**Homeword 7**

***Thành viên nhóm:***

|  |  |  |
| --- | --- | --- |
| STT | MSSV | *Họ và tên* |
| 1 | 18521320 | *Đoàn Thục Quyên* |
| 2 | 18521554 | *Nguyễn Thành Trung* |
| 3 | 18520454 | *Nguyễn Đức Anh* |

***Phân công công việc:***

|  |  |  |  |
| --- | --- | --- | --- |
|  | Đoàn Thục Quyên  18521320 | Nguyễn Thành Trung  18521554 | Nguyễn Đức Anh  18520454 |
|  | | | |
| Câu 1.1 | x |  |  |
| Câu 1.2 |  | x |  |
| Câu 2 |  |  | x |

## Câu 1.1 Cách xác định phương trình hồi quy có phù hợp không?

Mục tiêu đầu tiên của một nhà kinh tế lượng là làm sao sử dụng dữ liệu thu thập được để ước lượng hàm hồi quy của tổng thể. Thực vậy, đây là một đòi hỏi cao, vì không có một mô hình hoàn hảo như thế.

Nếu chúng ta muốn đưa vào mọi yếu tố riêng lẻ có ảnh hưởng đến một đối tượng nghiên cứu cụ thể, thì mô hình sẽ quá phức tạp không có giá trị ứng dụng thực tế. Xác định đúng dạng mô hình có nghĩa là chúng ta đạt được một hoặc nhiều điều kiện sau đây:

1. Mô hình không loại trừ bất kỳ các biến ‘cốt lõi’ nào.

2. Mô hình không đưa vào các biến không cần thiết.

3. Không có các lỗi đo lường trong biến phụ thuộc và các biến giải thích.

5. Các quan sát bất thường trong dữ liệu, nếu có, đều được tính đến.

6. Phân phối xác suất của hạng nhiễu được xác định đúng.

7. Điều gì xảy ra nếu các biến giải thích là ngẫu nhiên?

8. Vấn đề phương trình đồng thời: chệch do tính đồng thời.

Nếu số liệu sai ( không thỏa 1 trong 8 điều kiện trên ) thì kết quả sẽ sai theo 1 trong 3 dạng:

- Đa cộng tuyến

- Phương sai của sai số thay đổi

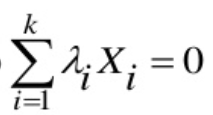
- Tự tương quan

1. ĐA CỘNG TUYẾN
2. ĐA CỘNG TUYẾN LÀ GÌ?

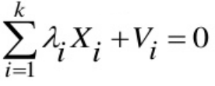
* Trong mô hình hồi quy bội, ta có giả thiết: các biến độc lập trong mô hình không có quan hệ tuyến tính. Nếu giả thiết này bị vi phạm thì sẽ dẫn đến hiện tượng đa cộng tuyến.
* Đa cộng tuyến là hiện tượng các biến độc lập trong mô hình phụ thuộc tuyến tính lẫn nhau, thể hiện được dưới dạng hàm số

1. PHÂN LOẠI ĐA CỘNG TUYẾN

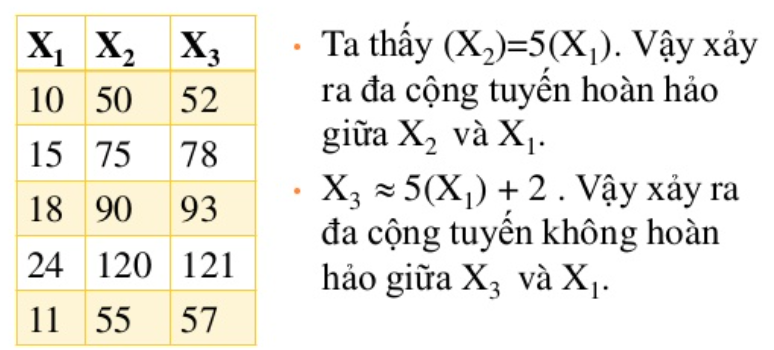
* Đa cộng tuyến hoàn hảo:  không đồng thời bằng 0 sao cho:



* Đa cộng tuyến không hoàn hảo:



* Ví dụ:



1. HẬU QUẢ CỦA ĐA CỘNG TUYẾN

* Khi xảy ra đa cộng tuyến hoàn hảo: không xác định được các hệ số β. (Thông thường không xảy ra trường hợp này)
* Khi xảy ra đa cộng tuyến không hoàn hảo:

