

---

---

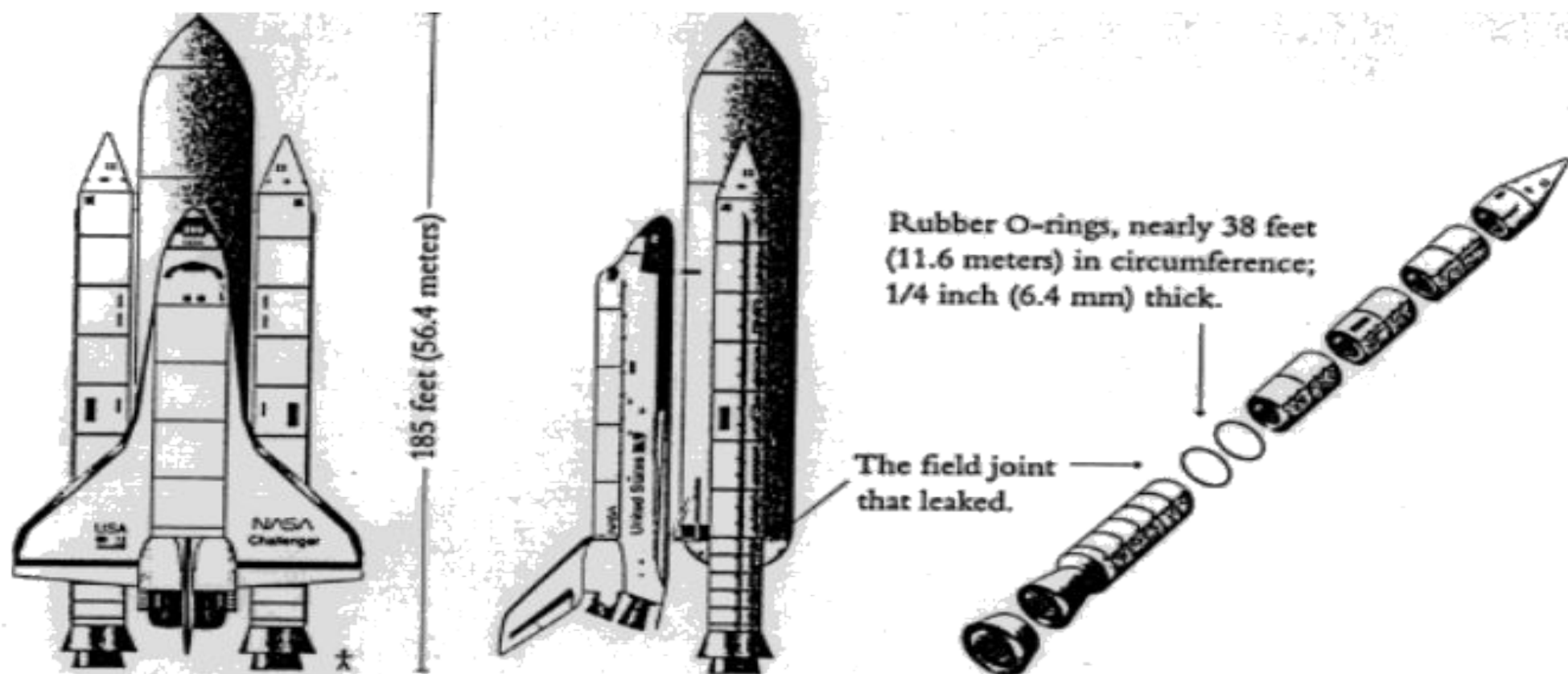
# Mô hình hồi quy logistic

— TS. Lưu Phúc Lợi —  
Trưởng phòng Nghiên cứu khoa học  
Viện ARiHA —

---

---

## Challenger O-Rings



# The *Challenger* shuttle disaster

Flight	Temp	Damage
STS-1	66	0
STS-2	70	1
STS-3	69	0
STS-4	80	.
STS-5	68	0
STS-6	67	0
STS-7	72	0
STS-8	73	0
STS-9	70	0
STS 41B	57	1
STS 41C	63	1
STS 41D	70	1
STS 41G	78	0
STS 51A	67	0
STS 51C	53	1
STS 51D	67	0

Flight	Temp	Damage
STS 51B	75	0
STS 51G	70	0
STS 51F	81	0
STS 51I	76	0
STS 51J	79	0
STS 61A	75	1
STS 61B	76	0
STS 61C	58	1

```
Temp = c(66, 70, 69, 80, 68, 67,  
72, 73, 70, 57, 63, 70, 78, 67,  
53, 67, 75, 70, 81, 76, 79, 75,  
76, 58)
```

```
Damage = c(0, 1, 0, ., 0, 0, 0,  
0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0,  
0, 0, 0, 1, 0, 1)
```

```
df = read.csv("content/Stroke_Data.csv")
```

```
> head(df)
```

