CEF

Manoel Galdino

2023-05-05

Esperança Condicional

One of our most fundamental goals as data scientists is to produce predictions that are *good.* In this week's async, we make a statement of performance that we can use to evaluate how good a job a predictor is doing, choosing Mean Squared Error.

With the goal of minimizing MSE, then we then present, justify, and prove that the conditional expectation function (*the CEF*) is the globally best possible predictor. This is an incredibly powerful result, and one that serves as the backstop for **every** other predictor that you will ever fit, whether that predictor is a "simple" regression, or that predictor is a machine learning algorithms (e.g. a random forest) or a deep learning algorithm. Read that again:

> Even the most technologically advanced machine learning algorithms *cannot possibly* perform better than the conditional expectation function at making a prediction.

Why does the CEF do so well? Because it can contain a *vast* amount of complex information and relationships; in fact, the complexity of the CEF is a product of the complexity of the underlying probability space. If that is the case, then why don't we just use the CEF as our predictor every time?

Well, this is one of the core problems of applied data science work: we are never given the function that describes the behavior of the random variable. And so, we're left in a world where we are forced to produce predictions from simplifications of the CEF. A very strong simplification, but one that is useful for our puny human brains, is to restrict ourselves to predictors that make predictions from a linear combination of input variables.

Why should we make such a strong restriction? After all, the conditional expectation function might be a fantastically complex combination of input features, why should we entertain functions that are only linear combinations? Essentially, this is because we're limited in our ability to reason about anything more complex than a linear combination.

Thunder Struck

Learning Objectives

At the end of this weeks learning, which includes the asynchronous lectures, reading the textbook, this live session, and the homework associated with the concepts, student should be able to

1. **Recognize** that the conditional expectation function, the *CEF*, is a the pure-form, best-possible predictor of a target variable given information about other variables.
2. **Recall** that all other predictors, be they linear predictors, non-linear predictors, branching predictors, or deep learning predictors, are an attempt to approximate the CEF.
3. **Produce** the conditional expectation function as a predictor, given joint densities of random variables.
4. **Appreciate** that the best linear predictor, which is a restriction of predictors to include only those that are linear combinations of variables, can produce reasonable predictions, and **anticipate** that the BLP forms the target of inquiry for regression.

Class Announcements

Test 1 is releasing to you today.

The first test is releasing today. There are review sessions scheduled for this week, practice tests available, and practice problems available. The format for the test is posted in the course discussion channel. In addition to your test, your instructor will describe your responsibilities that are due next week.

Roadmap

Rearview Mirror

- Statisticians create a population model to represent the world.
- $E[X], V[X], Cov[X, Y]$ are "simple" summaries of complex joint distributions, which are hooks for our analyses.
- They also have useful properties – for example, $E[X + Y] = E[X] + E[Y]$.

This week

- We look at situations with one or more "input" random variables, and one "output."
- Conditional expectation summarizes the output, given values for the inputs.
- The conditional expectation function (CEF) is a predictor – a function that yields a value for the output, give values for the inputs.
- The best linear predictor (BLP) summarizes a relationship using a line / linear function.

Coming Attractions

- OLS regression is a workhorse of modern statistics, causal analysis, etc
  - It is also the basis for many other models in classical stats and machine learning
- The target that OLS estimates is exactly the BLP, which we're learning about this week.

Conditional Expectation Function (CEF),

Part I

Think back to remember the definition of the expectation of Y:

$$E[Y] = \int_{-\infty}^{\infty} y \cdot f_Y(y) dy$$

This week, in the async reading and lectures we added a new concept, the conditional expectation of Y given $X = x \in \text{Supp}[X]$:

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

Part II

1. What desirable properties of a predictor does the expectation possess (note, this is thinking *back* by a week)? What makes these properties desirable?
2. Turning to the content from this week, how, if at all, does the conditional expectation improve on these desirable properties?

Part III

- Compare and contrast $E[Y]$ and $E[Y|X]$. For example, when you look at how these operators are "shaped", how are their components similar or different?[1]
- What is $E[Y|X]$ a function of? What are "input" variables to this function?
- What, if anything, is $E[E[Y|X]]$ a function of?