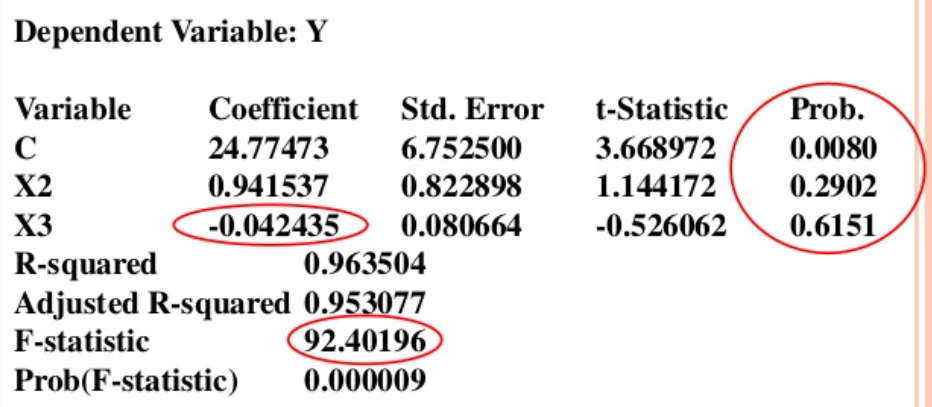
1. Phương sai var(β^) lớn => khoảng tin cậy các hệ số β lớn => vẫn xác được các β nhưng không chính xác, dấu có thể sai.
2. Mâu thuẫn giữa kiểm định t và F:

* Se (β^) lớn t nhỏ, tăng khả năng chấp nhận Ho => Biến X không ảnh hưởng đến Y
* R2 cao => F lớn => bác bỏ Ho của thống kê F => các biến X có ảnh hưởng đến Y

1. Khi thay đổi độ lớn của dữ liệu hoặc khi thêm vào (bớt đi) biến cộng tuyến, thì mô hình sẽ thay đổi về dấu hoặc độ lớn các hệ số ước lượng.
2. CÁCH PHÁT HIỆN ĐA CỘNG TUYẾN
3. **Phương pháp 1**: Dấu của β^ mâu thuẫn với lý thuyết kinh tế

Mâu thuẫn : R2 lớn, t nhỏ ( hậu quả thứ 2 của đa cộng tuyến)

* T-Statistic: Một thống kê tuân theo phân phối T. Thống kê T thường được dùng trong kiểm định giả thiết để xác định MỨC Ý NGHĨA THỐNG KÊ của các tham số trong các mô hình kinh tế lượng, và được tính bằng tỷ số giữa giá trị tham số ước tính và sai số chuẩn của nó.



* VÍ DỤ 1: Khảo sát chi tiêu, thu nhập và mức độ giàu có, người ta thu được kết quả sau:
* Trong đó: Y: chi tiêu ($), X2: thu nhập ($), X3: mức độ giàu có ($)
* Hệ số hồi quy của biến X3 là (-0,042435) không phù hợp lý thuyết kinh tế vì …

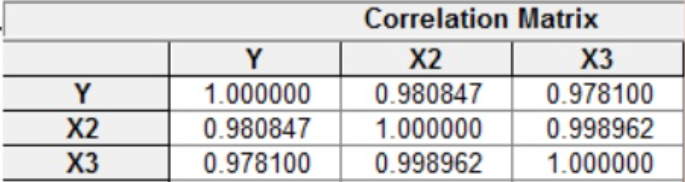
P\_value ứng với X2 = 0,2902

P\_value ứng với X3 = 0,6151

⇒ X2 và X3 không có ý nghĩa (1)

* R2 = 0,963504 , khá lớn ⇒ mô hình có ý nghĩa (2)
* (1) và (2) mâu thuẫn với nhau. Các dấu hiệu này cho thấy có thể có hiện tượng cộng tuyến xảy ra trong mô hình trên

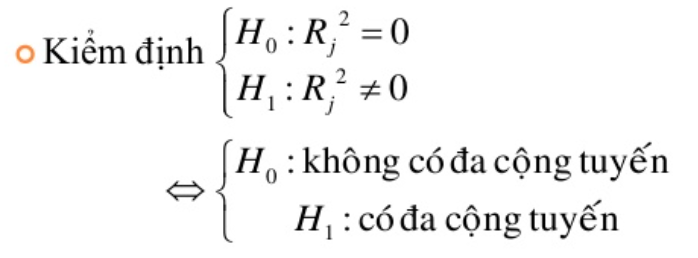
1. **Phương pháp 2**: Hệ số tương quan giữa cặp biến giải thích cao (> 0,7) là xem như có đa cộng tuyến) Ví dụ 1.1:



