**Homework 8**

Sinh viên thực hiện

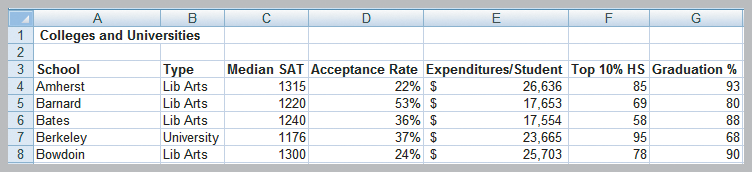
họ tên: đoàn thục quyên

MSSV: 18521320

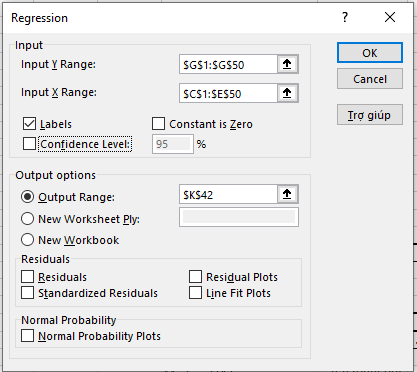
Câu 1. Dùng MS Excel thực hiện Example 9.10. Tính lại và giải thích tất cả các giá trị trong bảng kết quả.

Mô tả bài toán *Example 9.10*

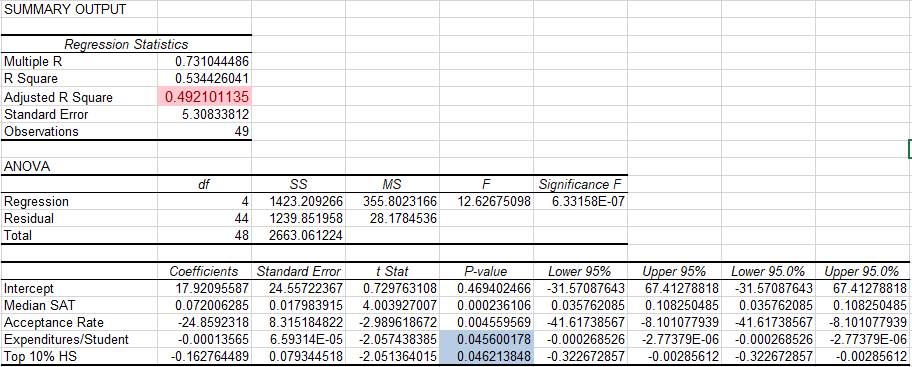
* Phân tích trên file *Colleges and Universities* Data.xlsx
* Trường đại học cố gắng dự đoán học tỷ lệ học sinh tốt nghiệp dựa trên các yếu tố sau:
  + Median SAT
  + Acceptance rate
  + Expenditures/student
  + Top 10% of HS class



Dùng Analysis của Excel:

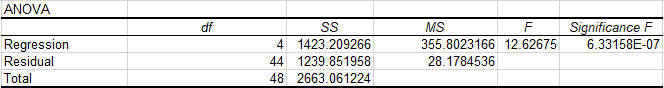


Ta có được bảng phân tích như sau:

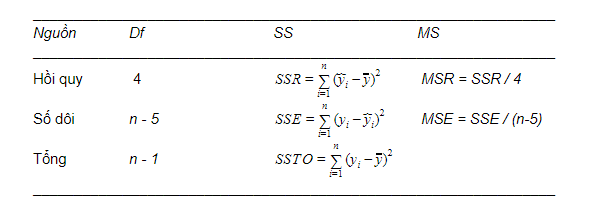


## GIẢI THÍCH CÁC GIÁ TRỊ CÓ TRONG BẢNG PHÂN TÍCH TRÊN:

# **Bảng Anova (bảng 2)**



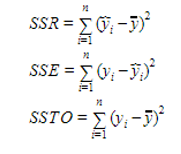
* Bảng trên được tính dựa vào công thức như hình bên dưới:



