Logistic Regression

Jong Yih Kuo

jykuo@ntut.edu.tw
Department of Computer Science and Information Engineering
National Taipei University of Technology

數量分析

- 透過數理模型描述觀察結果
  - 觀察現象=模型+誤差；$y = f(x) + error$；觀察值 = 訊號 + 雜訊。
- 量化模型的關鍵
  - 量化目標值 y ：定義問題
  - 選取關鍵變數：$x_1, x_2, \ldots, x_p$
  - 建立量化模型：統計學習、機器學習。
- 資料分析策略:「 觀察 」、「 推論 」、「 驗證 」三步驟
  - 檢查資料品質，避免 Garbage in, garbage out 。
  - 進行探索性資料分析(Exploratory Data Analysis, EDA) 找出關鍵變數(或特性/特徵)。
  - 驗證性資料分析 (Confirmatory Data Analysis, CDA)

資料分析類型

- 統計觀點
  - 探索性資料分析 (Exploratory Data)
  - 驗證性資料分析 (Confirmatory Data)
- 機器學習觀點
  - 敘述性分析(Descriptive): what's happen in my business?
  - 診斷性分析(Diagnostic): why is it happening?
  - 預測性分析(Predictive): what's likely to happen?
  - 建議性分析(Prescriptive): what do I need to do?

資料分析方法

- 分類 (Classification) 與群聚分析(Cluster Analysis)
- 羅吉士迴歸 Logistic Regression
- 分類/回歸/決策樹 Classification and Regression Tree CART
- 類神經網絡 Neural Networks NN
- 支持向量機 Support Vector Machine SVM
- 無母數迴歸 Nonparametric Regression
- 時間序列 Times Series

Regression Analysis

- 許多重要研究主題，依變數是「有限的」，數據資料不是連續的或呈常態分佈。
  - 投票意向，使用二元 Logistic Regression，
    - 是一種 Regression分析。
    - 依變數是虛擬變數：未投票(編碼 0)、或投票(編碼 1)。
  - 發病率：未發病(編碼 0) 、發病(編碼 1)。

- 線性迴歸中：$Y=\beta_0+\beta_1 X+\varepsilon$; 其中 $Y=(0, 1)$
  - 若存在問題
    - 殘差/誤差項，是異質變異數(Heteroscedasticity)，亦即，不一致。
    - e 不服從常態分佈，因為 Y 只有兩個值(0 或 1)。
    - 預測機率可能出現大於 1 或小於 0。

異質變異數(Heteroscedasticity)

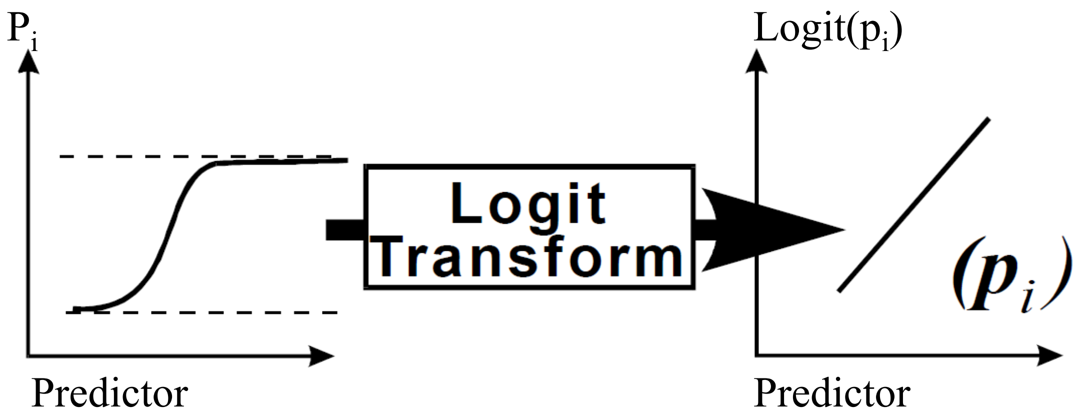
- OLS (Ordinary Least Squares regression) 問題
  - 迴歸分析中，最常用估計β(迴歸係數)的方法是普通最小平方法(Ordinary Least Squares, OLS)，它基於誤差值計算。
  - 用這種方法估計β，首先要計算殘差平方和(Residual Sum of Square, RSS)，RSS是指將所有誤差值的平方加起來。
  - 進行迴歸模型時，為進行有效統計推論，需對模型做若干假設。其中一個: 誤差項具有同質 (homoskedasticity) 變異數。
    - 同質變異數表示給定不同解釋變數(自變數)之值，此時誤差項或被解釋變數(依變數)有相同變異程度。
      - 例如，若被依變數為薪資，自變數為不同教育程度，則同質變異數表示在不同教育程度，薪資變異程度相同，這個假設明顯與實際不符。
      - 實際資料顯示，教育程度越高，薪資變異程度越高。合理解釋，不同學歷的最低薪資相同，但獲高薪機將隨學歷越高而增加。
    - 薪資—教育程度模型中，異質變異數問題幾乎必定存在且需解決。

