://ocw.mit.edu/courses/mathematics/18-06sc-linear-algebra-fall-2011/least-squares-determinants-and-eigenvalues/projections-onto-subspaces/MIT18_06SCF11_Ses2.2sum.pdf

Jeffrey Chasnov

If you understand Strang's proof,
Chasnov repeats it in the first 4 minutes.



Ref: https://www.youtube.com/watch?v=WABC6wmuLOk

Alex Townsend

Townsend's slide:
https://math.mit.edu/classes/18.085/summer2016/handouts/LeastSquares.pdf

Another way to show the same thing:

- https://theclevermachine.wordpress.com/2012/09/01/derivation-of-ols-normal-equations/
- https://sites.math.washington.edu/~burke/crs/308/LeastSquares.pdf

Online tool to perform differentiation of matrix:
http://www.matrixcalculus.org/

Matrix calculus reference (Advance):

- https://atmos.washington.edu/~dennis/MatrixCalculus.pdf
- https://www.comp.nus.edu.sg/~cs5240/lecture/matrix-differentiation.pdf
- https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

NORMAL EQUATIONS:

$$A^T Ax = A^T b$$

Why the normal equations? To find out you will need to be slightly crazy and totally comfortable with calculus.

In general, we want to minimize[1]

$$f(x) = \|b - Ax\|_2^2 = (b - Ax)^T(b - Ax) = b^T b - x^T A^T b - b^T Ax + x^T A^T Ax.$$

If x is a global minimum of f, then its gradient $\nabla f(x)$ is the zero vector. Let's take the gradient of f remembering that

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}.$$

We have the following three gradients:

$$\nabla(x^T A^T b) = A^T b, \quad \nabla(b^T Ax) = A^T b, \quad \nabla(x^T A^T Ax) = 2A^T Ax.$$

To calculate these gradients, write out $x^T A^T b$, $b^T Ax$, and $x^T A^T Ax$, in terms of sums and differentiate with respect to $x_1, \ldots, x_n$ (this gets very messy).

Thus, we have

$$\nabla f(x) = 2A^T Ax - 2A^T b,$$

just like we saw in the example. We can solve $\nabla f(x) = 0$ or, equivalently $A^T Ax = A^T b$ to find the least squares solution. Magic.

Is this the global minimum? Could it be a maximum, a local minimum, or a saddle point? To find out we take the "second derivative" (known as the Hessian in this context):

$$Hf = 2A^T A.$$

Next week we will see that $A^T A$ is a positive semi-definite matrix and that this implies that the solution to $A^T Ax = A^T b$ is a global minimum of $f(x)$. Roughly speaking, $f(x)$ is a function that looks like a bowl.

