Logistic Regression

王彥雯

13 July 2015

表格的分析與解釋

The New York Times | http://nyti.ms/19l8bbY

The Opinion Pages | OP-ED CONTRIBUTOR

My Medical Choice

By ANGELINA JOLIE MAY 14, 2013

LOS ANGELES

MY MOTHER fought cancer for almost a decade and died at 56. She held out long enough to meet the first of her grandchildren and to hold them in her arms. But my other children will never have the chance to know her and experience how loving and gracious she was.

We often speak of "Mommy's mommy," and I find myself trying to explain the illness that took her away from us. They have asked if the same could happen to me. I have always told them not to worry, but the truth is I carry a "faulty" gene, BRCA1, which sharply increases my risk of developing breast cancer and ovarian cancer.

My doctors estimated that I had an 87 percent risk of breast cancer and a 50 percent risk of ovarian cancer, although the risk is different in the case of each woman.

Only a fraction of breast cancers result from an inherited gene mutation. Those with a defect in BRCA1 have a 65 percent risk of getting it, on average.

Once I knew that this was my reality, I decided to be proactive and to minimize the risk as much I could. I made a decision to have a preventive double mastectomy. I started with the breasts, as my risk of breast cancer is higher than my risk of ovarian cancer, and the surgery is more complex.

列聯表
Contingency Table

	有BRCA1	沒有 BRCA1	合計
得到乳癌			
沒得到乳癌			
合計			

- 如何分析？
  - 相關性 (correlation) ?
  - 檢定 (hypothesis test)?

勝算比

Odds Ratio (OR)

勝算
Odds

- P/(1-P)
  - P: 某種事件發生的機率
- 例子
  - 賭博輸贏的勝算是10 -> Pr(賭博贏) = 1/(1+10)
  - 抽菸得肺癌的勝算是5 -> Pr(抽菸得肺癌) = 5/(5+1)
  - 已知丟兩個公正骰子出現7點的機率為1/6，請問不出現7點的勝算為？
    - odds = Pr(不出現7點)/Pr(出現7點) = (5/6)/(1/6) = 5

勝算比

	有暴露 (E)	沒暴露 (non-E)	合計
有病 (D)	A	B	A+B
沒病 (non-D)	C	D	C+D
合計	A+C	B+D	

$odds_1 = Pr(D|E)/Pr(non\text{-}D|E) = A/C$

$odds_2 = Pr(D|non\text{-}E)/Pr(non\text{-}D|non\text{-}E) = B/D$

$OR = odds_1/odds_2 = (A/C)/(B/D) = (AD)/(BC)$

如果有其他的變數想一起考慮相關性呢？

例子

- 帶有BRCA1突變的人，是否有比較高的機會得乳癌？

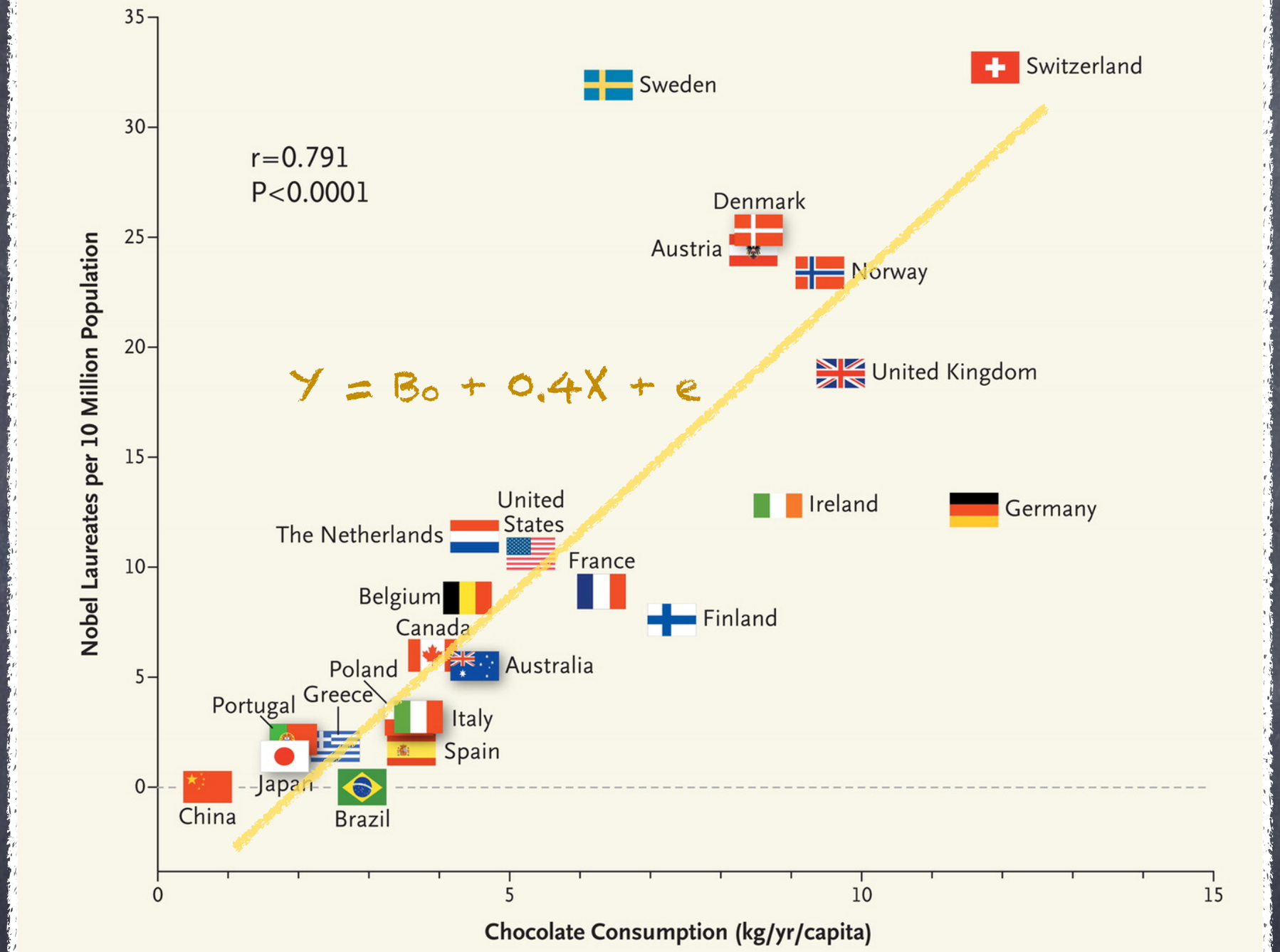
- Y: 有乳癌 vs. 沒有乳癌

- X: BRCA1, 年齡, 抽菸, 喝酒, 家族病史, 生過幾個小孩, 餵母乳, 服用避孕藥, 接受賀爾蒙治療, 體重, 運動習慣, ...