	<u>id</u>	<u>gender</u>	<u>age</u>	<u>hypertension</u>	<u>heart.disease</u>	<u>ever.married</u>	<u>work.type</u>	<u>Residence.type</u>	<u>glucose.level</u>	<u>bmi</u>	<u>smoking</u>	<u>stroke</u>
1	67	Female	17	0	0	No	Private	Urban	92.97	NA	formerly smoked	0
2	77	Female	13	0	0	No	children	Rural	85.81	18.6	Unknown	0
3	84	Male	55	0	0	Yes	Private	Urban	89.17	31.5	never smoked	0
4	91	Female	42	0	0	No	Private	Urban	98.53	18.5	never smoked	0
5	99	Female	31	0	0	No	Private	Urban	108.89	52.3	Unknown	0
6	121	Female	38	0	0	Yes	Private	Urban	91.44	NA	Unknown	0

# Đặc tính của các nghiên cứu

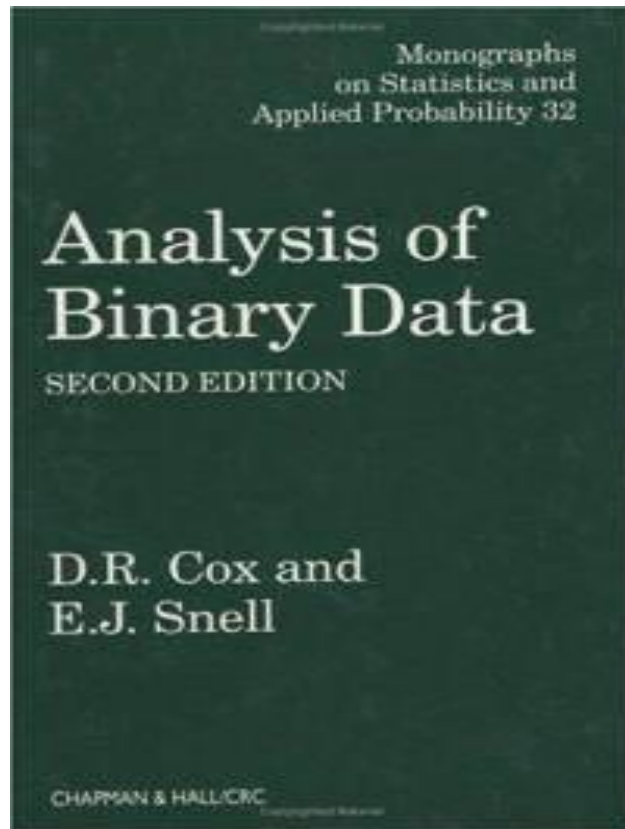
- **Outcome (dependent) variable:** biến nhị phân (binary variable), chỉ có 2 giá trị
- **Predictor (independent) variables:** đa dạng (nhị phân, biến liên tục)

**Không thể dùng mô hình hồi qui tuyến tính!**

# Ứng dụng của mô hình hồi qui logistic

- Mô tả mối liên quan giữa biến outcome và biến tiên lượng
- Kiểm soát các biến nhiễu (Controlling for confounders)
- Phát triển mô hình tiên lượng (Developing prognostic models)

# Ông "tổ" của mô hình hồi qui logistic



1970



Professor David R. Cox  
Imperial College, London

# Khi nào cần sử dụng mô hình hồi qui logistic

- Logistic regression:
  - outcome là biến phân loại (thường có 2 giá trị yes/no)
  - biến tiên lượng có thể là biến phân loại hay liên tục
- Mô hình hồi qui tuyến tính (Linear regression)
  - biến outcome là biến liên tục
  - biến tiên lượng có thể là biến phân loại hay liên tục



**Khái niệm cơ bản: odds và logit**

# Risk, probability và odds

- **Risk**: probability (P) of an event [during a period] – xác suất của một biến cố trong một thời gian
- **Odds**: xác suất biến cố xảy ra chia cho xác suất biến cố không xảy ra:

$$Odds = \frac{P}{1-P}$$

$n = 5$  bệnh nhân, 1 bệnh nhân bị đột quỵ:

$$P = \frac{1}{5} = 0.2$$

$$Odds = \frac{0.2}{0.8} = 0.25$$

# Odds ratio (tỉ số odds)

- Odds là khái niệm tương đối mới (từ J Cornfield, Johns Hopkins University, Chủ tịch American Statistical Association (1974))
- Odds là tỉ số 2 xác suất
- Odds ratio (OR hay tỉ số odds) là tỉ số 2 odds
- Khái niệm OR có thể áp dụng để so sánh tỉ lệ giữa 2 nhóm