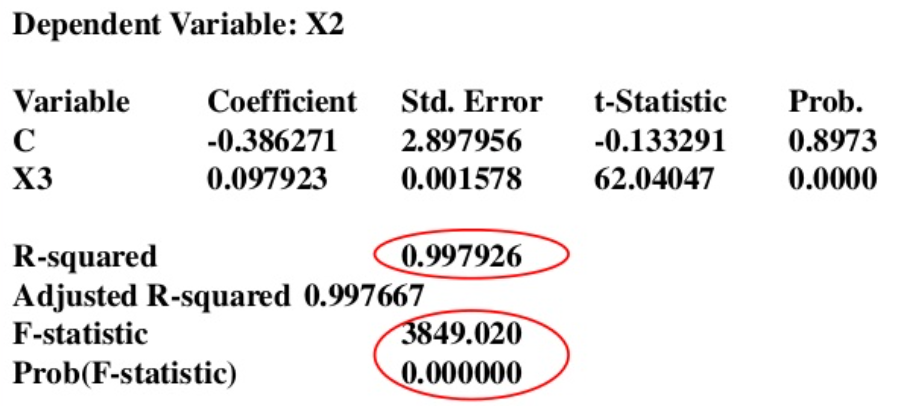
* Hệ số tương quan giữa X2 và X3 là 0,9989 (rất cao) => đa cộng tuyến gần hoàn hảo

1. **Phương pháp 3**. Dùng mô hình hồi quy phụ

* Hồi quy giữa 1 biến giải thích Xj với các biến giải thích còn lại. Tính Rj^2 và Fj của mô hình hồi quy phụ.

****

* Nếu bác bỏ Ho thì có xảy ra đa cộng tuyến.
* Ví dụ 1.2: Xét mô hình hồi quy phụ sau:

****

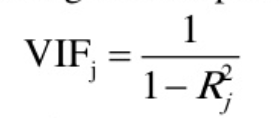
* Trong mô hình hồi quy phụ X2 theo X3 ta thấy:

Dựa vào F=3849,02 và Prob(F-statistic)=0,000

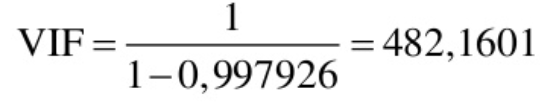
⇒ bác bỏ giả thiết Ho : R2 = 0.

Vậy X3 có ảnh hưởng đến X2 ⇒ có xảy ra hiện tượng đa cộng tuyến.

1. **Phương pháp 4**: Dùng nhân tử phóng đại phương sai.



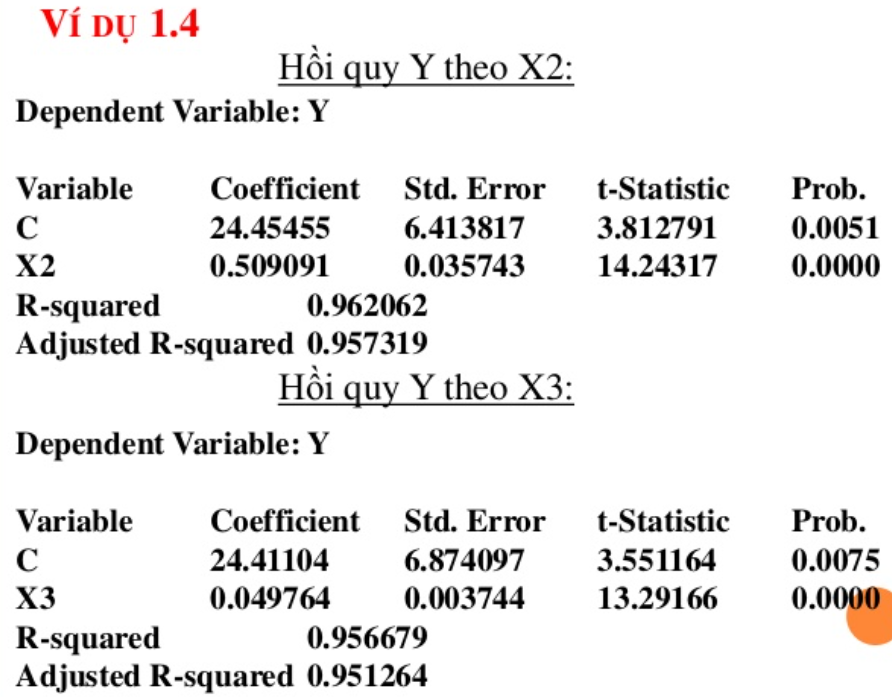
* Trong đó Rj^2 là hệ số xác định trong hàm hồi quy phụ
* Xj theo các biến giải thích còn lại trong mô hình.
* Nếu VIF > 10 thì có đa cộng tuyến.
* Ví dụ 1.3: lấy dữ liệu có được từ ví dụ 1.2



