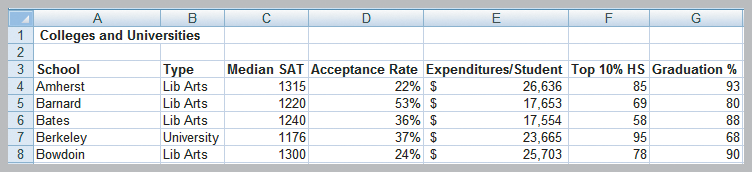
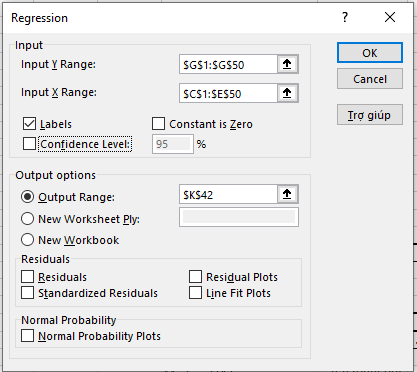
2. Dùng MS Excel và ngôn ngữ R thực hiện các Example 9.10, 9.11, 9.12 và tính lại tất cả các giá trị trong bảng kết quả

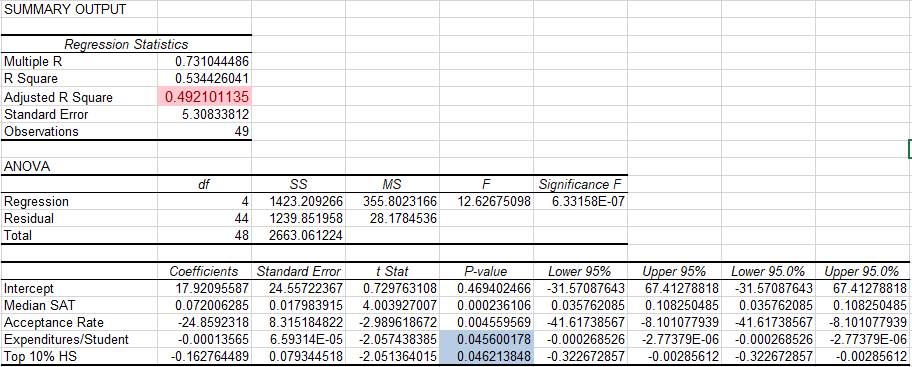
1. MS Excel
2. *Example 9.10*
3. Mô tả bài toán

* 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

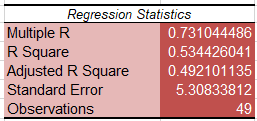


1. Dùng Analysis của Excel



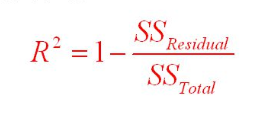


1. Giải thích các con số bảng 1



* 1. Multiple R: | r | đâu là hệ số tương quan mẫu r thay đổi từ -1 đến +1 (r âm nếu độ dốc hướng xuống)
  2. R Square: hệ số xác định( coefficient of detemination). [1]
* Giá trị R bình phương dao động từ 0 đến 1. R bình phương càng gần 1 thì mô hình đã xây dựng càng phù hợp với bộ dữ liệu dùng chạy hồi quy. R bình phương càng gần 0 thì mô hình đã xây dựng càng kém phù hợp với bộ dữ liệu dùng chạy hồi quy. Trường hợp đặt biệt, phương trình hồi quy đơn biến ( chỉ có 1 biến độc lập) thì R2 chính là bình phương của hệ số tương quan r giữa hai biến đó.
* Công thức:

Toàn bộ sự biến thiên của biến phụ thuộc được chia làm hai phần: phần biến thiên do hồi quy và phần biến thiên không do hồi quy( còn gọi là phần dư).

****

Regression Sum of Squares(RSS): tổng các độ lệch bình phương.

Residual Sum of Squares(ESS): tổng các độ lệch bình phương phần dư.

Total Sum of Squares(TSS): tổng các độ lệch bình phương toàn bộ.

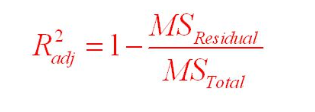
* Ý nghĩa:

Ví dụ R2 = 0.60, thì mô hình hồi quy tuyến tính này phù hợp với tập dữ liệu ở mức 60%. Nói cách khác, 60% biến thiên của biến phụ thuộc được giải thích bởi các biến độc lập.( còn 40% còn lại là do sai số đo lường, do cách thu thập dữ liệu, …).

* Hạn chế

Càng đưa thêm nhiều biến vào mô hình, mặc dù chưa xác định biến đưa vào có ý nghĩa hay không thì giá trị R2 sẽ tăng. Lý do là khi càng đưa thêm biến giải thích vào mô hình thì sẽ càng khiến phần dư giảm xuống (vì bản chất những gì không giải thích được đều nằm ở phần dư), do vậy tăng thêm biến sẽ khiến tổng bình phương phần dư(Residual Sum of Squares) giảm, trong khi Total Sum of Squares không đổi, dẫn tới R2 luôn luôn tăng.

* 1. Adjusted R Square: R2 hiệu chỉnh hoặc R2 hiệu chỉnh theo bậc tự do.[2]
* Cũng là chỉ số phản ánh mức độ phù hợp của mô hình như R bình phương, **ý** nghĩa của R bình phương hiệu chỉnh có chút khác biệt bởi R bình phương hiệu chỉnh không nhất thiết tăng lên khi chúng ta đưa thêm các biến độc lập vào mô hình.
* Công thức



n= số lượng mẫu quan sát.

k= số tham số của mô hình, bằng số lượng biến độc lập cộng 1

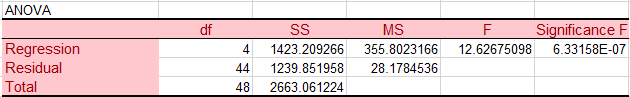
R2: hệ số R bình phương

R2 hiệu chỉnh có thể nhỏ hơn không. R bình phương hiệu chỉnh luôn nhỏ hơn hoặc bằng R bình phương.

* Ý nghĩa

Việc thêm vào một biến dẫn đến tăng R2 nhưng cũng làm giảm đi một bậc tự do, bởi vì chúng ta đang ước lượng thêm một tham số nữa. R2 hiệu chỉnh là một phép đo độ thích hợp tốt hơn bởi vì nó cho phép đánh đổi giữa việc tăng R2 và giảm bậc tự do.

