Cross Validated

Problem with proof of Conditional expectation as best predictor

Asked 6 years, 1 month ago Active 4 months ago Viewed 11k times

19

16

I have an issue with the proof of

> $$E(Y|X) \in \arg\min_{g(X)} E\Big[\big(Y-g(X)\big)^2\Big]$$

which very likely reveal a deeper misunderstanding of expectations and conditional expectations.

The proof I know goes as follows (another version of this proof can be found here)

$$\begin{aligned}
&\arg\min_{g(X)} E\Big[\big(Y-g(x)\big)^2\Big]\\
=&\arg\min_{g(X)} E\Big[\big(Y-E(Y|X)+E(Y|X)-g(X)\big)^2\Big]\\
=&\arg\min_{g(x)} E\Big[\big(Y-E(Y|X)\big)^2+2\big(Y-E(Y|X)\big)\big(E(Y|X)-g(X)\big)+\big(E(Y|X)-g(X)\big)^2\Big]\\
=&\arg\min_{g(x)} E\Big[2\big(Y-E(Y|X)\big)\big(E(Y|X)-g(X)\big)+\big(E(Y|X)-g(X)\big)^2\Big]
\end{aligned}$$

The proof then typically continues with an argument showing that $2E\Big[\big(Y-E(Y|X)\big)\big(E(Y|X)-g(X)\big)\Big]=0$, and hence

$$\arg\min_{g(x)} E\Big[\big(Y-g(x)\big)^2\Big]=\arg\min_{g(x)} E\Big[\big(E(Y|X)-g(X)\big)^2\Big]$$

which can be seen to be minimized when $g(X)=E(Y|X)$.

My puzzles about the proof are the following:

1. Consider

> $$E\Big[2\big(Y-E(Y|X)\big)\big(E(Y|X)-g(X)\big)+\big(E(Y|X)-g(X)\big)^2\Big].$$

It seems to me that, independently of any argument showing that the first term is always equal to zero, one can see that setting $g(X) = E(Y|X)$ minimizes the expression as it implies $\big(E(Y|X) - g(X)\big) = 0$ and hence

> $$E\Big[2\big(Y - E(Y|X)\big)\big(E(Y|X) - g(X)\big) + \big(E(Y|X) - g(X)\big)^2\Big] = E(0 + 0) = 0.$$

But if this is true, then one might repeat the proof replacing $E(Y|X)$ by any other function of X, say $h(X)$, and get to the conclusion that it is $h(X)$ that minimizes the expression. So there must be something I misunderstand (right?).

1. I have some doubts about the meaning of $E[(Y - g(X))^2]$ in the statement of the problem. How should the notation be interpreted? Does it mean

> $E_X[(Y - g(X))^2]$, $E_Y[(Y - g(X))^2]$ or $E_{XY}[(Y - g(X))^2]$?

mathematical-statistics | conditional-probability | proof | conditional-expectation

edited Oct 4 '13 at 23:15

asked Oct 4 '13 at 0:24

Martin Van der Linden
609 1 5 12

4 Answers

11

(This is an adaptation from Granger & Newbold(1986) "Forecasting Economic Time Series").

By construction, your *error cost function* is $[Y - g(X)]^2$. This incorporates a critical assumption (that the error cost function is symmetric around zero) -a different error cost function would not necessarily have the conditional expected value as the $\arg\min$ of its expected value. You cannot minimize your error cost function because it contains unknown quantities. So you decide to minimize its expected value instead. Then your objective function becomes

$$E[Y - g(X)]^2 = \int_{-\infty}^{\infty} [y - g(X)]^2 f_{Y|X}(y|x)dy$$

which I believe answers also your second question. It is intuitive that the expected value will be of Y conditional on X, since we are trying to estimate/forecast Y based on X. Decompose the square to obtain

$$E[Y - g(X)]^2 = \int_{-\infty}^{\infty} y^2 f_{Y|X}(y|x)dy - 2g(X)\int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy + \Big[g(X)\Big]^2 \int_{-\infty}^{\infty} f_{Y|X}(y|x)dy$$

The first term does not contain $g(X)$ so it does not affect minimization, and it can be ignored. The integral in the second term equals the conditional expected value of Y given X, and the integral in the last term equals unity. So

$$\arg\min_{g(x)} E[Y-g(X)]^2 = \arg\min_{g(x)} \Big\{ -2g(X)E(Y \mid X) + \Big[g(X)\Big]^2 \Big\}$$

The first derivative w.r.t $g(X)$ is $-2E(Y \mid X) + 2g(X)$ leading to the first order condition for minimization $g(X) = E(Y \mid X)$ while the second derivative is equal to $2 > 0$ which is sufficient for a minimum.