Regression

Review -

Linear Regression



吃越多巧克力的國家有比較多的諾貝爾獎得主？

Messerli, F.H. (2012) The New England Journal of Medicine

- $Y = B_0 + B_1X_1 + B_2X_2 + \dots + B_pX_p + e$
- 估計係數 $(B_0, B_1, \dots, B_p)$
  - 最小平方法: $\min \sum_{i=1}^{n} \left(Y_i - E(Y_i)\right)^2 = \min \sum_{i=1}^{n} \left[Y_i - \left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}\right)\right]^2$
  - 最大概似估計法: $Y_i \sim N\left(\mu, \sigma^2\right),\ \mu = E\left(Y_i\right)$

$$\max L\left(\beta \mid \mathbf{Y}\right) = \max \prod_{i=1}^{n} f\left(Y_i \mid \beta\right)$$

- 變數選擇
- 模式配適值 (goodness of fit)
- 迴歸診斷：殘差診斷、影響點分析、共線性檢查

Logistic Regression

- $Y = \{0,1\}$: yes vs. no
- X: 連續變數（身高、體重、血壓...）、類別變數（性別、教育程度、居住地...）

- $Y \overset{?}{=} B_0 + B_1X_1 + B_2X_2 + \ldots + B_kX_k + e$

Y: $\{0,1\}$

$B_0 + B_1X_1 + B_2X_2 + \ldots + B_kX_k + e$: $(-\infty, \infty)$

- $p=Pr(Y=1) \quad \longleftrightarrow \quad p = E(Y)$
- $odds = p/(1-p)$
- $logit(p) = ln(p/(1-p))$

link function

$$= \underbrace{B_0 + B_1X_1 + B_2X_2 + \ldots + B_kX_k}_{X^TB}$$

$X=(1, X_1, \ldots, X_k) \qquad B=(B_0, B_1, \ldots, B_k)$

- $p = exp(X^TB)/[1+exp(X^TB)]$

有其他的Link function嗎？

- The probit link (or the inverse Normal link): $\Phi^{-1}(P)$
- The complementary log-log link: ln [-ln(1 - P)]
- The log-log link: -ln [-ln(P)]

係數的解釋-1

- $\ln(p/(1-p)) = B_0 + B_1X_1$
- 連續變數
- $odds_A = p/(1-p) = \exp(B_0 + B_1X_1)$
- $odds_B = p/(1-p) = \exp[B_0 + B_1(X_1+1)]$
- $OR = odds_B/odds_A = \exp(B_1)$
- X每增加一個單位，odds增加$\exp(B_1)$倍

係數的解釋-2

- $\ln(p/(1-p)) = B_0 + B_1X_1$
- 離散變數
- $X_1=1 \Rightarrow odd_1 = p/(1-p) = \exp(B_0 + B_1)$
- $X_1=0 \Rightarrow odd_2 = p/(1-p) = \exp(B_0)$
- $OR = odds_1/odds_0 = \exp(B_1)$
- $X_1=1$的odds是$X_1=0$的$\exp(B_1)$倍

係數估計

- 最大概似估計法

$$Y_i \sim Bernoulli(p_i)$$

$$f(y_i) = p_i^{y_i}(1-p_i)^{1-y_i}$$

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} = \prod_{i=1}^{n}\left(\frac{\exp(\mathbf{x}_i^T\beta)}{1+\exp(\mathbf{x}_i^T\beta)}\right)^{y_i}\left(1-\frac{\exp(\mathbf{x}_i^T\beta)}{1+\exp(\mathbf{x}_i^T\beta)}\right)^{1-y_i}$$

$$l(\beta) = \ln L(\beta) = \sum_{i=1}^{n}\left\{y_i \ln\left(\frac{\exp(\mathbf{x}_i^T\beta)}{1+\exp(\mathbf{x}_i^T\beta)}\right) + (1-y_i)\ln\left(1-\frac{\exp(\mathbf{x}_i^T\beta)}{1+\exp(\mathbf{x}_i^T\beta)}\right)\right\}$$

$$\frac{\partial l(\beta)}{\partial \beta} = ?$$

- 牛頓法解參數

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \frac{U^{(k)}}{I^{(k)}}$$

$$U^{(k)} = \left.\frac{\partial l(\beta)}{\partial \beta}\right|_{\beta=\hat{\beta}^{(k)}}$$

$$I^{(k)} = \left.\frac{\partial^2 l(\beta)}{\partial \beta^2}\right|_{\beta=\hat{\beta}^{(k)}}$$

變數選擇

- 為什麼要做？
  - Underfitting
    - Biasedness: 迴歸係數的估計值會是有偏的
  - Overfitting
  - Inefficiency: 迴歸係數估計值的變異數會變大
- 方法
  - All possible subsets
    - AIC = -2l(B) + 2k
  - Sequential variable selection
    - Forward
    - Backward
    - Stepwise

模式配適值

1. χ^2_D or χ^2_P or χ^2_{HW}	p-value > 0.05 (not reject H_0)
2. Concordance pairs (%)	>= 80%
Disconcordance pairs (%)	
Tie pairs (%)	
3. c = AUC	
Somers' D = 2(c-0.5)	
Kendall's Tau	

- Likelihood ratio test (X^2_D)

$$\chi^2_D = -2l(\beta) = 2\sum_{i=1}^{n}\left[y_i \ln\frac{y_i}{\hat{p}_i} + (1-y_i)\ln\frac{1-y_i}{1-\hat{p}_i} \right] \sim \chi^2_{n-(k+1)}$$

- Pearson chi-squared goodness-of-fit test:

$$\chi^2_P = \sum_{i=1}^{n}\frac{(y_i-\hat{p}_i)^2}{\hat{p}_i(1-\hat{p}_i)} \sim \chi^2_{n-(k+1)}$$

- Hosmer-Lemeshow goodness-of-fit test (X^2_{HW})

$$\chi^2_{HW} = \sum_{j=1}^{g}\frac{\left(\sum_i y_{ji} - \sum_i \hat{p}_{ji}\right)^2}{\sum_i \hat{p}_{ji}\left(1-\sum_i \hat{p}_{ji}/n_j\right)} \sim \chi^2_{g-2}$$

迴歸診斷

- 殘差分析 (residual analysis)
- 影響點分析 (influence analysis)
- 共線性檢查 (check for multicollinearity)
- Detection of separation (or high discrimination)

殘差分析

1. The $\Delta\chi^2_D$ vs. $\hat{P}$ plot	Poorly fitted i's?
2. The $\Delta\chi^2_P$ vs. $\hat{P}$ plot	Poorly fitted i's?

