# STA 355 Lecture 10

## Regression.

Predict response $Y$ using covariate $X$.

$(X_i, Y_i) \overset{iid}{\sim} \mathbb{P}$.

Want to estimate conditional expectation

$$f^*(x) = \mathbb{E}[Y \mid X = x].$$

Method: use a class $\mathcal{F}$ of functions.

search $f^*$ (or its approximation)

with $\mathcal{F}$.

($f^*$ does not have to be in $\mathcal{F}$.)

(We do not need to assume that

$Y_i = f^*(X_i) + \text{noise}$

(no need to model the entire data-generating mechanism).

— Linear regression.

Feature vector $X \mapsto \phi(X) \in \mathbb{R}^d$.

$$\mathcal{F} = \{ x \mapsto \beta^T \phi(x) : \beta \in \mathbb{R}^d \}.$$

— Sparse linear regression.

where we consider a subset of $\mathcal{F}$ in LR, with $\beta$ having only a small subset of non-zero entries.

— Nonparametric smooth function classes.

eg. $\mathcal{F} = \{ f : [0,1] \to [0,1] \mid |f(x) - f(y)| \le |x-y| \; \forall x, y \in [0,1] \}$

eg. $\mathcal{F} = \{ f : [0,1] \to [0,1] \mid |f^{(k)}(x)| \le L, \; \forall x \in [0,1] \}$

— Neural nets.

Linear regression.

$$\hat{\beta}_n := \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \beta^T X_i \right)^2 \right\}$$

(for simplicity, we just let $\phi(X) = X$).

In many cases, we may want to consider the model class $x \mapsto \beta_0 + \beta^T x$.

In such a case, we simply let $\phi(x) = \begin{bmatrix} 1 \\ x \end{bmatrix}$

"Lease-square estimator"

(In general, we fit

$$\hat{f}_n = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - f(X_i) \right)^2 .)$$

Fact. Suppose that $Y_i = f^*(X_i) + \varepsilon_i$

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2). \qquad (\sigma > 0)$$

then the lease square estimator is MLE.

(and therefore consistent, asymptotically normal, asymptotically optimal, under certain assumptions)

Proof of the fact.

$$\text{log-likelihood} = \sum_{i=1}^{n} \left( -\frac{(Y_i - f(X_i))^2}{2\sigma^2} - \log(2\pi\sigma^2) \right)$$

note: this holds true for any regression problems.

---

For linear regression, let's start with models under strong assumptions.

$$Y_i = \beta_*^T X_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} P.$$

$$\mathbb{E}[\varepsilon_i] = 0, \quad \text{var}(\varepsilon_i) = \sigma^2.$$

only need ← well-specified linear model

Fact. $\mathbb{E}\left[\hat{\beta}_n \mid X_1 \cdots X_n\right] = \beta_*$.

$$\text{cov}\left(\hat{\beta}_n \mid X_1 \cdots X_n\right) = \sigma^2 \cdot (XX^T)^{-1}.$$

↑ requires iid noises w/ finite var.

where $X = [X_1, X_2, \cdots X_n] \in \mathbb{R}^{d \times n}$.

Proof.

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \beta^T X_i\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \frac{2}{n}\sum_{i=1}^{n} Y_i X_i^T \beta$$

$$+ \beta^T \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T\right)\beta.$$

Solving the minimization problem

$$\hat{\beta}_n = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \cdot \left(\frac{1}{n}\sum_{i=1}^{n} Y_i X_i\right).$$

(this holds true regardless of probabilistic assumptions)

Now using prob assumptions.

$$Y_i = \beta_*^T X_i + \varepsilon_i$$

$$\sum_{i=1}^{n} Y_i X_i = \sum_{i=1}^{n} X_i X_i^T \beta_* + \sum_{i=1}^{n} \varepsilon_i X_i.$$

$$\hat{\beta}_n = \left(\frac{1}{n}\sum_{i}^{n} X_i X_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i}^{n} X_i X_i^T \beta_* + \frac{1}{n}\sum_{i}^{n} \varepsilon_i X_i\right)$$

$$\mathbb{E}\left[\hat{\beta}_n \mid X_1, X_2, \cdots X_n\right] = \beta_*.$$

(for this step, we only use the fact that $\mathbb{E}[Y_i \mid X_i = x] = \beta_* x$ is actually linear, i.e. $\mathbb{E}[\varepsilon_i \mid X_i = x] = 0$).

