

Chương Trình Giảng Dạy Kinh tế Fulbright

Học kỳ Thu năm 2013

Các Phương Pháp Phân Tích Định Lượng

Gợi ý giải bài tập 8

HỎI QUY ĐA BIẾN

Ngày Phát: Thứ ba 3/12/2013

Ngày Nộp: Thứ ba 7/12/2013

Bản in nộp lúc **8h20 sáng**, tại Hộp nộp bài tập trong phòng Lab

Bản điện tử gửi theo hướng dẫn của trợ giảng.

Bài 1 (50 điểm)

Bạn hãy sử dụng file WOODY's DATA ở bài tập 6, thực hiện các thao tác cần thiết trên Eviews để trả lời các câu hỏi sau:

a. Nếu xem Y là biến đo lường lượng cầu về dịch vụ nhà hàng, bạn kỳ vọng như thế nào về mối quan hệ giữa Y và biến N, P, I và vì sao?

Cần có sự giải thích rõ ràng về mối quan hệ giữa các biến mà học viên kỳ vọng.

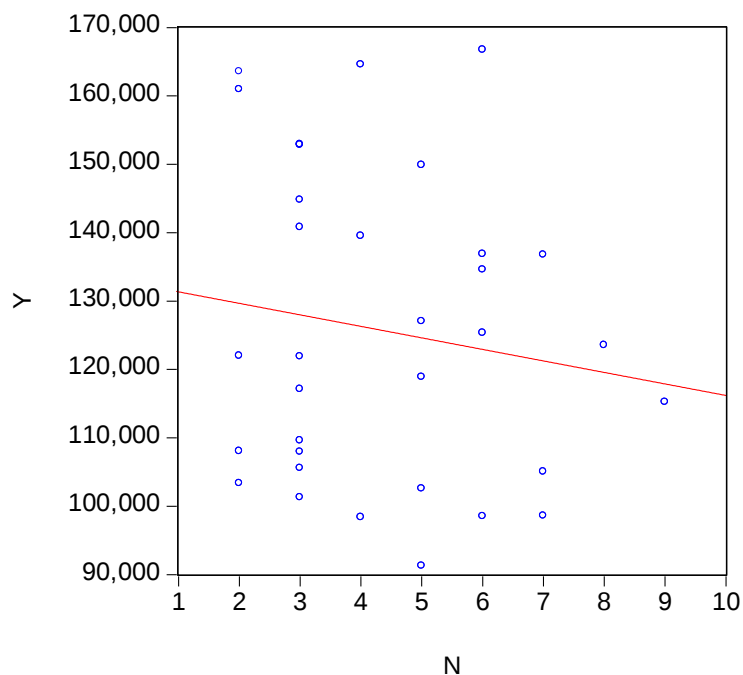
Trong môn Kinh tế vi mô, lượng cầu của một sản phẩm có thể phụ thuộc vào giá cả sản phẩm, giá cả sản phẩm có liên quan, thu nhập, thị hiếu, quy mô thị trường...

Nếu ta cho rằng dịch vụ nhà hàng là dịch vụ bình thường thì lượng cầu (Y) có thể chịu ảnh hưởng thuận bởi I, chịu ảnh hưởng nghịch bởi N, chịu ảnh hưởng thuận bởi P.

Biến N, biến P cho ta biết quy mô thị trường, khi số đối thủ cạnh tranh nhiều (quy mô thị trường nhỏ) sẽ làm cho lượng cầu dịch vụ nhà hàng mà ta quan tâm giảm, ngược lại khi dân số tăng (tức quy mô thị trường lớn hơn) sẽ làm cho lượng cầu dịch vụ nhà hàng tăng.

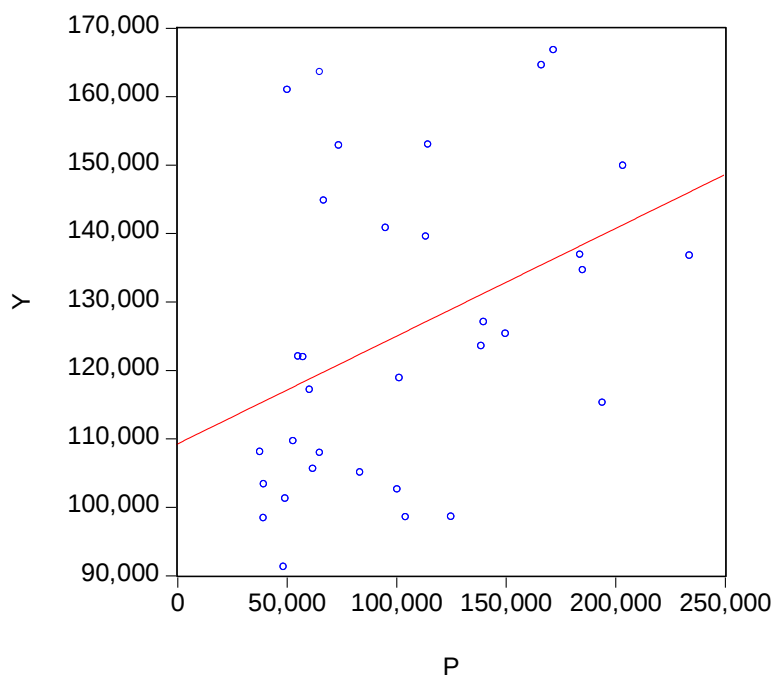
b. Vẽ ba đồ thị phân tán giữa Y và N, giữa Y và P, giữa Y và I và cho biết nhận xét của bạn.