$$\Delta\chi^2_D = \frac{d_i^2}{1-h_i},\ d_i = 2\left[y_i \ln\frac{y_i}{\hat{p}_i} + (1-y_i)\ln\frac{1-y_i}{1-\hat{p}_i} \right]$$

$$\Delta\chi^2_P = \frac{r_i^2}{1-h_i},\ r_i^2 = \frac{(y_i-\hat{p}_i)^2}{\hat{p}_i(1-\hat{p}_i)}$$

$$h_i = \hat{p}_i(1-\hat{p}_i)x_i\left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1}x_i^T$$

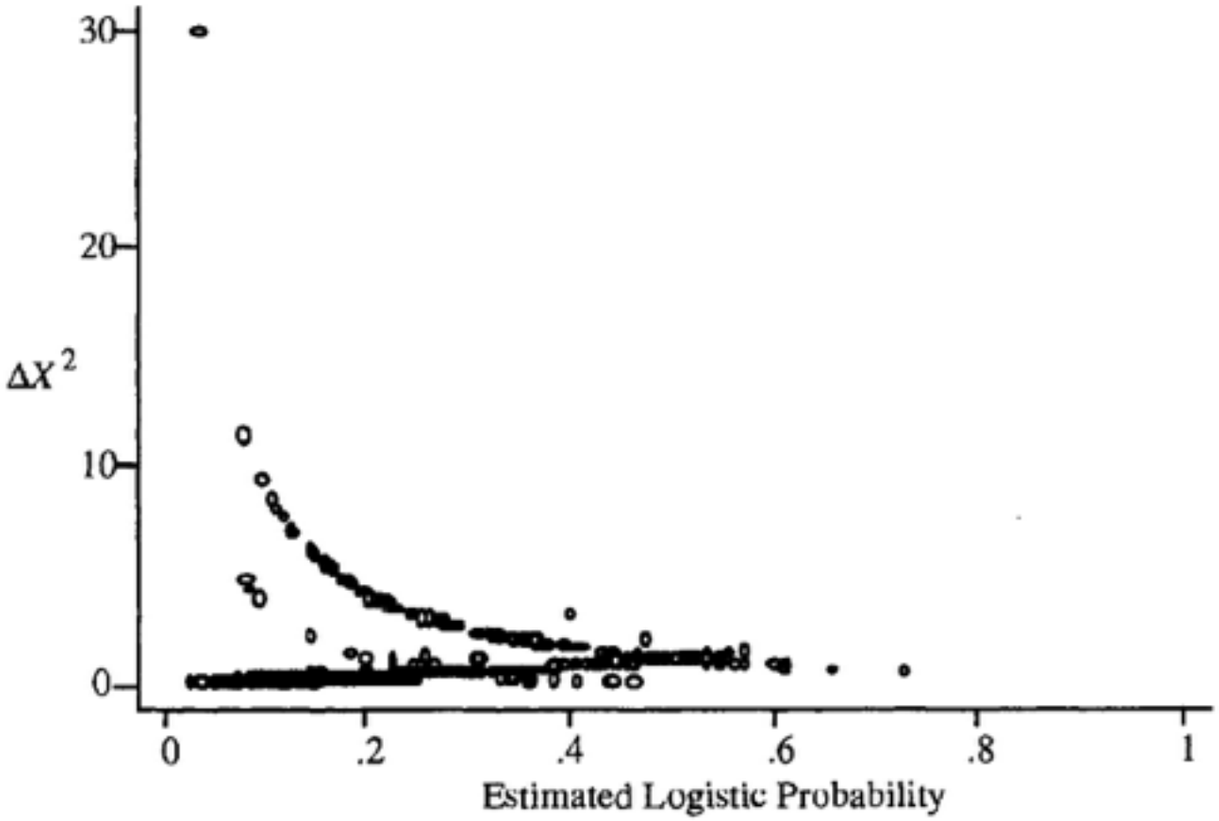


Figure 5.5 Plot of ΔX^2 versus the estimated probability from the fitted model in Table 4.9, UIS $J = 521$ covariate patterns.

Hosmer and Lemeshow (2000).