Jerome Cornfield  
(1912 – 1979)

# Xác suất, odds, logit

- **Xác suất**: từ 0 đến 1
- **Odds**: biến liên tục  
Khi  $P = 0.5$ , odds = 1
- **Logit = log odds**

$$\text{logit}(P) = \log\left(\frac{p}{1-p}\right)$$

# Mô hình hồi qui logistic dựa trên logit

- Gọi  $X$  là biến tiên lượng
- Gọi  $P$  là xác suất của một biến cố (outcome)
- Mô hình hồi qui logistic phát biểu rằng:

$$\text{logit}(P) = \alpha + \beta X$$

hay

$$\log\left(\frac{P}{1-P}\right) = \alpha + \beta X$$

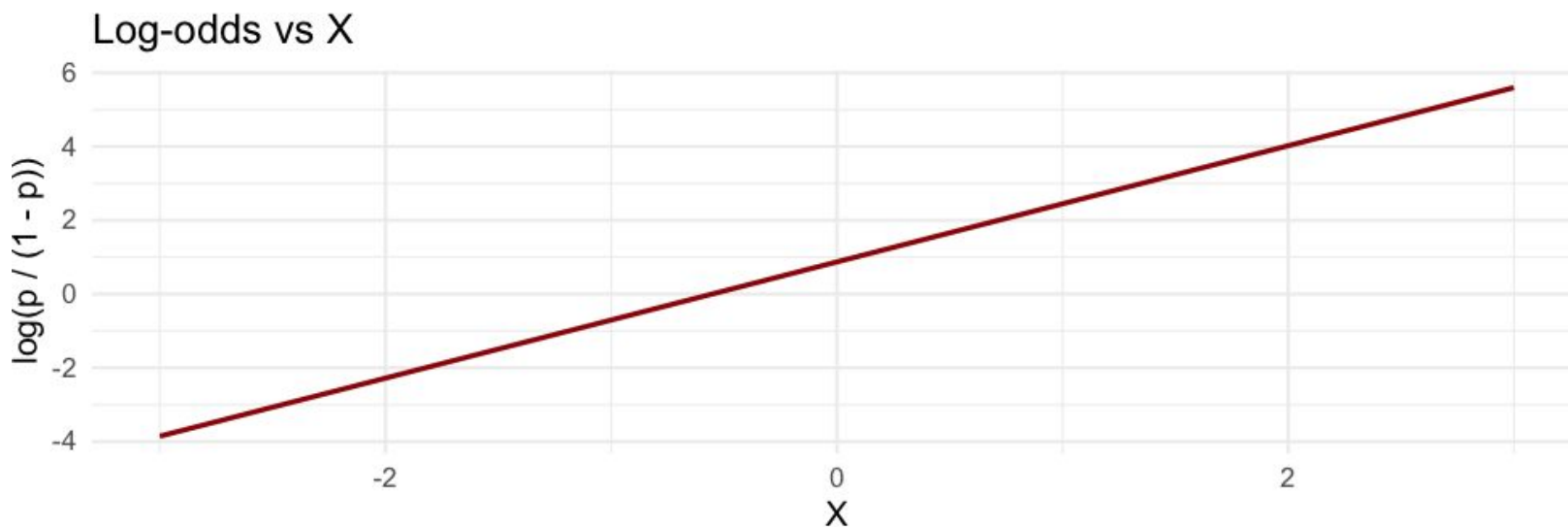
# Mô hình hồi qui logistic

$$\log \left( \frac{P}{1-P} \right) = \alpha + \beta X$$

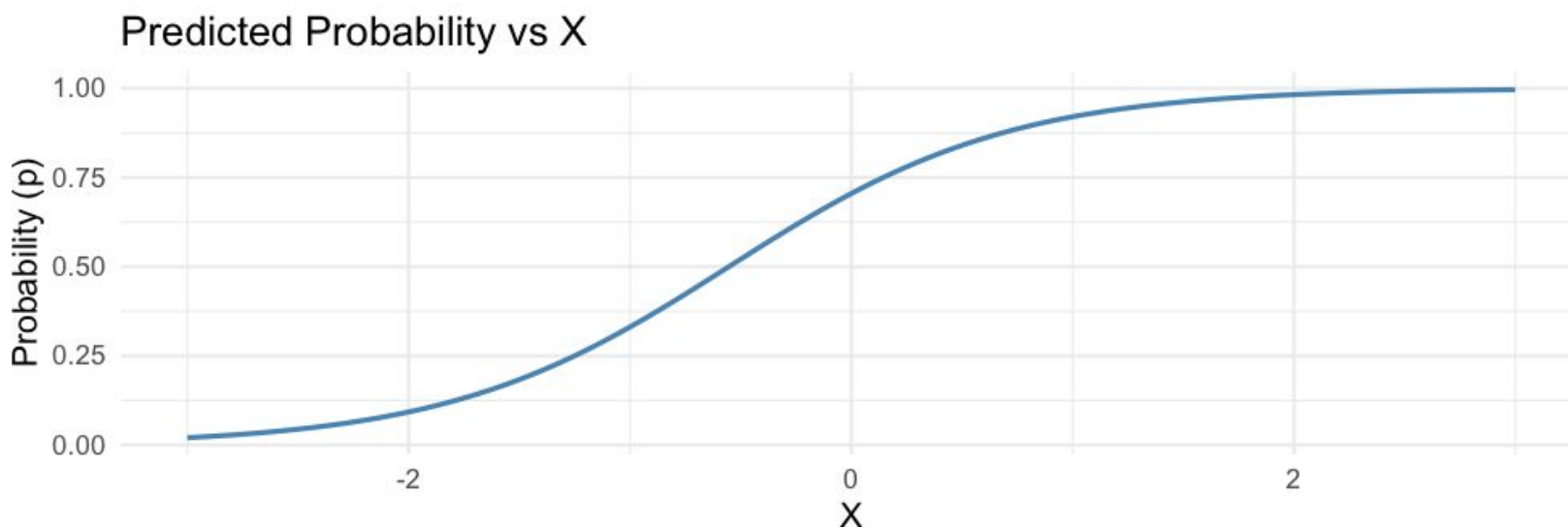
Điều này cũng có nghĩa là:

$$P = \frac{\exp(\alpha + \beta X)}{1 + \exp(\alpha + \beta X)}$$

# Mối liên quan giữa $X$ , $\text{logit}(p)$ , và $p$



$$\text{logit}\left(\frac{P}{1 - P}\right) = \alpha + \beta X$$



$$P = \frac{\exp(\alpha + \beta X)}{1 + \exp(\alpha + \beta X)}$$

# Ý nghĩa của tham số mô hình logistic

$$\log \left( \frac{P}{1-P} \right) = \alpha + \beta X$$

- $\alpha$  là log odds của biến outcome khi  $X = 0$
- $\beta$  là log odds ratio (tỉ số) liên quan với một đơn vị tăng của  $X$
- Odds ratio =  $e^{\beta}$



# Ưu điểm của mô hình hồi qui logistic

- Xác suất của outcome có thể thay đổi với giá trị của biến tiên lượng
- Hệ số có thể diễn giải như là log odds ratio