1. Giải thích các con số bảng 2:



df: Bậc tự do

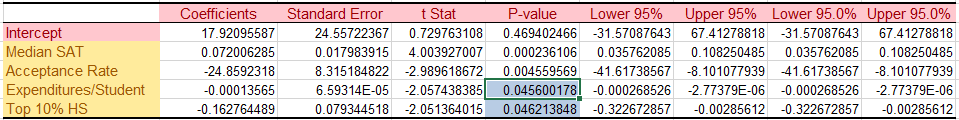
SS: Tổng bình bình phương

MS: Trung bình bình phương

F: F-test (MS của regression / MS của residual)

Significance F: p-value

1. Giải thích các con số bảng 3:



Theo công thức:



Coefficients của Intercept: b0

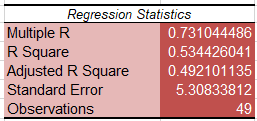
Coefficients của Median SAT, Acceptance Rate, Expenditures/Student, Top 10% HS lần lượt là b1, b2, b3, b4

Standard Error: sai số chuẩn

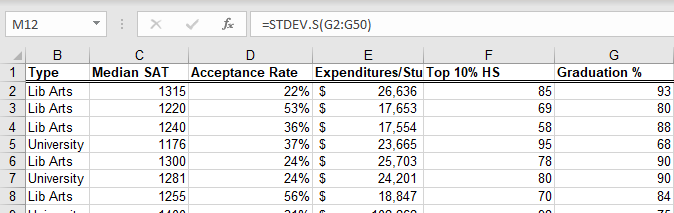
T Stat (t – test) = Coefficients/Standard Error

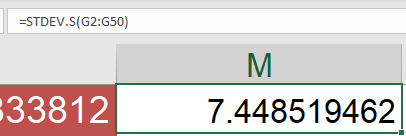
p-value: chỉ số kiểm tra giá trị thống kê

1. Diễn giải
   * + - 1. Bảng 1

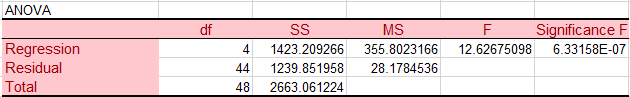


* - 53% những sự khác biệt về tỷ lệ tốt nghiệp có thể giải thích bằng độ khác biệt của các yếu tố (Median SAT, Acceptance Rate, Expenditures/Student, Top 10% HS)
* - Sai số chuẩn(Standard Error) là 5.31 thấp hơn phương sai của tỷ lệ tốt nghiệp là 7.45.



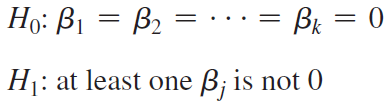


* + - * 1. Bảng 2: Phân tích hồi quy của phương sai

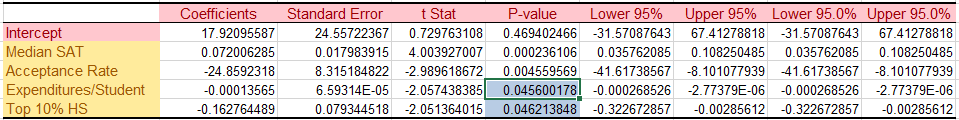


ANOVA tiến hanh F-test để xác định xem sự thay đổi trong Y có phải là do các mức X khác nhau hay không. ANOVA được sử dụng để kiểm tra ý nghĩa của hồi quy.

Giả sử:



* Do p-value(*Significance F)* trong bảng ANOVA rất nhỏ nên ta từ chối giả thuyết H0. Độ đốc (the slope) không bằng 0 và chỉ ra rằng các biến phụ thuộc X (Median SAT, Acceptance Rate, Expenditures/Student, Top 10% HS) giải thích được sự thay đổi của tỷ lệ tốt nghiệp
  + - * 1. Bảng 3: Giải thích các bài kiểm tra giả thuyết cho hệ số hồi quy



Từ bảng ta có phương trình hồi quy tuyến tính

Graduation% = 17.92 + 0.072\*Media SAT – 24.86\*Acceptance rate – 0.00013\*Expenditure – 0.163\*Top10%

Tất các hệ số p-value đều nhỏ hơn 0.05 nên ta từ chối giả thuyết H0 và cả 5 hệ số trên đều có ý nghĩa khác biệt khác 0.

Ước tính khoảng tin cậy 95% của

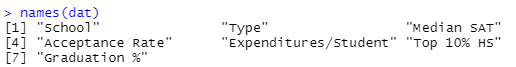
Intercept là giữa -31.5% và 67.41%

**Dùng R**

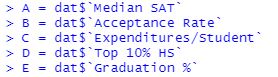
* Ta import file vào R

****

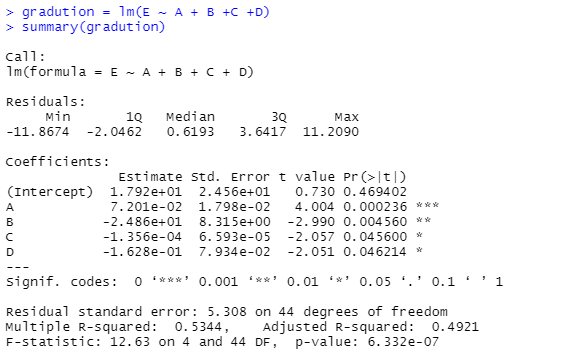
* Dữ liệu có các cột dữ liệu sau

****

* Đặt các biến theo từng cột dữ liệu

****

* Dùng câu lệnh lm() (viết tắt từ linear model) của R ta được kết quả như hình

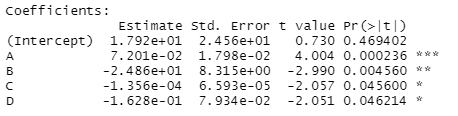


* Lệnh thứ hai, summary(graduation), yêu cầu R liệt kê các thông tin tính toán trong graduation
* Phần kết quả chia làm 3 phần:
  + Phần 1 mô tả phần dư (residuals) của mô hình hồi qui(Min, quantiles 25% (1Q), trung vị, quantiles 75% (3Q) và Max)



Mô hình càng tốt khi phần dư càng gần về không, ở đây trung vị phần dư 0.6, cũng không xa 0 là bao nhiêu, cho thấy phần dư của phương trình này tương đối cân đối.

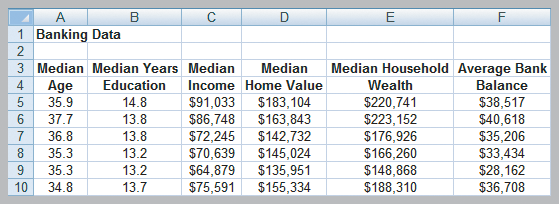
* + Phần hai trình bày ước số của b0, b1, b2, b3, b4 cùng với sai số chuẩn và giá trị của kiểm định t. Giá trị kiểm định t cho b1 là 7.201e-02 với trị số p = 0.000236, cho thấy b1 không phải bằng 0. Nói cách khác, chúng ta có bằng chứng để cho rằng có một mối liên hệ giữa Median SAT và tỷ lệ tốt nghiệp, và mối liên hệ này có ý nghĩa thống kê. Tương tự với các b2,b3,b4 , ta cũng kết luận một mối liên hệ giữa Acceptance Rate, Expenditure, Top 10% HS với tỷ lệ tốt nghiệp là những mối quan hệ có ý nghĩa thống kê.