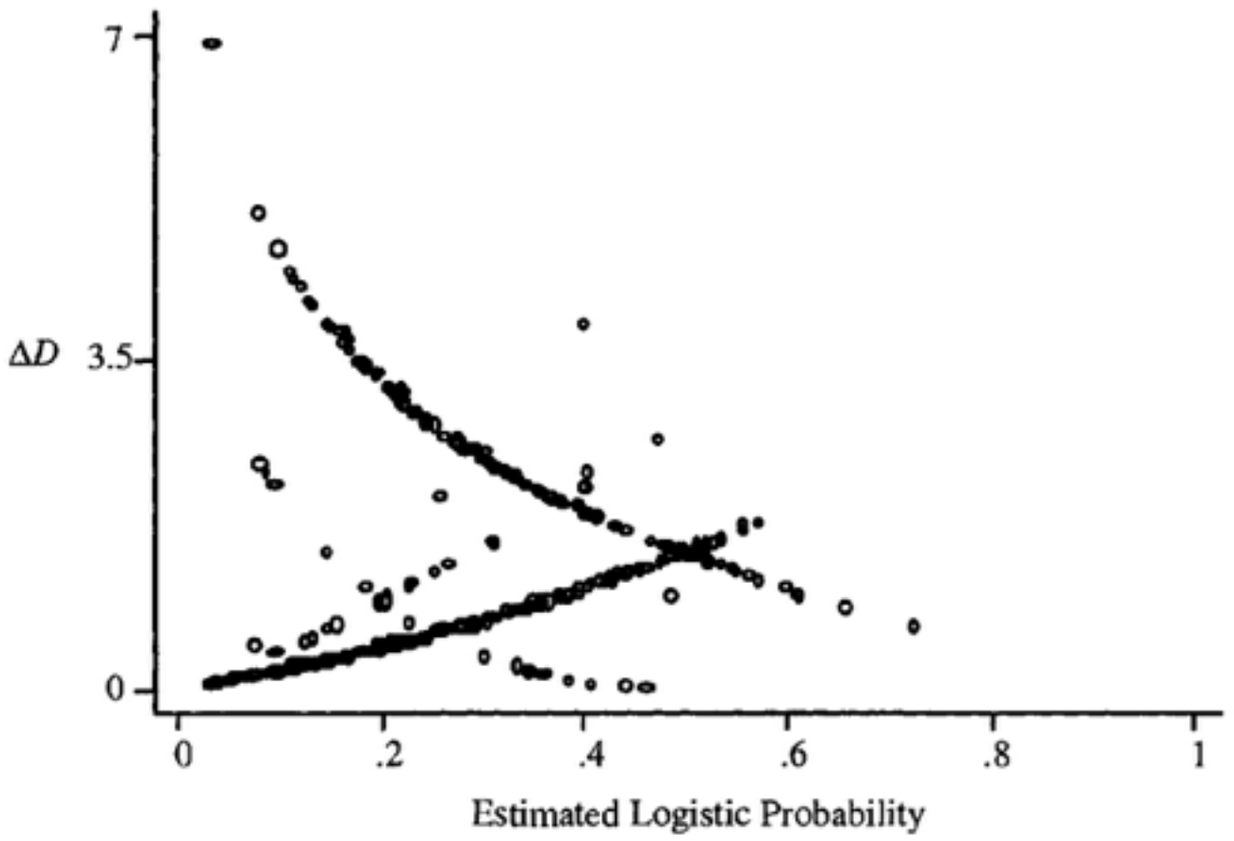
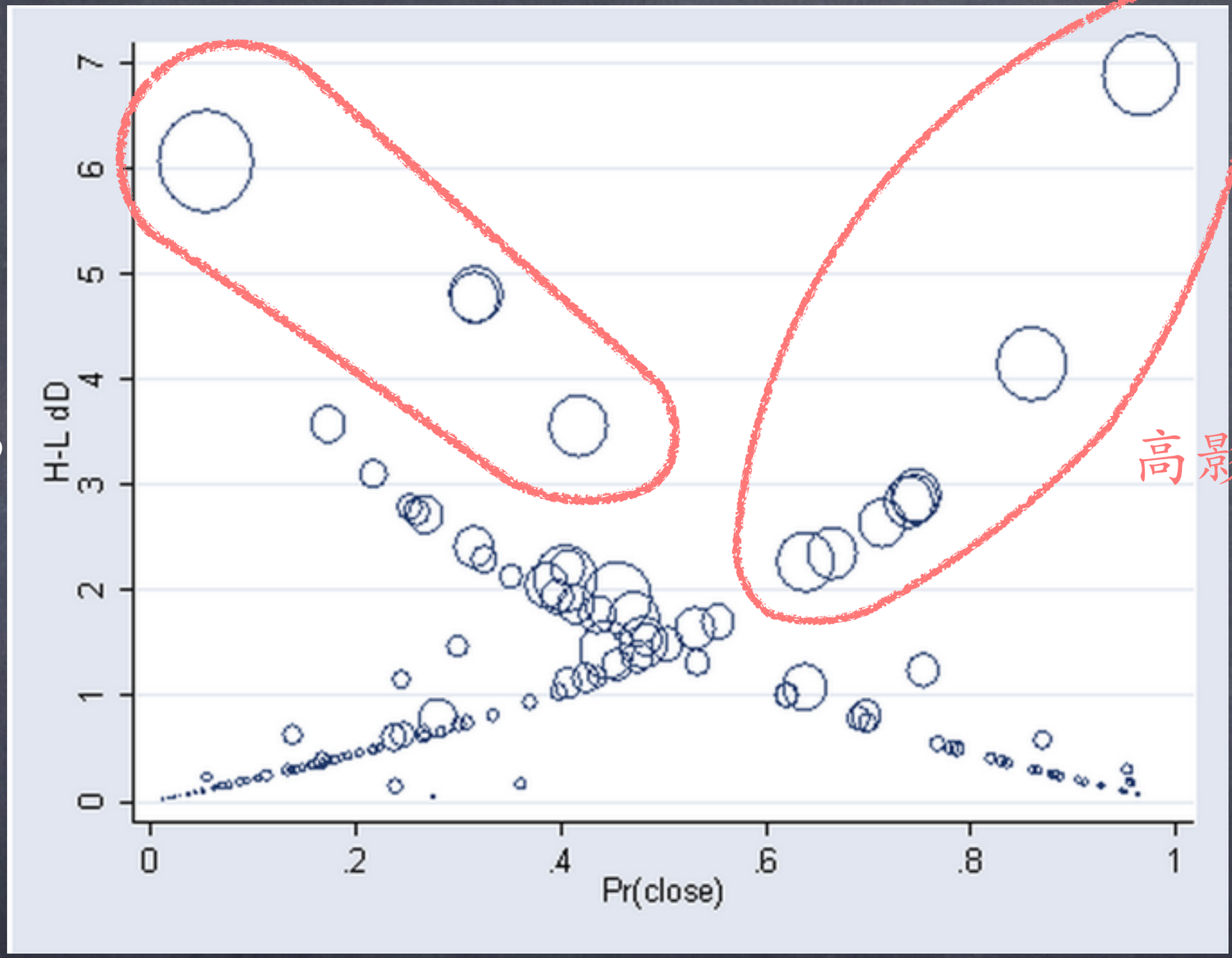


Figure 5.6 Plot of ΔD versus the estimated probability from the fitted model in Table 4.9, UIS $J = 521$ covariate patterns.

影響點分析

1. ΔB_j	Problem: Relatively large
2. $\lvert DFBETAS_{jk} \rvert$	Problem: Relatively large

$$\Delta B_j = \left(\hat{\beta} - \hat{\beta}_{(-j)}\right)^T \mathbf{X}^T \mathbf{V} \mathbf{X} \left(\hat{\beta} - \hat{\beta}_{(-j)}\right) = \frac{r_j^2 h_j}{1 - h_j}$$



$\Delta\chi^2_D$

$\hat{P}$

高影響力的點

圓形的面積代表 ΔB_j

Figure 7.10 in Hamilton (1992)

共線性檢查

1. 檢查X的相關係數矩陣 (correlation matrix of X)	Problem: >= 0.9 (or 0.8)
2. 檢查迴歸係數的相關係數矩陣 (correlation matrix of $\hat{\beta}$)	Problem: >= 0.9 (or 0.8)
3. R_k^2 for X_k	Problem: >= 0.9 (or 0.8)
or Tolerance$_k$ (= $1-R_k^2$)	Problem: <= 0.1 (or 0.2)
or VIF$_k$ (=1/Tolerancek)	Problem: >= 10 (or 0.5)

Detection of separation

- 探索性資料分析（描述性統計）
- 是否有格子的數字趨近於0
- 也就是某一子分類幾乎沒有樣本

Logistic Regression in R

Example

A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, admit/don't admit, is a binary variable.

http://www.ats.ucla.edu/stat/r/dae/logit.htm

- 建立logistic regression model
  - glm(Y ~ X1 + X2 + ..., data = mydata, family = "binomial")
- 讀報表
  - summary(glm.object)

