OLS & Lasso Regression

羅健華、劉德翎

Economics, NTU

Dec 8, 2017

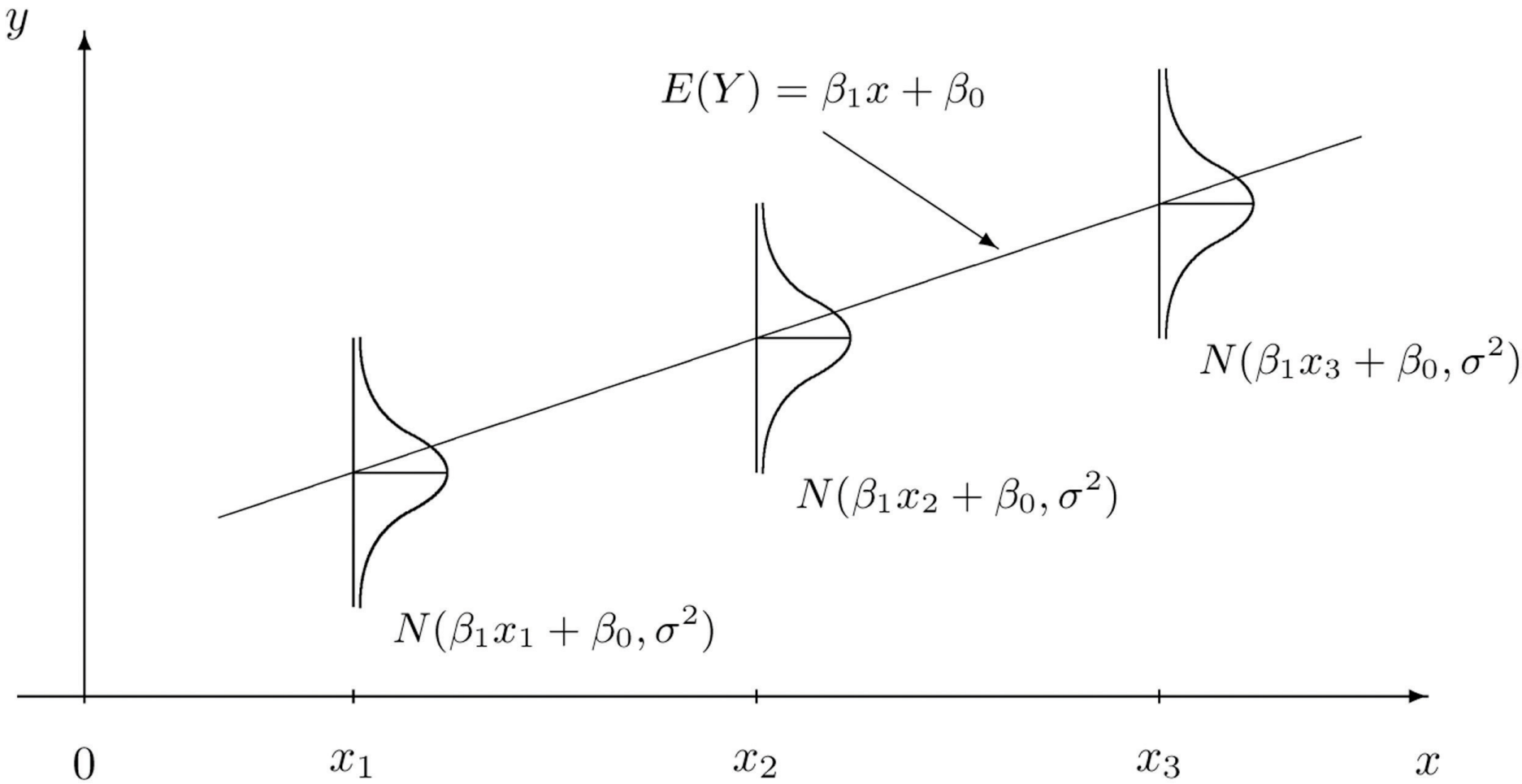
Outline

1. OLS
   - 線性迴歸
   - T-test
   - Dummy Variable
   - Interaction Term
2. Lasso
   - Why Lasso?
   - Penalization

Outline

線性迴歸

- 給定一迴歸函數 $Y_i = \alpha + \beta X_i + u_i$ 及未知的兩個參數 α 和 β
- 我們使用最小平方法 ordinary least squares (OLS) 去估計 $\hat{\alpha}$, $\hat{\beta}$, 及 $\hat{u}_i$
- min $Q = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$
- $Y_i = \hat{Y}_i + \hat{u}_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$
- u_i 是一個隨機變數，我們假設其期望值為 0，且假設變異數都一樣。
- β 衡量的是當 X 增加一單位時，Y 增加多少單位。但這只是相關性，而非因果關係。
- 也可以同時加入很多自變數 (X) 在迴歸式中。


y
$E(Y) = \beta_1 x + \beta_0$
$N(\beta_1 x_3 + \beta_0, \sigma^2)$
$N(\beta_1 x_2 + \beta_0, \sigma^2)$
$N(\beta_1 x_1 + \beta_0, \sigma^2)$
0
x_1
x_2
x_3
x