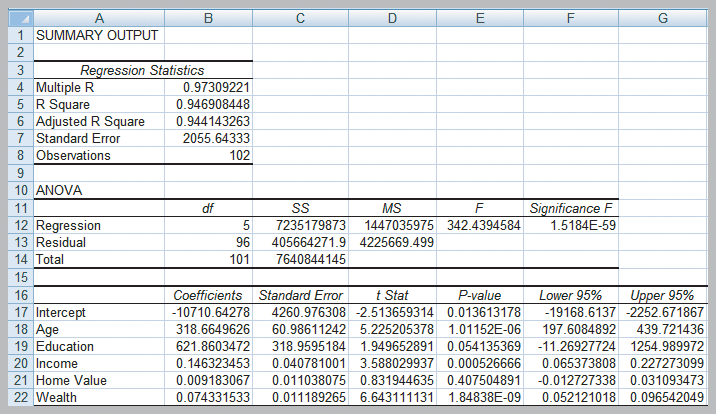
* + Phần ba của kết quả cho chúng ta thông tin về phương sai của phần dư (residual mean square)( Residual standard error). Ở đây, s^2 = 5.308. Trị số R2 hay hệ số xác định (coefficient of determination) là 0.5344, có nghĩa là phương trình tuyến tính giải thích khoảng 53,44% các khác biệt về tỷ lệ tốt nghiệp giữa các cá nhân. Hệ số điều chỉnh xác định (mà trong kết quả trên R gọi là “Adjusted R-squared”) là 0.4921. Đây là hệ số cho chúng ta biết mức độ cải tiến của phương sai phần dư (residual variance) do yếu tố các yếu tố như Median SAT, Top 10%HS,… có mặt trong mô hình tuyến tính. Giá trị p-value là 6.332e-07.

1. Example 9.11
2. Mô tả bài toán

* Tất cả các biến độc lập trong mô hình hồi quy tuyến tính không phải lúc nào cũng đáng kể. Hãy xây dựng các mô hình hồi quy tốt bao gồm bộ biến "tốt nhất".
* Dữ liệu lấy từ file “Banking Data.xlsx”
* Dữ liệu ngân hàng bao gồm thông tin nhân khẩu học của khách hàng trên thị trường hiện tại của ngân hàng.



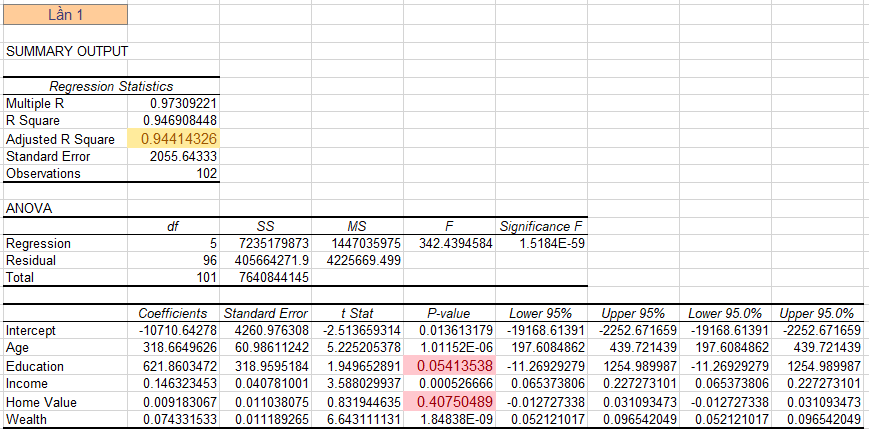
* Dự đoán số dư ngân hàng trung bình bằng cách sử dụng hồi quy



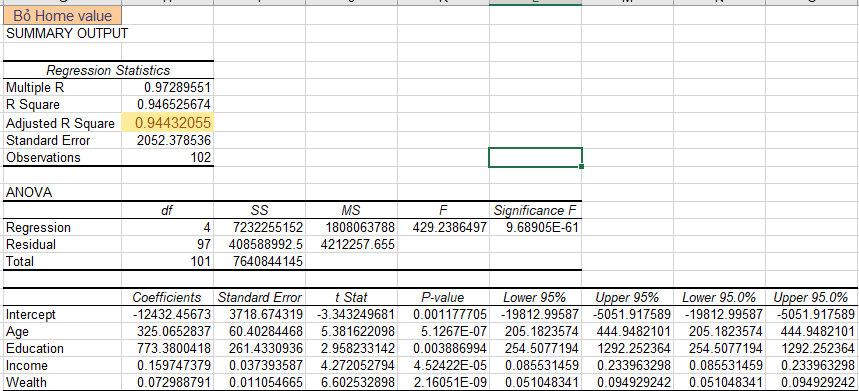
* Qua phân tích ta thấy p-value của Home Value và Education không có ý nghĩa thống kê

1. Cách xây dựng mô hình hồi quy đa biến tốt
   1. Xây dựng mô hình với tất các biến phụ thuộc và kiểm tra từng biến có ý nghĩa thống kê hay không
   2. Các giá trị p-value lớn hơn α, chọn giá trị p-value lớn nhất
   3. Loại bỏ biến phụ thuộc đó, tính lại R2 đã điều chỉnh ( Adjusted R square)
   4. Tiếp tục cho đến khi tất cả các biến đều có ý nghĩa thống kê.( có p-value < 0.05). Tìm mô hình có R2 đã hiệu chỉnh cao nhất.
2. Thực hiện

Thực hiện hồi quy tuyến tính bằng công cụ Data Analysis của Excel trên tất cả các biến phụ thuốc của file dữ liệu.



* Qua bảng kết quả, ta thấy p-value của home value và education lớn hơn 0.05 => 2 biến này không có ý nghĩa thống kê đối với biến độc lập số dư ngân hang
* Chọn giá giá trị p-value lớn nhất: biến Home value
* Ta bỏ biến Home value và thực hiện thống kê lại bảng hồi quy.



* Kết quả cho thấy tất cả các giá trị p-value đều nhỏ hơn 0.05. Điều đó chứng tỏ 4 biến phụ thuộc trên đều ý nghĩa đối với biến độc lập số dư ngân hàng\
* Thêm vào đó, giá trị R2 đã hiệu chỉnh sau khi tính lại lần thứ 2 lớn hơn lần thứ 1 (0.9443 > 0.9441)
* Như vậy ta đã tìm được mô hình hồi quy tuyến tính tốt:

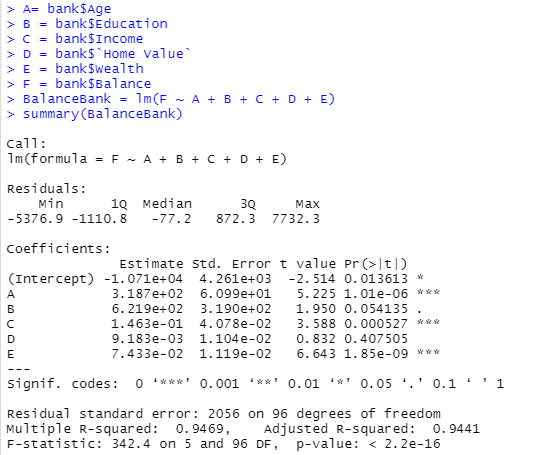
Y^ = -12432.45673 + 325.0652837\*Age + 773.3800418\*Education + 0.159747379\*Income + 0.072988791\*Wealth

**Dùng R**

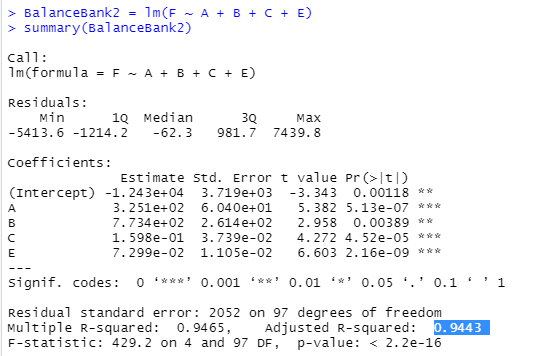
* Ta import file Banking Data.