VIF rất lớn => xảy ra cộng tuyến giữa X2 và X3

1. KHẮC PHỤC ĐA CỘNG TUYẾN
2. Sử dụng thông tin tiên nghiệm
3. Loại trừ 1 biến giải thích ra khỏi mô hình
4. Thu thập thêm số liệu hoặc lấy mẫu mới
5. Sử dụng sai phân cấp 1

* **Loại trừ 1 biến giải thích ra khỏi mô hình**
* Xác định các biến cộng tuyến với nhau
* Tìm hệ số xác định hiệu chỉnh‾R2 của mô hình: có tất cả biến, không có mặt 1 biến cộng tuyến.
* Dùng hệ số xác định hiệu chỉnh‾R2 để xác định biến cộng tuyến nào là biến cần loại bỏ khỏi mô hình.

****

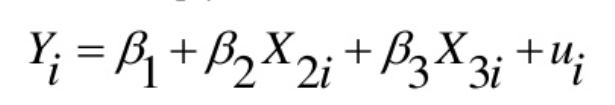
1. PHƯƠNG SAI THAY ĐỔI
2. Phương sai của sai số thay đổi là gì?

- Theo giả thiết A2: var(Ui) = σ2 = const : phương sai của sai số không đổi

- Khi vi phạm giả thiết : var(Ui) = σi 2 ≠ const: phương sai của sai số thay đổi

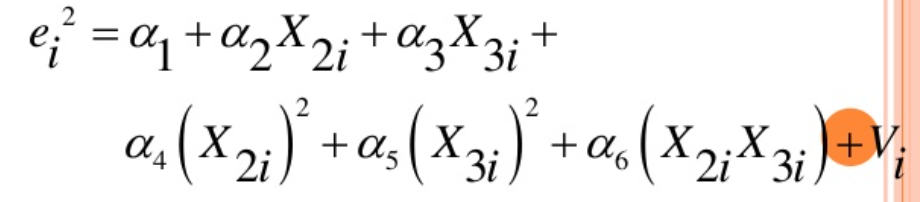
1. Phương pháp kiểm định White

* Xét mô hình hồi quy 3 biến

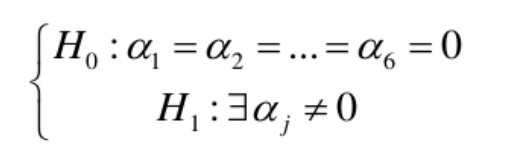


Bước 1: Ước lượng hàm hồi quy, thu được ei

Bước 2: Ước lượng mô hình sau



Bước 3: kiểm định giả thiết



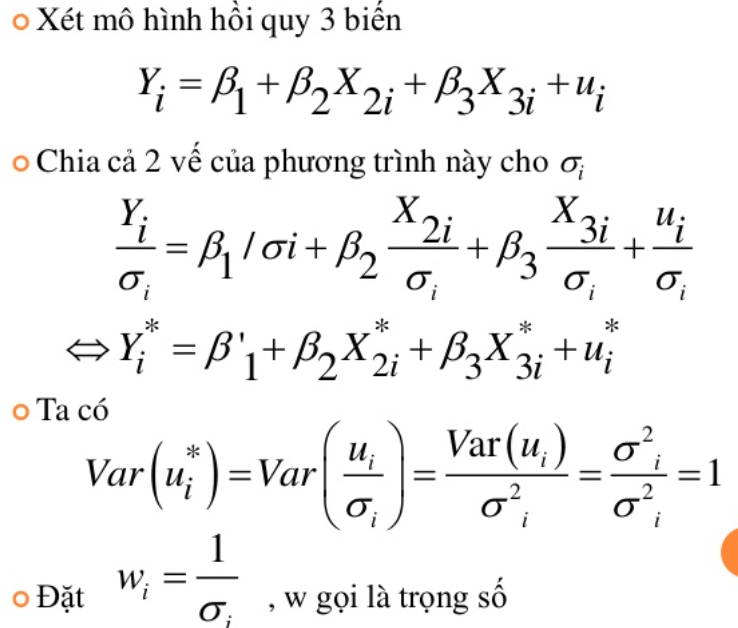
Ho: “ không có hiện tượng phương sai thay đổi” ;

H1: “ có hiện tượng phương sai thay đổi”,

* Tính giá trị thống kê (n.R2), trong đó n là cỡ mẫu và R2 là hệ số xác định của mô hình hồi quy phụ ở bước 2
* Tra bảng phân phối Chi bình phương, 
* Nếu  thì bác bỏ giả thiết Ho, kết luận có hiện tượng phương sai thay đổi.