異質變異數(Heteroscedasticity)

- 異質變異數對迴歸模型主要影響
  - 1. OLS 估計式(方程式模型)失去有效性
    - 由於係數有效性影響較小，只要具有不偏性與一致性即可，因此只要樣本數夠大通常不會考慮估計式的有效性問題。
  - 2. 迴歸係數的 t 檢定失效
    - 此問題影響小樣本的假設檢定，可使用 Heteroskedasticity(HC) 標準誤解決。只要誤差項的變異數隨解釋變數值增加而增加，則 HC 標準誤會比同質變異數下的標準誤還大，使得 t 檢定值較小，造成迴歸係數不顯著。
    - 2 個可能解決顯著性方案
      - 用較大資料集合：顯著性會隨樣本數增加，只要樣本數夠大，雖 HC 標準誤會使 t 檢定統計量變小，但檢定統計量通常可大於檢定臨界值。
      - 將依變數取自然對數：迴歸係數的標準誤會與殘差 (residual) 平方和成正比，而殘差表示被解釋變數偏離迴歸線的程度。若依變數為指數成長，則與迴歸線間的距離可能很大。對依變數取自然對數，指數成長將為線性，偏離迴歸線情況會改善。

Concept

- 對數機率模型(Logit model)
  - 屬於多變量分析，是社會學、生物統計學、計量經濟學、等統計實證分析常用方法。
  - 透過事件的對數發生率(Log-odd)，建構一個或多個自變數的線性組合，對事件發生的機率(依變數)進行建模。
    - 將對數發生率轉換為機率的函數，即Logistic Function。對數發生率單位稱為logit (logistic unit)。



logit (pi) = log (odds) =$\alpha + \beta x$
logit(pi): logit transformation of the probability of the event
α: intercept of the regression line
β: slope of the regression line

Concept

- 對數機率模型(Logit model)
  - 變數
    - 自變數 X 可以是類別變數，或是連續變數。
    - 依變數 Y 主要為類別變數，特別是分兩類的變數，例如：是或否、有或無、同意或不同意、成功或失敗等。
  - 根據輸入對輸出機率進行建模，不進行統計分類變數與分佈。
    - Logistic分佈中，自變數對依變數的影響以指數方式變動，不需常態分佈的假設。

Concept

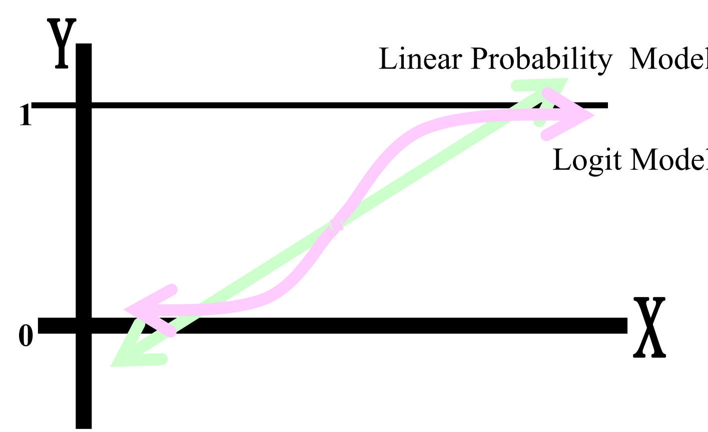
- 二元Logistic Regression
  - 自變數每個都可以是二元值(兩個類)，或連續值。
  - 依變數是二元，編碼值標記為「0」和「1」。
    - 「1」的值的相應機率可以在0和1之間變化；
  - 統計學中廣泛用於對某一類別或事件發生機率的建模，例如團隊獲勝機率、患者健康機率等，
  - 當存在兩個以上可能值(例如圖像是否貓、狗、獅子等)時，二元變數擴充為多分類變數。
    - 二元Logistic Regression擴充為多項Logistic Regression 。
    - 若多個類別是有序的，則可使用序數Logistic Regression 。