Alexey Grigorev

Alexey Grigorev

A personal page

ML Wiki

Home Page Info

Normal Equation

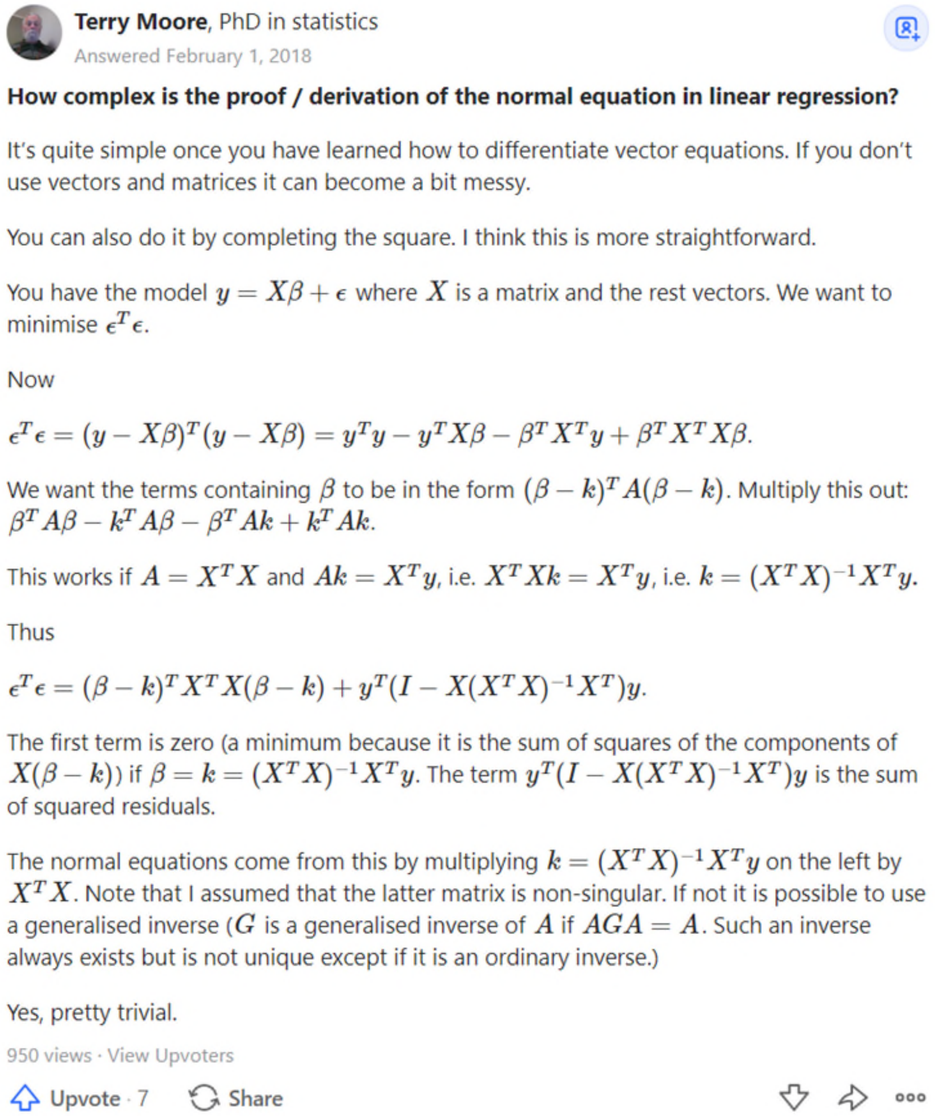
Contents [hide]

Projection onto Subspaces

Contents [hide]

Ref: http://mlwiki.org/index.php/Normal_Equation

Sometimes Quora has the Answer!



Terry Moore, PhD in statistics

Answered February 1, 2018

How complex is the proof / derivation of the normal equation in linear regression?

It's quite simple once you have learned how to differentiate vector equations. If you don't use vectors and matrices it can become a bit messy.

You can also do it by completing the square. I think this is more straightforward.

You have the model $y = X\beta + \epsilon$ where X is a matrix and the rest vectors. We want to minimise $\epsilon^T\epsilon$.

Now

$$\epsilon^T\epsilon = (y - X\beta)^T(y - X\beta) = y^Ty - y^TX\beta - \beta^TX^Ty + \beta^TX^TX\beta.$$

We want the terms containing β to be in the form $(\beta - k)^TA(\beta - k)$. Multiply this out: $\beta^TA\beta - k^TA\beta - \beta^TAk + k^TAk$.

This works if $A = X^TX$ and $Ak = X^Ty$, i.e. $X^TXk = X^Ty$, i.e. $k = (X^TX)^{-1}X^Ty$.

Thus

$$\epsilon^T\epsilon = (\beta - k)^TX^TX(\beta - k) + y^T(I - X(X^TX)^{-1}X^T)y.$$

The first term is zero (a minimum because it is the sum of squares of the components of $X(\beta - k)$) if $\beta = k = (X^TX)^{-1}X^Ty$. The term $y^T(I - X(X^TX)^{-1}X^T)y$ is the sum of squared residuals.

The normal equations come from this by multiplying $k = (X^TX)^{-1}X^Ty$ on the left by X^TX. Note that I assumed that the latter matrix is non-singular. If not it is possible to use a generalised inverse (G is a generalised inverse of A if $AGA = A$. Such an inverse always exists but is not unique except if it is an ordinary inverse.)

