Math 342W Lecture 2

$y = t(z_1, z_2, \ldots, z_t)$

y:
- response
- outcome
- endpoint
- D.V

t: Unknowable Function (True)

$z_1, \ldots, z_t$: Unknown Causal Drivers

Binary Response

$y \in \{0, 1\} = \mathcal{Y}$ → Output Space

0: Does not pay loan; 1: Pay loan

Assume 3 Casual Drivers

z_1: has \$ at maturity $\in \{0, 1\}$

z_2: Unforseen Emergency $\in \{\text{Yes}, \text{No}\}$

z_3: Criminal Intent $\in \{\text{Yes}, \text{No}\}$

$t(z_1, z_2, z_3) = z_1(1 - z_2)(1 - z_3)$

We can obtain information that is related to z_i's, denoted x_i's

Ex: x_1: Yearly Salary @ time of loan $\in \mathbb{R}$ [Continuous] type
x_2: Missed previous credit card [Binary] type
x_3: Previous criminal record/charges [Binary] type

x_i's are called variables / Features / attributes / characteristics / regressors / co-variants

Collect $x_1 \ldots x_p$ where p is # of regressors

Selecting which features is called feature selection ← IMPORTANT!

Collect measurements

$\vec{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}] \in \mathcal{X}$ => Covariate/input space

Consider.... Levels = 4

$x_3 \in \{$ none $<$ infraction $<$ misdemeanor $<$ felony $\}$

↳ This is no longer Binary ⇒ Ordinal Categorical Variable / Ordinal Factor

↳ How to Represent this numerically (strategies)

1) Code the levels as numbers respecting their order
⇒ none = 0, infraction = 1, misdemeanor = 2, Felony = 3

↳ Downside is these values are subjective

2) Dummification

$x_{3a} \in [0,1]$
$x_{3b} \in [0,1]$
$x_{3c} \in [0,1]$

⇒ All dummies = 0, refrence value/level

⇒ Binaries computed from categerical variables

Nominal Categorical Variables must be dummified

Can we say $y = f(x_1, x_2, x_3)$ ⇒ No, b/c we haven't captured all features

↳ so we say $y = f(x_1, x_2, x_3) + \delta$,
where $\delta = t - f$, and $y = t$ ⇒ This is an error source due to ignorance (#1)

How to minimize δ?

1) Collect more *relevant* regressors

f : the most accurate way of combining your features

How do we do this?

↳ Learn from the data (Supervised Learning), training data

① $\mathbb{D} = \{\langle \vec{x}_1, y_1 \rangle, \langle \vec{x}_2, y_2 \rangle, \ldots, \langle \vec{x}_n, y_n \rangle\}$, Sample size n

$\mathbb{D} = \langle \vec{X}, \vec{y} \rangle$

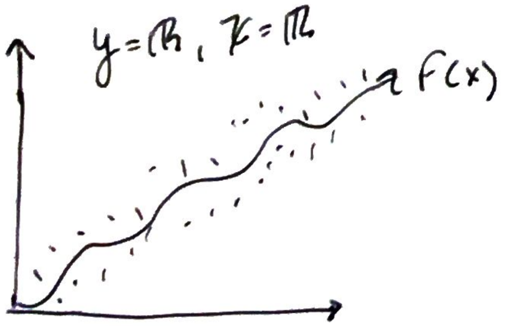
$$\vec{X} = \begin{bmatrix} \leftarrow \vec{x}_1 \rightarrow \\ \leftarrow \vec{x}_2 \rightarrow \\ \leftarrow \vec{x}_n \rightarrow \end{bmatrix}, \quad \vec{y} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

② Let $\mathcal{H}$ = set of candidate functions h that aproximate f

Do we expect $f \in \mathcal{H}$? ⇒ NO ⇒ This is an error source (#2) ⇒ ε = "noise"

↳ Let $h^* \in \mathcal{H}$ ⇒ aproximation to f

$$y = h^*(x_1, x_2, x_3) + (f - h^*) + \delta$$



$\mathcal{H} = \{a + bx : a, b \in \mathbb{R}\}$

Employ a richer $\mathcal{H}$ space

⇓

$\mathcal{H} = \{a + bx + cx^2 : a, b, c \in \mathbb{R}\}$

f(x) is unknowable, we never see it.

③ Algorithm $\mathcal{A}$, where $\mathcal{A}(\mathbb{D}, \mathcal{H}) = g \in \mathcal{H}$

$\Rightarrow h^* - g$ is an error source (#3)

And we are done, g is our model

$$y = \underbrace{g(x_1, x_2, x_3) + \underbrace{(h^* - g) + (f - h^*) + \delta}_{\varepsilon \Rightarrow \text{"noise"}}}$$

$$\underbrace{(h^* - g) + (f - h^*) + \delta}_{e \Rightarrow \text{"residual"}}$$

Error Sources

1) Ignorance Error
2) Misspecification Error
3) Estimation Error

How do we minimize error (estimation)?

1) Get a better $\mathcal{A}$
2) Get/Collect more observations n

$y = \hat{y} + e$, $y = g(\vec{x}) + e$, $y = h^*(\vec{x}) + \varepsilon$, $y = f(\vec{x}) + \delta$

$\Downarrow$

$e = y - \hat{y}$