Logistic Regression Model

- Logit 可解決上述問題

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta x + e$$

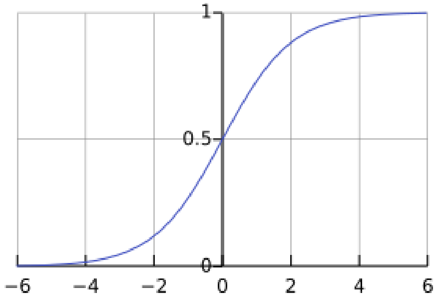
  - p 是事件 Y發生的機率，p(Y=1)
  - p/(1-p) 是勝算比"odds ratio" ，每增加一個單位對整體Y增加/減少的機率
  - ln[p/(1-p)] 是log odds ratio, "logit"
  - Logistic distribution限制評估機率在0~1之間。
  - 評估機率 $p = \dfrac{1}{1 + e^{-(\alpha+\beta x)}}$
  - 若$\alpha + \beta$X =0， 估則p = .50
    - 當$\alpha + \beta$X很大，p趨近1。
    - 當$\alpha + \beta$X很小， p趨近0。



Logistic Regression Model

- Logistic Curve (*Sigmoid function*)

$$f(x) = \frac{exp(x)}{1+exp(x)} = \frac{1}{exp(-x)+1} = \frac{1}{1+e^{-x}}$$



0.5為分界門檻值(Threshold)

- Logistic function
  - 10 場牌局贏 4 場，贏的機率(4/6)，
  - 若贏機率p，輸機率$1\text{-}p$，贏的**勝算比(勝率**Odds ratio) = $\left(\frac{p}{1-p}\right)$
  - **勝算比(勝率**Odds ratio)，取自然對數

$$Logit = \ln\left(\frac{p}{1-p}\right)$$

  - 若 $$p(x) = \frac{exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} = \frac{1}{1+e^{-(\alpha+\beta x)}}$$
  - 則 $$Logit(p(x)) = \ln\left(\frac{p(x)}{1-p(x)}\right) = \alpha+\beta x \ \ or \ \ \beta_0+\beta_1 x_1+\beta_1 x_1+\cdots+\beta_k x_k$$

Maximum Likelihood Estimation

- 最大概似估計maximum likelihood estimation (MLE)
  - 又稱「極大概似估計」，估計一個機率模型的母數(parameters)的方法。
- MLE找出coefficients (α, β)
  - 使Likelihood function對數 (LL < 0) 盡可能大，
  - 或，找出Likelihood function對數的 -2 倍（-2LL）盡可能小。
  - 解下列問題
    - $\{Y - p(Y=1)\}X_i = 0$，對所有觀測值加總，$i = 1,\ldots, n$。

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \alpha + \beta x \text{ ，} \quad p(x) = \frac{1}{1+e^{-(\alpha+\beta x)}}$$

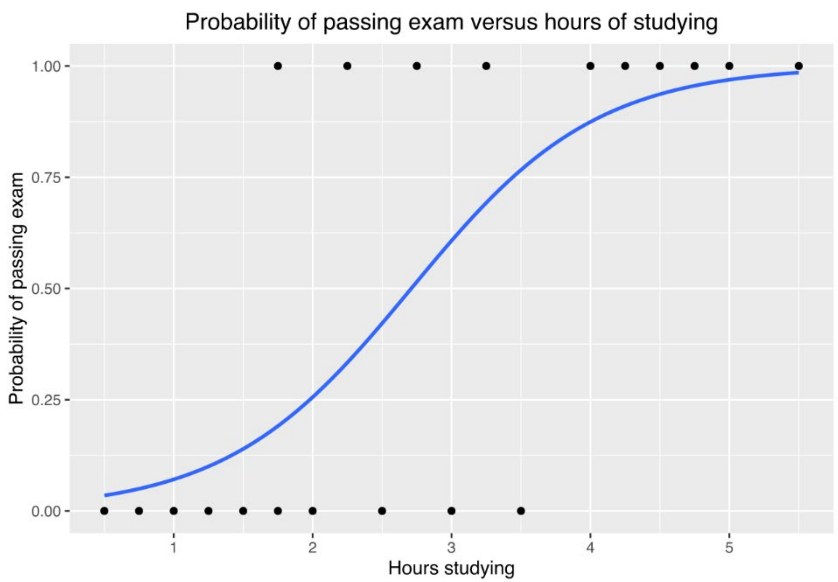
Odds ratio ==> $\frac{p(x)}{1-p(x)} = e^{(\alpha+\beta x)}$