- Có thể áp dụng cho nhiều mô hình nghiên cứu
- Nhiều software có thể dùng để ước tính tham số

# Ước tính tham số

Triển khai trong hàm R “`glm`” và “`lrm`” (trong package `rms`)

# Hàm **glm** trong base R

- Công thức chung:

```
m = glm(outcome ~ riskfactor, family = binomial)
```

outcome có giá trị (0, 1)

riskfactor – bất cứ biến nào

- Có có khoảng tin cậy 95% OR:

```
library(epiDisplay)
```

```
logistic.display(m)
```

# Hàm **lrm** trong rms package

- Package **rms** – Frank Harrell (regression modeling strategies)
- Hàm **lrm** – logistic regression model

```
lrm(outcome ~ factor1 + factor2 + ...)
```

# Hàm **Logit** trong lessR package

Package `lessR` cũng có thể dùng để ước tính tham số mô hình hồi qui logistic qua hàm `Logit`

```
library(lessR)
```

```
Logit(outcome ~ factor1 + factor2 + ...)
```

**Phân tích với package**

`lessR`

# Nghiên cứu đột quỵ: khác biệt giữa nam và nữ?

- **Câu hỏi nghiên cứu:** nguy cơ đột quỵ có khác biệt giữa nam và nữ?
- **Mô hình hồi qui logistic:** gọi  $p$  là xác suất bị đột quỵ, và gender là yếu tố nguy cơ, mô hình là

$$\log \left( \frac{p}{1-p} \right) = \alpha + \beta \times gender$$

```
install.packages("lessR")  
  
df = read.csv("/content/Stroke_Data.csv")  
  
library(lessR)  
  
fit = Logit(stroke ~ gender, brief=T, data=df)
```



