Week 9: predict CLTV

Suppose we want to predict $CLTV_i$ (transaction in one year)

using recency R_i

R_i: days since last transaction.

R_i captures a customer is active or not.

$E[CLTV_i | R_i]$ is the best predictor.

Q: how to estimate it?

Sample analogue

$\hat{E}[CLTV | R_i=r]$

= sample average of CLTV with $R_i=r$

Quiz: how would you calculate $\hat{E}[CLTV | R_i=r]$?

e.g. $\hat{E}[CLTV | R_i=7]$

subsample last transact on 2022/10/24

Y_i = total transaction between 2022/11/31 ~ 2023/10/31.

We can estimate $E[CLTV | R_i]$ by OLS regression:

(β_n: sample avg. of $R_i=n$; $\mathbb{1}$ is its indicator function.)

$$CLTV_i = \beta_1 \mathbb{1}\{R_i=1\} + \beta_2 \mathbb{1}\{R_i=2\} + \cdots + \beta_{90} \mathbb{1}\{R_i=90\}$$

① convenient package
② easy-to-communicate
③ inference

Ex: Is vaccine effective?

Covid: infective or not?
Vaccine: Vaccinated or not?
Sup. vaccine is randomly given

We can compare

$$\overline{Covid}_{Vacc=1} - \overline{Covid}_{Vacc=0} \quad \cdots ①$$

You can also:

$Covid = \beta_0 + \beta_1 Vaccine + \varepsilon$
and ① will be your $\hat{\beta}_1$

$Covid = \gamma_0 + \gamma_1 Vaccine + \gamma_2 Male + \gamma_3 Vaccin.Male$
$\gamma_1 + \gamma_3$ is the effect of vaccine on male.

Very often, estimators in economics are equivalent to an OLS w/

certain specification OLS is a work horse.

We know:
$CLTV = \beta_1 \mathbb{1}\{R_i=1\} + \beta_2 \mathbb{1}\{R_i=2\} + \cdots + \beta_{90} \mathbb{1}\{R_i=90\}$
estimate the conditional mean.

The equation has 90 parameters. if $n = 1800$ (small) When p large, we call it "high dimensional"
$\frac{n}{p} = \frac{1800}{90} = 20 \Rightarrow$ Use 20 obs. to estimate one mean $\Rightarrow$ noisy

The ratio $\frac{n}{p}$ determines whether the model is high-dim.

$\frac{n}{p}$: $c>1$ c moderate
<1 very high
$\gg 1$ low-dim

Alternative model:

$CLTV_i = r_1 \mathbb{1}\{1 \le R_i \le 7\}$
$+ r_2 \mathbb{1}\{8 \le R_i \le 15\}$
$+ \cdots +$
$+ r_{14} \mathbb{1}\{85 \le R_i \le 90\}$

Now, $\frac{n}{p} = \frac{1000}{14} > 100$

easier to estimate, but biased $\because \neq$ conditional mean.

Alternatively,

$CLTV = b_0 \mathbb{1}\{1 \le R_i \le 30\}$
$+ b_1 \mathbb{1}\{31 \le R_i \le 60\}$
$+ b_2 \mathbb{1}\{61 \le R_i \le 90\}$

easier to estimate. but more biased.

Daily, $p=90$, $\frac{n}{p} = 20$, non-biased

Weekly $p=14$ $\frac{n}{p} > 100$, medium-biased

Montly $p=3$ $\frac{n}{p} > 500$, highly. biased.

Facing a trade-off problem:
bias - variance tradeoff.

bias: whether the model is similar to best predictor.

variance: whether the model is easy to estimate.

Can be chosen by cross-validation.

"All models are wrong, but some are useful"

Penalized regression

last class "smoothing" exploiting the continuity of $f(r) = E[CLTV \mid R_i = r]$

Penalization

Penalize complicated models (model that have many non-zero coefficients)

$Y_i = \beta_0 + \beta_1 x_i + \beta_2 X_2 \cdots + \cdots + \beta_p X_p$.

if only a few β_j is not zero. $\Rightarrow$ Simple model.

Penalized regression.

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \sum (y_i - x_i^T \beta)^2 + \underbrace{\lambda \|\beta\|_q}_{\text{penalization.}}$$

where $\lambda > 0$ is the penalty term.

$$\|\beta\|_q = \sqrt[q]{\sum \|\beta_j\|^q}$$

Remark:

λ is a "tuning parameter" specified by the user.

"There are many penalized regression methods $q=1$. LASSO (Least Absolute Selection and Shrinkage Operator)

Why $q=1$?

