Causal Machine Learning

Stats/'metrics recap

Michael Knaus

WiSe 25/26

Plan for today

This is not the sexy part, but we will need these components at several points of the lecture

1. Notation
2. Conditional expectation function
3. How to model and estimate CEF?
4. Convergence rates

Notation

Notation

There is a zoo of different notations between and within fields

I try to consistently use the following notation:

- Capital letters describe random variables (RV), e.g. X
- Small letters describe realizations of RVs, e.g. x
- Capital letters with subscript i represent the RV value of observation i, e.g. X_i
- Greek letters are used to denote unknown population parameters, e.g. α
- Hats are used to indicate estimated parameters, e.g. $\hat{\alpha}$
- $:=$ defines a symbol, e.g. $\mu := \mathbb{E}[Y]$
- p-dimensional RV are represented as column vectors, e.g. RVs X_1 to X_p are collected into $X = (X_1, ..., X_p)'$

Example and crucial refresher

The law of iterated/total expectations (LIE) tells us that the unconditional expectation of RV Y can be obtained as taking the expectation of the conditional expectations of Y given X:

$$\mu := \mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$$

The standard estimator of μ based on a sample of size N is

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

In contrast, for a conditional expectation at a fixed value x, taking the expectation makes no difference b/c it is a constant

$$m(x) := \mathbb{E}[Y|X = x] = \mathbb{E}[\mathbb{E}[Y|X = x]]$$

Conditional expectation function

Conditional expectation function

We call the function that provides the expected value of Y given X the CONDITIONAL EXPECTATION FUNCTION (CEF)

$$m(X) := \mathbb{E}[Y|X] \tag{1}$$

A RV can be decomposed into CEF and a mean independent residual

$$Y = \mathbb{E}[Y|X] + \varepsilon = m(X) + \varepsilon \tag{2}$$

with

$$\mathbb{E}[\varepsilon|X] = \mathbb{E}[Y - m(X)|X] = m(X) - m(X) = 0 \tag{3}$$

Important

This is not an assumption! It follows from probability theory under mild regularity conditions ensuring that $\mathbb{E}[Y|X]$ is well-defined.

Why conditional expectation functions?

The decomposition allows us to show why we like the CEF

Note that any function $g(X)$ produces an error $Y - g(X)$

CEF is the function that minimizes the expected squared error (proof on next slide):

$$m(X) = \underset{g(X)}{\arg\min}\, \mathbb{E}[(Y - g(X))^2] \tag{4}$$

$\Rightarrow$ The CEF delivers the best possible guess for the outcome value in the squared error sense (L^2-norm)

Remark: While there are other loss function we could care about, the squared error loss is the most important for our purposes

Proof that CEF minimizes expected squared error

$$
\begin{aligned}
\mathbb{E}[(Y-g(X))^2] &= \mathbb{E}[(Y-m(X)+m(X)-g(X))^2] \\
&= \mathbb{E}[(Y-m(X))^2] + \underbrace{\mathbb{E}[2(Y-m(X))(m(X)-g(X))]}_{=0\text{, shown below}} + \mathbb{E}[(m(X)-g(X))^2] \\
&\overset{(2)}{=} \mathbb{E}[(m(X)+\varepsilon-m(X))^2] + \mathbb{E}[(m(X)-g(X))^2] \\
&= \mathbb{E}[\varepsilon^2] + \mathbb{E}[(m(X)-g(X))^2] \\
&= Var[\varepsilon] + \mathbb{E}[(m(X)-g(X))^2]
\end{aligned}
$$

b/c $Var[\varepsilon] = \mathbb{E}[\varepsilon^2] - \mathbb{E}[\varepsilon]^2 \overset{(3)}{=} \mathbb{E}[\varepsilon^2]$

$\Rightarrow$ expected squared error minimized if $g(X) = m(X)$ b/c 2nd term becomes zero $\blacksquare$

$$
\begin{aligned}
\mathbb{E}[2(Y-m(X))(m(X)-g(X))] &\overset{LIE,(2)}{=} 2\,\mathbb{E}[\mathbb{E}[(m(X)+\varepsilon-m(X))(m(X)-g(X))|X]] \\
&= 2\,\mathbb{E}[\underbrace{\mathbb{E}[\varepsilon|X]}_{=0}(m(X)-g(X))] = 0
\end{aligned}
$$

How to model and estimate CEF?