1. KHẮC PHỤC PHƯƠNG SAI THAY ĐỔI

* Phương pháp bình phương nhỏ nhất có trọng số
* Ý tưởng của phương pháp: Bộ dữ liệu thu được ban đầu có hiện tượng phương sai thay đổi, do đó các ước lượng  không còn chính xác nữa. Để tìm được hàm hồi quy, ta cần tìm  bằng cách tạo một bộ dữ liệu mới bằng dữ liệu cũ chia cho trọng số.
* Trọng số là số được tạo ra từ nguyên nhân gây ra phương sai thay đổi.



## Câu 1.2 Cách xác định hệ số hồi qui có phù hợp không?

Mô hình phát sinh (P) và thu hút (A) chuyến đi được xây dựng là các hàm hồi quy tuyến tính 3 biến với các biến là số lượng dân số, số lượng lao động và số lượng học sinh sinh viên. Cụ thể kết quả tính toán như sau:

Mô hình phát sinh chuyến đi: Pi = ai\*Population + bi\*Workers + ci\*Students

Mô hình thu hút chuyến đi: Aj = xj\*Population + yj\*Workers + zj\*Students

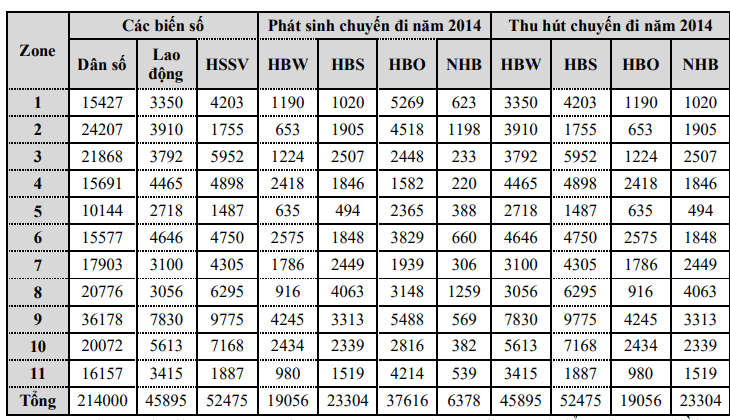
Trong đó:

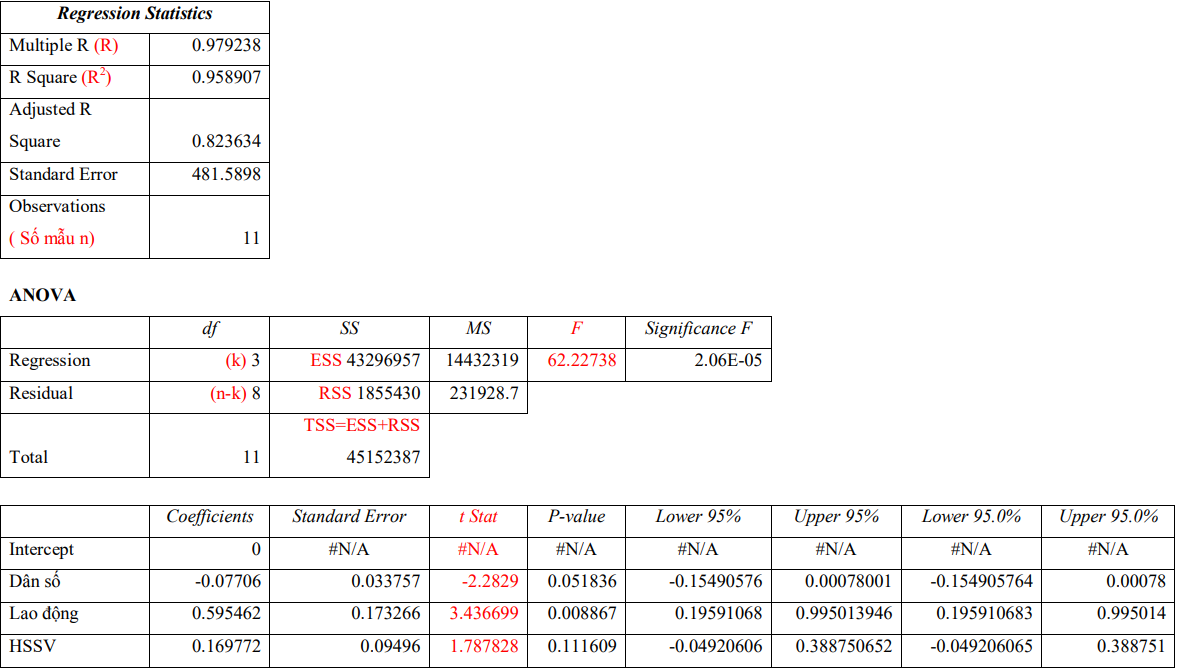
Population : Dân số

Workers : Số lao động

Students : Số HSSV

ai , bi, ci, xj, yj, zj là tham số, hệ số hồi quy tương ứng



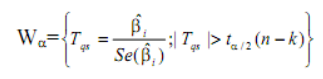


Kiểm định giả thiết với hệ số hồi quy Dân số.