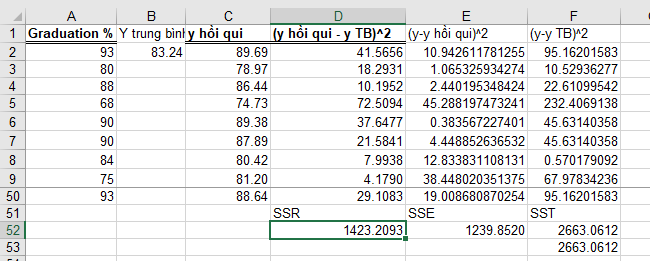
* Ta dùng công thức trên và thực hiện bằng excel lần lượt như sau:
* Ta có 49 mẫu kiểm định => n = 49



* df của residual: 49-5 = 44
* df của rotal: 49-1= 48
* Thông số SS
* Theo công thức:



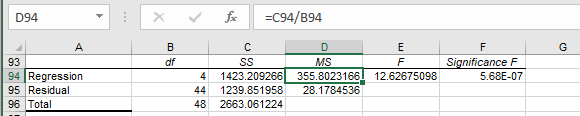
* Từ y và y hồi qui, ta tình được 3 cột cuối, tổng của từng cột, ta được:



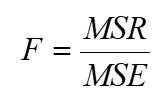
* Thông số MS



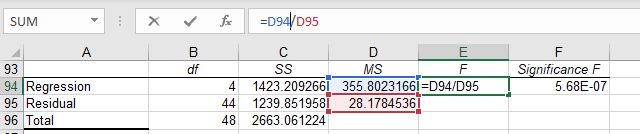
* Ta lấy SS chia cho df của hàng tương ứng:



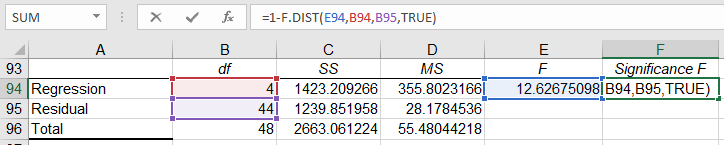
* Giá trị kiểm định F
* Theo công thức:



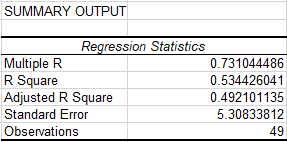
* Ta được:



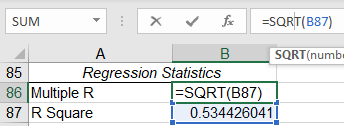
* Significance F



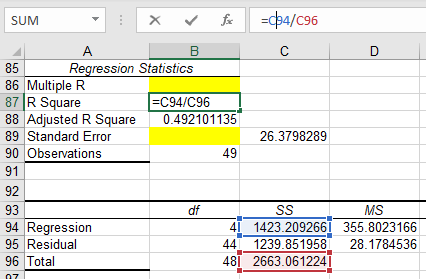
# **Bảng 1**



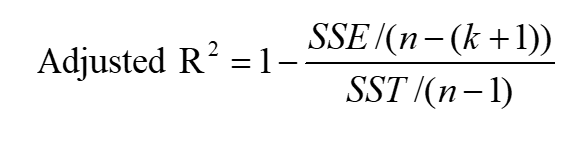
* Multiple R: hay căn bậc 2 của R square, ta có được kết quả như trong bảng phân tích



* Hệ số xác định R square

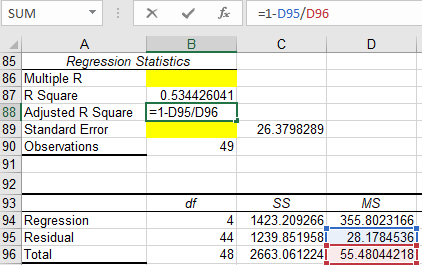


* Adjusted R square
* Dựa vào công thức sau:

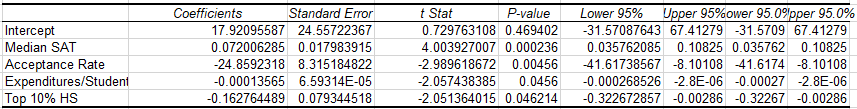


Suy ra R^2 = 1 – (MSR/MST)