Different options

We can choose different ways to model the CEF:

- Parametric model
- Nonparametric model
- Semiparametric model

Parametric models

Assume $m(X) = m(X; \beta)$ with $m(\cdot)$ a known function and $\beta \in \mathbb{R}^p$ a finite vector of parameters

Example 1: Linear model

$$m(X; \beta) = X'\beta$$

$\Rightarrow \beta$ most often estimated with Ordinary Least Squares (OLS)

Example 2: Probit model for binary $Y \in \{0, 1\}$

$$\mathbb{P}[Y = 1|X] = \mathbb{E}[Y|X] = m(X; \beta) = \Phi(X'\beta)$$

where $\Phi(\cdot)$ is the normal cdf

$\Rightarrow \beta$ most often estimated via Maximum Likelihood

OLS in expectations

OLS identifies the population parameters of the linear CEF model by minimizing the expected squared error

$$\beta = \arg\min_b \mathbb{E}[(Y - X'b)^2]$$

Important

Even if the CEF is not really linear, OLS provides in expectation the best linear approximation of the CEF (proof e.g. in Angrist & Pischke (2009), Thrm. 3.1.6):

$$\beta = \arg\min_b \mathbb{E}[(\mathbb{E}[Y|X] - X'b)^2]$$

OLS estimation

Assuming a sample of N i.i.d. observations, we can estimate the parameter by minimizing the sum of / mean squared error:

$$\begin{aligned}
\hat{\beta} &= \arg\min_{\beta} \sum_{i=1}^{N} \left(Y_i - X_i'\beta\right)^2 = \arg\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - X_i'\beta\right)^2 \\
&= \arg\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{N} \left(Y_i - \beta_0 - X_{i1}\beta_1 + \ldots + X_{ip}\beta_p\right)^2 \\
&= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}
\end{aligned}$$

I hope at least one of these equivalent representations looks familiar

Punchline: OLS is one way to approximate the unknown and potentially non-linear CEF

OLS illustration

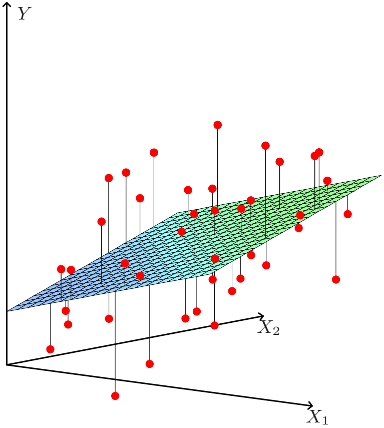

FIGURE 3.1. *Linear least squares fitting with* $X \in \mathbb{R}^2$. *We seek the linear function of* X *that minimizes the sum of squared residuals from* Y.

Hastie, Tibshirani & Friedman (2017)

Nonparametric methods

The linear CEF is a functional form assumption

$\Rightarrow$ OLS will never uncover the CEF in the likely case that the world is not linear

In contrast, nonparametric methods leave the functional form of CEF unspecified and aim to learn $m(X) = \mathbb{E}[Y|X]$ completely from the data

As we do not tell them how the world looks like, they are more data-hungry than parametric models, but consistently estimate the CEF

Classic nonparametric methods:

- Kernel regression (nice bookdown of Eduardo García-Portugués here)
- Series regression

Kernel regression - illustration

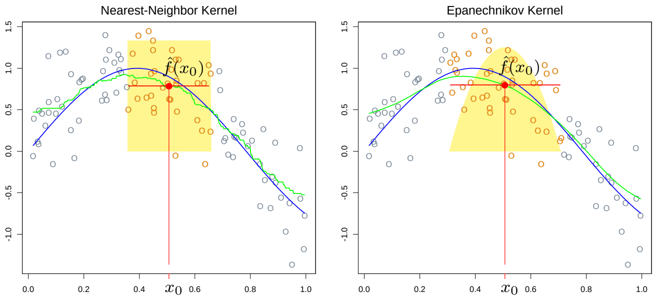

FIGURE 6.1. *In each panel* 100 *pairs* x_i, y_i *are generated at random from the blue curve with Gaussian errors:* $Y = \sin(4X) + \varepsilon$, $X \sim U[0, 1]$, $\varepsilon \sim N(0, 1/3)$. *In the left panel the green curve is the result of a* 30*-nearest-neighbor running-mean smoother. The red point is the fitted constant* $\hat{f}(x_0)$, *and the red circles indicate those observations contributing to the fit at* x_0. *The solid yellow region indicates the weights assigned to observations. In the right panel, the green curve is the kernel-weighted average, using an Epanechnikov kernel with (half) window width* $\lambda = 0.2$.