Sau khi tiến hành kiểm định sự phù hợp của mô hình ta cũng cần phải kiểm tra từng hệ số cụ thể trong mô hình hồi quy để khẳng định sự tồn tại hoặc có thể nhận một giá trị cụ thể nào đó của các hệ số.

Kiểm định cặp giả thuyết: H0 : β1 = β0 và H1 : β1 ≠ β0 trong đó β1 là hệ số hồi quy của biến độc lập Dân số, β0 = 0.

Tiêu chuẩn kiểm định: 

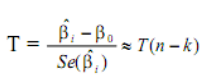
Miền bác bỏ đối với H0 là: 

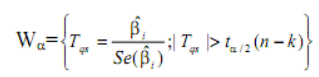
Với tα/2 (n-k) có thể tính được bằng cách sử dụng hàm TINV trong Excel.

ta có: |Tqs| = 3.4366 và tính được t0.025(8) theo công thức =TINV(0.025,8) = 2.7515 nên Tqs > t0.025(8). Do đó bác bỏ giả thiết H0 tức là ở mức ý nghĩa 5% có thể khẳng định có sự ảnh hưởng của Lao động đến Số chuyến đi Phát sinh với mục đích HBW.

Kiểm định giả thiết với hệ số hồi quy Số HSSV.

Kiểm định cặp giả thuyết: H0 : β3 = β0 và H3 : β3 ≠ β0 trong đó β3 là hệ số hồi quy của biến độc lập HSSV, β0 = 0.

Tiêu chuẩn kiểm định: 

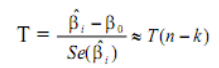
Miền bác bỏ đối với H0 là: 

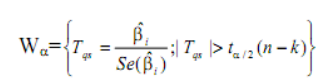
Với tα/2 (n-k) có thể tính được bằng cách sử dụng hàm TINV trong Excel.

ta có: |Tqs| = 1.7878 và tính được t0.025(8) theo công thức =TINV(0.025,8) = 2.7515 nên – t0.025(8) < t0.025(8). Do đó, chưa có cơ sở bác bỏ H0 tức là ở mức ý nghĩa 5% chưa thể khẳng định có sự ảnh hưởng của Dân số đến Số chuyến đi Phát sinh với mục đích HBW

Kiểm định giả thiết với hệ số hồi quy\_Sô Lao động.

Tương tự như kiểm định hệ số hồi quy Dân số ta kiểm định hệ số hồi quy Số lao động. Kiểm định cặp giả thuyết: H0 : β2 = β0 và H2 : β2 ≠ β0 trong đó β2 là hệ số hồi quy của biến độc lập Lao động, β0 = 0.

Tiêu chuẩn kiểm định: 

Miền bác bỏ đối với H0 là: 

Với tα/2 (n-k) có thể tính được bằng cách sử dụng hàm TINV trong Excel.

ta có: |Tqs| = 3.4366 và tính được t0.025(8) theo công thức =TINV(0.025,8) = 2.7515 nên Tqs > t0.025(8). Do đó bác bỏ giả thiết H0 tức là ở mức ý nghĩa 5% có thể khẳng định có sự ảnh hưởng của Lao động đến Số chuyến đi Phát sinh với mục đích HBW.

VÌ VẬY CÁCH XÁC ĐỊNH HỆ SỐ HỒI QUY PHÙ HỢP

## 2. Hiện nay có các phương pháp nào để đánh giá độ chính xác của các phương trình hồi qui tìm được.

**t – test, kiểm định t [i]**

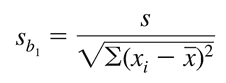
Đặt các giả thuyết như sau, tùy theo mục đích kiểm định:

H0: β1 = 0                                        H0: β1 ≤ 0                                   H0: β1≥ 0

H1: β1 ≠ 0                                        H1: β1 > 0                                   H1: β1 < 0

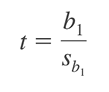
Tuy nhiên để kiểm định cho giá trị β1cho tổng thể phải dựa vào giá trị b1tìm được từ bộ dữ liệu mẫu, do đó phải xem xét phân phối mẫu của b1 có đủ điều kiện để thực hiện kiểm định hay không. Phân phối mẫu của b1 có các tính chất như sau:

* Giá trị kỳ vọng E(b1) = β1
* Độ lệch chuẩn:



* Phân phối của mẫu thuộc dạng phân phối chuẩn

Như vậy sau cùng công thức tổng quát của giá trị kiểm định t có dạng như sau:



Nguyên tắc bác bỏ H0:

Với kiểm định 2 phía: H0: β1 = 0 được bác bỏ khi trị tuyệt đối của t lớn hơn t tra bảng (tα/2, n-2)

Với kiểm định bên phải: H0: β1≤ 0 được bác bỏ khi giá trị t dương lớn hơn giá trị dương của t tra bảng (tα, n-2)

Với kiểm định bên trái: H0: β1 ≥ 0 được bác bỏ khi giá trị t < giá trị âm của t tra bảng (tα, n-2)

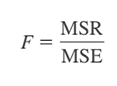
Nếu xét trên giá trị p-value, bác bỏ H0 khi p-value < α. (p-value là mức ý nghĩa nhỏ nhất mà tại đó H0 bị bác bỏ).

**F** – **test, kiểm định F [i]**

Tương tự như kiểm định t, kiểm định F, dựa trên phân phối F – một dạng phân phối xác suất, cũng sẽ xác định bác bỏ hay không bác bỏ giả thuyết H0: β1 = 0. Tuy nhiên khác với kiểm định t, kiểm định F có thể kết luận có hay không có mối quan hệ tuyến tính giữa một biến phụ thuộc y và nhiều biến độc lập x một cách tổng quát**.** Trong bài viết này em chỉ xét trường hợp một biến x.

Kiểm định F ngoài sử dụng MSE, ước lượng phương sai của sai số, mà còn sử dụng thêm **MSR** **(mean square regression)**được tính bằng cách lấy SSR chia cho bậc tự do của phương trình hồi quy, là số biến độc lập có trong phương trình, ở đây em chỉ xét 1 biến x nên bậc tự so bằng 1. MSR chính là giá trị ước lượng phương sai σ2dựa trên SSR

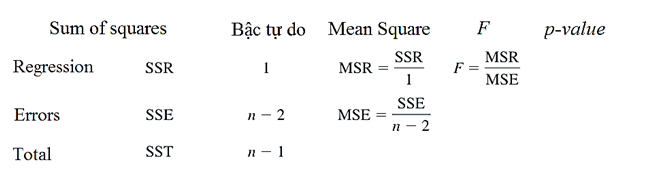
Lưu ý, kiểm định F cho hệ số hồi quy β1chủ yếu là **kiểm định 1 phía**, sử dụng mức ý nghĩa α cho trước. Công thức tổng quát của giá trị kiểm định F là:



Cơ sở bác bỏ H0:

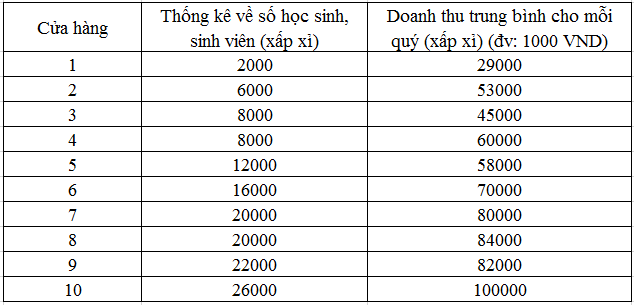
* p – value nhỏ hơn mức ý nghĩa α
* giá trị F tính được phải lớn hơn giá trị F tra bảng phân phối F (với bậc tự do thứ nhất là 1 ở hàng trên cùng, và bậc tự do thứ hai là n – 2 ở cột ngoài cùng, α ở cột thứ 2 tính từ cột ngoài cùng)

Bảng ANOVA (phân tích phương sai) tổng quát như sau:



Ví dụ:

Giá sử một chuỗi cửa hàng gà rán có 10 cửa hàng nằm tại các quận khác nhau trên thành phố Hồ Chí Minh, tại từng khu vực của từng cửa hàng sẽ có số lượng học sinh, sinh viên sinh sống và học tập, chuỗi cửa hàng này muốn biết rằng doanh thu của từng cửa hàng có mối liên hệ nào với số lượng học sinh, sinh viên này không (dựa trên dữ liệu doanh thu trung bình theo quý của mỗi cửa hàng, và dữ liệu thống kê về số lượng học sinh, sinh viên)