****

* Ta đặt các biến để lấy dữ liệu từng cột dữ liệu
* Dùng câu lệnh lm() để tính và summary() để xem kết quả.

****

* Kết quả giá trị Ajusted R-square là 0.9441 và giá trị p-value của B (Education) và D(Home Value) lớn hơn 0.05. Ta bỏ giá trị D và tính lại.

****

* Giá trị Ajusted R-square tăng lên 0.9443 và giá trị p-value tất cả các biến đều nhỏ hơn 0.05
* Dù dùng Excel hay R ta đều có kết quả tương tự nhau

1. Example 9.12

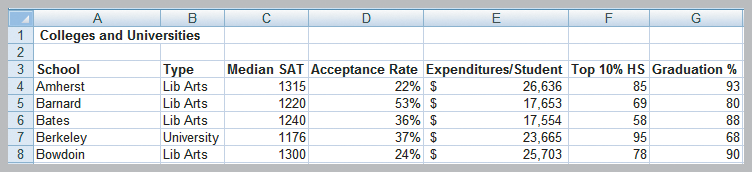
Một số cách khác để xây dựng một mô hình hồi quy tốt dựa vào 3 cách sau:

Khái niệm: Đa đối chiếu(Multicollinearity):

* + Xảy ra khi có mối tương quan (correlations) mạnh mẽ giữa các các biến độc lập.
  + Khó trách các tác động của biến độc lập
  + Dấu hiệu của hệ số dốc(slope) có thể dương giá trị đúng và p-value có thể tang.
* Mối tương quan vượt ±0,7 là một dấu hiệu cho thấy đa đối chiếu (Multicollinearity) có thể tồn tại.
* Các yếu tố làm phương sai tăng (Variance Inflation Factors) là một cách tốt hơn.
* Parsimony là một nguyên tắc lâu đời áp dụng ở đây.

1. Mô tả bài toán:

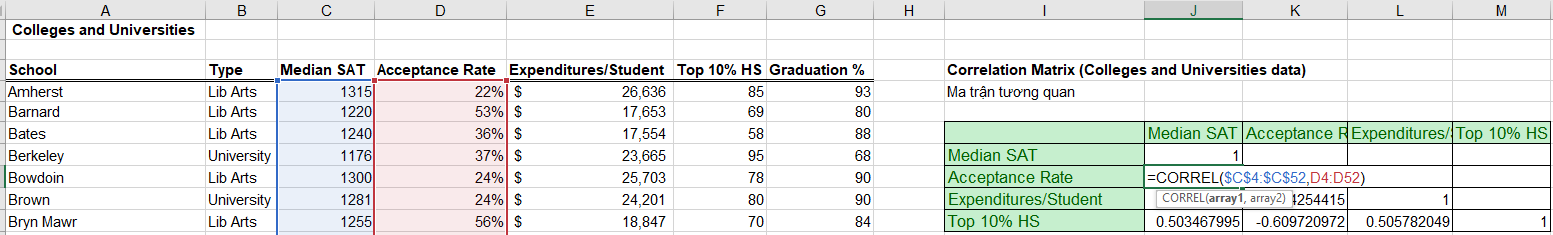
* Phân tích trên file *Colleges and Universities* Data.xlsx
* Yêu cầu: tính giá trị Multicollinearity tiềm năng

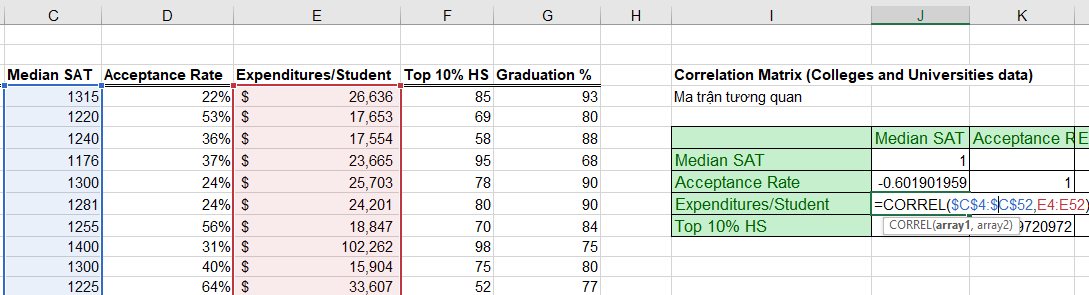


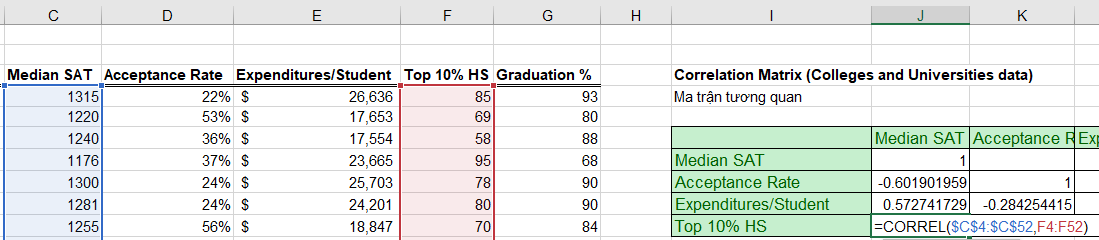
1. Thực hiện

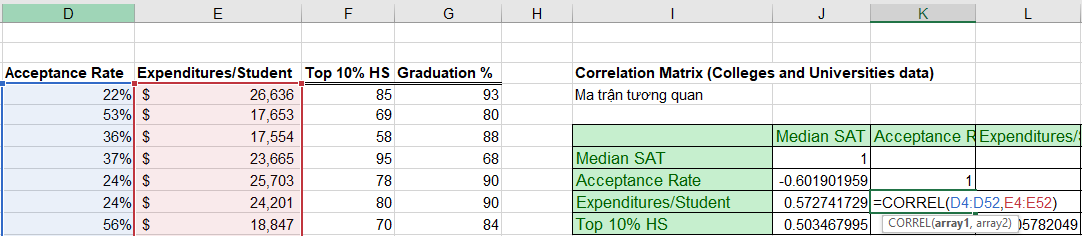
Để tính độ tương quan giữa 2 biến, ta dùng hàm CORREL() trong execl:

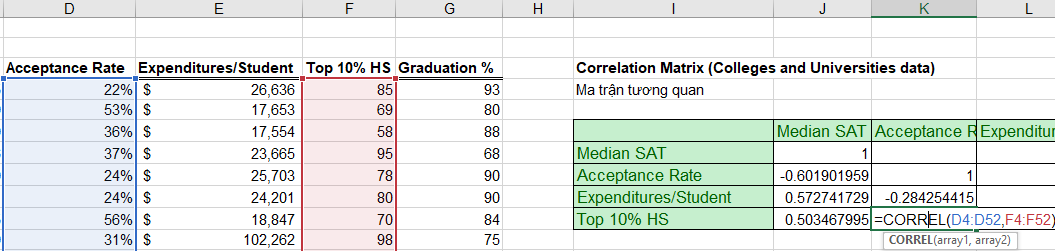
|  |
| --- |
| **CORREL(array1, array2)**  Cú pháp hàm CORREL có các đối số sau đây:   * **array1**    Bắt buộc. Một phạm vi giá trị ô. * **array2**    Bắt buộc. Phạm vi thứ hai chứa các giá trị ô |