類別變數如不設 dummy variable 要用 factor()

rank1 = 1 當做比較基準組

```
> # logistic regression
> mydata$rank1 <- factor(mydata$rank)
> res <- glm(admit ~ gre + gpa + rank1, data = mydata, family = "binomial")
> summary(res)

Call:
glm(formula = admit ~ gre + gpa + rank1, family = "binomial",
    data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6268  -0.8662  -0.6388   1.1490   2.0790

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.989979   1.139951  -3.500 0.000465 ***
gre          0.002264   0.001094   2.070 0.038465 *
gpa          0.804038   0.331819   2.423 0.015388 *
rank12      -0.675443   0.316490  -2.134 0.032829 *
rank13      -1.340204   0.345306  -3.881 0.000104 ***
rank14      -1.551464   0.417832  -3.713 0.000205 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 458.52  on 394  degrees of freedom
AIC: 470.52

Number of Fisher Scoring iterations: 4
```

- 估計係數的信賴區間
  - confint(glm.object)
- 計算OR及OR的信賴區間
  - exp(glm.object$coefficients)
  - exp(confint(glm.object))

```
> # CIs using profiled log-likelihood
> confint(res)
Waiting for profiling to be done...
                    2.5 %       97.5 %
(Intercept) -6.2716202334 -1.792547080
gre          0.0001375921  0.004435874
gpa          0.1602959439  1.464142727
rank12      -1.3008888002 -0.056745722
rank13      -2.0276713127 -0.670372346
rank14      -2.4000265384 -0.753542605
>
> # ORs and corresponding CIs
> exp(res$coefficients)
(Intercept)         gre         gpa      rank12      rank13      rank14
  0.0185001   1.0022670   2.2345448   0.5089310   0.2617923   0.2119375
> exp(confint(res))
Waiting for profiling to be done...
                   2.5 %    97.5 %
(Intercept) 0.001889165 0.1665354
gre         1.000137602 1.0044457
gpa         1.173858216 4.3238349
rank12      0.272289674 0.9448343
rank13      0.131641717 0.5115181
rank14      0.090715546 0.4706961
```

- 變數篩選
  - step(glm.object, direction = c("both", "backward", "forward"))

```
> # variable selection
> ## stepwise
> step(res, direction = "both")
Start:  AIC=470.52
admit ~ gre + gpa + rank1

        Df Deviance    AIC
<none>       458.52 470.52
- gre    1   462.88 472.88
- gpa    1   464.53 474.53
- rank1  3   480.34 486.34

Call:  glm(formula = admit ~ gre + gpa + rank1, family = "binomial",
    data = mydata)

Coefficients:
(Intercept)          gre          gpa       rank12       rank13       rank14
  -3.989979     0.002264     0.804038    -0.675443    -1.340204    -1.551464

Degrees of Freedom: 399 Total (i.e. Null);  394 Residual
Null Deviance:       500
Residual Deviance: 458.5  AIC: 470.5
```

- 模式配適值
  - Likelihood ratio test (X^2_D):

    glm.object$deviance

  - Hosmer-Lemeshow goodness-of-fit test (X^2_{HW}):

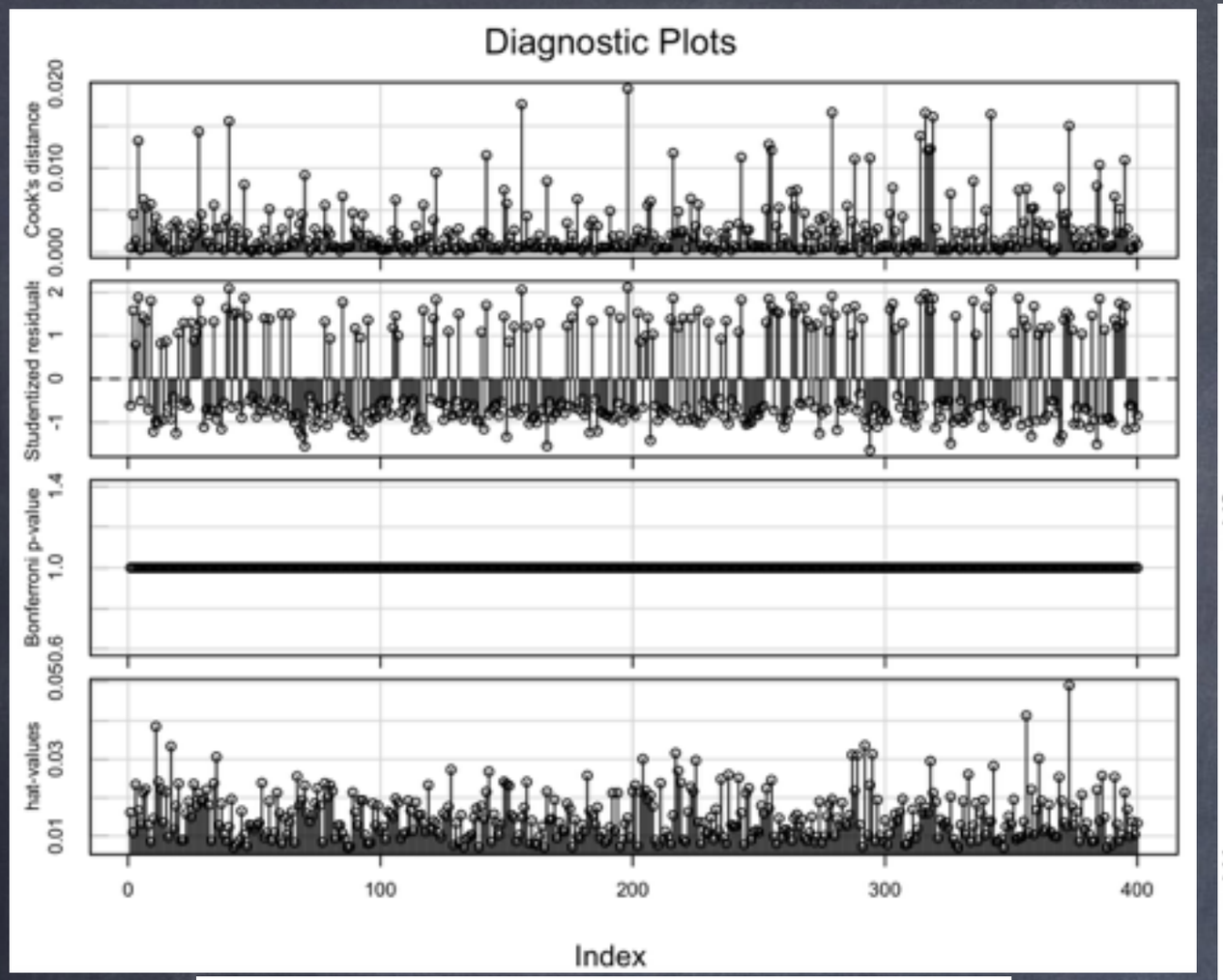
    hoslem.test(Y, fitted(glm.object))