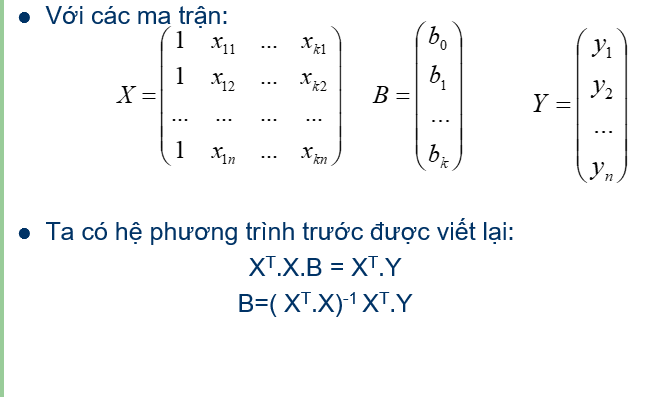
* Áp dụng công thức trong excel, ta được kết quả như trong bảng phân tích

****

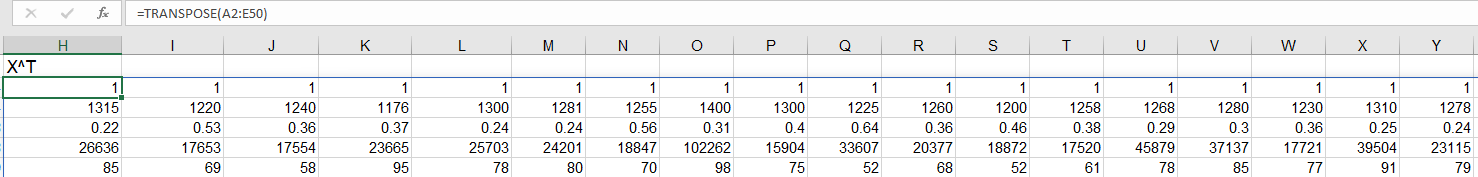
# **Bảng 3:**



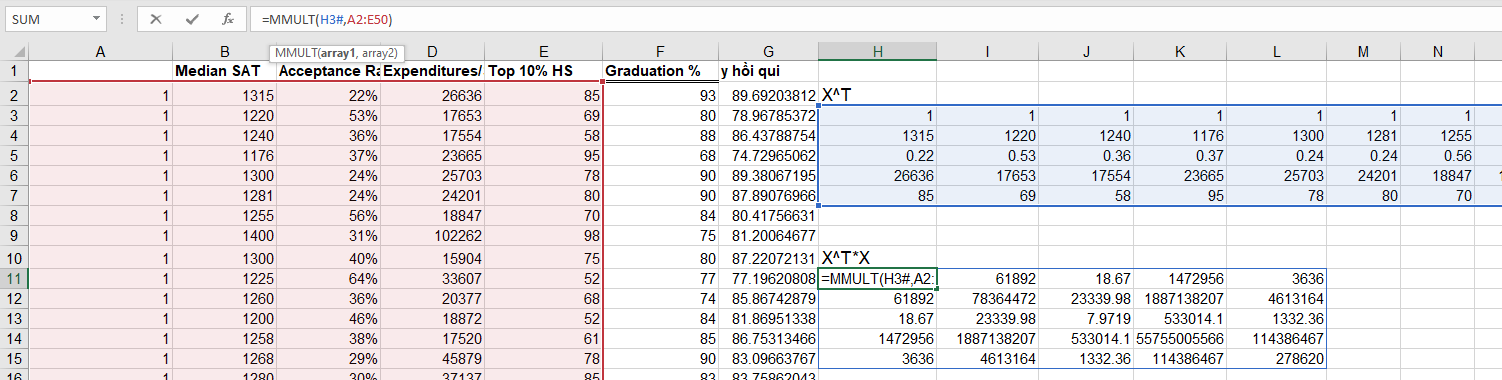
* Thông số coefficients:
* Ta có công thức:

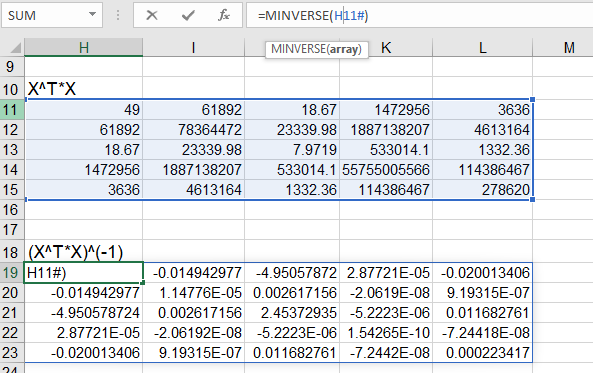
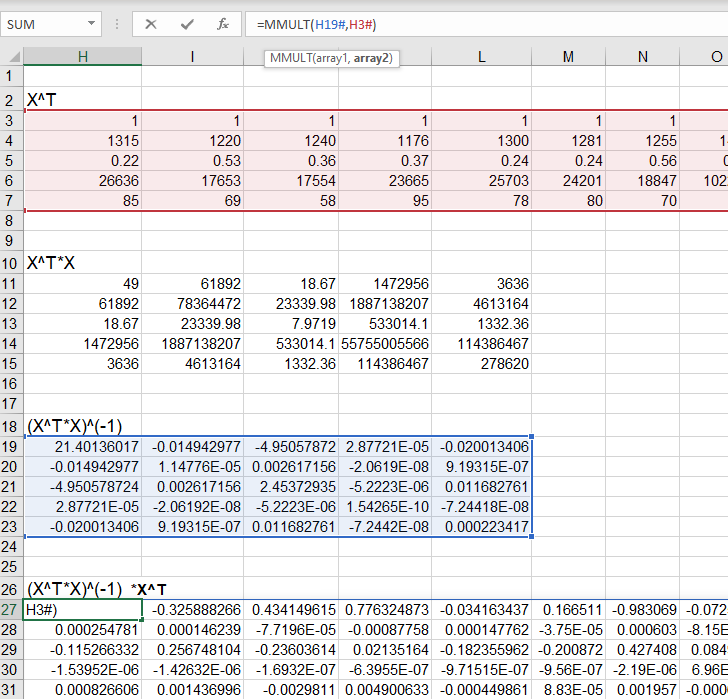