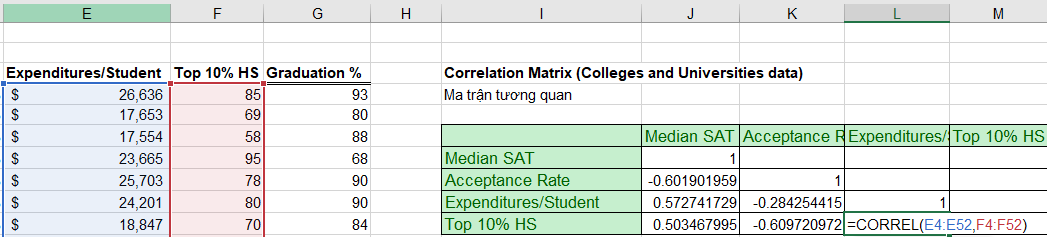




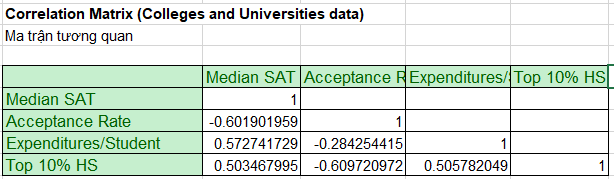








Ta được bảng hoàn chỉnh:



Hoặc

Ta áp dụng công thức

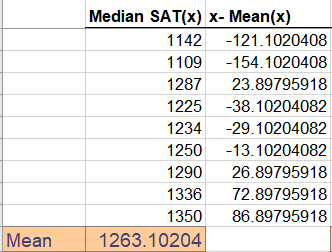
* Phương trình cho hệ số tương quan là:

Phương trình

trong đó  và là các giá trị trung bình của biến X và biến Y.

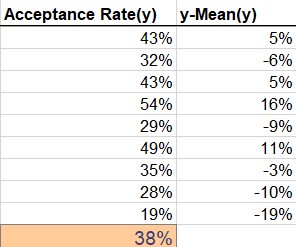
Ta tính giá trị tương quan giữa Median SAT và Acceptance Rate

* Tính giá trị trung bình của Median SAT (X) và lấy từng giá trị X trừ cho X trung bình

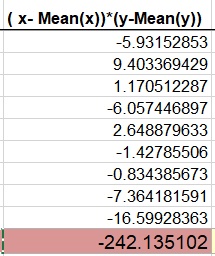


* Tính giá trị trung bình của Acceptance Rate

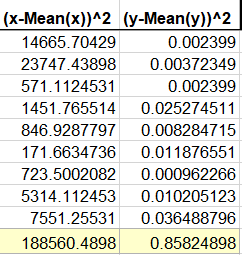
và lấy từng giá trị X trừ cho X trung bình



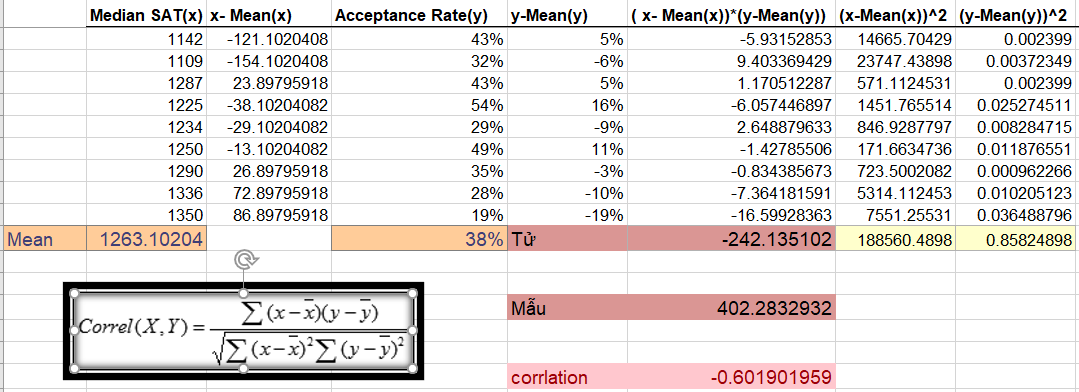
* Lấy từng giá trị X nhân với giá trị Y tương ứng, sau đó tính tổng, ta được giá trị tử số.



* Tính bình phương của hiệu X và X trung bình; của Y và Y trung bình. Sau đó tính tổng

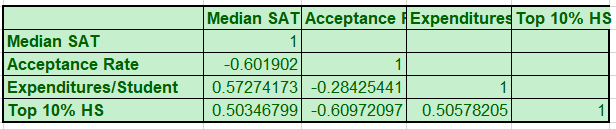


* Lấy 2 tổng vừa tính nhân lại, ta được giá trị mẫu



* Lấy tử chia mẫu, ta mức độ tương quan giữa median SAT và Acceptance Rate. So sánh với số tính bằng công thức Correl() thì bằng nhau

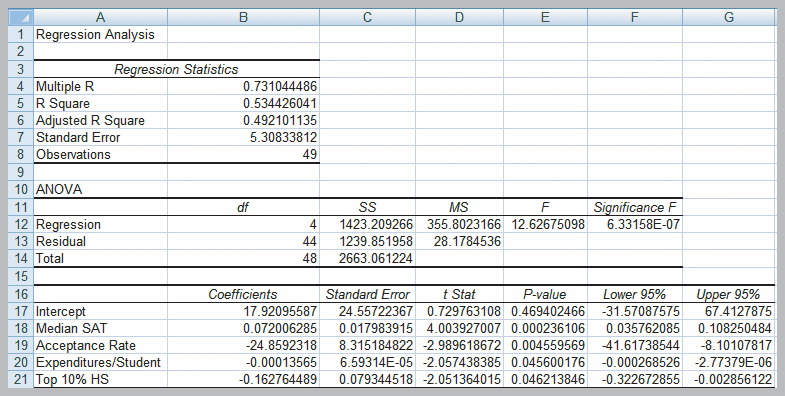
1. Nhận xét



Ta thấy có tất cả giá trị mối tương quan đều nằm trong khoảng [-0.7,0.7]. Suy ra có dấu hiệu của có tính đa đối chiếu