ADDENDUM:The logic of the "add and subtract" proof approach.

The OP is puzzled by the approach stated in the question, because it seems tautological. It isn't, because while using the tactic of adding and subtracting makes a *specific part* of the objective function zero for an arbitrary choice of the term that is added and subtracted, it does NOT equalize the *value function* , namely the value of the objective function evaluated at the candidate minimizer.

For the choice $g(X) = E(Y \mid X)$ we have the value function $V\left(E(Y \mid X)\right) = E\Big[(Y-E(Y \mid X))^2 \mid X\Big]$ For the arbitrary choice $g(X) = h(X)$we have the value funtion $V\left(h(X)\right) = E\Big[(Y-h(X))^2 \mid X\Big]$.

I claim that

$$V\left(E(Y \mid X)\right) \leq V\left(h(X)\right)$$

$$\Rightarrow E(Y^2 \mid X) - 2E\Big[(YE(Y \mid X)) \mid X\Big] + E\Big[(E(Y \mid X))^2 \mid X\Big]$$

$$\leq E(Y^2 \mid X) - 2E\Big[(Yh(X)) \mid X\Big] + E\Big[(h(X))^2 \mid X\Big]$$

The first term of the LHS and the RHS cancel out. Also note that the outer expectation is conditional on X. By the properties of conditional expectations we end up with

$$\ldots \Rightarrow -2E(Y \mid X)\cdot E\Big(Y \mid X\Big) + \Big[E(Y \mid X)\Big]^2 \leq -2E(Y \mid X)h(X) + \Big[h(X)\Big]^2$$

$$\Rightarrow 0 \leq \Big[E(Y \mid X)\Big]^2 - 2E(Y \mid X)h(X) + \Big[h(X)\Big]^2$$

$$\Rightarrow 0 \leq \Big[E(Y \mid X) - h(x)\Big]^2$$

which holds with strict inequality if $h(x) \neq E(Y \mid X)$. So $E(Y \mid X)$ is the global and unique minimizer.

But this also says that the "add-and-subtract" approach is not the most illuminating way of proof here.

edited Oct 4 '13 at 10:26

answered Oct 4 '13 at 1:05

Alecos Papadopoulos

45.8k 2 103 208

Thanks for you answer. It helps clarifying my second question. As I tried to convey in the title of the question, my main issue (the first one in the post) was more about the proof mechanism. My main concern is about my understanding of the proof I presented in the question. As I explained, my understanding of the proof leads me to blatantly problematic statement. So I would like to understand were my mistake is as it might reveal some deeper misunderstandings about concepts of expectaction and conditional expectation. Any thoughts about this? – Martin Van der Linden Oct 4 '13 at 4:00

1 I added some explanation on the "add and subtract" approach to proof. – Alecos Papadopoulos Oct 4 '13 at 10:27

Took me some time to understand it, but I finally got my basic mistake : true enough $E\Big[-2\big(Y-h(X)\big)\big(h(X)-g(X)\big)+\big(h(X)-g(X)\big)^2\Big]=0$ when $g(X)=h(X)$, but by no means does it imply that $h(X)$ minimizes the expression. There is no reason which the bracketed expression could not be lower than zero. Because of the minus sign in front of $\big(Y-h(X)\big)\big(h(X)-g(X)\big)$ one could find some $g(X)$ such that $E\Big[-2\big(Y-h(X)\big)\big(h(X)-g(X)\big)+\big(h(X)-g(X)\big)^2\Big]<0$. – Martin Van der Linden Oct 4 '13 at 16:24

1 Hmmm... the minus sign in the expression you refer to is a mistake - it should be a plus sign. You could of course then rearrange the terms to obtain again a minus sign... does this hurt the intuition you gained? – Alecos Papadopoulos Oct 4 '13 at 21:03

