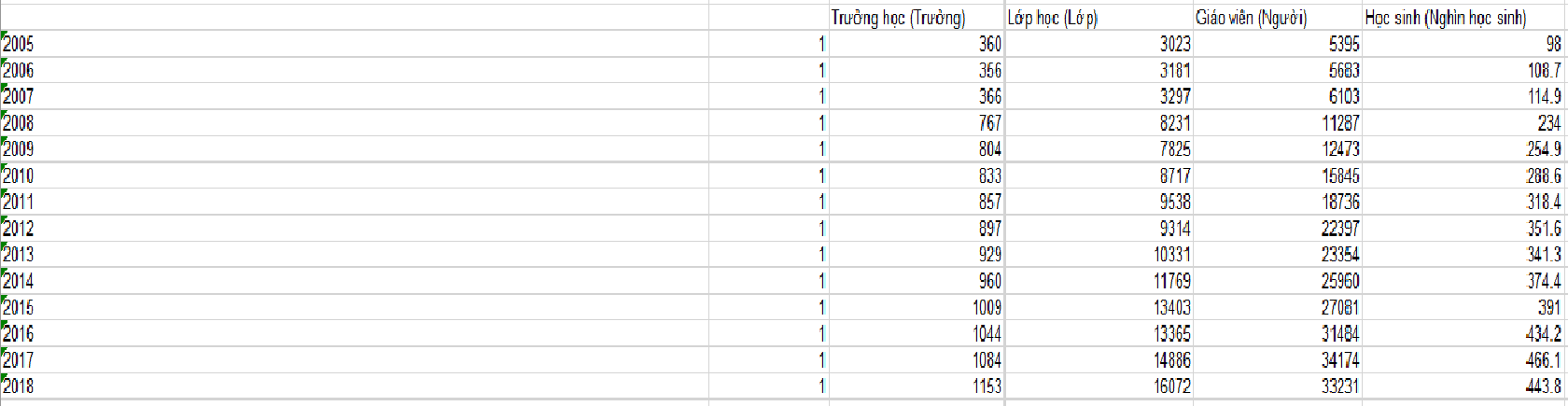
**3. Thực hiện câu 2 truy vấn tùy chọn dựa vào dữ liệu tùy chọn Việt Nam**

Dữ liệu về Số trường học, lớp học, giáo viên và học sinh mẫu giáo của thành phố Hà Nội từ năm

2005 – 2018

* Bộ dữ liệu gồm 14 dòng tương ứng 14 năm từ 2005 – 2018 quan sát trường học, lớp học, giáo viên mẫu giáo của thành phố Hà Nội
* Các cột thuộc tính gồm:
* Số trường học
* Số lớp học
* Số giáo viên
* Số học sinh (nghìn người)



* Bài toán đưa ra: Từ năm 2005 – 2018, số lượng trường học, lớp học, giáo viên mần

non của của thành phố Hà Nội ảnh hưởng đến số lượng học sinh mầm non

Em đặt:

+ Trường học là biến độc lập X1

+ Lớp học là biến độc lập X2

+ Giáo viên là biến độc lập X3

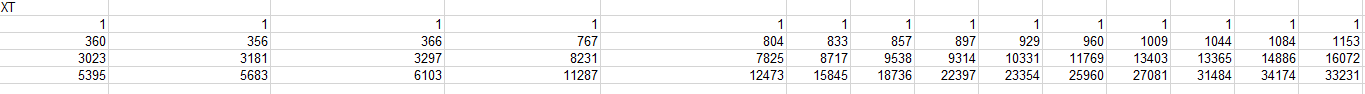
Phương trình hồi quy bội có dạng:

Y = b0 + b1X1 + b2X2 + b3X3

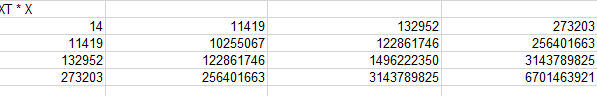
* **Tính các hệ số b bằng công thức ma trận**