# Output from lessR

-- Estimated Model of stroke for the Logit of Reference Group Membership

	Estimate	Std Err	z-value	p-value	Lower 95%	Upper 95%
(Intercept)	-3.0074	0.0863	-34.862	0.000	-3.1764	-2.8383
genderMale	0.0846	0.1311	0.645	0.519	-0.1724	0.3416

-- Odds Ratios and Confidence Intervals

	Odds Ratio	Lower 95%	Upper 95%
(Intercept)	0.0494	0.0417	0.0585
genderMale	1.0883	0.8416	1.4072

exp(0.0846) = 1.088

$$\log\left(\frac{p}{1-p}\right) = -3.01 + 0.0846 \times \text{gender}(\text{male} = 1, \text{female} = 0)$$

**Diễn giải:** Odds bị đột quị ở nam giới cao hơn nữ giới 1.09 lần, với khoảng tin cậy 95% dao động từ 0.84 đến 1.41.

# Nghiên cứu đột quị: ảnh hưởng của tuổi?

- **Câu hỏi nghiên cứu:** nguy cơ đột quị có tăng theo độ tuổi?
- Mô hình hồi qui logistic bây giờ là

$$\log \left( \frac{p}{1-p} \right) = \alpha + \beta \times age$$

```
fit = Logit(stroke ~ age, brief=T, data=df)
```

```
-- Estimated Model of stroke for the Logit of Reference Group Membership
```

	Estimate	Std Err	z-value	p-value	Lower 95%	Upper 95%
(Intercept)	-7.2314	0.3350	-21.588	0.000	-7.8880	-6.5749
age	0.0747	0.0049	15.181	0.000	0.0651	0.0844

```
-- Odds Ratios and Confidence  
Intervals
```

	Odds Ratio	Lower 95%	Upper 95%
(Intercept)	0.0007	0.0004	0.0014
age	1.0776	1.0672	1.0880

**$\exp(0.0747) = 1.078$**

$$\log\left(\frac{p}{1-p}\right) = -7.23 + 0.0747 \times age$$

**Diễn giải:** Mỗi 1 tuổi tăng lên có liên quan đến tăng odds bị đột quỵ 8% (OR 1.08; khoảng tin cậy 95% dao động từ 1.07 đến 1.09).

# Cách báo cáo

Variable	Unit	OR (G5% CI)	P-value
Sex	Men vs women	1.09 (0.84 – 1.41)	0.519
Age	+1 yr	1.08 (1.07 – 1.09)	<0.0001

# Ứng dụng ChatGPT

**Prompt:** Tôi có hai biến số: biến tiên lượng là age và gender, và biến phụ thuộc stroke trong dataset tên là df. Tôi muốn tìm hiểu sự ảnh hưởng của hai yếu tố này đến stroke qua package `lessR`.

```
fit = Logit(stroke ~ age + gender, data = df)
```

# BASIC ANALYSIS

-- Estimated Model of stroke for the Logit of Reference Group Membership

	Estimate	Std Err	z-value	p-value	Lower 95%	Upper 95%
(Intercept)	-7.2834	0.3418	-21.308	0.000	-7.9533	-6.6134
age	0.0748	0.0049	15.167	0.000	0.0651	0.0844
genderMale	0.1153	0.1374	0.839	0.401	-0.1539	0.3845

## -- Odds Ratios and Confidence Intervals

	Odds Ratio	Lower 95%	Upper 95%
(Intercept)	0.0007	0.0004	0.0013
<b>age</b>	<b>1.0776</b>	<b>1.0673</b>	<b>1.0881</b>
<b>genderMale</b>	<b>1.1222</b>	<b>0.8573</b>	<b>1.4689</b>

## -- Model Fit

Null deviance: 1990.373 on 5109 degrees of freedom

Residual deviance: 1615.595 on 5107 degrees of freedom

AIC: 1621.595

# Thông điệp mang về

- Mô hình:  $\text{logit} \left( \frac{p}{1-p} \right) = \alpha + \beta X$
- Mô hình áp dụng cho biến outcome là nhị phân (yes/no)
- Có thể ứng dụng cho các nghiên cứu cắt ngang, đoàn hệ, bệnh chứng
- Triển khai trong R: hàm glm, hàm Logit trong package lessR