Hastie, Tibshirani & Friedman (2017)

Semiparametric models: a compromise

Nonparametric CEF models have - as the name suggests - no interpretable parameters

However, (linear) parameters can serve as useful condensation of information

Assume that we are interested in the linear parameter θ of variable X_1, but are not willing to commit to functional forms of $\tilde{X} = (X_2, ..., X_p)'$ such that $X = (X_1, \tilde{X}')'$

The so-called partially linear model assumes

$$m(X) = m(X; \theta, f) = X_1\theta + f(\tilde{X})$$

The assumed CEF has a parametric and a nonparametric part $\Rightarrow$ semiparametric

These types of models play an important role in Causal ML $\Rightarrow$ they will be back

Convergence rates

$\sqrt{N}$-convergence of OLS

I guess/hope you all have seen something like this in your econometrics/stats education

$$\sqrt{N}(\hat{\beta}_N - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

i.e. the difference between estimated and population parameters blown up by $\sqrt{N}$ converges to a multivariate normal distribution as $N \to \infty$

If we would not blow it up, it converges to zero (consistency)

This implies that also the fitted/predicted value for a fixed value of x converges at $\sqrt{N}$:

$$\sqrt{N}(x'\hat{\beta}_N - x'\beta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Convergence of OLS predictions

$\sqrt{N}$-convergence of predicted values implies that we expect the root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{N}\sum_i (X_i'\hat{\beta}_N - X_i'\beta)^2}$$

also to converge at $\sqrt{N}$

$\Rightarrow$ We expect the RMSE to halve if we have access to four times more observations

Important

This means convergence to the best linear approximation of the CEF and does not imply convergence to the CEF unless it is actually linear.

Convergence of OLS predictions

The squared error needs to be blown up by N to converge to a distribution:

$$
\begin{aligned}
[\sqrt{N}(x'\hat{\beta}_N - x'\beta)]^2 &\xrightarrow{d} [\mathcal{N}(0, \sigma^2)]^2 \\
N(x'\hat{\beta}_N - x'\beta)^2 &\xrightarrow{d} [\mathcal{N}(0, \sigma^2)]^2 \\
&\xrightarrow{d} \Gamma(1/2, 2\sigma^2)
\end{aligned}
$$

$\Rightarrow$ We expect MSE to converge at rate N

Finally, the square root of the squared error converges with $\sqrt{N}$:

$$
\sqrt{N(x'\hat{\beta}_N - x'\beta)^2} = \sqrt{N}[(x'\hat{\beta}_N - x'\beta)^2]^{1/2} \xrightarrow{d} \mathit{Nakagami}(1/2, \sigma^2)
$$

$\Rightarrow$ We expect RMSE to converge at rate $\sqrt{N}$

No structure, slower convergence

Non-parametric estimators have substantially slower convergence rates

For example, an optimal Kernel Regression with one X variable can achieve $N^{2/5} < N^{1/2}$ convergence (see e.g. Cameron & Trivedi, Ch. 9.5 or Li & Racine Ch. 2)

This means that we need $\sim$ 6 times the sample size to halve RMSE

This becomes worse with higher dimensions of $X \Rightarrow$ curse of dimensionality 😱

	OLS	**Kernel regression**						
$dim(X)$	$d < N$	1	2	3	4	6	8	10
Convergence rates	$N^{1/2}$	$N^{2/5}$	$N^{1/3}$	$N^{2/7}$	$N^{1/4}$	$N^{1/5}$	$N^{1/6}$	$N^{1/8}$
Sample size for 1/2 RMSE	4	$\sim$6	8	$\sim$11	16	32	64	128

How to make sense of convergence rates?

There are probably different ways, but this is the one that works for me

We want to find the value α such that the "blow up factor" doubles, which means in turn that the thing that is blown up halves

For any convergence rate N^δ we can therefore write

$$
\begin{aligned}
\frac{(\alpha N)^\delta}{N^\delta} &= 2 \\
\alpha^\delta N^\delta &= 2N^\delta \\
\alpha^\delta \cancel{N^\delta} &= 2\cancel{N^\delta} \\
\alpha &= 2^{1/\delta}
\end{aligned}
$$

For example for $N^{1/4} \Rightarrow \delta = 1/4 \Rightarrow \alpha = 2^{1/(1/4)} = 2^4 = 16 \Rightarrow$ we need 16 times more observations to halve RMSE

Simulation Notebook: Basics: Convergence rates