**B=( X ^T .X) ^-1 X^T .Y**

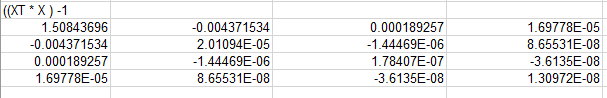
* Tính X^T



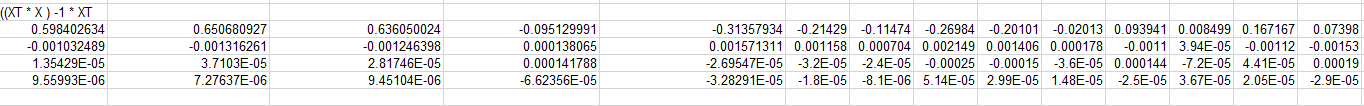
* Tính X^T .X



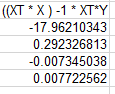
* Tính (X^T.X) ^ -1



* Tính ( X ^T .X) ^-1 X^T



* Tính ( X ^T .X) ^-1 X^T .Y



Từ kết quả trên được :

b0 = -17.96210343

b1 =  0.292326813

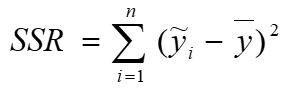
b2 = -0.007345038

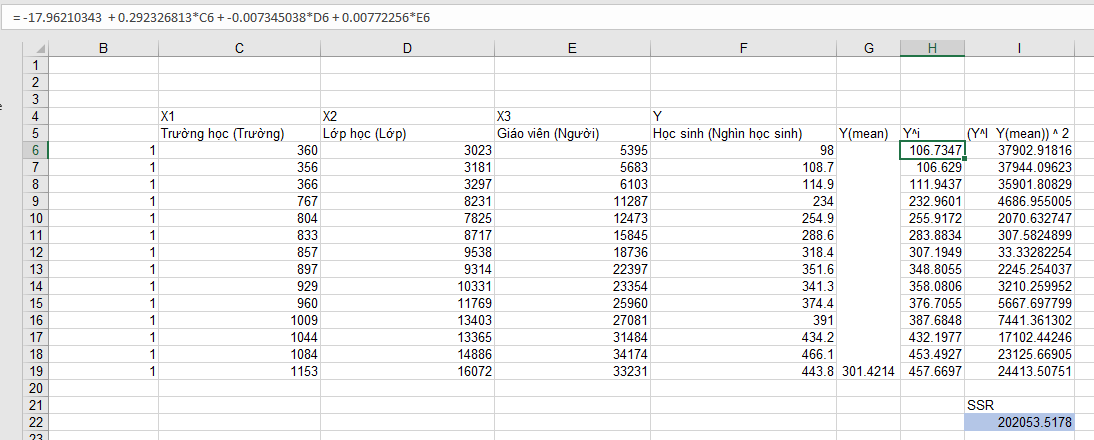
b3 = 0.007722562

🡺Phương trình hồi quy theo em tính được:

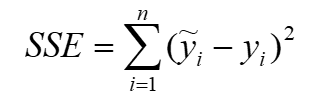
***Y =  -*17.96210343*+* 0.292326813*X1 +* -0.007345038*X2 +* 0.007722562*X3***