For the conditional variance.

$$\mathrm{Cov}\left(\hat{\beta}_n \mid X_1, \cdots X_n\right) = \mathrm{Cov}\left(\left(\sum_{i=1}^n X_i X_i^T\right)^{-1} \cdot \left(\sum_{i=1}^n \varepsilon_i X_i\right)\right)$$

$$= \sum_{j=1}^n \mathrm{Cov}\left(\left(\sum_{i=1}^n X_i X_i^T\right)^{-1} X_j \varepsilon_j\right).$$

$$= \sigma^2 \cdot \sum_{j=1}^n \left(XX^T\right)^{-1} X_j X_j^T \left(XX^T\right)^{-1}$$

$$= \sigma^2 \cdot \left(XX^T\right)^{-1}.$$

(this step requires $\varepsilon_i$'s to be i.i.d).

---

Asymptotics.

Fact. under above assumptions.

assuming that $\mathbb{E}[XX^T] \succ 0$;

then we have

$$\sqrt{n}\left(\hat{\beta}_n - \beta^*\right) \xrightarrow{d} N\left(0, \sigma^2 \cdot \left(\mathbb{E}[XX^T]\right)^{-1}\right).$$

Proof: $\hat{\beta}_n - \beta^* = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i X_i\right)$.

(by LLN) $\downarrow p$ $\mathbb{E}[XX^T]$

$\downarrow d$ $N\left(0, \sigma^2 \mathbb{E}[XX^T]\right)$ (after $\sqrt{n}$ scaling).

by Slutsky thm, we have the convergence.

---

We can perform inference based on asymptotics.

- Straightforward approach.

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$$

$$\hat{\sigma}^2 = \frac{1}{n-d}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_n^T X_i\right)^2.$$

This $\frac{1}{n-d}$ normalization makes $\hat{\sigma}^2$ a conditionally unbiased estimator for $\sigma^2$

(conditioned on $X_1, X_2, \cdots X_n$)

(when $n \gg d$, this does not make much difference)

Putting them together

$$\sqrt{n} \cdot \frac{\hat{\Sigma}_n^{1/2}}{\hat{\sigma}_n} \left( \hat{\beta}_n - \beta^* \right) \xrightarrow{d} N(0, I_d).$$

. We can also use bootstrap etc.

---

Removing assumptions. ~~$\varepsilon_i$ i.i.d~~

$$\mathbb{E}[Y | X = x] = \beta_*^\top x.$$

$$\text{var}(Y | X = x) = \sigma^2(x).$$

Fact. $\sqrt{n}\left(\hat{\beta}_n - \beta^*\right) \xrightarrow{d} N\left(0, \Sigma_X^{-1} \Sigma^* \Sigma_X^{-1}\right)$

where $\Sigma_X = \mathbb{E}[XX^T]$.

$$\Sigma^* = \mathbb{E}\left[\sigma(X)^2 \cdot XX^T\right].$$

Proof. $\sqrt{n}(\hat{\beta}_n - \beta^*) = \left(\frac{1}{n}\sum_i^n X_i X_i^T\right)\left(\frac{\sqrt{n}}{n}\sum_{i=1}^n \varepsilon_i X_i\right)$.

$\left(\frac{1}{n}\sum_i^n X_i X_i^T\right) \xrightarrow{p} \Sigma_X$

$\left(\frac{\sqrt{n}}{n}\sum_{i=1}^n \varepsilon_i X_i\right) \xrightarrow{d} N(0, \Sigma^*)$.