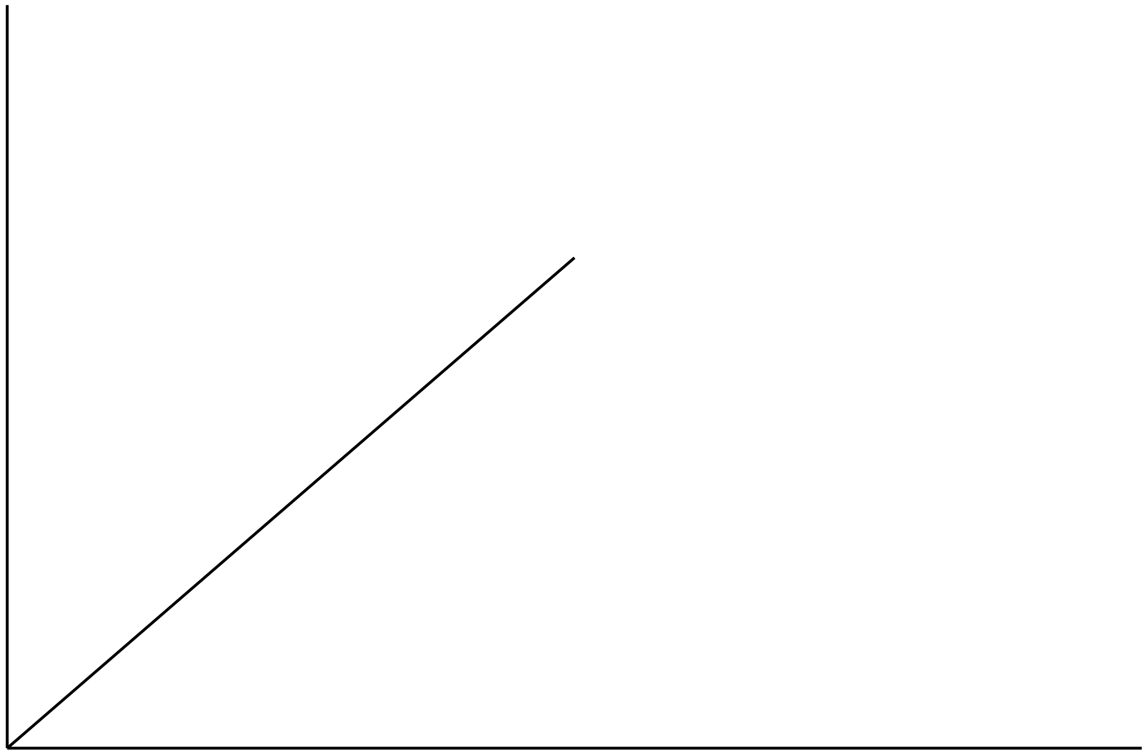
Computing the CEF

- Suppose that random variables X and Y are jointly continuous, with joint density function given by,

$$f(x,y) = \begin{cases} 2, & 0 \leq x \leq 1, 0 \leq y \leq x \\ 0, & otherwise \end{cases}$$

What does the joint PDF of this function look like?

Joint PDF of X,Y



Simple Quantities

To begin with, let's compute the simplest quantities:

- What is the expectation of X?
- What is the expectation of Y?

[1] Note, when we say "shaped" here, we're referring to the deeper concept of a statistical functional. A statistical functional is a function of a function that maps to a real number. So, if T is the functional that we're thinking of, $\mathcal{F}$ is a family of functions that it might operate on, and $\mathbb{R}$ is the set of real numbers, a statistical functional is just $T : \mathcal{F} \to \mathbb{R}$. The Expectation statistical functional, $E[X]$ always has the form $\int x f_X(x) dx$.)

- How would you compute the variance of X? (We're not going to do it live).

Conditional Quantities

Conditional Expectaiton And then, let's think about how to compute the conditional quantities. To get started, you can use the fact that in week two, we already computed the conditional probability density function:

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{x}, & 0 \leq y \leq x \\ 0, & \text{otherwise.} \end{cases}$$

With this knowledge on hand, compute the $CEF[Y|X]$.

{r, echo = FALSE, results = 'asis'} # blank_lines(n=10)

Once you have computed the $CEF[Y|X]$, use this function to answer the following questions:

- What is the conditional expectation of Y, given that $X = x = 0$?
- What is the conditional expectation of Y, given that $X = x = 0.5$?
- What is the conditional expectation of X, given that $Y = y = 0.5$?