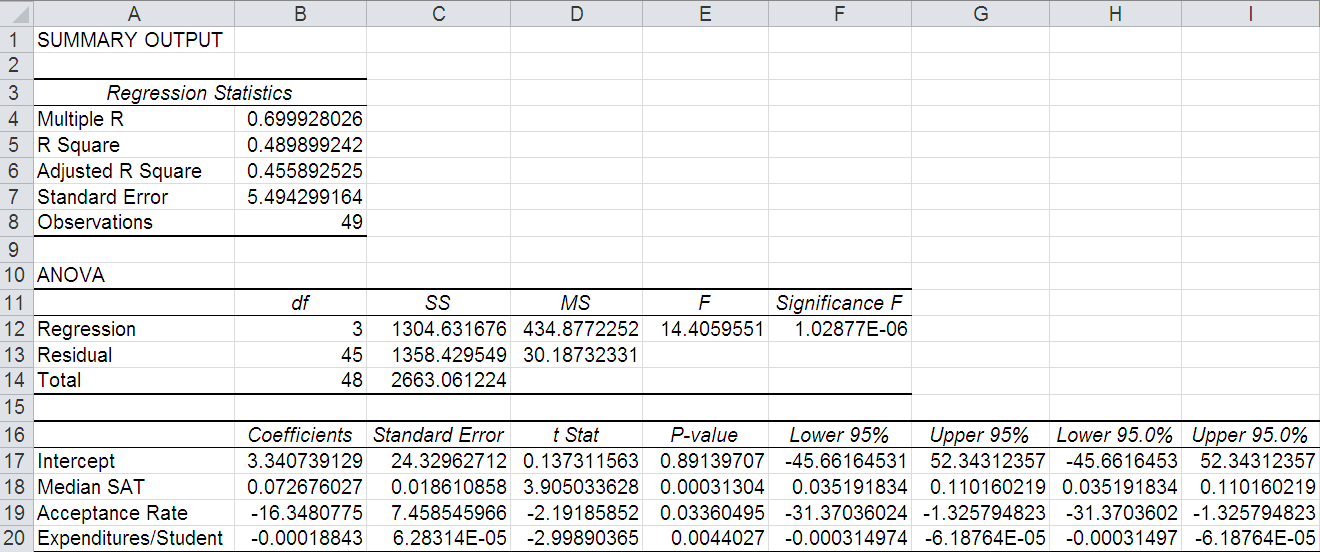
1. Theo ví dụ slide yêu cầu có 4 mô hình

Figure 9.13



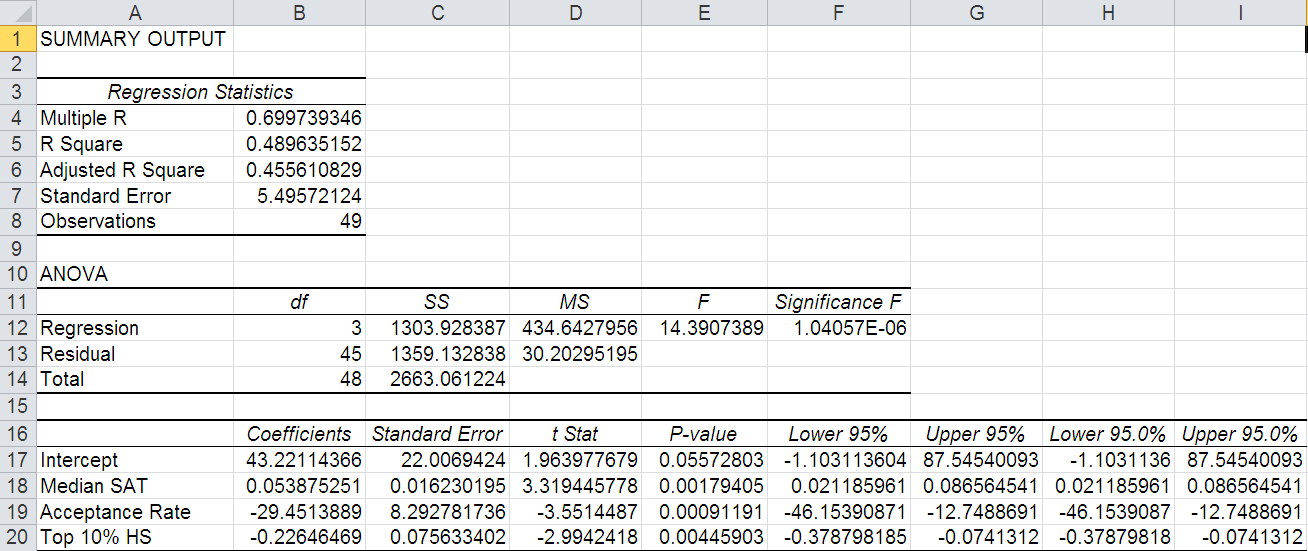
Full model

Adjusted *R*2 = 0.4921

****

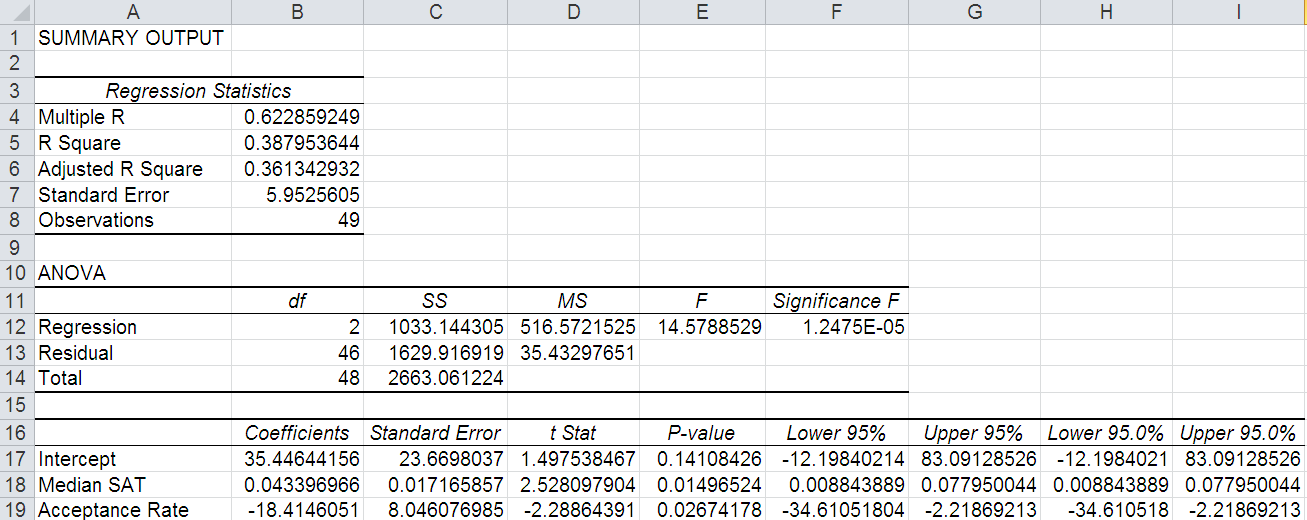
Dropping Top 10%

Adjusted *R*2 drops to 0.4559



Dropping Expenditures

Adjusted *R*2 drops to 0.4556



Dropping Expenditures and Top 10%

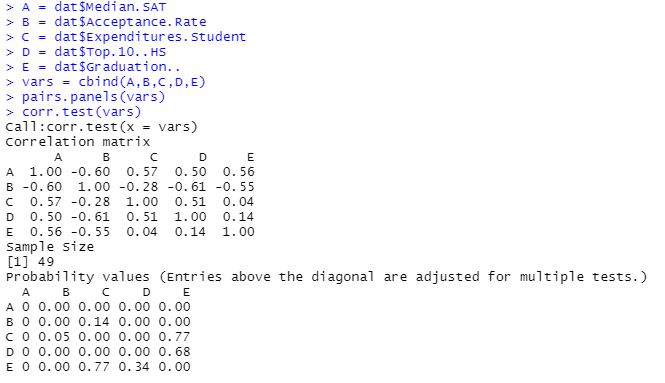
Adjusted *R*2 drops to 0.3613

Câu hỏi đặt ra ở đây là chọn mô hình nào?