. We can also perform inference by estimating $\Sigma_X$ and $\Sigma^*$.

$$\hat{\Sigma}_X = \frac{1}{n}\sum_i^n X_i X_i^T$$

$$\hat{\Sigma}^* = \frac{1}{n}\sum_i^n \left(Y_i - \hat{\beta}_n X_i\right)^2 X_i X_i^T.$$

Remark - this method is more robust

. We can always use bootstrap.

---

Removing more assumptions. ~~$\mathbb{E}[Y|X=x] = \beta_x^T x$.~~

. Fact. under only moment assumptions on $X, Y$

We have that

$$\hat{\beta}_n \xrightarrow{p} \bar{\beta} = \underset{\beta \in \mathbb{R}^d}{\arg\min}\ \mathbb{E}\left[(Y - X^T\beta)^2\right].$$

$$\left(= \underset{\beta \in \mathbb{R}^d}{\arg\min}\ \mathbb{E}\left[|f^*(X) - \beta^T X|^2\right]\right)$$

(Best linear approximation of the conditional expectation function.

$$f^*(x) = \mathbb{E}[Y \mid X = x]\ )$$

(we can also study $\sqrt{n}(\hat{\beta}_n - \bar{\beta}) \to$ sth)

Proof:

$$\hat{\beta}_n = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i^T\right)^{-1} \cdot \left(\frac{1}{n}\sum_{i=1}^n X_i Y_i\right)$$

$\frac{1}{n}\sum_{i=1}^n X_i X_i^T \xrightarrow{p} \mathbb{E}[XX^T]$

$\frac{1}{n}\sum_{i=1}^n X_i Y_i \xrightarrow{p} \mathbb{E}[XY]$

$$\bar{\beta} = \underset{\beta \in \mathbb{R}^d}{\arg\min}\left\{\beta^T \mathbb{E}[XX^T]\beta - 2\beta^T \mathbb{E}[XY]\right\}$$

$$= \mathbb{E}[XX^T]^{-1} \cdot \mathbb{E}[XY].$$

Other extensions / procedures.

- Prediction interval.

  — Learn a model using data $(X_i, Y_i)_{i=1}^n$

  — Given a new data $X_0$

  the goal is to construct an interval $I$ (based on data)

  s.t. $\mathbb{P}(Y \in I \mid X = x_0) \geq 1-\alpha$.

  (both $Y$ and $I$ are random)

- Naive idea: use CI for $\beta_*^T x_0$ (using $\hat{\beta}_n^T x_i$)

  Problem: $Y$ is also a r.v.

  $Y$ is indp of $I$.

  Assume $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

then we can construct prediction interval.

$$\hat{\beta}_n^T x_0 \pm \sqrt{\hat{\sigma}^2 + \frac{\hat{\sigma}^2 \cdot x_0^T \hat{\Sigma}_X^{-1} x_0}{n}} \; z_{\alpha/2}$$

$$\sqrt{n}\left(\hat{\beta}_n^T x_0 - \beta_*^T x_0\right) \xrightarrow{d} N(0, \sigma^2 x_0^T \Sigma_X^{-1} x_0)$$

$$(\Sigma_X = \mathbb{E}[XX^T])$$

$$Y - \beta_*^T x_0 \sim N(0, \sigma^2) \longrightarrow \text{independent}$$

Taking difference

$$Y - \hat{\beta}_n^T x_0 \sim N\left(0, \sigma^2 + \frac{\sigma^2 x_0^T \Sigma_X^{-1} x_0}{n}\right)$$

---

Model selection in linear regression.

- Training error.

$$\hat{R}_{tr} = \sum_{i=1}^{n} \left(Y_i - \hat{f}_n(X_i)\right)^2.$$

- Prediction risk (in sample)

$$R = \sum_{1}^{n} \left(Y_i^* - \hat{f}_n(X_i)\right)^2.$$

where $Y_i^*$ is a fresh indp sample from $Y_i | X_i$.