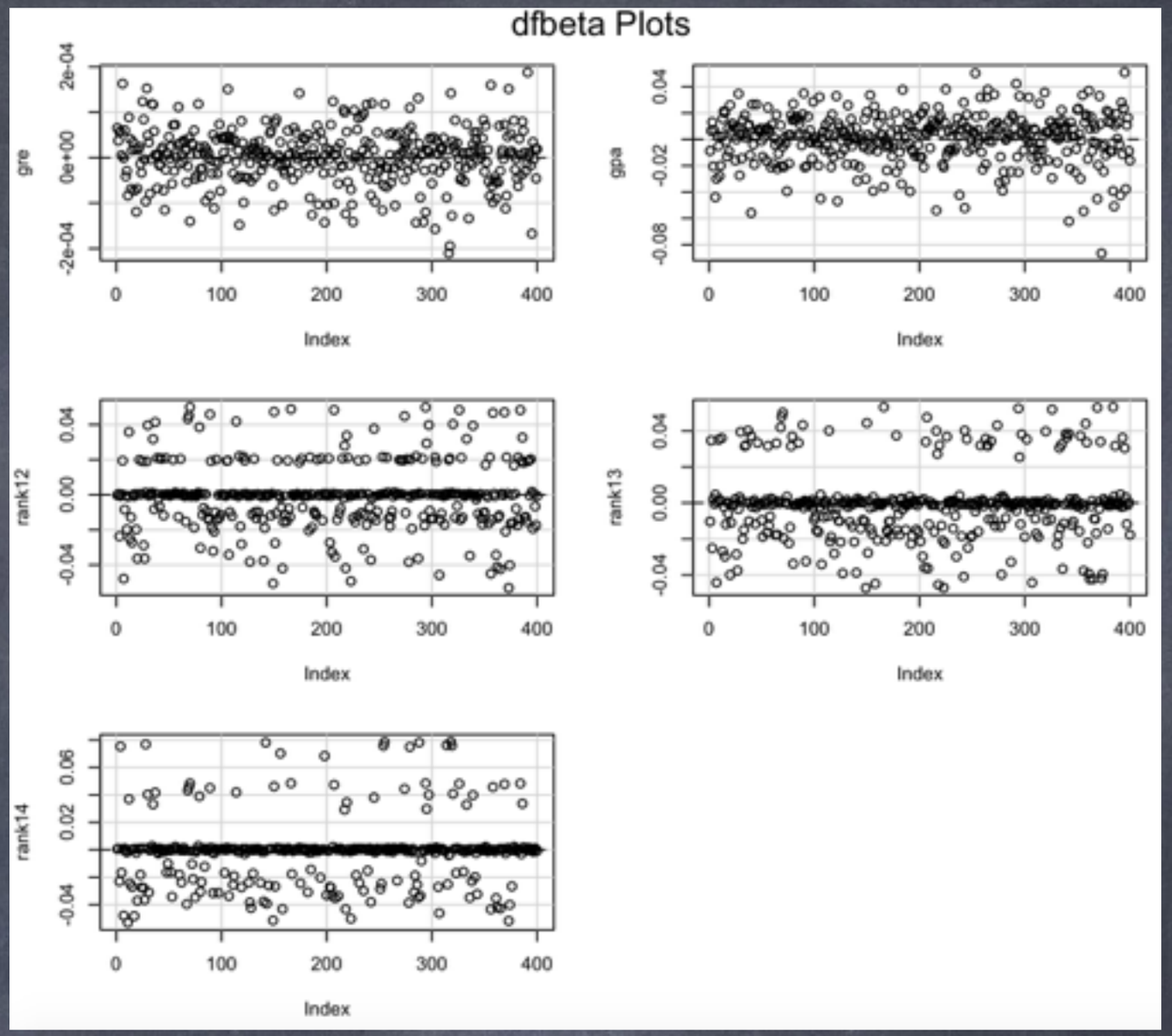
  - 先安裝package: ResourceSelection

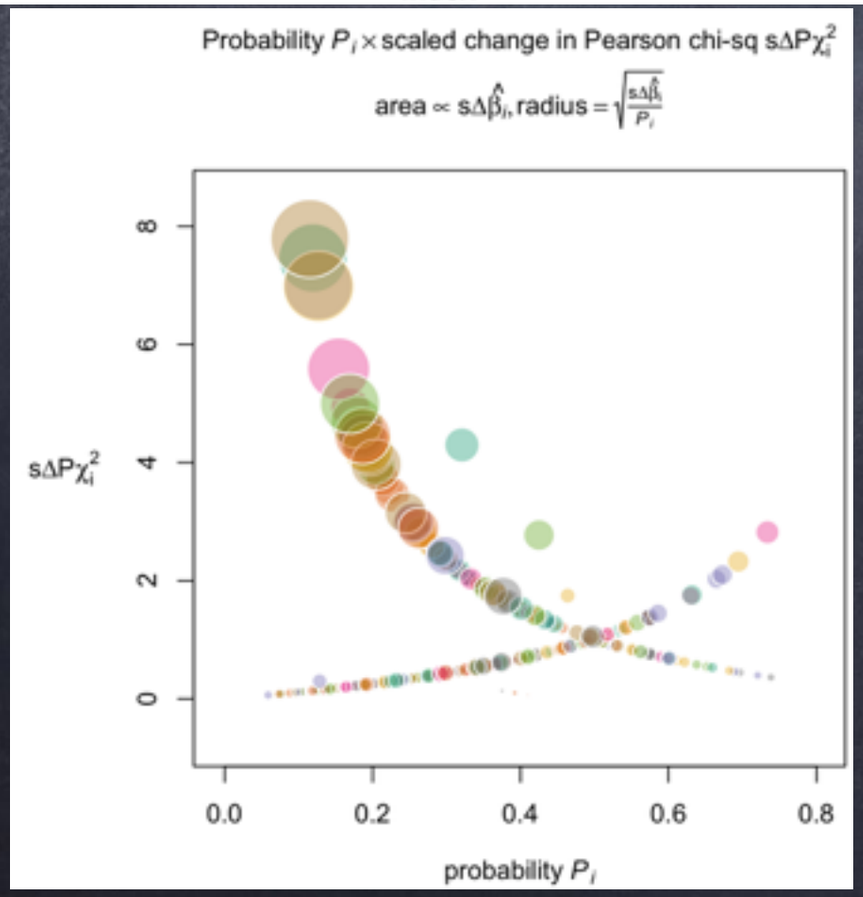
```
> # goodness-of-fit
> ## Likelihood ratio test (X^2_D)
> res$deviance
[1] 458.5175
> pchisq(res$deviance, df = res$df.residual, lower.tail = F)
[1] 0.01365347
>
> ## Hosmer-Lemeshow goodness-of-fit test (X^2_HW)
> install.packages("ResourceSelection")
trying URL 'http://cran.csie.ntu.edu.tw/bin/macosx/mavericks/contrib/3.1/ResourceSelection_0.2-4.tgz'
Content type 'application/x-tgz' length 439444 bytes (429 Kb)
opened URL
==================================================
downloaded 429 Kb


The downloaded binary packages are in
    /var/folders/0d/m3w2dzdd5t9gff0kqw_6twj00000gn/T//RtmpZNeaPG/downloaded_packages
> library(ResourceSelection)
ResourceSelection 0.2-4   2014-05-19
Warning message:
package ‘ResourceSelection’ was built under R version 3.1.2
> hoslem.test(mydata$admit, fitted(res))

    Hosmer and Lemeshow goodness of fit (GOF) test

data:  mydata$admit, fitted(res)
X-squared = 11.0855, df = 8, p-value = 0.1969
```

