Lecture 10.

$$\ln(\underbrace{\text{lograte}}_{y} \sim \text{lag1} + \underbrace{\text{lag2} + \text{lag3}}_{X=(x_1, x_2, \dots, x_p)} + \text{dens} + \text{vax}, \dots)$$

$$y \sim \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Model: $\mathbb{E}(y|x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

a collection of distns. $\quad y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon, \quad \mathbb{E}(\varepsilon|x) = 0$

---

* Survey $n = 500$ students

  X = # students in favor of policy A.

→ $X \sim \text{Binom}(500, 0.6)$, $P(X > 300) = ?$

→ ϕ = fraction of all students in favor of policy A.

$X \sim \text{Binom}(500, \phi)$, $\phi \in (0,1)$ ← statistical model

Record $X = 260$

It's not one single distribution explaining the data. Instead, we have an entire collection of distns fighting against each other to explain the same data.

---

① There are infinite $(\beta_0, \dots, \beta_p)$ we can take

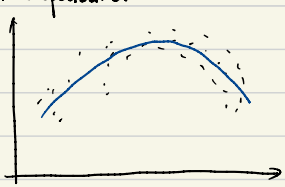
② Distn of ε? Candidates:

* $\varepsilon \sim N(0, \sigma^2)$, $\varepsilon \perp\!\!\!\perp x$
* ε is cont, $\mathbb{E}(\varepsilon|x) = 0$, $\text{Var}(\varepsilon|x) = \sigma^2$ (a huge number of distns in model space)

There is a finite collection just by changing the coefficients

---

Nonparametric



$\mathbb{E}(Y|X) = \eta(x)$ restriction of models

$Y = \eta(x) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$

Each distribution is determined by $\eta(x)$ and ε.

$Y_i \overset{\text{iid}}{\sim} N(\eta(x_i), \sigma^2)$,

$\eta \in$ Some collection of functions

So $p(y) = \prod_{i=1}^{n} N(y_i | \eta(x_i), \sigma^2)$

On the other hand, $\mathbb{E}(Y|X) = \beta_0 + x^T\beta$, $\mathbb{E}(\varepsilon|x) = 0$ is also non-parametric.

$Y = \beta_0 + \beta^T x + \varepsilon, \quad \varepsilon \sim f(\varepsilon|x)$ w/ $\mathbb{E}(\varepsilon|x) = 0$, $\text{Var}(\varepsilon|x) < \infty$.

Sample space is (β_0, β, f)

β_0: $\mathbb{R}$; β: $\mathbb{R}^p$; f: ∞-dim space

dimension is larger than any positive integer.

Concepts:

→ Data: $X \in \mathbb{R}^d$ Data can be a matrix $\underset{n \times d}{X} = (x_1, \dots, x_n)^T$, $x_j \in \mathbb{R}^d$

Model: $X \sim P$, $P \in \mathcal{P}$

↓ index by some simpler quantities

↘ model space a collection of distributions that are trying to explain the same data

$P(\cdot|\theta)$, $\theta \in \Theta$ → parameter space

$\text{Binom}(500, \phi)$, $\phi \in (0,1)$

$\bigotimes_{i=1}^{n} N(\eta(x_i), \sigma^2)$, $\eta \in$ function space, $\sigma > 0$

$Y_i \overset{iid}{\sim} N(\eta(x_i), \sigma^2)$

$\dim(\Theta)$ = finite or infinite

→ Three goals of Data analysis:

① Hypothesis testing:

$H_0: P \in \mathcal{P}_0 \subset \mathcal{P}$

$X_1, \dots, X_{22}$ (self-reported weekly expense on food)

Assume $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$P = \bigotimes_{i=1}^{n} N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma > 0$,

"$H_0: \mu \geq 175$"

$\mathcal{P}_0 = \{ \bigotimes_{i=1}^{n} N(\mu, \sigma^2), \mu \geq 175, \sigma > 0 \}$

* This is not exactly a formal hypothesis testing.

* This is the ealier version: want to prove it wrong

Formal Hypothesis, $\mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1$, $\mathcal{P}_0 \cap \mathcal{P}_1 = \phi$

$H_0: P \in \mathcal{P}_0$, $H_1: P \in \mathcal{P}_1$

② Prediction

Data: X $(X, X^*) \sim P$, $P \in \mathcal{P}$

Future obs: X^* ↙ $X^* \sim P(\cdot|X)$

Clinical: Control $X_1, \dots, X_n, X^* \overset{iid}{\sim} N(\mu_1, \sigma^2)$

Treatment $Y_1, \dots, Y_m, Y^* \overset{iid}{\sim} N(\mu_2, \sigma^2)$

Typical $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$

$Y^* - X^*$, $P(Y^* > X^*)$ ← real interests

③ Estimation

$X_1, \dots, X_n, X_1^*, \dots, X_N^* \overset{iid}{\sim} N(\mu_1, \sigma^2)$
$Y_1, \dots, Y_n, Y_1^*, \dots, Y_N^* \overset{iid}{\sim} N(\mu_2, \sigma^2)$
} $\bar{Y}^* - \bar{X}^* \approx \mu_2 - \mu_1$

Est = large scale avg prediction.
↓
(1 future obs)