Yes, pretty trivial.

950 views · View Upvoters

Upvote · 7 Share

Ref: https://www.quora.com/How-complex-is-the-proof-derivation-of-the-normal-equation-in-linear-regression

Why the normal equation has a solution?

Does $A^T b$ lie in the column space of $A^T A$? If it does, why?

Asked 4 years, 11 months ago Modified 4 years, 11 months ago Viewed 729 times

1

This problem comes from the normal equations of linear least-squares:

$$(A^T A)x = A^T b$$

My question is whether or not $A^T b$ lies in the column space of $A^T A$ regardless of $A^T A$'s rank will be.

linear-algebra vector-spaces least-squares

Share Cite Edit Follow

asked Nov 8, 2017 at 3:45
Finley
993 ●1 ■5 ▲17

Add a comment

3 Answers

Sorted by: Highest score (default)

1

Yes, if A is a real $m \times n$ matrix and $b \in \mathbb{R}^m$, then $A^T b$ is in the column space of $A^T A$. Here is a proof:

Let $\hat{b}$ be the projection of b onto $R(A)$. Notice that $b - \hat{b}$ is orthogonal to every vector in $R(A)$, and in particular $b - \hat{b}$ is orthogonal to each column of A. This tells us that

$$A^T(b - \hat{b}) = 0. \qquad (\spadesuit)$$

Because $\hat{b} \in R(A)$, there exists x such that $Ax = \hat{b}$. Plugging in for $\hat{b}$ in equation $(\spadesuit)$, we see that $A^T Ax = A^T b$. This shows that $A^T b$ is in the range of A.

Share Cite Edit Follow

edited Nov 8, 2017 at 5:13

answered Nov 8, 2017 at 5:08
littleO
48.6k ●8 ■90 ▲157

Nice writeup by Micah Fogel – Illinois Maths and Science

Ref: http://staff.imsa.edu/~fogel/LinAlg/PDF/33%20Least%20Squares.pdf

Where he discussed 3 ways to find least squares: geometry, algebra and calculus

He discussed: Why is $A^TAx = A^Tb$ always solvable?

So why is $A^TA\hat{\mathbf{x}} = A^T\mathbf{b}$ always solvable? Well, we use our Fundamental Theorem of Linear Algebra. The column space $\boldsymbol{C}(A^TA)$ is the orthogonal complement of the left nullspace of A^TA. Well, this is easier in symbols: $\boldsymbol{C}(A^TA) = \left(\boldsymbol{N}(A^TA)^T\right)^{\perp} = \left(\boldsymbol{N}(A^TA)\right)^{\perp} = \left(\boldsymbol{N}(A)\right)^{\perp} = \boldsymbol{C}(A^T)$ (we've seen that A and A^TA have the same nullspace because if $A\mathbf{x} = \mathbf{0}$ certainly $A^TA\mathbf{x} = \mathbf{0}$, but if $A^TA\mathbf{x} = \mathbf{0}$, we multiply on both sides by $\mathbf{x}^T$ and find the $\|A\mathbf{x}\| = 0$, so $A\mathbf{x} = \mathbf{0}$). But since the column spaces of A^TA and A^T are the same, and $A^T\mathbf{b}$ is in the column space of A^T we can certainly always solve $A^TA\hat{\mathbf{x}} = A^T\mathbf{b}$.

He discussed: When A has dependent column , i.e (A^TA) $inverse$ does not exist

We know that if the columns of A are independent then A^TA is invertible, and there is a unique solution to this system. What if they aren't?

It turns out that we can still always solve the system, but now the solution won't be unique. We can add any null vector of A to a particular $\hat{\mathbf{x}}$ and obtain another solution. *This will be the point of the pseudoinverse*. We will pick out the particular solution that is in the row space of A. That way, any other solution will be this plus a null vector, which is orthogonal to the row space. So any other solution will be *longer* than the row space solution. Thus not only will the pseudoinverse make the error as small as possible, it will make the choice of $\hat{\mathbf{x}}$ as small as possible, too!