Conditional Variance

- What is the conditional variance function?[2]

{r, echo=FALSE, results='asis'} # blank_lines(20)

- Which of the two of these has a lower conditional variances?
  - $V[Y|X = 0.25]$; or,
  - $V[Y|X = 0.75]$.
- How does $V[Y]$ compare to $V[Y|X = 1]$? Which is larger?

Conditional Expectation

Minimizing the MSE

Minimizing MSE

Theorem 2.2.20 states,

> The CEF $E[Y|X]$ is the "best" predictor of Y given X, where "best" means it has the smallest mean squared error (MSE).

Oh yeah? As a breakout group, *ride shotgun* with us as we prove that the conditional expectation is the function that produces the smallest possible Mean Squared Error.

Specifically, **you group's task** is to justify every transition from one line to the next using concepts that we have learned in the course: definitions, theorems, calculus, and algebraic operations.

The pudding (aka: "Where the proof is")

We need to find such function $g(X) : \mathbb{R} \to \mathbb{R}$ that gives the smallest mean squared error.

First, let MSE be defined as it is in **Definition 2.1.22**.

[2] Take a moment to strategize just a little bit before you get going on this one. There is a way to compute this value that is easier than another way to compute this value.

For a random variable X and constant $c \in \mathbb{R}$, the *mean squared error* of X about c is $E[(x-c)^2]$.

Second, let us note that since $g(X)$ is just a function that maps onto $\mathbb{R}$, that for some particular value of $X = x$, $g(X)$ maps onto a constant value.

- Deriving a Function to Minimize MSE

$$\begin{aligned}
E[(Y - g(X))^2|X] &= E[Y^2 - 2Yg(X) + g^2(X)|X] \\
&= E[Y^2|X] + E[-2Yg(X)|X] + E[g^2(X)|X] \\
&= E[Y^2|X] - 2g(X)E[Y|X] + g^2(X)E[1|X] \\
&= (E[Y^2|X] - E^2[Y|X]) + (E^2[Y|X] - 2g(X)E[Y|X] + g^2(X)) \\
&= V[Y|X] + (E^2[Y|X] - 2g(X)E[Y|X] + g^2(X)) \\
&= V[Y|X] + (E[Y|X] - g(X))^2
\end{aligned}$$

Notice too that we can use the *Law of Iterated Expectations* to do something useful. (This is a good point to talk about how this theorem works in your breakout groups.)

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

- $E[V[Y|X]]$ doesn't depend on g; and,
- $E[(E[Y|X] - g(X))^2] \geq 0$.

$\therefore g(X) = E[Y|X]$ gives the smallest $E[(Y - g(X))^2]$

The Implication

If you are choosing some g, you can't do better than $g(x) = E[Y|X = x]$.

Working with the BLP

Why Linear?

- In some cases, we might try to estimate the CEF. More commonly, however, we work with linear predictors. Why?
- We don't know joint density function of Y. So, it is "difficult" to derive a suitable CEF.
- To estimate *flexible* functions requires considerably more data. Assumptions about distribution (e.g. a linear form) allow you to leverage those assumptions to learn 'more' from the same amount of data.
- Other times, the CEF, even if we *could* produce an estimate, might be so complex that it isn't useful or would be difficult to work with.
- And, many times, linear predictors (which might seem trivially simple) actually do a very good job of producing predictions that are 'close' or useful.

##Joint Distribution Practice

Professorial Mistakes (Discrete RVs)

- Let the number of questions that students ask be a RV, X.

- Let X take on values: $\{1, 2, 3\}$, each with probability 1/3.

- Every time a student asks a question, the instructor answers incorrectly with probability $1/4$, independently of other questions.
- Let the RV Y be number of incorrect responses.
- **Questions:**
  - Compute the expectation of Y, conditional on X, $E[Y|X]$
  - Using the law of iterated expectations, compute $E[Y] = E\big[E[Y|X]\big]$.

Continuous BLP

- Recall the PDF that we worked with earlier to produce the $CEF[Y|X].

$$f(x,y) = \begin{cases} 2, & 0 \leq x \leq 1, 0 \leq y \leq x \\ 0, & otherwise \end{cases}$$

Find the BLP for Y as a function of X. What, if anything, do you notice about this BLP and the CEF?

`{r, echo = FALSE} # blank_lines(20) #`