BÁO CÁO 1 - NHÓM D

Table of Contents

Thành viên:  
1. Đỗ Thị Thanh Thảo (23C23009)  
2. Nguyễn Kim Anh (23C23004)  
3. Nguyễn Bích Trâm (23C23010)  
4. Trần Thị Thuận (23C23002)

## Bài toán thực tế:

Giả sử chúng ta muốn tính xác suất 1 học sinh hút thuốc (1) hoặc không (0) dựa vào việc xét xem ba/ mẹ của học sinh đó có hút thuốc (1) hay không (0).

# Tạo dữ liệu  
data <- data.frame(  
 `Parent Smoking Status` = c("1–2 parents smoke", "Neither parent smokes"),  
 `Student smokes` = c(816, 188),  
 `Student does not smoke` = c(3203, 1168)  
)  
  
# Xuất bảng  
knitr::kable(data,   
 caption = "Student Smoking vs. Parent Smoking Status",  
 col.names = c("Parent Smoking Status",   
 "Student smokes",   
 "Student does not smoke"))

Student Smoking vs. Parent Smoking Status

| Parent Smoking Status | Student smokes | Student does not smoke |
| --- | --- | --- |
| 1–2 parents smoke | 816 | 3203 |
| Neither parent smokes | 188 | 1168 |

Ta xác định:  
- Biến phản hồi: khi học sinh thứ hút thuốc, và ngược lại, khi học sinh thứ không hút thuốc.  
- Biến giải thích. : Học sinh có phụ huynh (1–2 người) hút thuốc, và ngược lại, : Học sinh có phụ huynh không hút thuốc. Cần chú ý, là biến định danh (dummy variable)  
Như vậy, đối với từng học sinh, ta xét biến , trong đó là xác suất học sinh hút thuốc. # BÁO CÁO 1

## Kiểm tra ~ có thuộc họ phân phối mũ phân tán

Ta có hàm phân phối nhị thức:

$\hspace{2cm} f\_Y(y\_i) = C\_{m\_i}^{y\_i}p\_i^{y\_i}(1-p\_i)^{m\_i-y\_i}\\\\$

Lấy log 2 vế, ta được:

$\hspace{2cm} log[f\_Y(y\_i)] = log(C\_{m\_i}^{y\_i}) + y\_ilog(p\_i) + (m\_i-y\_i)log(1-p\_i)\\\\$

$\hspace{2cm} log[f\_Y(y\_i)] = log(C\_{m\_i}^{y\_i}) + y\_i[log(p\_i)-log(1-p\_i)] + m\_ilog(1-p\_i)\\\\$

$\hspace{2cm} log[f\_Y(y\_i)] = log(C\_{m\_i}^{y\_i}) + y\_ilog\frac{p\_i}{1-p\_i} + m\_ilog(1-p\_i)\\\\$

Đưa về dạmg lũy thừa, ta được:

$\hspace{2cm} f\_Y(y\_i) = exp\{log(C\_{m\_i}^{y\_i}) + y\_ilog\frac{p\_i}{1-p\_i} + m\_ilog(1-p\_i)\}\\$ **(\*)**

Đặt:

$\hspace{3cm}\Leftrightarrow e^{\theta\_i} = \frac{p\_i}{1-p\_i} \\$

$\hspace{3cm}\Leftrightarrow p\_i = (1-p\_i)e^{\theta\_i} \\$

$\hspace{3cm}\Leftrightarrow p\_i(1+e^{\theta\_i}) = e^{\theta\_i} \\$

**(1)**

Với:

$\hspace{2cm} p\_i = \frac{e^{\theta\_i}}{1+e^{\theta\_i}}\\$

$\hspace{3cm} \Rightarrow {1-p\_i} = 1-\frac{e^{\theta\_i}}{1+e^{\theta\_i}}\\$

$\hspace{3cm} \Leftrightarrow {1-p\_i} = \frac{1+e^{\theta\_i}-e^{\theta\_i}}{1+e^{\theta\_i}}\\$