Đồ thị phân tán giữa Y và N



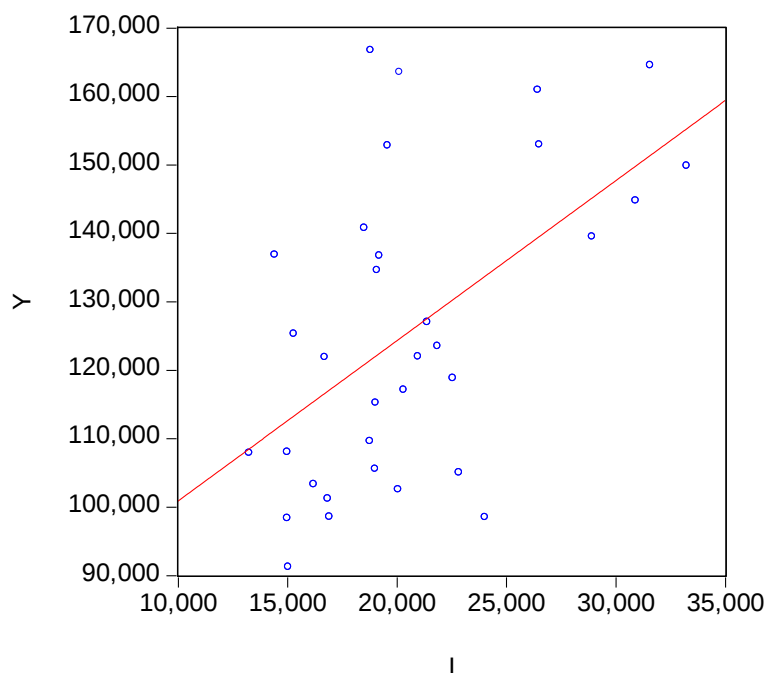
Đồ thị phân tán giữa Y và N cho thấy hai biến này có tương quan nghịch, mối liên hệ tuyến tính giữa hai biến không chặt

Đồ thị phân tán giữa Y và P



Đồ thị phân tán giữa Y và P cho thấy hai biến này có mối tương quan thuận, mối liên hệ tuyến tính giữa hai biến ở mức khá.