① convexity (for $q \geq 1$)
if $q<1$, objective function isn't convex.

eg. $q = \frac{1}{2}$. $p=1$ $\|\beta\| = \sqrt{\beta}$, not convex.

② $q=1$ has sparsity 稀疏 $\frac{T}{-}$ $\frac{P}{x}$

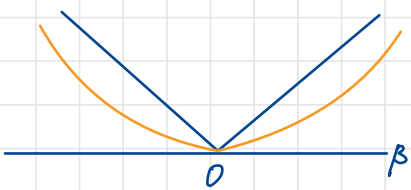
$\hat{\beta}_{Lasso}: \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$ most $\hat{\beta}_j = 0$

Advantage of sparsity:

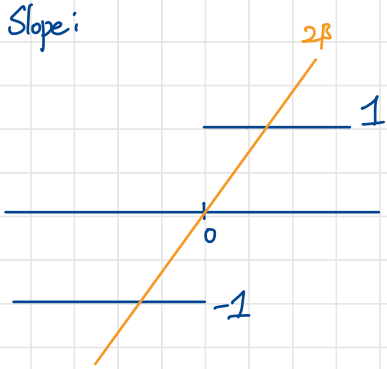
① easy to interpret.

② Can be used for variable selection.
(only choose X_j $|\hat{\beta}_j| \neq 0$)

Penalty ($p=1$)
$\|\beta\|_q = |\beta|$
β^2



Slope:



Marginal penalty:

move ε away from 0
marginal penalty $= 1 \Rightarrow q = 1$
" $\cong 0 \Rightarrow q = 2$

LASSO regression likes to stay in 0, different from Ridge.

$p=2$, $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$

What is a circle?

$q=1$ $\|\beta\| = |\beta_1| + |\beta_2|$ $\{X \mid \|X\| = 1\}$

$q=2$ $\|\beta\|_2 = \sqrt{\beta_1^2 + \beta_2^2}$ $\{X \mid \|X\|_2 = 1\}$

LASSO

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_1, \quad \|\beta\|_1 = |\beta_1| + |\beta_2| + \cdots + |\beta_p|$$

is differentiable everywhere.

$\hat{\beta}_{LASSO} = ?$

In general, LASSO does not have analytic solution.

Can be written as func. of data.

($\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$)

⇒ has to use numerical optimization.

(versions of gradient descent)

Special Case:

$x_{1i}, x_{2i}, \ldots, x_{pi}$ are dummy variables

$$\sum_{i=1}^{p} x_{1i} = 1$$

e.g.

x_{1i}: $\mathbb{1}\{Female\}$

x_{2i}: $\mathbb{1}\{Male\}$

⇒ (As group membership indicators)

Objective function

$$\sum (y_i - x_i'\beta)^2 + \lambda \|\beta\|_1$$

$$= \sum_{x_{i1}=1} (y_i - \beta_1)^2 + \sum_{x_{i2}=1} (y_i - \beta_2)^2 + \cdots \sum_{x_{ip}=1} (y_i - \beta_p)^2 + \lambda(|\beta_1| + |\beta_2| + |\beta_3| + |\beta_4| + \cdots + |\beta_p|)$$