- 迴歸診斷
  - 先安裝package: car & LogisticDx
  - 共線性檢查
    - cor(X)
    - vif(glm.object)
- 影響點分析
  - influenceIndexPlot(glm.object)
  - dfbetaPlots(glm.object)
  - plot(glm.object)
    - 殘差分析
      - plot(glm.object)

```
> # regresion diagnosis
> install.packages("car")
trying URL 'http://cran.csie.ntu.edu.tw/bin/macosx/mavericks/contrib/3.1/car_2.0-25.tgz'
Content type 'application/x-tgz' length 1386352 bytes (1.3 Mb)
opened URL
==================================================
downloaded 1.3 Mb


The downloaded binary packages are in
    /var/folders/0d/m3w2dzdd5t9gff0kqw_6twj00000gn/T//Rtmpd4cDCL/downloaded_packages
> install.packages("LogisticDx")
trying URL 'http://cran.csie.ntu.edu.tw/bin/macosx/mavericks/contrib/3.1/LogisticDx_0.2.tgz'
Content type 'application/x-tgz' length 888711 bytes (867 Kb)
opened URL
==================================================
downloaded 867 Kb


The downloaded binary packages are in
    /var/folders/0d/m3w2dzdd5t9gff0kqw_6twj00000gn/T//Rtmpd4cDCL/downloaded_packages
> library(car)
> library(LogisticDx)
>
> ## multicollinearity
> cor(mydata[,2:4])
             gre         gpa        rank
gre   1.0000000  0.38426588 -0.12344707
gpa   0.3842659  1.00000000 -0.05746077
rank -0.1234471 -0.05746077  1.00000000
> vif(res)
          GVIF Df GVIF^(1/(2*Df))
gre   1.134377  1        1.065071
gpa   1.155902  1        1.075129
rank1 1.025759  3        1.004248
>
> ## influence analysis
> infIndexPlot(res)
> dfbetaPlots(res)
>
> ## residual analysis
> plot(res, cex.main = 1)
```

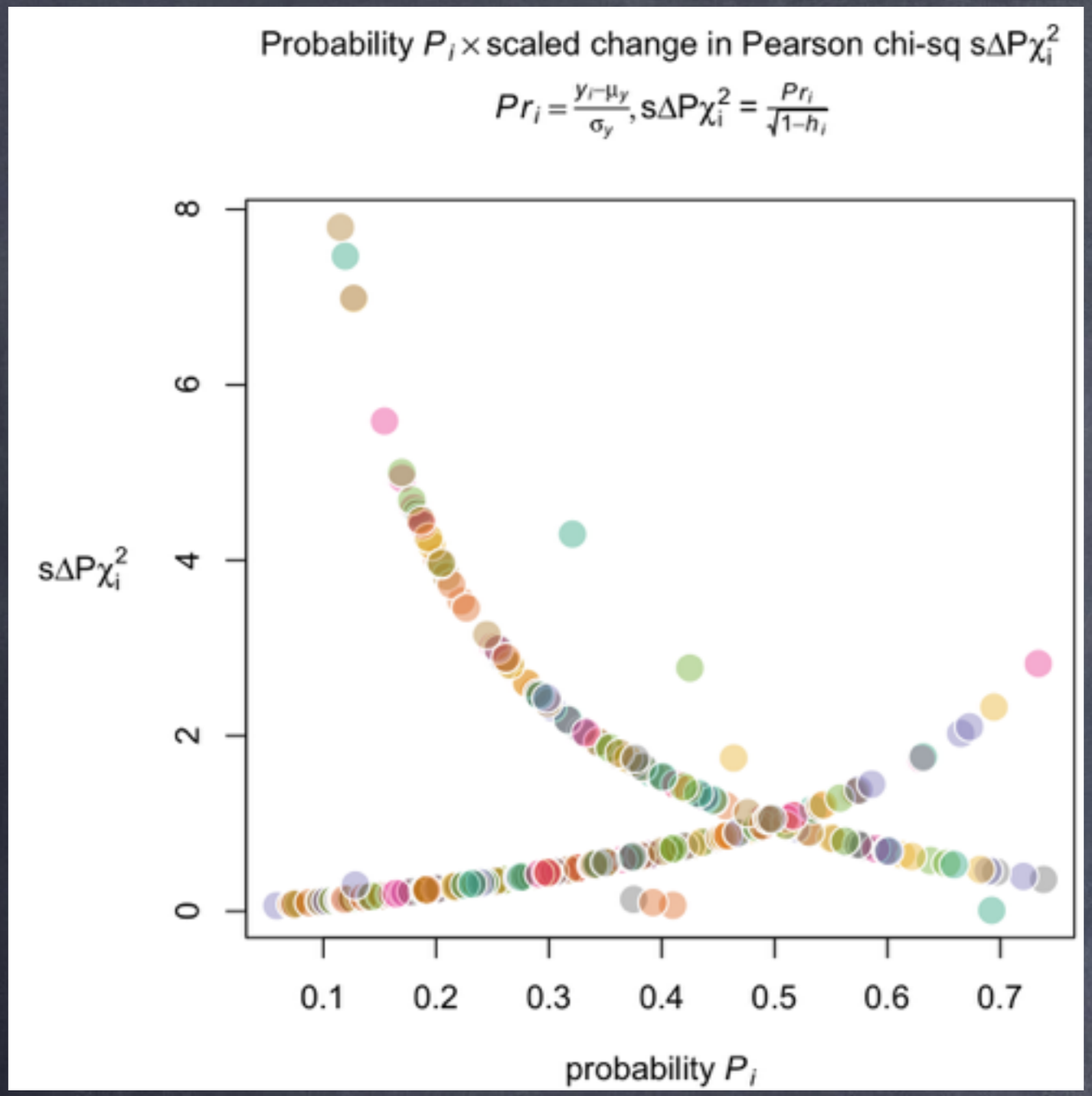

Diagnostic Plots
Cook's distance
Studentized residual
Bonferroni p-value
hat-values
Index