例子：租金與房子大小。

- 也可以同時加入很多自變數 (X) 在迴歸式中。
- 模型將會被以向量與矩陣的方式表示，寫成 $Y = X\beta + u$，其中 y 是具有 N 個元素的向量，X 是 $N \times k$ 矩陣，β 是 k 個元素的向量。
- $\hat{\beta} = (X^T X)^{-1}(X^T Y)$

Fitness

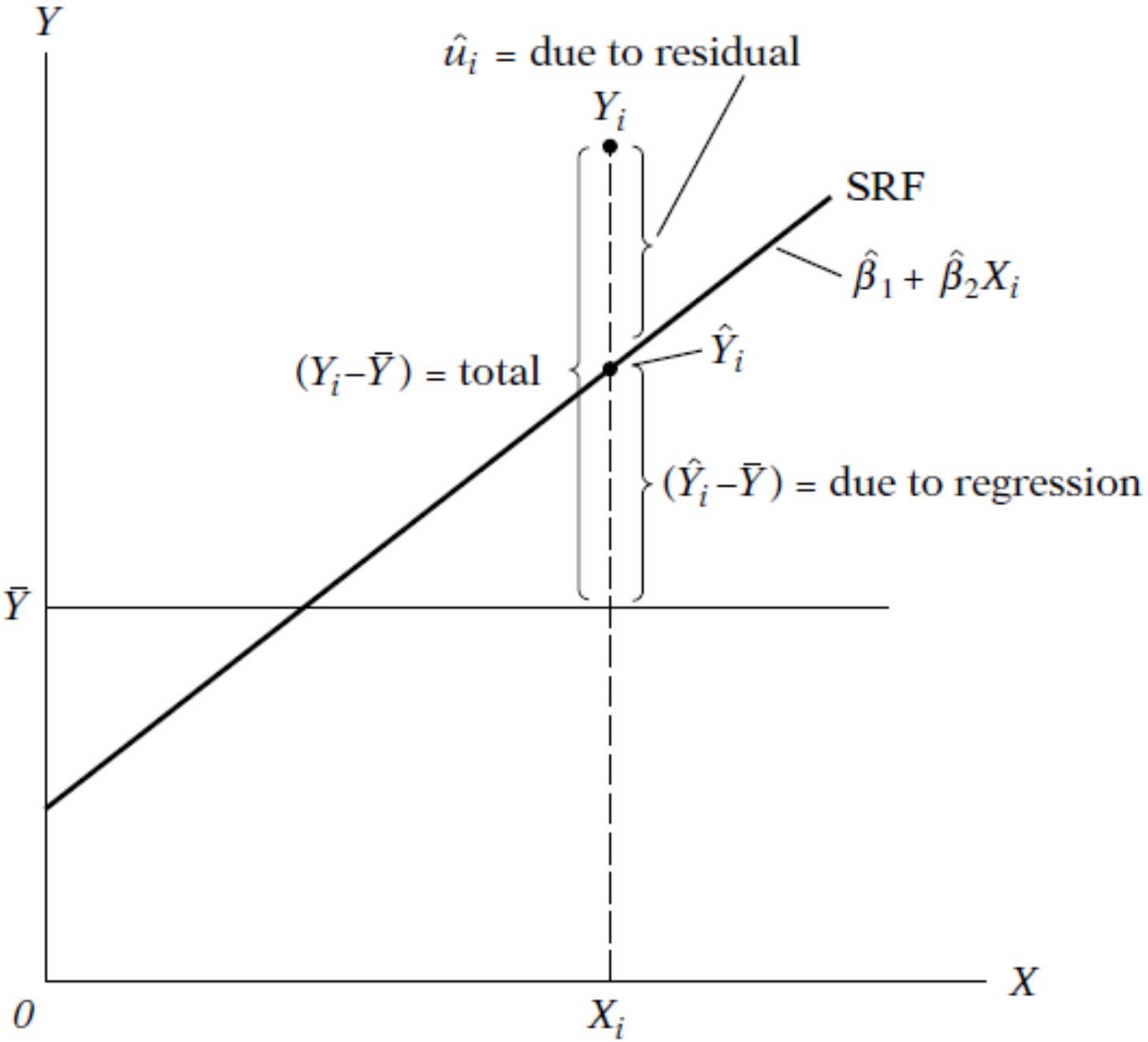
定義

- $TSS = \sum(Y_i - \bar{Y})^2$
- $RSS = \sum(Y_i - \hat{Y}_i)^2$
- $ESS = \sum(\hat{Y}_i - \bar{Y})^2$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