$m = 1$

$m = 2$

Define

$$L_j(\beta_j) = \sum_{x_{ij}=1} (y_i - \beta_j)^2 + \lambda|\beta_j|$$

$$\text{LASSO Obj.} = \sum_{j=1}^{n} L_j(\beta_j)$$

$$\min_{\beta \in \mathbb{R}^p} \sum (y_i - x_i'\beta)^2 + \lambda \|\beta\|_1$$

$\Leftrightarrow$ $\min_{\beta_1} L_1(\beta_1)$

$\min_{\beta_2} L_2(\beta_2)$

$\vdots$

$\min_{\beta_3} L_p(\beta_p)$

Define

$n_j = \sum_i \mathbb{1}\{x_{ij}=1\}$

$=$ # observations $x_{ij}=1$

$\bar{y}_j = \frac{1}{n_j} \sum_{i: x_{ij}=1} y_i$

$=$ avg. of y among $x_{ij}=1$

$$L_j(\beta_j) = \sum_{i: x_{ij}=1} (y_i - \beta_j)^2 + \lambda |\beta_j|$$

$$\sum (y_i - \beta_j)^2 = \sum (y_i^2 - 2\beta_j y_i + \beta_j^2)$$

$$= \sum y_i^2 - 2\beta_j \sum y_i + n\beta_j^2$$

$$= \beta_j^2 - 2n\bar{y}_j \beta_j + c$$

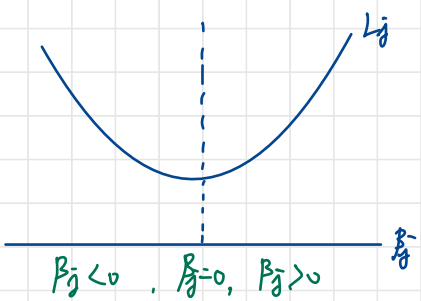
$\hat{\beta}_j^{OLS} = \bar{y}_j$ since

$n\beta_j^2 - 2n\bar{y}_j\beta + c$

$= n(\beta_j - \bar{y}_j)^2 + c$

$\hat{\beta}_j^{\lambda=0} = \hat{\beta}_j^{OLS} = \bar{y}_j$

$$L_j(\beta_j) = n(\beta_j - \bar{y}_j)^2 + \lambda|\beta_j|$$



At $\beta_j = 0$

$L_j(\beta_j)|_{\beta_j=0} = n\bar{y}_j^2$

At $\beta_j > 0$

$\frac{dL_j}{d\beta_j} = 2n(\beta_j - \bar{y}_j) + \lambda = 0$

$2n\beta_j = 2n\bar{y}_j - \lambda$

$\beta_j^* = \bar{y}_j - \frac{\lambda}{2n}$, if $\bar{y}_j - \frac{1}{2n}\lambda > 0$

If $\beta_j < 0$

$L_j = n(\beta_j - \bar{y}_j) - \lambda\beta_j$ $\frac{dL_j}{d\beta_j} = 0$

$$\Rightarrow \beta_j = \begin{cases} \bar{y}_j + \frac{1}{2}\lambda, & \text{if } \bar{y}_j + \frac{\lambda}{2n} < 0 \\ 0 & \end{cases}$$

$$\Rightarrow \hat{\beta}_j = \begin{cases} \bar{y}_j - \frac{1}{2n}\lambda, & \text{if } \bar{y}_j - \frac{1}{2n}\lambda > 0 \\ 0 & \text{o.w.} \\ \bar{y}_j + \frac{1}{2n}\lambda, & \text{if } \bar{y}_j + \frac{1}{2n} < 0 \end{cases}$$

Compare OLS & LASSO

$\hat{\beta}^{OLS} > \frac{1}{2n}\lambda \Rightarrow \hat{\beta}^{LASSO} = \hat{\beta}^{OLS} - \frac{1}{2n}\lambda$

$\hat{\beta}^{OLS} < \frac{1}{2n}\lambda \Rightarrow \hat{\beta}^{LASSO} = \hat{\beta}^{OLS} + \frac{1}{2n}\lambda$

$\frac{-1}{2n}\lambda \le \hat{\beta}^{OLS} \le \frac{1}{2n}\lambda \Rightarrow \hat{\beta}^{LASSO} = 0$

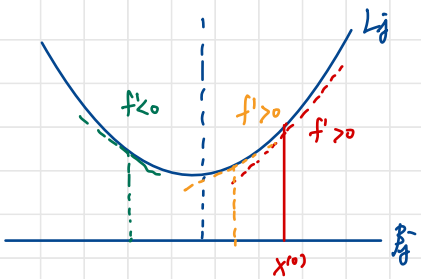
Selection: $\hat{\beta} \mapsto 0$

only keep coefficients with large $|\hat{\beta}_j^{OLS}|$

Shrinkage: shrink non-zero coefficients toward 0.

Solving LASSO with numerical optimization.

LASSO can be solved by versions of gradient descent.



Start with initial point $X^{(0)}$

Iterate

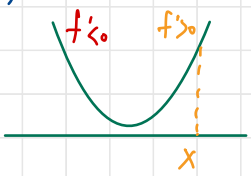
$X^{(t+1)} = X^t - \eta^{(t)} Df(X^{(t)})$

$\eta^{(t)}$ step size.

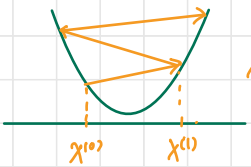
Eg.

$f(x) = 2(x-1)^2 + 5$

$f'(x) = 4(x-1)$

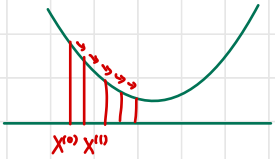


η too large $\eta_t \to \infty$



NOT Converge!

η too small



TOO SLOW!

How to choose $\eta^{(t)}$?

Newton's method.

$\eta^{(t)} = (\nabla^2 f)^{-1} = (f''(x^{(t)}))^{-1}$, when $p=1$

→ More conservative (smaller step) when slope is changing fast.

Challenges.

non-convexity: converges to local minima.