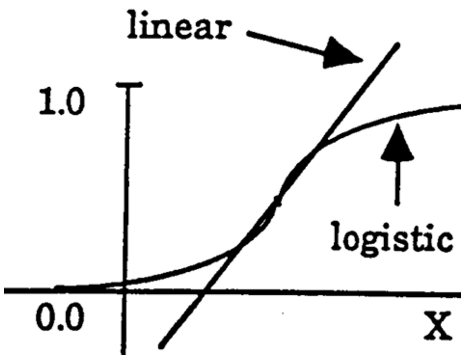
$e^{(\beta)}$ is the effect of the independent variable on the "odds ratio"

Model

- 某班20名學生，各自花費0~6小時準備考試，不同學習時數與通過考試的資料(1-pass/0-fail)。
  - 將對數發生率轉換為機率的函數。

小時（x_k）	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
通過（y_k）	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1





檢定

- 假設檢定
  - 模型的顯著性檢定(F test)：探討模型中的β係數是否全部為0。當係數不全為0時，迴歸模型才具有預測力。
    - 虛無假說(Null Hypothesis)→ H0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
    - 對立假說(Alternative Hypothesis)→ H1: $\beta_1, \beta_2, \ldots, \beta_k$ *不全為 0*
  - 統計值(Statistics)→ $F = \frac{MSR}{MSE}$

Example 1

- 實驗設計

	事件成功	事件失敗	總和
實驗組	4	16	20
對照組	1	19	20

  - 實驗組的勝算(Experimental event odds) = 4/16 = 0.25
  - 控制組的勝算(Control event odds) = 1/19=0.053
  - 勝算比(odds ratio) = (0.25) / (0.053) = 4.72
- Logistic function $Logit(p(x)) = Logit(odds)$

$$Logit(p(x)) = \ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_1 x_1 + \cdots + \beta_k x_k$$

  - 係數值β，計算當 x 的值增加一單位，勝算的改變量(Δ odds)。
    - Δ odds > 1，表示當x_i 增加時，事件Y發生的勝算會提高
    - Δ odds < 1，表示當x_i 增加時，事件Y 發生的勝算會降低
    - Δ odds 又稱**OR值**，以*exp(β)*表示

Example 2-1

- 某公司根據過去「溫度」與「零件測試成功與否」的資料，建立以 溫度預測零件測試成功機率之迴歸模式。
  - 以連續型自變數X(溫度)，預測Y(零件測試成功與否)。
  - 二元迴歸，Y=1表示零件測試成功，0表示零件測試失敗。
  - 模式係數Omnibus測試
    - 相當於線性迴歸ANOVA-F檢定，探討模型的β係數是否全部為0。
    - 本例顯著性p值<0.001，拒絕虛無假說。模型顯著，具有預測能力。
  - 模式摘要：呈現解釋力的值為參考
    - -2對數概似: 16.292 (參數估計值變化<0.01，工作疊代數約7停止)
    - Cox-Snell R平方: .570
    - Nagelkerke R 平方: .760

Example 2-2

- 分類表：呈現預測值的準確度。

觀察次數			預測次數 零件測試成功與否 0	1	百分比修正
步驟 1	零件測試成功與否	0	13	2	86.7
		1	0	15	100.0
	概要百分比				93.3

a. 分割值為 .500

- 方程式變數：Δ odds(OR值)

		B 之估計值	S.E,	Wals	df	顯著性	Exp(B)
步驟 1[a]	溫度	.468	.161	8.432	1	.004	1.597
	常數	-30.295	10.604	8.162	1	.004	.000

- 根據上表得出迴歸式：
  - exp(β)=1.597，即Δ odds=1.597>1，表示溫度每上升一度，零件測試成功機率會比零件測試失敗機率多出1.597倍。