在有截距項 α 的情況下，$TSS = ESS + RSS$ 且 $0 \leq R^2 \leq 1$

R^2 越高，代表該模型解釋能力越好。


Y
$\hat{u}_i$ = due to residual
Y_i
SRF
$\hat{\beta}_1 + \hat{\beta}_2 X_i$
$\hat{Y}_i$
$(Y_i - \bar{Y})$ = total
$(\hat{Y}_i - \bar{Y})$ = due to regression
$\bar{Y}$
X
0
X_i

- 然而，放入越多自變數 X 時，該模型解釋能力只升不降。
- 但放入越多自變數 X 不一定是比較好的。缺點：(1) 收集資料需要成本。(2) 在預測上會 overfitting。
- 我們通常比較 Adjusted R-squared

$$\bar{R}^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - k - 1}$$

其中 k 代表自變數的個數，N 代表樣本數。

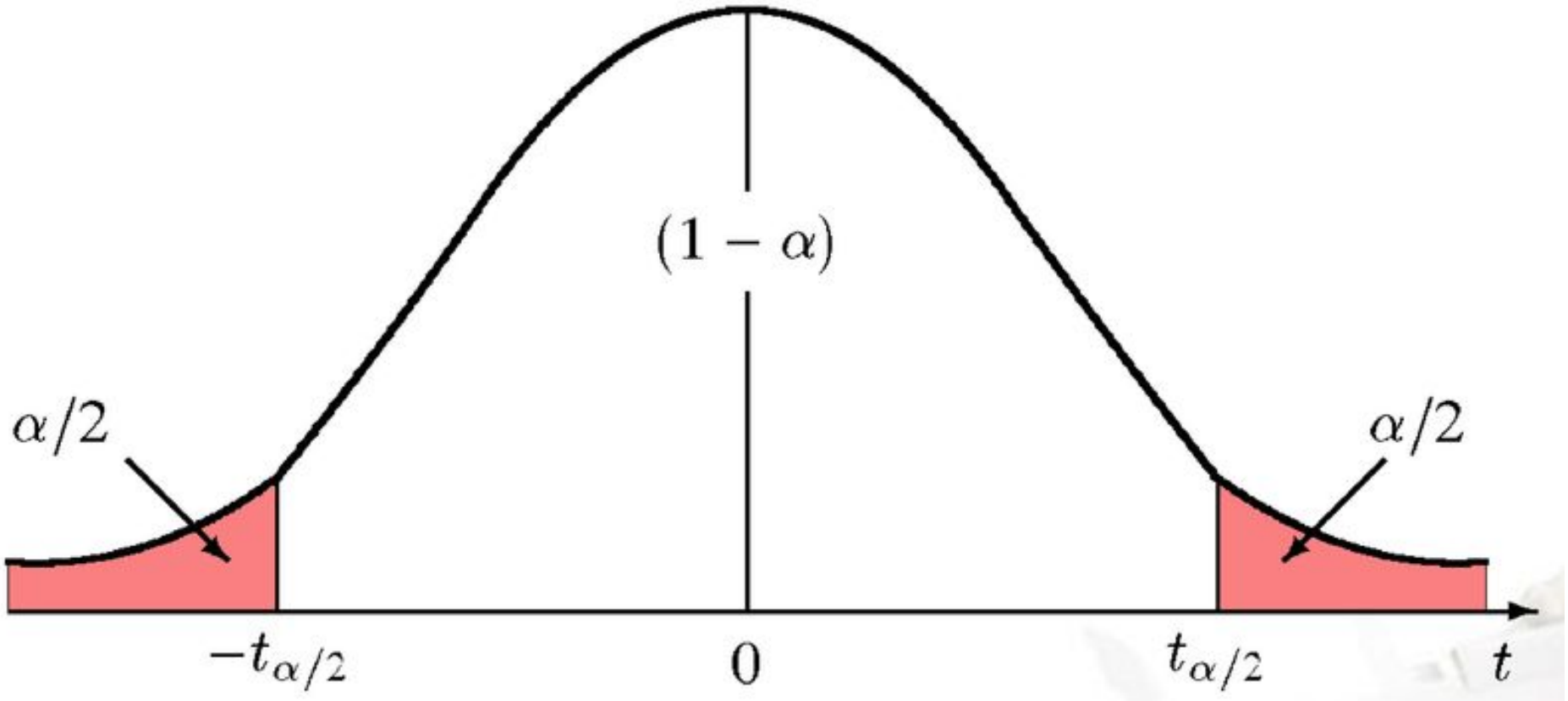
隨機變數

- 隨機變數指的是由某狀態空間映射到實數線的函數。ex: 擲骰子。
- 期望值是試驗中每次可能的結果乘以其結果機率的總和，記做 $E(X)$。
- 變異數 (Variance) 則是描述的是隨機變數的離散程度，也就是離其期望值的距離。標準差 (standard error) 則為變異數開根號。

$$Var(X) = E((X - E(X))^2)$$

T-test

- 當我們估出一個係數非 0 後，並不能武斷地說該變數就是有效果，而是必須透過假設檢定才行。
- 每個估計出來的 $\hat{\beta}$ 其實是一個隨機變數，除以標準差後，分配是 t 分配。
- 我們假設真正的參數 (期望值) 為 0，看看在這個樣本之下，估計出來的數字與 0 距離多遠。
- 若很遠，則拒絕 (虛無) 假設 $\beta = 0$，並說這個係數是顯著的，也就是該變數是有效果的。
- 那要多遠呢？我們必須設定一個顯著水準 (significance level)。
- 若 p-value 比顯著水準小，則拒絕虛無假設。


$(1-\alpha)$
$\alpha/2$
$\alpha/2$
$-t_{\alpha/2}$
0
$t_{\alpha/2}$
t

Dummy Variable

- 有時候有些變數可能不是連續的數值，而是代表該樣本的”類別”，例如地區。
- 我們通常設定 Dummy Variable(虛擬變數) 來代表類別。所謂的 Dummy Variable 是一個結果只有 1 與 0 的隨機變數，用來表示是或不是。
- 若有 J 個類別，則只需放 J-1 個虛擬變數進入迴歸式中。因為對每個樣本來說，代表某種類的虛擬變數的值相加必為 1。(完全共線性)