Fact. $$\mathbb{E}\left[\hat{R}_{tr} \mid X_1 \cdots X_n\right] \le R$$

and $$R - \mathbb{E}\left[\hat{R}_{tr} \mid X_1 \cdots X_n\right]$$

$$= 2 \cdot \sum_{i=1}^{n} \mathrm{cov}\left(\hat{Y}_i, Y_i \mid X_1 \cdots X_n\right)$$

where $\hat{Y}_i = \hat{f}_n(X_i)$.

("overfitting" in ML).

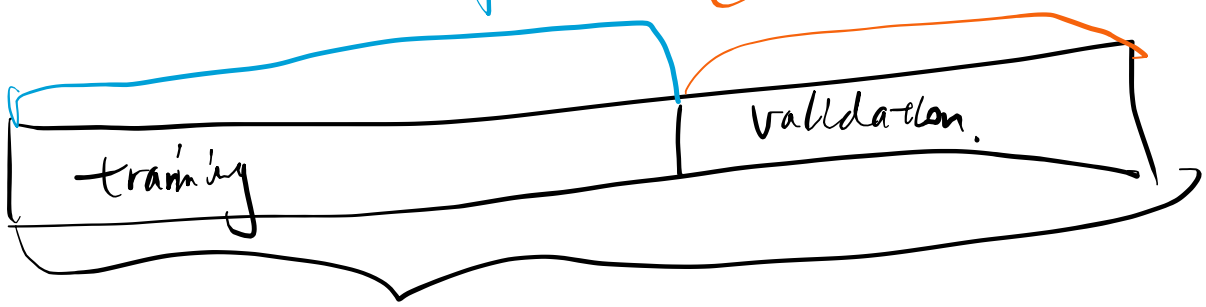
- Mallow's $C_p$ statistics.

  Consider linear regression using subset $S$ of the covariate variables.