Example 3-1

- 某醫療單位欲根據過去肺部疾病就診病患基本資料，建立以有無「吸菸」、有無「家族病史」預測「罹患肺癌」機率之迴歸模式。
  - 二元迴歸，y=1表示罹患肺癌，y=0表示沒有罹患肺癌。
  - 以兩個類別型的自變數(吸菸、家族病史)預測y(罹患肺癌與否)。
  - 模式係數Omnibus測試
    - 相當於線性迴歸ANOVA-F檢定，探討模型的β係數是否全部為0。
    - 本例顯著性p值<0.001，拒絕虛無假說。模型顯著，具有預測能力。

		卡方	df	顯著性
步驟 1	步驟	22.707	2	.000
	區塊	22.707	2	.000
	模式	22.707	2	.000

Example 3-2

- 模式摘要：呈現解釋力的值為參考
  - 參數估計值變化<0.01，工作疊代數約6停止。

步驟	-2 對數概似	Cox & Snell R 平方	Nagelkerke R 平方
1	39.980[a]	.365	.511

- 分類表：呈現預測值的準確度。

觀察次數			預測次數		
			罹患肺癌		百分比修正
			0	1	
步驟 1	罹患肺癌	0	24	10	70.6
		1	1	15	93.8
	概要百分比				78.0

a. 分割值為 .500

- 方程式中的變數：可Δ odds(OR值)。

		B 之估計值	S.E.	Wals	df	顯著性	Exp(B)	EXP(B) 的 95% 信賴區間 下界	EXP(B) 的 95% 信賴區間 上界
步驟 1[a]	吸菸(1)	-3.487	1.127	9.571	1	.002	.031	.003	.279
	家族病史(1)	-3.538	1.250	8.012	1	.005	.029	.003	.337
	常數	3.800	1.334	8.112	1	.004	44.714		

Example 3-3

- 根據上表得知
  - 吸菸exp(β) =0.031，即Δ odds=0.031，表示沒有吸菸的人(=0)罹患肺癌的機率是有吸菸的人(=1)罹患肺癌機率的0.031倍。
  - 家族病史Exp(B)=0.029，即Δ odds=0.029，表示沒有家族病史的人(=0)罹患肺癌的機率是有家族病史的人(=1)的0.029倍。
  - 由於上述兩個變數皆達顯著(p<.05)，故可推論此筆病患資料「罹患肺癌與否」與「吸菸」及「有無家族病史」有直接相關。

比較

- 線性迴歸
  - 透過一組制定自變數，預測連續的因變數。連續變數可具有一定範圍值，例如價格或年齡。可預測因變數的實際值，「10年後的米價多少？」等問題
- 二進制邏輯迴歸
  - 適用於只有兩個可能結果的二進制分類問題。因變數只能有兩個值，例如 yes 和 no 或 0 和 1。
  - 若邏輯函數計算 0 ~1範圍，四捨五入至最接近值。小於 0.5為0，大於 0.5為 1，以便邏輯函數傳回二進制結果。
- 多項邏輯迴歸
  - 分析幾種可能結果的問題，例如根據人口數據，預測房價是否會增加 25%、50%、75% 或 100%，但無法預測房屋確切價值。
  - 將結果值映射至 0~1的不同值，例如 0.1、0.11、0.12 等，因此，多項迴歸會將數組輸出至最接近的可能值。

Example 4-1

成績好壞	聰明 (x = 1)	笨 (x = 0)
好 (y = 1)	5	1
壞 (y = 0)	1	3
小計	6	4

- 「聰明的人成績好」對「聰明的人成績不好」勝算為 $\frac{5/6}{1/6} = 5$
- 「笨的人成績好」對「笨的人成績不好」勝算為 $\frac{1/4}{3/4} = 0.3333$
- 笨 (x = 0)
  - 「笨的人成績好」對「笨的人成績不好」勝算為(1/4)/(3/4)= 0.333

    $log(0.3333) = a =$ -1.098712
- 聰明(x=1)
  - 「聰明的人成績好」 對「聰明的人成績不好」勝算為(5/6)/(1/6)=5

    $log(5) = -1.0987 + \beta = 1.6094$ ，$\beta = 1.6094 + 1.0987 = 2.7081$

    $\boxed{\text{Logit}(p(x)) = -1.0987 + 2.7081x}$