Thanks for keeping up with the question. I edited the initial post to correct for this mistake. Fortunately, I think it does not hurt the gained intuition. Actually it helps me understand yet another mistake : I was assuming that the minus sign was important to guarantee that 0 was not necessarily the minimum of $E[-2(Y-h(X))(h(X)-g(X))+(h(X)-g(X))^2]$. But I realize this is not just about the sign before the 2. (Hopefully) What I really needed to understand is that, in general (i.e. for arbitrary $h(X)$) $E[2(Y-h(X))(h(X)-g(X))]$ needs not be minimized when $g(X)=h(X)$ (right?). – Martin Van der Linden Oct 4 '13 at 23:36

|

5

Note that to prove the answer, you really only need to show that

$$E\Big[-2\big(Y-E(Y|X)\big)\big(E(Y|X)-g(X)\big)\Big]=0$$

As for which expectation to take, you take it conditionally, otherwise the term

$$\arg\min_{g(X)} E\Big[\big(Y-g(X)\big)^2\Big]$$

Doesn't make sense, as $g(X)$ is a random variable if E is E_{XY} and not $E_{Y|X}$. Show you should really write $E\Big[\big(Y-g(X)\big)^2|X\Big]$ or $E_{Y|X}\Big[\big(Y-g(X)\big)^2\Big]$ to make this clear. Now given this clarification, the term $\big(E(Y|X)-g(X)\big)$ is a constant, and can be pulled outside the expecation, and you have:

$$-2\big(E(Y|X)-g(X)\big)E\Big[\big(Y-E(Y|X)\big)|X\Big]=$$
$$-2\big(E(Y|X)-g(X)\big)\Big[E(Y|X)-E\big[E(Y|X)|X\big]\Big]=$$
$$-2\big(E(Y|X)-g(X)\big)\Big[E(Y|X)-E(Y|X)\Big]=0$$

Hence you can write the objective function as:

$$E_{Y|X}\left[\left(Y-g(X)\right)^2\right]=E_{Y|X}\left[\left(Y-E_{Y|X}(Y|X)\right)^2\right]+\left(E_{Y|X}(Y|X)-g(X)\right)^2$$

The minimiser is obvious from here. Note that if you were to average over X as well, then a very similar argument can be used to show:

$$E_X\left[\left(E(Y|X)-g(X)\right)^2\right]=E_X\left[\left(E_{Y|X}(Y|X)-E_X\left[E_{Y|X}(Y|X)\right]\right)^2\right]$$
$$+\left(E_X\left[E_{Y|X}(Y|X)\right]-E_X\left[g(X)\right]\right)^2$$

This shows that if you set $g(X)=E_{Y|X}(Y|X)$ for each X, then you also have a minimiser over this function as well. So in some sense it doesn't really matter whether E is E_{YX} or $E_{Y|X}$.

answered Oct 4 '13 at 4:57

probabilityislogic
20.6k 3 66 89

3

There's a mathematical point of view that is very simple. What you have is a projection problem in a Hilbert space, much like projecting a vector in $\mathbb{R}^n$ onto a subspace.

Let $(\Omega,\mathcal{F},\mu)$ denote the underlying probability space. For the problem to make sense, consider the random variables with finite second moments, that is, the Hilbert space $L^2(\Omega,\mathcal{F},\mu)$. The problem now is this: given $X,Y\in L^2(\Omega,\mathcal{F},\mu)$, find the projection of Y onto the subspace $L^2(\Omega,\mathcal{F}_X,\mu)$, where $\mathcal{F}_X$ is the σ-subalgebra of $\mathcal{F}$ generated by X. (Just as in the finite dimensional case, minimizing L^2-distance to a subspace means finding the projection). The desired projection is $E(X|Y)$, by construction. (This actually characterizes $E(X|Y)$, if one inspects the proof of existence).

answered Nov 3 '13 at 3:17

Michael
658 3 10

This is a beautiful response. – jll Mar 18 '15 at 20:36

0

Regarding your last question, the expectation can be either w.r.t. $p(x,y)$ (the unconditional error) or w.r.t. $p(y\mid x)$ (the conditional error at each value $X=x$). Happily, minimizing the conditional error at each value $X=x$ also minimizes the unconditional error, so this is not a crucial distinction.

answered Jun 25 at 2:36

Ulisses Braga-Neto
101 1