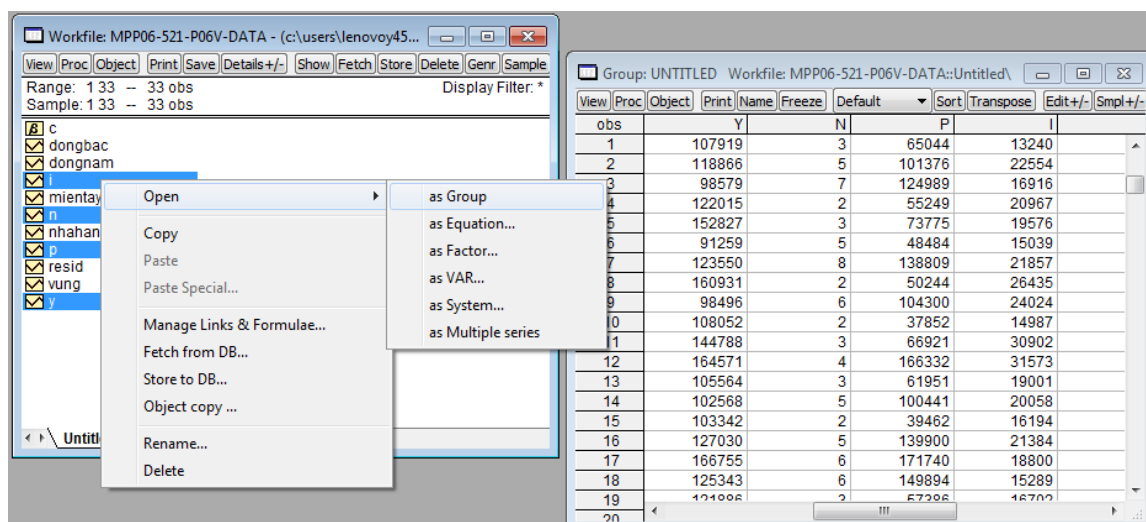
Đồ thị phân tán giữa Y và I



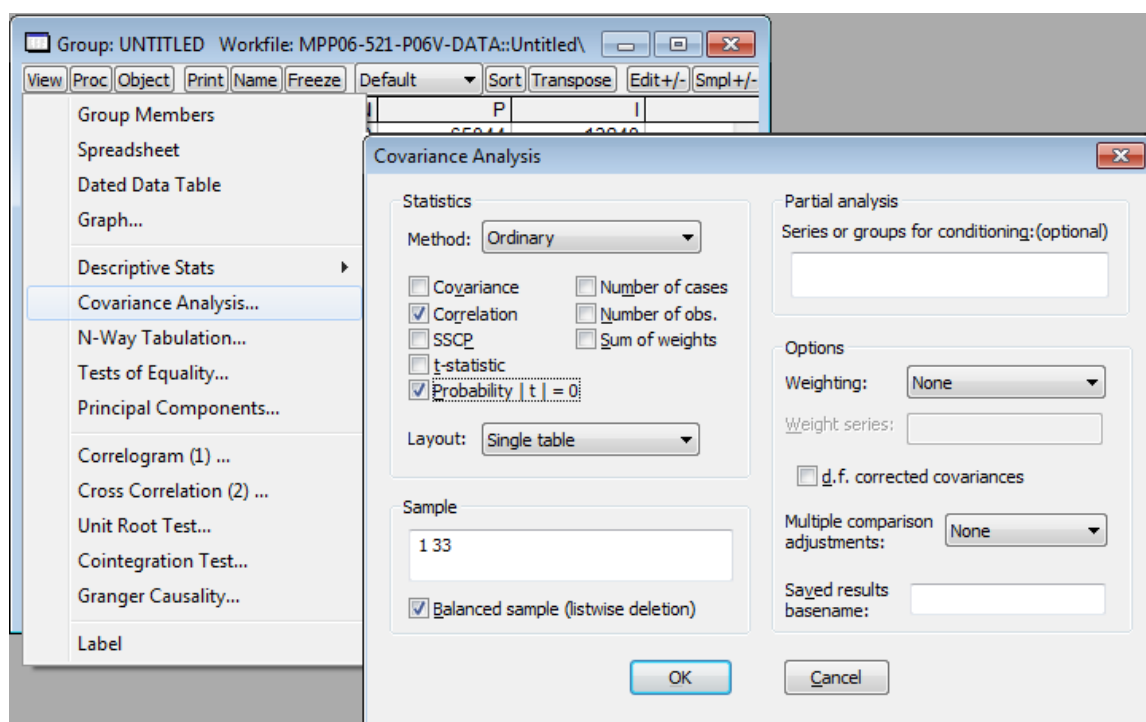
Đồ thị phân tán giữa Y và I cho thấy 2 biến này có mối tương quan thuận, có tương quan tuyến tính giữa hai biến khá chặt chẽ.

c. Tính hệ số tương quan cặp giữa Y và N, giữa Y và P, giữa Y và I, đồng thời kiểm định xem từng hệ số tương quan trên có ý nghĩa thống kê ở độ tin cậy 90% hay không? Kết quả tính toán của bạn có ủng hộ những kỳ vọng ban đầu của bạn ở câu a hay không?

Trong Eview, ta chọn đồng thời các biến Y,N,P,I, bấm chuột phải->Open->As group



Trong cửa sổ Group, chọn View->Covariance Analysis->Đánh dấu vào Correlation và Probability [t]=0 -> OK



Ta được kết quả như sau:

Covariance Analysis: Ordinary
Date: 12/04/13 Time: 21:18
Sample: 1 33
Included observations: 33

Correlation Probability	Y	N	P	I
Y	1,000000 -----			
N	-0,144225 0,4233	1,000000 -----		
P	0,392568 0,0238	0,726251 0,0000	1,000000 -----	
I	0,537022 0,0013	-0,031534 0,8617	0,245198 0,1690	1,000000 -----

Kết quả cho thấy hệ số tương quan giữa các biến, cũng như P-value trong kiểm định giả thuyết về sự bằng 0 của hệ số tương quan.

$r_{YI}=0,537$ (>0) và có ý nghĩa thống kê ở mức ý nghĩa 10% (P-value=0.0013) kết quả này cũng cho thấy I và Y có mối tương quan thuận.¹

$r_{YN}=-0,14425$ cho thấy hai biến này có mối tương quan nghịch, và mối liên hệ tuyến tính yếu. hệ số tương quan này không có ý nghĩa thống kê ở mức ý nghĩa 10% do P-value=0,4233 (>0.1)

¹ Có thể lập cặp giả thuyết: $H_0 (r=0)$; $H_a(r\neq0)$ để kiểm định.

