Math 390.4 1/31/19 Lec 2

---

the equals sign!

$y = t(z_1, \ldots, z_t)$ Not a model... the truth

the causal drivers

the relationship

the presumed phenomenon of interest

Impossible since z's unknown.

Next best thing: $x_1, \ldots, x_p$ which hopefully captures a lot of the information in the z's.

let $\vec{x}_i := [x_{i1}, \ldots x_{ip}] \in \mathcal{X}$ "input space" or "covariate space"

each one variables, features, attributes, characteristics, regressors, covariates

observation, setting, record, objects, inputs, subjects,

x_1 is called a covariate variable

x_1 was salary... $\# \in \mathbb{R}$ could be negative if debt > income

yesterday... x_2 represented criminal history

$x_2 \in \{$'has past crime record', 'does not have past crime record'$\}$

↓ 1 ↓ 0

x_2 is called a binary variable.

We can also define as worst crime so far committed...

$X_2 \in \{$ none, infraction, misdemeanor, felony $\}$

or factor with $L = 4$ levels.

this is called a categorical / nominal / factor variable. But mathematical models need numbers. How do we make this a numeric? Two choices:

(a) Code it with an inherent order

$X_2 \in \{0, 1, 2, 3\}$ ordinal categorical var.
ordinal factor var.
ordinal var.

Major downside:
Coding is arbitrary

(b) Code it without an order. How?

nominal categorical var.
nominal factor var.
nominal var.

$X_{2a} \in \{0, 1\}$ infraction or not?
$X_{2b} \in \{0, 1\}$ misdemeanor ... ?
$X_{2c} \in \{0, 1\}$ felony ... ?

if $X_{2a} = X_{2b} = X_{2c} = 0 \Rightarrow$ "none"

price to pay $p = 3 \longrightarrow p = 5$ since this single feature

Because $L - 1 = 3$ binary features.

Can we say

$$y = f(x_1, x_2, x_3) \quad \text{if} \quad y = t(z_1, z_2, z_3)$$

No... x_1, x_2, x_3 does not contain the perfect information of z_1, z_2, z_3

Q: How can y be different if same inputs??

Instead...

a model

$$y = f(x_1, x_2, x_3) + \delta \quad \text{or} \quad y \approx f(x_1, x_2, x_3)$$

where $\delta = t - f$ and we call this "error due to ignorance". Why? You don't have access to all the information needed to construct y in a given setting.

How to decrease δ? Collect more relevant x's!

How to get f? → Analytic solution? No...

The approach we will use is called "learning from data", an "empirical approach"! (from measurements)

There are many flavors. We will use "supervised learning" from historical data.

Supervised learning requires 3 ingredients

(1) Training data $\mathbb{D} = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, y_n)\}$
historical samples of inputs/outputs

$\vec{x}_1$ may be Bob's situation, $y_1 = 1$ (he repaid his loan)
$\vec{x}_2$ " Jill's " $y_2 = 1$ (she " her loan)
$\vec{x}_3$ " Bill's " $y_3 = 0$ (he did not repay)
⋮

Alternate notation $\mathbb{D} = (X, \vec{y})$ where $X := \begin{bmatrix} \leftarrow \vec{x}_1 \rightarrow \\ \leftarrow \vec{x}_2 \rightarrow \\ \vdots \\ \leftarrow \vec{x}_n \rightarrow \end{bmatrix}$ dim $n \times p$

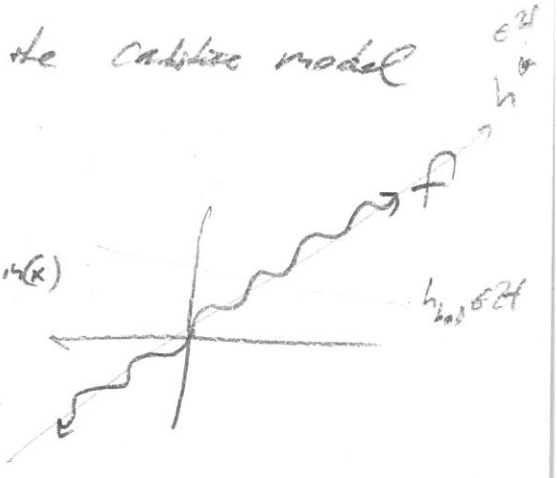
$X \in \mathcal{X}^n$, $\vec{y} \in \mathcal{Y}^n$

(2) $\mathcal{H}$:= a set of candidate functions / the approx. f.
f is an unknown and arbitrary relationship between $\mathbb{R}^p \to \mathbb{R}$.
You cannot learn / compute arbitrary functions. You need to simplify the candidate set.

(3) $\mathcal{A}$:= an "algorithm" that takes $\mathbb{D}$ and $\mathcal{H}$ as input and returns g, an approximation to f,
$g = \mathcal{A}(\mathbb{D}, \mathcal{H})$.

Would $f \in \mathcal{H}$? Usually not! f is complex and usually unknown and $\mathcal{H}$ contains simple functions thus can be fit in A.

However, there is $h^* \in \mathcal{H}$ which is the closest model that most closely approximates f.



e.g. $\mathcal{H} = \{\text{all linear models}\}$ $\quad f(x) = x + 0.1\sin(x)$

$$\Rightarrow h^*(x) = x$$

$$y = h^*(\vec{x}) + \underbrace{f(\vec{x}) - h^*(\vec{x})}_{\text{model misspecification error}} + \underbrace{\varepsilon(\vec{x}) - f(\vec{x})}_{\varepsilon\text{: error due to ignorance}}$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{\varepsilon \text{ again}}$$

How to minimize model misspecification error? Make $\mathcal{H}$ richer, more flexible to capture f more exactly.

Just because h^* exists in $\mathcal{H}$, doesn't mean you will find it! A won't be perfect and the ε's will appear random. So A does its best to produce $g \in \mathcal{H}$.

e.g. $y = f(x) + \varepsilon = x + 0.15\sin(x) + \varepsilon$

prediction
the model we "use"

↓

$$y = g(\vec{x}) + \underbrace{h^*(\vec{x}) - g(\vec{x})}_{\text{estimation error}} + \underbrace{f(x) - h^*(x)}_{\text{misspec. error}} + \underbrace{t(\vec{z}) - f(\vec{x})}_{\text{ignorance error}}$$

e

↑

"residual"
(what's leftover)

It has three components, the three sources of error.

How to reduce est. err? Increase n! Better $\mathcal{A}$?