$\hspace{3cm} \Leftrightarrow {1-p\_i} = (1+e^{\theta\_i})^{-1}\\\\$

**(2)**

Thế **(1), (2)** vào **(\*)**. Ta được:

:

thuộc họ phân phối mũ phân tán

## Xác định hàm liên kết chính tắc, hàm liên kết không chính tắc

Ta có:

Đặt , ta có:

Với:

Tính kỳ vọng của Z:

Tính kỳ vọng của Y:

(Do **1**) **(3)**

Ta có:

(Do **3**)

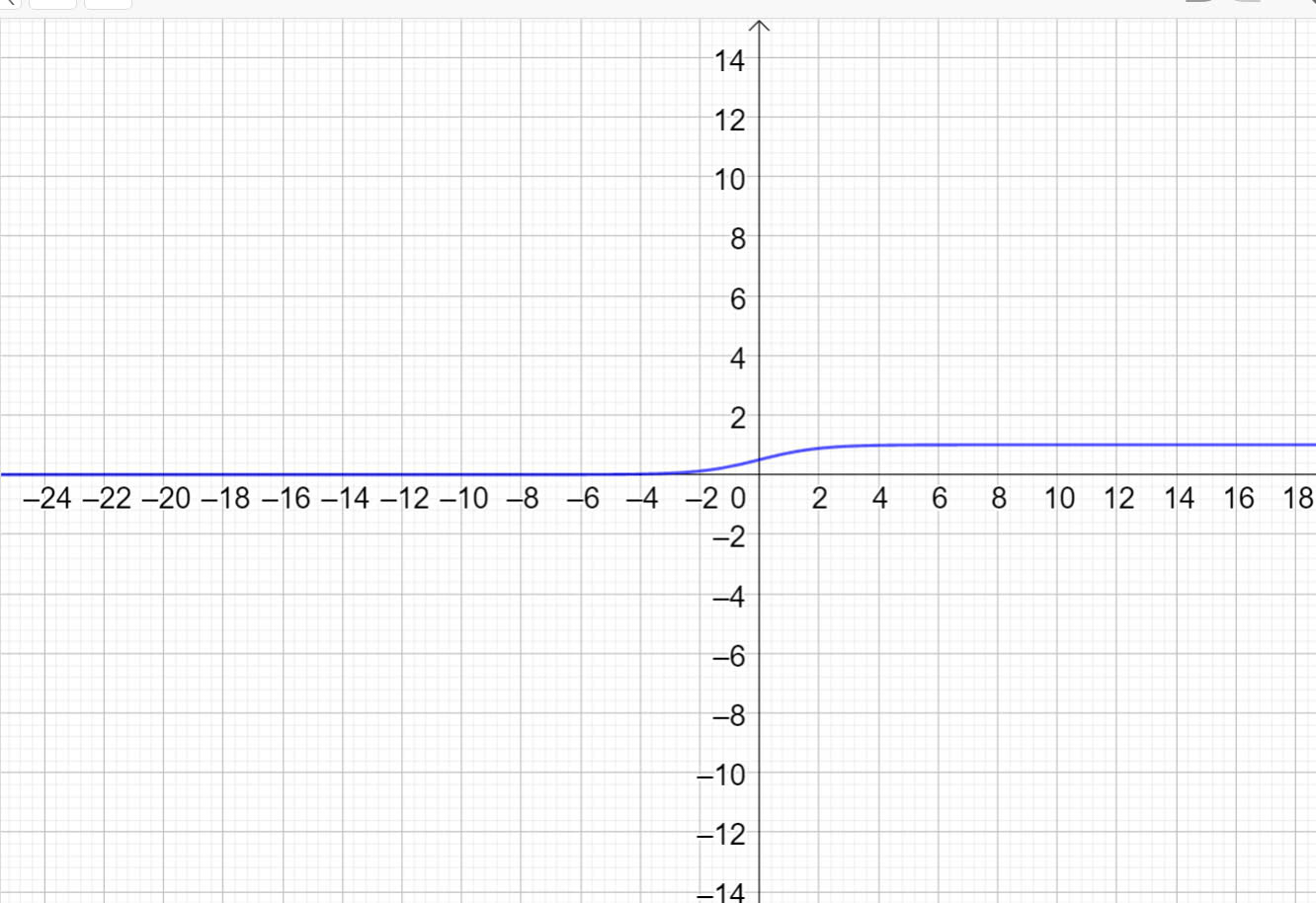
Mà:

Vậy hàm liên kết chính tắc là

Vậy hàm liên kết là

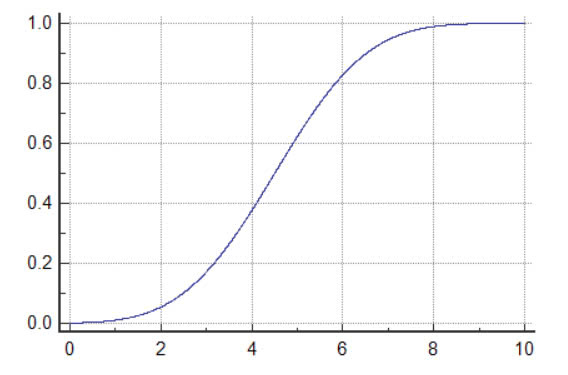
**Hàm liên kết khác:**

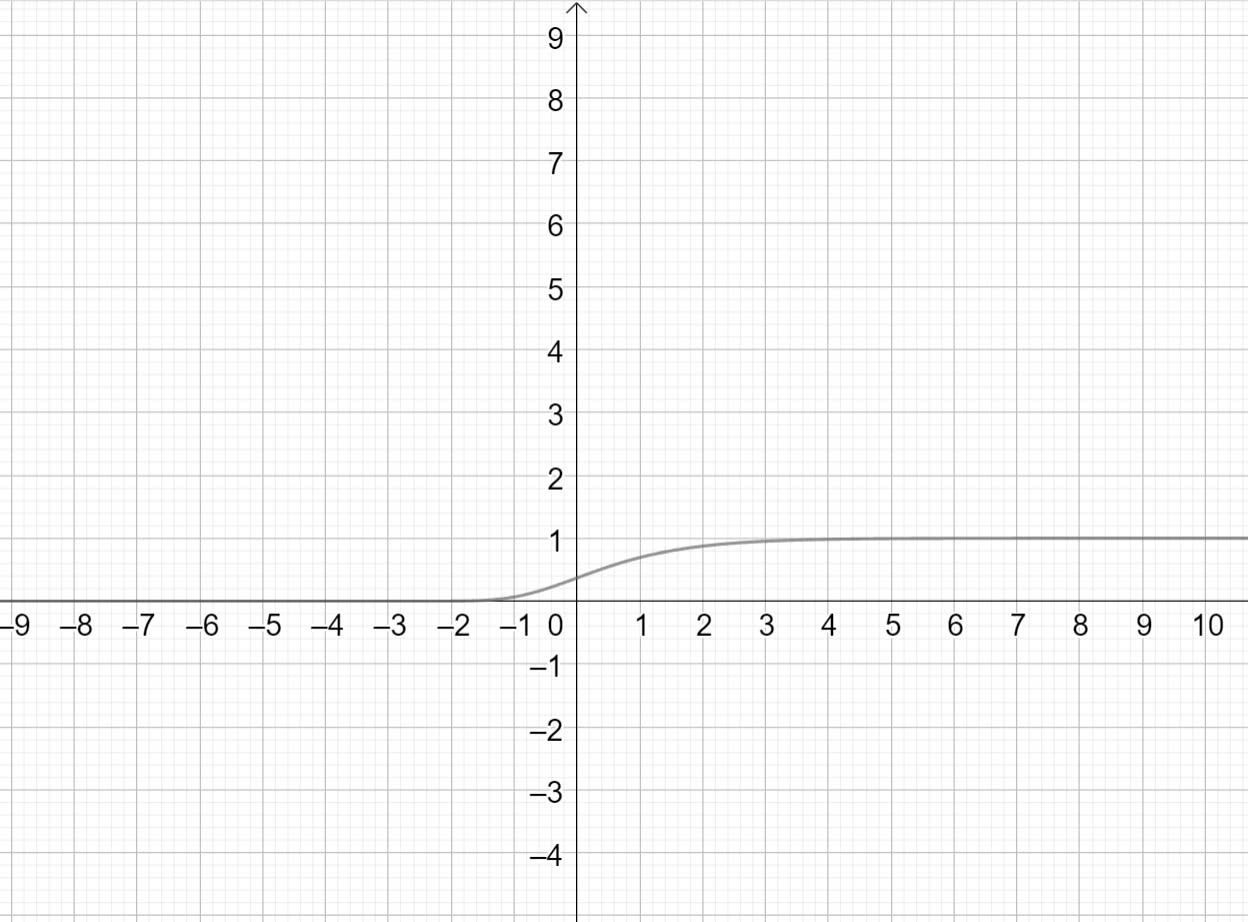
Hàm



Hàm liên kết 1

Hàm

  
 Hàm



Hàm liên kết 3

## Xây dựng mô hình:

Khi đó, với:

.

.

Thành phần linear predictor được xác định là:

.

Ta sử dụng hàm liên kết chuẩn tắc, tức là .

Do đó, mô hình tuyến tính cần ước lượng sẽ là:

#### define the explanatory variable with two levels:  
#### 1=one or more parents smoke, 0=no parents smoke  
  
parentsmoke=as.factor(c(1,0))  
parentsmoke

## [1] 1 0  
## Levels: 0 1

#### create a response vector so that it has counts for both "success" and "failure"  
  
response<-cbind(yes=c(816,188),no=c(3203,1168))  
response

## yes no  
## [1,] 816 3203  
## [2,] 188 1168

#### fit the logistic regression model  
smoke.logistic<-glm(response~parentsmoke,family=binomial(link=logit))  
  
#### OUTPUT  
  
smoke.logistic

##   
## Call: glm(formula = response ~ parentsmoke, family = binomial(link = logit))  
##   
## Coefficients:  
## (Intercept) parentsmoke1   
## -1.8266 0.4592   
##   
## Degrees of Freedom: 1 Total (i.e. Null); 0 Residual  
## Null Deviance: 29.12   
## Residual Deviance: 5.631e-13 AIC: 19.24

summary(smoke.logistic)

##   
## Call:  
## glm(formula = response ~ parentsmoke, family = binomial(link = logit))  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.82661 0.07858 -23.244 < 2e-16 \*\*\*  
## parentsmoke1 0.45918 0.08782 5.228 1.71e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2.9121e+01 on 1 degrees of freedom  
## Residual deviance: 5.6311e-13 on 0 degrees of freedom  
## AIC: 19.242  
##   
## Number of Fisher Scoring iterations: 2

anova(smoke.logistic)

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: response  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 1 29.121   
## parentsmoke 1 29.121 0 0.000 6.801e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Ước lượng :

Mô hình tuyến tính cần ước lượng sẽ là:

.

Khi , phương trình trở thành:

.

Nghĩa là, là log-odds của xác suất hút thuốc khi không có phụ huynh nào hút thuốc.

Xác suất hút thuốc khi không có phụ huynh nào hút thuốc ():

.

Log-odds khi là:

.

### Mô hình ước lượng:

Cuối cùng, chúng ta có mô hình ước lượng:

.

# THAM KHẢO

* <https://www.youtube.com/watch?v=7mNrsFr7P_A&si=3CSLdQa-Sg4aeKwj>
* <https://www.youtube.com/watch?v=MfynxwNqgfU>