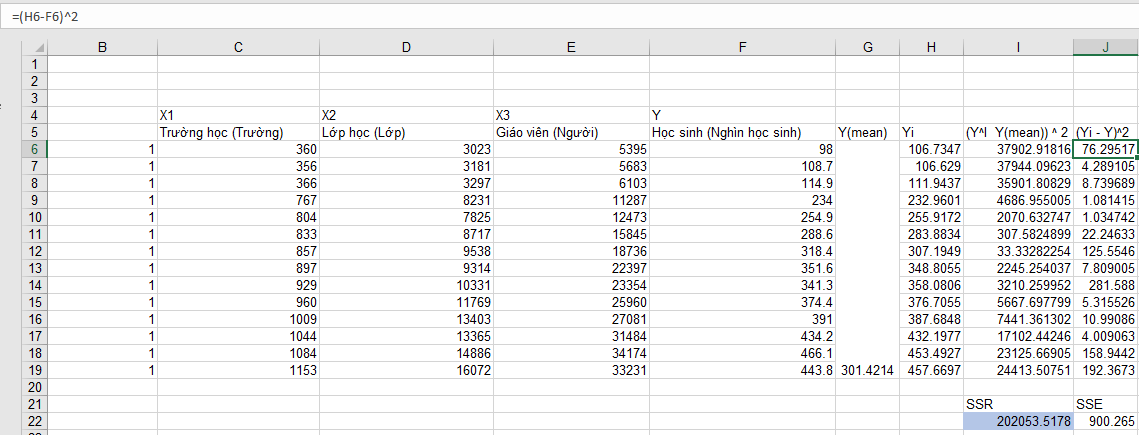
* **SSR: Sum of Square for Regression**



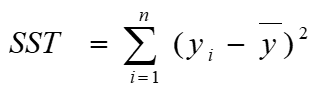


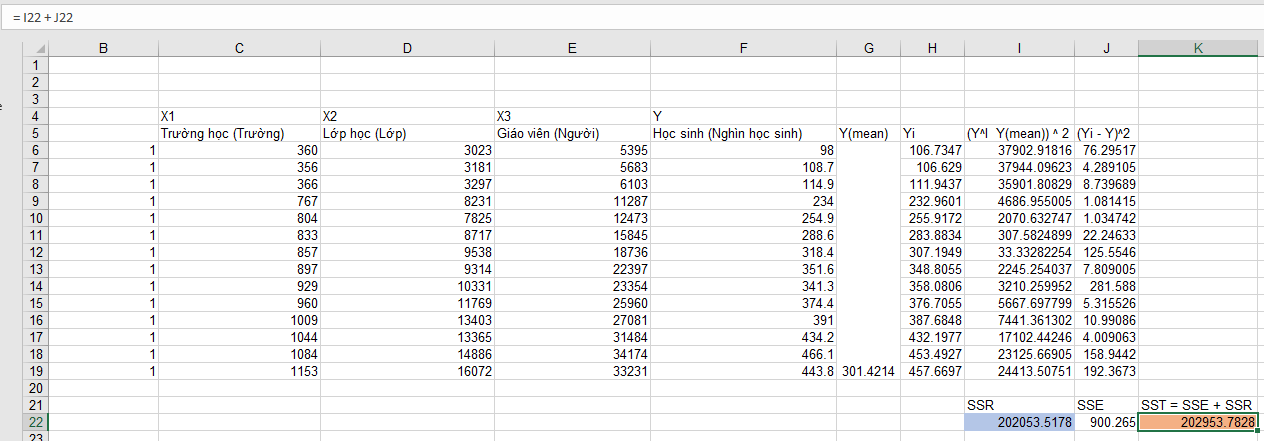
* **SSE: Sum of Square for Error (Residual)**