Trả lời: theo như lý thuyết đã trình bày ở trên “R2 hiệu chỉnh chỉ số phản ánh mức độ phù hợp của mô hình” và R2 hiệu chỉnh càng lớn thì mô hình hồi quy đó càng tốt nên ta chọn mô hình 1 (“Full model”) với R2 hiệu chỉnh cao nhất.

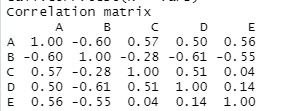
**Dùng R**

Ta dùng import file dữ và đặt các biến cho các cột dữ liệu.

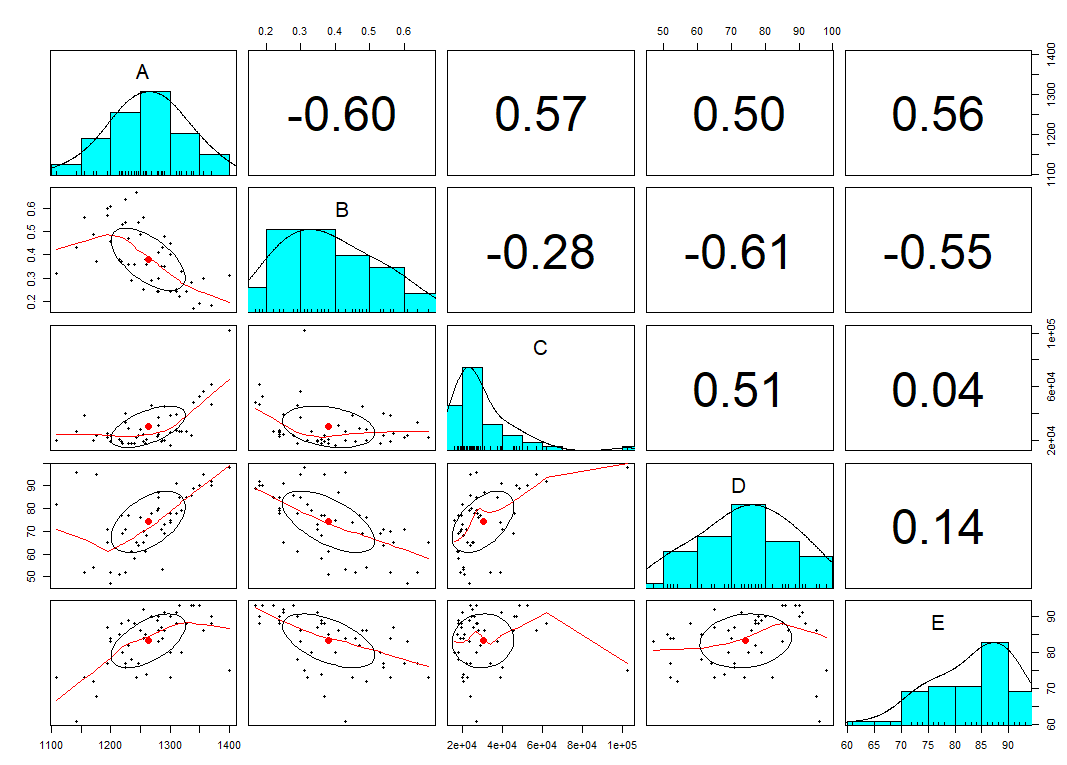


Câu lệnh corr.test() của R cho ta thấy được ma trận tương quan, kích thước mẫu và giá trị p-value.

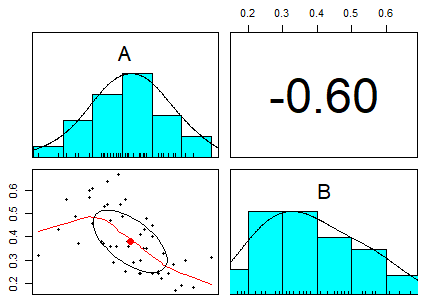
* Ma trận tương quan có giá trị tương như làm bằng excel



* Câu lệnh paris.panels cho ta thấy mức độ tương quan theo từng cặp biến A,B,C,D,E

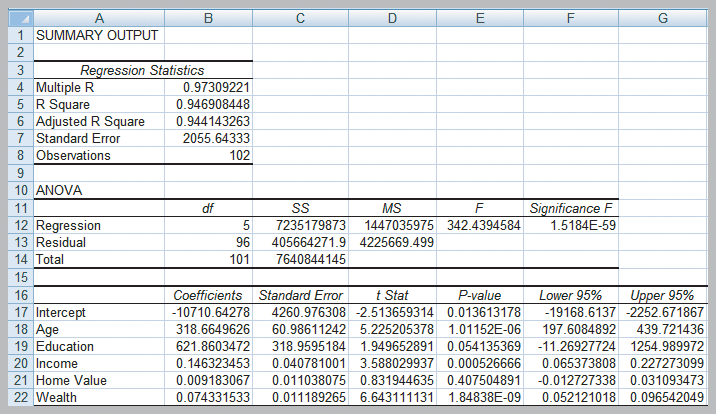


* A và B có r = -0.6 và hình vẽ tương ứng như sau:



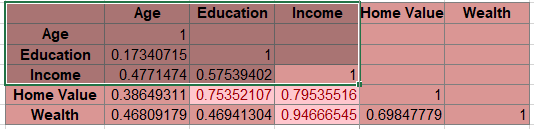
1. Example 9.12(continued)
2. Mô tả bài toán:

* Dùng file *Banking* Data.xlsx
* Yêu cầu xác định Multicollinearity

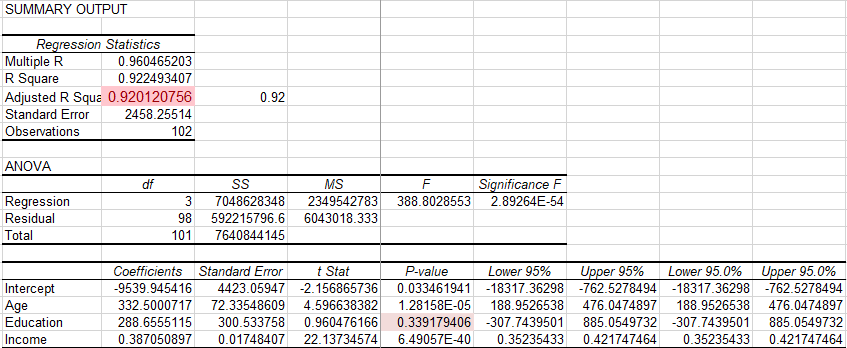


1. Thực hiện:

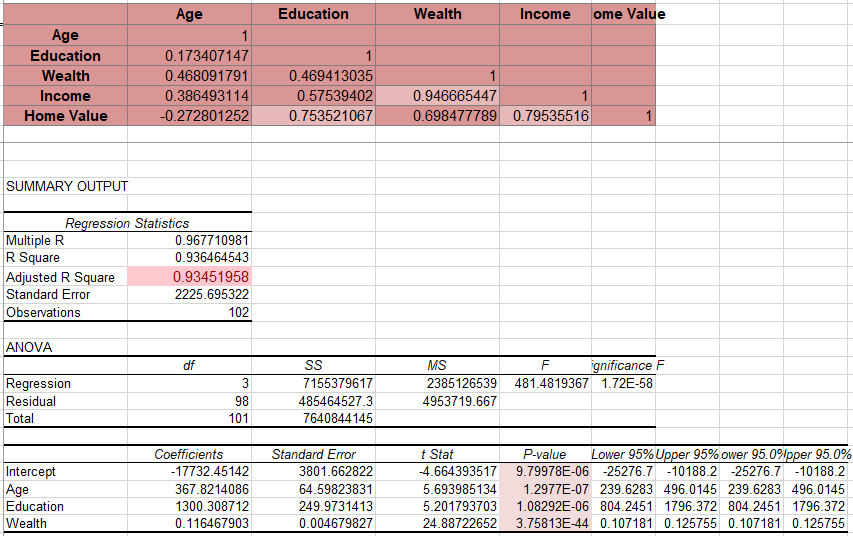
Dùng công thưc correl tương tự như trên ta được bảng như hình:



* Ta thấy có vài giá trị nằm ngoài vùng [-0.7,0.7] cho biến Home Value và biến Wealth.
* Loại bỏ biến Home Value và biến Wealth, tính mô hình hồi quy .



* Tính hồi quy tuyến tính, giá trị Education có p-value >0.05, biến này không có ý nghĩa thống kê
* Ta thử đổi vị trí các biến trong ma trận tương quan và tính lại ma trận tương quan. Ở ma trận tương quan vừa tạo, ta thấy ở biến Home Value và Income có vài giá trị vượt quá khoảng [-0.7,0.7]. Loại bỏ biến Home Value và Income, tính lại mô hình hồi quy



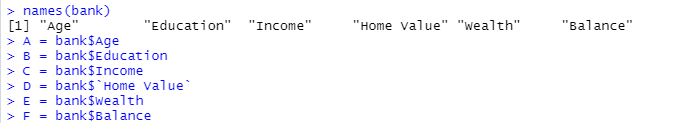
* Kết quả cho ta thấy tất cả các biến đều có ý nghĩa thống kê( vì tất các giá trị biến p-value đều lớn hơn 0.05)

**Dùng R**

* Đọc dữ liệu từ fiel csv.



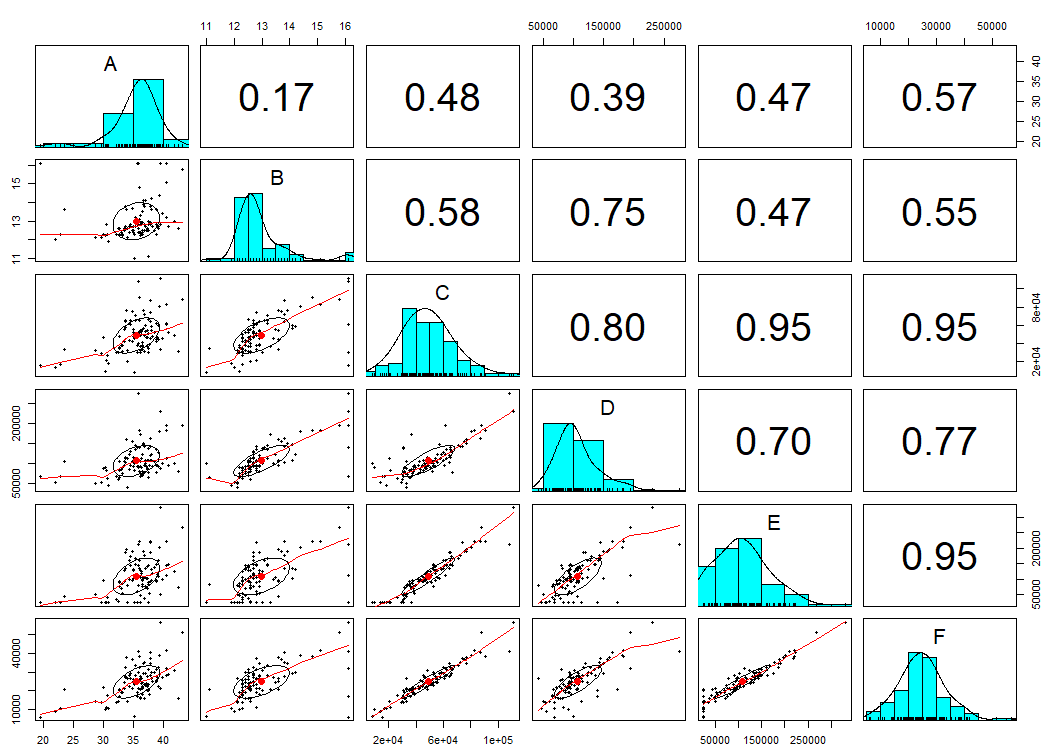
* Đặt các biến A đến F để lưu các cột dữ liệu



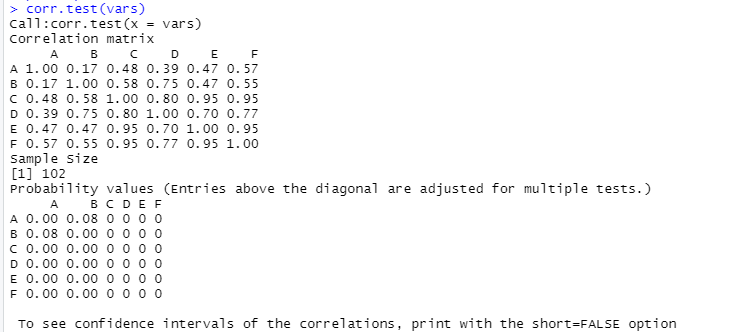
* Gọi hàm pairs.panels để sao mức độ tương quan giữa các biến bằng hình vẽ.



* Đối mỗi hình hoặc một con số, ta giống ngang và giống dọc để xem 2 biến đang xét. Ví dụ: mức độ tương quan giữa biến A và biến B là 0.17.



* Dùng cậu lệnh corr.test để xem ma trận tương quan.



* Ma trận tương ở đây thayfsw tương tự với ma trận tương quan tính trong Excel.

**Tài liệu tham khảo:**

Silde chương 9-Regression Analysis.pptx của thầy Nguyễn Đình Thuân

## [1] Ý nghĩa R bình phương

<https://phantichspss.com/r-binh-phuong-r-binh-phuong-hieu-chinh-cong-thuc-y-nghia-cach-tinh-thu-cong-va-cach-tinh-bang-spss.html>

[2] <https://xulydinhluong.com/y-nghia-cua-gia-tri-r-binh-phuong-hieu-chinh-trong-hoi-quy/#2_Gia_tri_R_binh_phuong_hieu_chinh_la_gi>

Tham khảo cách dùng R:

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

<https://www.youtube.com/watch?v=bTzGlI2Kjtg&list=UU21dOPe-YHO3Gw6BRbyeotQ&index=88>

Công thức R2 và R2 hiệu chỉnh:

<https://www.greycampus.com/blog/data-science/linear-regression-with-python-scikit-learn>