$$\hat{R}(S) = \hat{R}_{tr}(S) + 2|S| \cdot \hat{\sigma}^2.$$

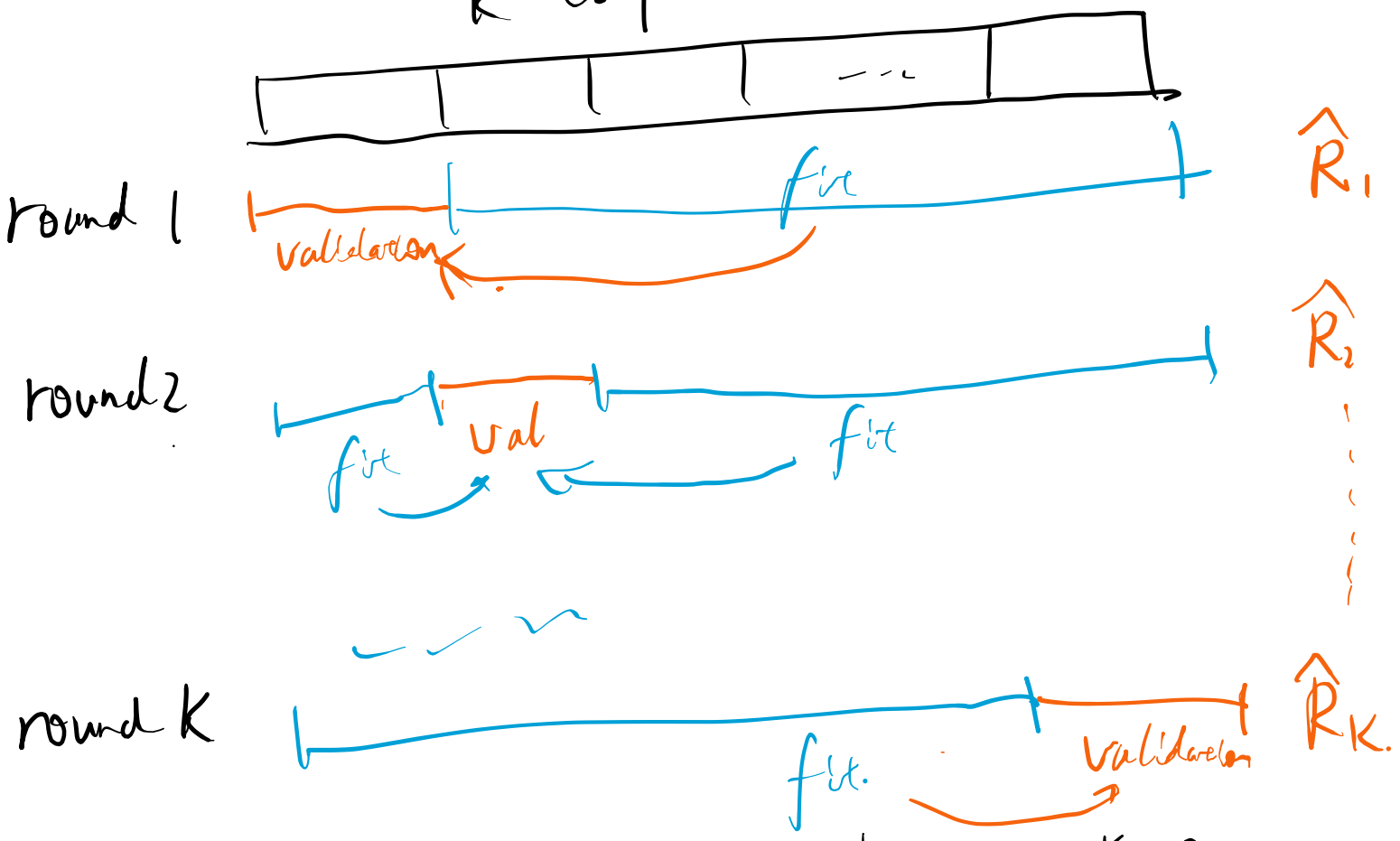
gives a better estimator for risk.

- cross validation. $\hat{f}_n$ → estimate loss.




unbiased estimation for prediction risk.

- K-fold CV.




Estimate the prediction risk by $\frac{1}{K}\sum_{i=1}^{K}\hat{R}_i$.

- Leave-one-out CV

(K-fold with $K = n$).

---

Logistic regression.

$Y_i \in \{0, 1\}$. $\quad Y_i \mid X_i = x \sim \text{Ber}(p^*(x))$

(Issue with linear model

$p^*(x) = x^T \beta_*$.

$x^T \beta_*$ may not always lie in $[0,1]$).

Natural idea: put a link function $\sigma(X^T \beta_*)$

Choice of $\sigma$:

- Smooth (for computational convenience)
- increasing
- mapping to $[0,1]$

Practical choice

$$\sigma(x) = \frac{e^x}{1+e^x} \quad \text{(sigmoid function)}$$

Assuming $Y_i | X_i \sim \text{Ber}\left(\sigma(X_i^T \beta_*)\right)$.

$$\text{log likelihood} = \log\left(\prod_{i=1}^{n} p_i^{Y_i} (1-p_i)^{(1-Y_i)}\right)$$

$$= \sum_{i=1}^{n} \left(Y_i \log p_i + (1-Y_i) \log(1-p_i)\right)$$

$$= \sum_{1}^{n} Y_i \log \sigma(X_i^T \beta) + (1-Y_i) \log\left(1-\sigma(X_i^T \beta)\right)$$

$$= \sum_{1}^{n} Y_i \cdot \left(X_i^T \beta - \log\left(1+e^{X_i^T \beta}\right)\right)$$

$$- (1-Y_i) \cdot \log\left(1+e^{X_i^T \beta}\right)$$

$$= \sum_{i=1}^{n} Y_i X_i^T \beta - \log\left(1+e^{X_i^T \beta}\right)$$