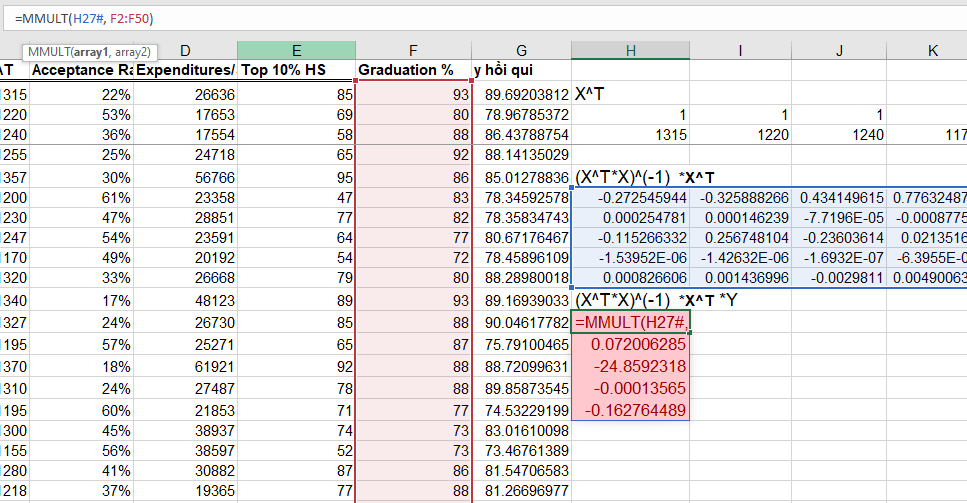
* Tính X^T



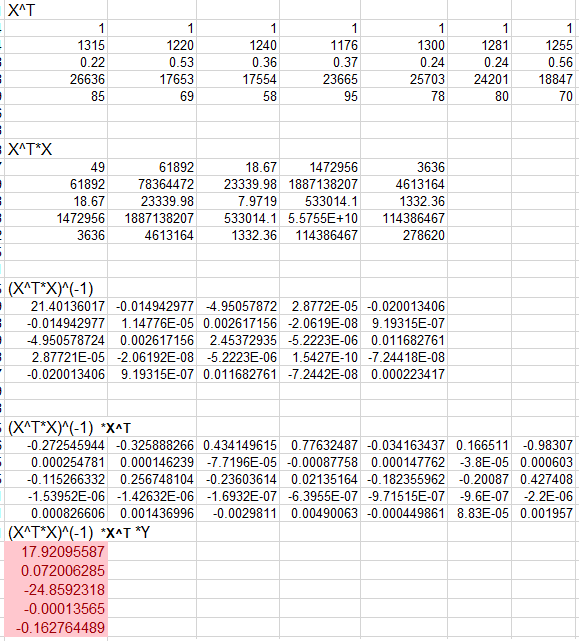
* Tính X^T\*X



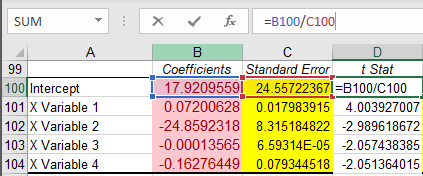
* Tính (X^T\*X)^(-1) 
* (X^T\*X)^(-1) \*X^T
* (X^T\*X)^(-1) \*X^T \*Y



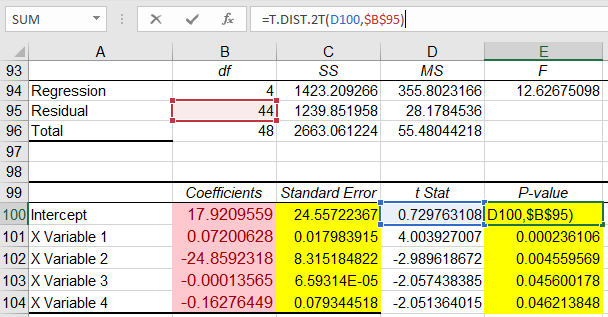
* Ta nhân lần lượt các số theo công thức, ta tính được B



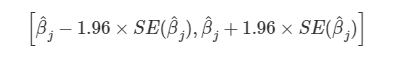
* T Stat



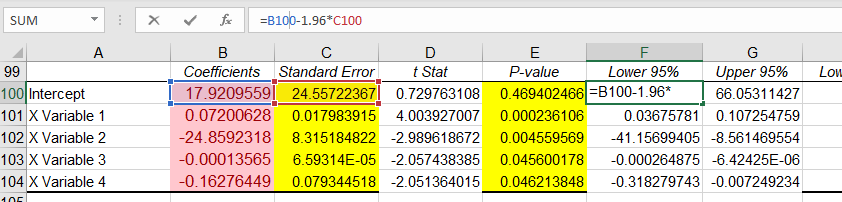
* P-value



* Khoảng tin cậy 95%
* Ta dùng công thức:



* Lần lượt tính, ta được lower 95% và upper 95%:



# **Tài liệu tham khảo:**

Slide Bài 1: Phân tích dữ liệu kinh doanh của giảng viên Nguyễn Đình Thuân

Công thức khoảng tin cậy:

<http://cameron.econ.ucdavis.edu/excel/ex61multipleregression.html>

Cách tính: p-value, signification F

<https://econometricsr.hieunguyenphi.com/kim-inh-gia-thuyt-va-khoang-tin-cy-hi-quy-a-bin.html>

Video bài giảng của thầy Nguyễn Văn Tuấn:

<https://www.youtube.com/watch?v=bTzGlI2Kjtg&list=UU21dOPe-YHO3Gw6BRbyeotQ>

<https://www.youtube.com/watch?v=Hc3Z1OjYAQA&list=UU21dOPe-YHO3Gw6BRbyeotQ>

<https://www.youtube.com/watch?v=LI3acanCvWA&list=UU21dOPe-YHO3Gw6BRbyeotQ>

Công thức standard error:

<https://www.statisticshowto.com/find-standard-error-regression-slope/#:~:text=Standard%20Error%20of%20Regression%20Slope%20Formula,i%20%E2%80%93%20x)2%20%5D>.

Câu 2. Tìm hiểu và cho ví dụ về cách giải bài toán hồi qui logistic

# **Khái niệm**

Phân tích hồi qui logistic là một kỹ thuật thống kê để xem xét mối liên hệ giữa biến độc lập với biến phụ thuộc là biến nhị phân. Trong hồi qui tuyến tính đơn, biến độc lập x và phụ thuộc y là biến số liên tục liên hệ qua phương trình:

y = α + βx + ε

Trong hồi qui logistic, biến phụ thuộc y chỉ có 2 trạng thái 1 ( ví dụ tử vong) và 0 ( ví dụ sống). Muốn đổi ra biến số liên tục người ta tính xác suất của 2 trạng thái này. Nếu gọi p là xác suất để một biến cố xảy ra (ví dụ: tử vong), thì 1-p là xác suất để biến cố không xảy ra (ví dụ: sống).