- 迴歸式中的係數指的是該類別與被忽略的類別之間的差距。
- Ex: 我們沒將代表士林區的 Dummy Variable 放入迴歸式中，則各個地區的 Dummy Variable 估出來的係數代表該地區的租金比起士林區平均來說高多少。

Interaction Term

- 有時候我們可能會懷疑兩個變數之間有交互效果，因此我們會使用交乘項 (Interaction Term)。
- Example: 我們可能懷疑教育對男女未來報酬的影響不一。
  Y ：薪水
  D ：性別; $D_i = 1$ 假如該樣本是男生。
  X ：受教育年數 (連續)

$$Y_i = \beta_1 + \beta_2 D_i + \beta_3 (D_i \times X_i) + \beta_4 X_i + u_i$$

- 對於女生來說，多受一年教育平均來說可以增加 β_4 的收入
- 對於男生來說，多受一年教育平均來說可以增加 $\beta_3 + \beta_4$ 的收入

A Little More

- 此外，也可以加入自變數的平方項，三次方項.....Ex: 薪水 v.s. 年齡
- 我們所說的”線性”迴歸所指的是 Specification 要是係數的線性組合。
- 更多假設?(變異數齊一)
- 如何設定一個好的 Specification?

Outline

1. OLS
   - 線性迴歸
   - T-test
   - Dummy Variable
   - Interaction Term

2. Lasso
   - Why Lasso?
   - Penalization

Why Lasso?

- 我們可能有時會蒐集到很多變數可以當自變數，此外這些自變數的次方項也可以當自變數。
- 如果把所有變數都放到迴歸式中，可能會導致 Overfitting。
- 我們最好先找出解釋能力較高的變數 (降維)。

Penalization

$$\hat{\beta}(\lambda) = \textit{argmin}\{||y - X\beta||^2 + \lambda \times \sum_{j=2}^{k} |\beta_j|\} : \beta \in \mathbb{R}^k$$

- $\lambda > 0$
- 我們在原本要極小化的式子後面，再加上一個懲罰項。
- 對於解釋能力較差的變數，其係數應該等於 0，才能使整個式子極小。
- 我們並不對截距項做懲罰。

- 並沒有 close form solution (not differentiable)，需要先給定 λ 才可以做估計。
- 如何決定 λ? cross validation 或理論值。
- Luckily, there are some packages of Lasso in R.