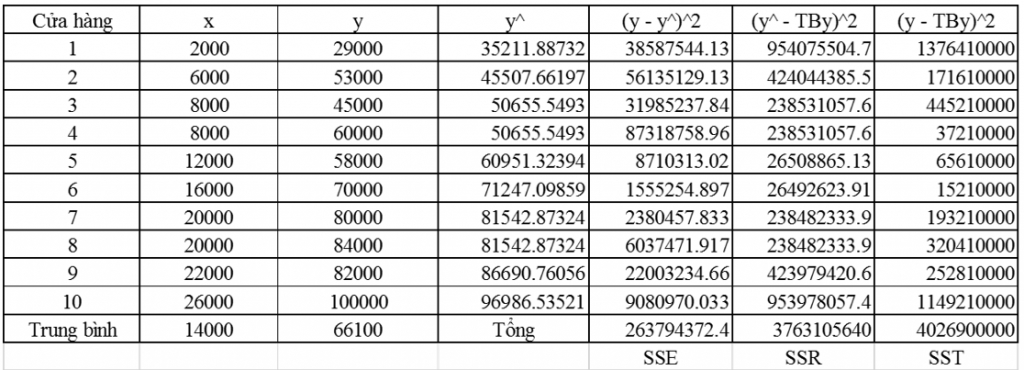


Gọi x là số lượng HS, SV, y là doanh thu, là biến mục tiêu dự báo, em sẽ sử dụng công thức tính bo và b1 để lập phương trình, chúng ta tính được trung bình x: TBx = 14000, trung bình của y: TBy = 66100 Significance F là p-value, các bạn cũng có thể sử dụng hàm F.DIST trong excel để tính.

Với b1= 2.57. Với số lượng HS, SV tăng 1000 thì doanh thu của 1 cửa hàng sẽ tăng 2570000 VND, và b0 = 30064.

Phương trình hồi quy tuyến tính đơn giản: Y^ = 2.57X + 30064

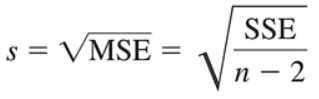
Thay từng giá trị x vào phương trình mới tìm được để tính giá trị Y^ dự báo, sau đó tính SSE, SSR dựa trên Y^.



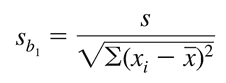
Đặt giả thuyết:     H0: β1 = 0 : Không có mối quan hệ giữa số HS, SV và doanh thu  mỗi cửa hàng

H1: β1 ≠ 0  Có mối quan hệ giữa số HS, SV và doanh thu  mỗi cửa hàng.

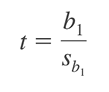
Kiểm định t:



S = √(SSE/n – 2) = √(263794372.4/8) = 5742.3



Sb = 0.24



t = b/Sb = 2.57/0.24 = 10.7

t tra bảng với mức ý nghĩa α = 5%, bậc tự do là 8, tα/2,8 = 2.306

Như vậy với t > tα/2,8 bác bỏ giả thuyết H0, tức có mối quan hệ giữa 2 biến số lượng học sinh sinh viên trong khu vực và doanh thu của mỗi nhà hàng trong khu vực ấy.

Tương tự giá trị p-value = 0.000005 (có thể dùng hàm T.DIST trong excel để tìm p-value) tính được nhỏ hơn rất nhiều so với mức ý nghĩa, nên em bác bỏ H0.

Còn với kiểm định F:

MSE = SSE/(n – 2) = 263794372.4/8 = 32974295.77

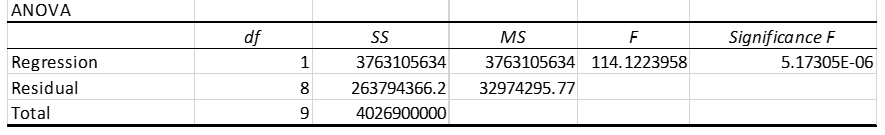
MSR = SSR/1 = 3763105640

F = MSR/MSE = 114

F tra bảng với mức ý nghĩa 5%, Fα,1,8 = 5.32

F > Fα,1,8, bác bỏ H0 tức có mối quan hệ giữa 2 biến số lượng học sinh, sinh viên trong khu vực và doanh thu của mỗi nhà hàng trong khu vực ấy.

Bảng ANOVA có được:



**Tham khảo:**

Câu 1.1:<https://www.slideshare.net/camthuninh/chuong6-2>

<https://vi.vnp.edu.vn/wp-content/uploads/securepdfs/2020/01/Gujarati-2011-Ch%C6%B0%C6%A1ng-7-_-Ch%E1%BA%A9n-%C4%91o%C3%A1n-h%E1%BB%93i-quy-C%C3%A1c-l%E1%BB%97i-x%C3%A1c-%C4%91%E1%BB%8Bnh-d%E1%BA%A1ng.pdf>

Câu 2:[PHƯƠNG PHÁP KIỂM ĐỊNH TRONG TƯƠNG QUAN VÀ HỒI QUY TUYẾN TÍNH ĐƠN BIẾN - Big Data Uni](https://bigdatauni.com/tin-tuc/kiem-dinh-trong-tuong-quan-va-hoi-quy-don-bien.html) [i]