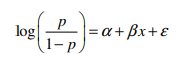
Odd có thể định nghĩa như sau:



Gọi kết quả là x, và x có hai giá trị: 0 (hoặc false), và 1 (hoặc true). Mô hình hồi qui logistic phát biểu rằng log(odd) tùy thuộc vào giá trị của x qua một hàm số tuyến tính gồm 2 thông số như sau:



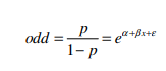
Hay

(1)

Trong đó:

* log(odd) hay  còn được gọi là logit(p) (và do đó, mới có tên logistic)
* α và β là hai thông số cần ước tính từ dữ liệu
* ε là phần dư (residual), tức là phần không thể giải thích bằng x; ε tuân theo luật phân phối chuẩn (normal distribution) với trung bình bằng 0 và phương sai bất biến (constant variance).

Lí do hoán chuyển từ p thành logit(p) là vì p có giá trị trong khoảng 0 và 1, trong khi đó logit(p) có giá trị vô giới hạn và do đó thích hợp cho việc phân tích theo mô hình hồi qui tuyến tính.

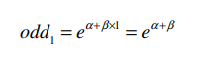
1. =>  (2)

Dựa vào phương trình [1]:

-Nhóm (x = 0) có odd0 là:

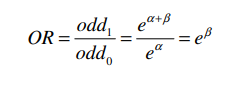
(3)

-Nhóm (x = 1) có odd1 là:

****(4)

Chỉ số thống kê quan trọng để phân tích là tỉ số nguy cơ (odds ratio hay OR). Odd là tỉ số của hai giá trị của một biến số nhị phân. Do đó, OR là tỉ số của hai odds.

Tỉ số của hai odds chính là odds ratio (và đó chính là lí do tại sao tôi dịch odds ratio là tỉ số nguy cơ). Tỉ số nguy cơ – OR – có thể ước tính từ [3] và [4] như sau:

****(5)

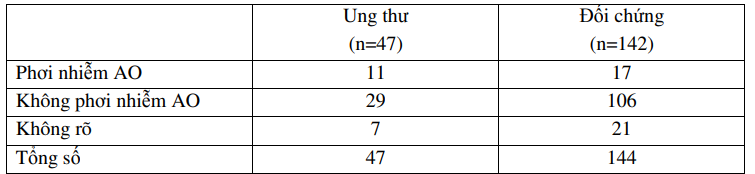
Trong thực tế, chúng ta không biết giá trị thật của hai thông số α và β, và phải ước tính từ số liệu quan sát được. Theo qui ước thống kê, ước số (estimates) của hai thông số này được kí hiệu hóa bằng dấu mũ:



# **Ví dụ: ( lấy ví dụ của thầy Nguyễn Văn Tuấn)**

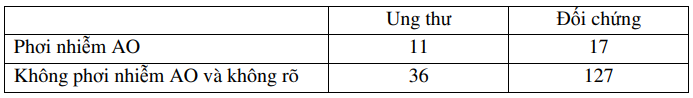
Nghiên cứu mối tương quan giữa phơi nhiễm chất độc da cam và ung thư tuyến tiền liệt.

Chọn ngẫu nhiên 144 cựu chiến binh nhập viện vì các lí do không liên quan đến ung thư. Gọi nhóm này là nhóm “Đối chứng” (control). Các nhà nghiên cứu tìm trong sơ bệnh lí: trong số 47 trường hợp ung thư, có 11 người từng bị phơi nhiễm AO, 29 người không từng bị phơi nhiễm, và 7 người không rõ tiền sử; trong nhóm đối chứng có 17 người không từng bị phơi nhiễm, 106 người không từng bị phơi nhiễm, và 21 người không thể xác định phơi nhiễm. Kết quả có thể tóm lược trong bảng số liệu sau đây:

****

*Nguồn số liệu: Giri VN, Cassidy AE, Beebe-Dimmer J, Ellis LR, Smith DC, Bock CH, Cooney KA. Association between Agent Orange and prostate cancer: a pilot case-control study. Urology. 2004 Apr;63(4):757-60; discussion 760-1. Correction in Urology. 2004 Jun;63(6):1213.*

Để minh họa cho phân tích hồi qui tuyến tính và đơn giản hóa vấn đề, ta sẽ gộp chung hai nhóm “Không phơi nhiễm AO” và “Không rõ” thành một nhóm chung. Bảng số liệu trên, do đó, có thể rút gọn như sau:



* odd mắc ung thư trong nhóm từng bị phơi nhiễm AO là: 11/17 = 0.647;
* odd mắc ung thư trong nhóm không từng bị phơi nhiễm AO là: 36/127 = 0.283;

odds ratio mắc bệnh ung thư trong nhóm từng bị phơi nhiễm so với nhóm không từng bị phơi nhiễm là: OR = 0.647 / 0.283 = 2.28 = e^β

* ln(e^β)= 0.824
* Ước số của thông số β là βˆ = 0.824. Do đó, OR phản ánh odd bị ung thư trong nhóm bị phơi nhiễm AO so với odd trong nhóm không từng bị phơi nhiễm AO. Nói cách khác, khi ao=1 thì nguy cơ ung thư tăng 2.28 lần so với nhóm ao=0.

# **Kiểm tra lại bằng phần mềm bằng R**

Ta nhập vào phần mềm R các dòng lệnh sau:

> ao = c(1,0)

> ntotal(28,163)

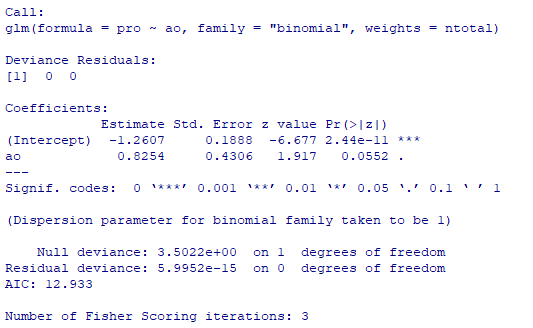
> cancer = c(11,36)

> pro = cancer/ntotal

> logistic = glm(pro ~ ao, family="binomial", weight=ntotal)

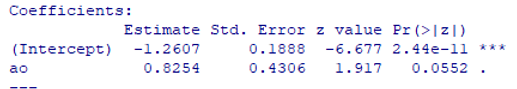
> summary(logistic)

Ta sẽ được kết quả như hình bên dưới:

****

(a) Phần “Call:” báo cho chúng ta biết mô hình phân tích;

(b) Deviance: phần thứ hai của kết quả cho biết qua về deviance, tức phần dư (hay residual trong mô hình [1]). Deviance như giải thích trên phản ánh độ khác biệt giữa mô hình và dữ liệu (tương tự như mean square residual trong phân tích hồi qui tuyến tính).

(c) Phần kế tiếp cung cấp ước số của α (mà R đặt tên là intercept) và β (ao) và sai số chuẩn (standard error) cho từng ước số: 

Qua kết quả này, chúng ta có αˆ = -1.2607 và β ˆ = -0.8254.

* Ước số β ˆ là số dương, cho thấy mối liên hệ giữa cancer và ao là mối liên hệ thuận: nguy cơ ung thư tăng khi giá trị của ao tăng.
* Kiểm định z (tính bằng cách lấy ước số chia cho sai số chuẩn)
* Vì trị số p = 0.055, cho chúng ta thấy ảnh hưởng của ao không hẳn có ý nghĩa thống kê.
* OR chính là e^0.8254 = 2.28 (phương trình [5]) mà chúng ta vừa có được qua phân tích thủ công trong phần trên. Nói cách khác, khi ao=1 thì nguy cơ ung thư tăng 2.28 lần so với nhóm ao=0.

# **Tài liệu tham khảo:**

<http://bomonnoiydhue.edu.vn/upload/file/lstk15_logistic.pdf>  
<https://bvag.com.vn/wp-content/uploads/2013/01/k2_attachments_PHAN-TICH-HOI-QUY-LOGISTIC-DON-VA-DA-BIEN.pdf>