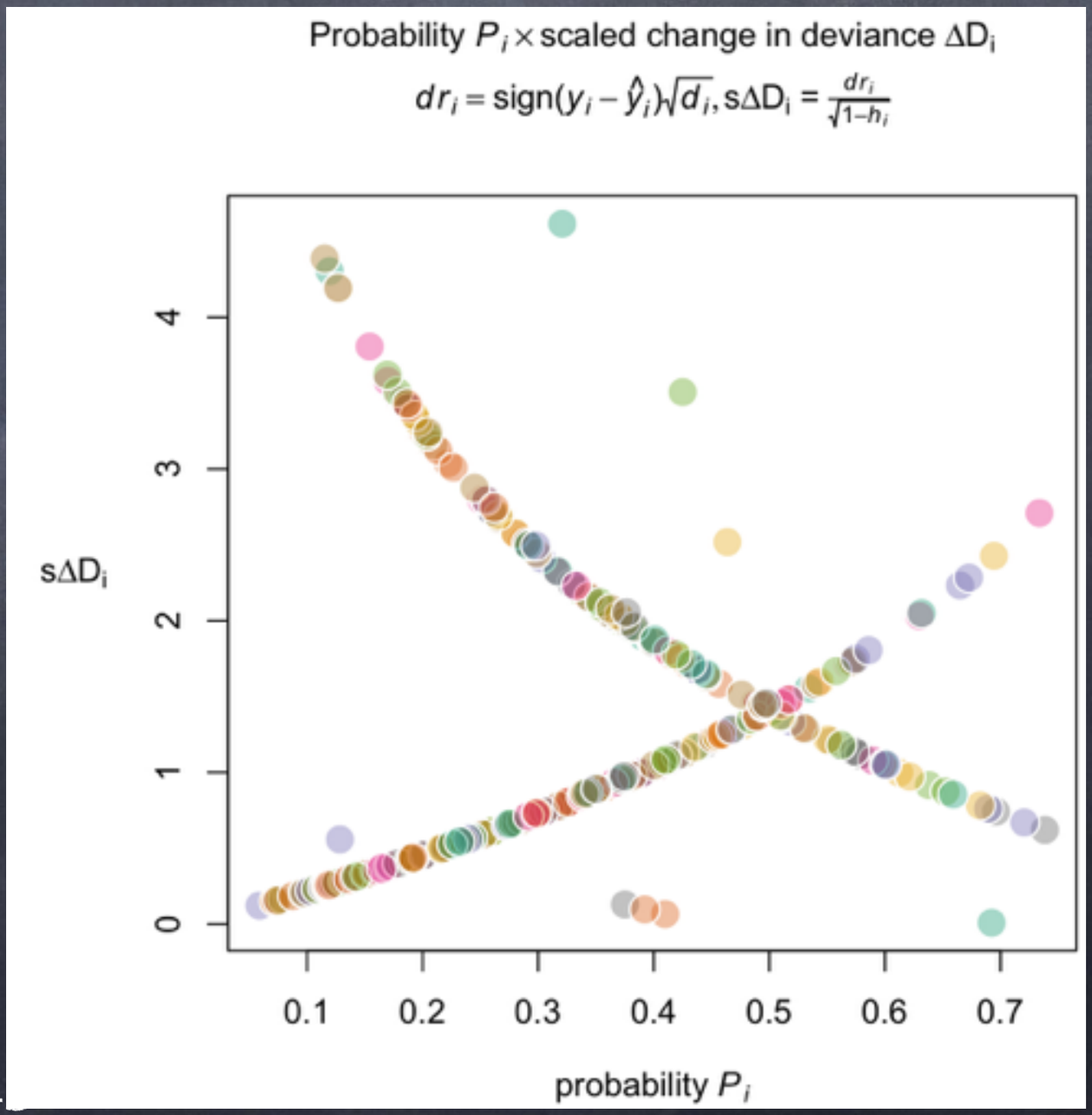

dfbeta Plots
gre
gpa
rank12
rank13
rank14
Index


Probability P_i × scaled change in Pearson chi-sq $s\Delta P\chi_i^2$
area ∝ $s\Delta\hat{\beta}_i$, radius = $\sqrt{\frac{s\Delta\hat{\beta}_i}{P_i}}$
$s\Delta P\chi_i^2$
probability P_i

影響點分析

残差分析


Probability P_i × scaled change in Pearson chi-sq $s\Delta P\chi_i^2$
$Pr_i = \frac{y_i - \mu_y}{\sigma_y}, s\Delta P\chi_i^2 = \frac{Pr_i}{\sqrt{1-h_i}}$
$s\Delta P\chi_i^2$
0
2
4
6
8
0.1
0.2
0.3
0.4
0.5
0.6
0.7
probability P_i


Probability P_i × scaled change in deviance ΔD_i
$dr_i = \text{sign}(y_i - \hat{y}_i)\sqrt{d_i}, s\Delta D_i = \frac{dr_i}{\sqrt{1-h_i}}$
$s\Delta D_i$
0
1
2
3
4
0.1
0.2
0.3
0.4
0.5
0.6
0.7
probability P_i

練習

- 請用健保資料庫進行logistic regression analysis
  - 先決定要問的問題是什麼？
  - Y = ? X = ?
  - 描述性統計：瞭解欲分析的變數的分佈
  - logistic regression: 模式估計、變數選擇、模式配適值、迴歸診斷、解釋結果

References

- Agresti, A. (2002).Categorical Data Analysis, 2nd Ed. Hoboken, NJ: John Wiley & Sons, Inc.
- Dobson, A. J. (2002) An Introduction to Generalized Linear Models, 2nd Ed. Boca Raton, FL: Chapman & Hall/CRC.
- Hamilton, L. C. (1992). Regression with Graphics: A Second Course in Applied Statistics. Belmont, CA: Duxbury Press.
- Hosmer, D. W. and Lemeshow, S. (2000). Applied Logistic Regression, 2nd Ed. New York, NY: John Wiley & Sons.
- McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models, 2nd Ed. London: Chapman & Hall.
- Rosner, B. (2000). Fundamentals of Biostatistics, 5th Ed. Pacific Grove, CA: Duxbury.