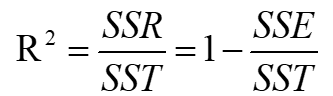


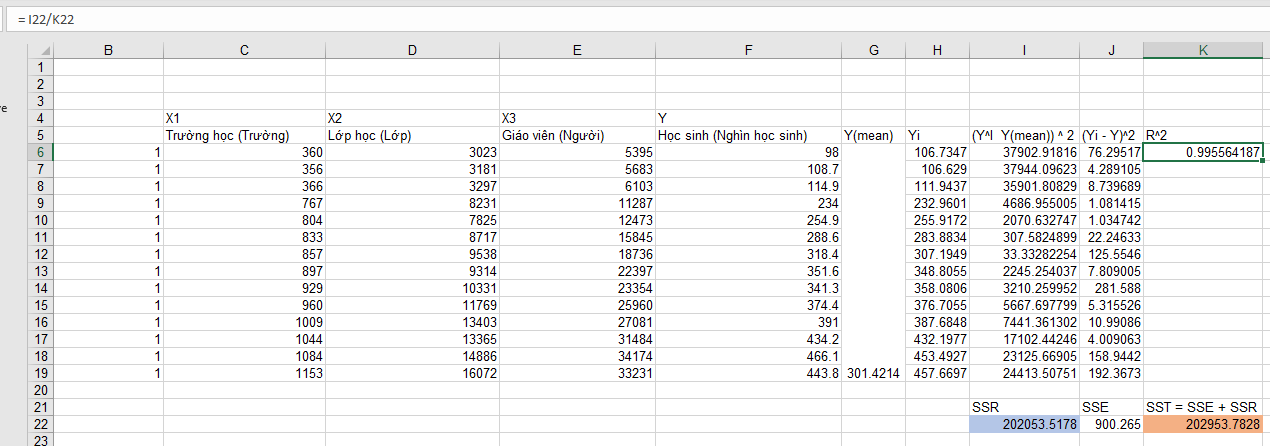
* **SST: Sum of Square for Total**



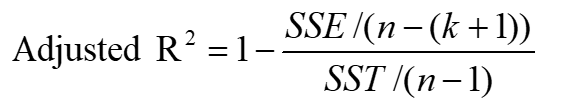


* **Hệ số R^2**



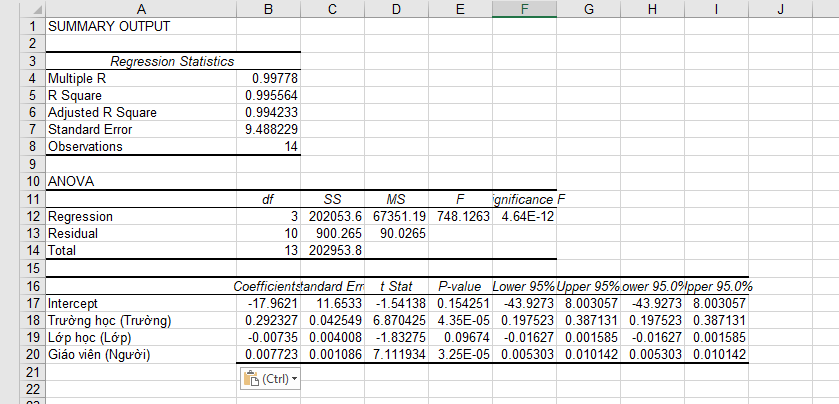


* **Adjusted R^2**



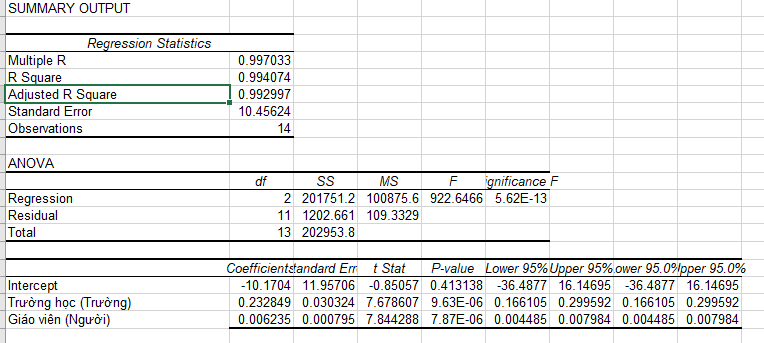
* ***MS Excel:***

Vào Data -> Data Analysis ->Regression . Chọn Y là cột tỷ suất di cư thuần, X là  4 cột còn lại. Ta được bảng kết quả như đã tính



Ta nhận thấy : p-value của các biến đọc lập Lớp học = 0.09674 >  α = 0.05

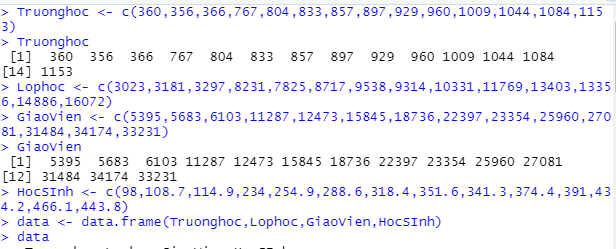
* Biến lớp học so với biến phụ thuộc Học sinh là không có ý nghĩa nên ta loại bỏ không xét tới biến lớp học
* Sau khi loại bỏ ta được bảng kết quả mới:



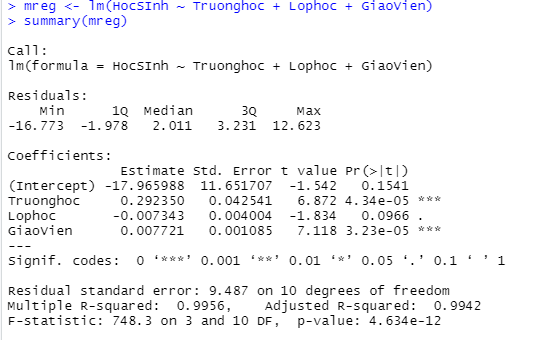
* Vậy phương trình hồi quy tốt nhất là:

***Y =  -*17.96210343*+* 0.292326813*X1 +* 0.007722562*X2***

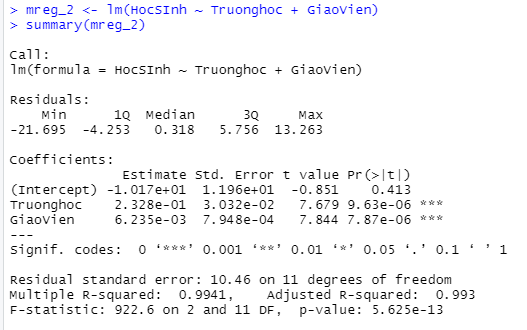
* ***Ngôn ngữ R***
* **Nạp dữ liệu vào R**



* **Phân tích hồi quy**



* Sau khi loại bỏ biến đi LopHoc



TÀI LIỆU THAM KHẢO

1. Slide môn học Phân tích dữ liệu kinh doanh. Khoa HTTT trường ĐH Công nghệ

Thông Tin – ĐHQGHCM.

2. Phân tích số liệu và biểu đồ bằng R – Nguyễn Văn Tuấn – Garvan Institute of

Medical Research Sydney, Australia.

3. Bộ dữ liệu: PX Web – General Statistics Office of Vietnam (gso.gov.vn)