$r_{Y,P} = 0,39256$ cho thấy hai biến này có mối tương quan thuận, và mức độ tương quan tuyến tính là khá. Hệ số này có ý nghĩa thống kê ở mức ý nghĩa 10% do $P\text{-value} = 0,0238 (< 0.1)$

Hệ số tương quan ủng hộ những kỳ vọng ban đầu đã được đưa ra trong câu a.

d. Tính thống kê mô tả (trung bình, độ lệch, min, max) của các biến Y, N, P, I

Tiếp tục từ cửa sổ Group, chọn View->Descriptive Stats->Common Sample

Date: 12/04/13
Time: 21:22
Sample: 1 33

	Y	N	P	I
Mean	125634,6	4,393939	103887,5	20552,58
Median	122015,0	4,000000	95120,00	19200,00
Maximum	166755,0	9,000000	233844,0	33242,00
Minimum	91259,00	2,000000	37852,00	13240,00
Std. Dev.	22404,09	1,919300	55884,51	5141,865
Skewness	0,355246	0,555101	0,672915	0,933694
Kurtosis	1,920334	2,359612	2,280488	3,161758
Jarque-Bera	2,296908	2,258639	3,202315	4,830791
Probability	0,317127	0,323253	0,201663	0,089332
Sum	4145942,	145,0000	3428287,	678235,0
Sum Sq. Dev.	1,61E+10	117,8788	9,99E+10	8,46E+08
Observations	33	33	33	33

e. Bạn hãy ước lượng hàm hồi quy mẫu sau:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 N_i + \hat{\beta}_3 P_i + \hat{\beta}_4 I_i + \hat{u}_i$$

Trong cửa sổ lệnh, gõ lệnh hồi quy: Ls Y c N P I

Dependent Variable: Y
Method: Least Squares
Date: 12/04/13 Time: 21:25
Sample: 1 33
Included observations: 33

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	102192,4	12799,83	7,983891	0,0000
N	-9074,674	2052,674	-4,420904	0,0001
P	0,354668	0,072681	4,879810	0,0000
I	1,287923	0,543294	2,370584	0,0246
R-squared	0,618154	Mean dependent var		125634,6
Adjusted R-squared	0,578653	S,D, dependent var		22404,09
S.E. of regression	14542,78	Akaike info criterion		22,12079
Sum squared resid	6,13E+09	Schwarz criterion		22,30218
Log likelihood	-360,9930	Hannan-Quinn criter,		22,18182
F-statistic	15,64894	Durbin-Watson stat		1,758193
Prob(F-statistic)	0,000003			

Phương trình hồi quy mẫu như sau: $Y_i = 102192,40 - 9074,67N_i + 0,35P_i + 1,29I_i + u_i$

Hay $\hat{Y}_i = 102192,40 - 9074,67N_i + 0,35P_i + 1,29I_i$

f. Từ hàm hồi quy ở câu e, hãy tính tác động biên của từng biến giải thích lên biến Y (về mặt trung bình) và cho biết những nhận xét của bạn.

Tác động biên của N lên Y là -9074,67. Trong điều kiện các yếu tố khác không đổi, khi N tăng (hoặc giảm) 1 đối thủ, thì trung bình Y giảm (hoặc tăng) 9074.67 lượt khách.

Tác động biên của P lên Y là 0,35. Trong điều kiện các yếu tố khác không đổi, khi P tăng (hoặc giảm) 1 người, thì trung bình Y tăng (hoặc giảm) 0.35 lượt khách.

Tác động biên của I lên Y là 1,29. Trong điều kiện các yếu tố khác không đổi, khi I tăng (hoặc giảm) 1 USD, thì trung bình Y tăng (hoặc giảm) 1.29 lượt khách.

g. Nếu xét tại giá trị trung bình của các biến giải thích, hãy tính hệ số co giãn của Y theo từng biến giải thích, và cho biết những nhận xét của bạn.

Tại giá trị trung bình của N, P, I ước lượng của Y là 125634,55

Hệ số co giãn của Y theo N là: $-9074,67 \cdot N_{tb} / Y_{tb} = -9074,67 \cdot 4,4 / 125634,55 = -0,3174$

Hệ số trên cho biết tại giá trị trung bình của N, trong điều kiện các yếu tố khác không đổi, khi N tăng 1% thì về mặt trung bình Y giảm 0,3174%, và ngược lại.

Tương tự, ta tính được hệ số co giãn của Y theo P là 0,2933, và hệ số co giãn của Y theo I là 0,2107 nhận xét.

h. Hãy ước lượng hàm hồi quy sau và cho biết ý nghĩa kinh tế của các hệ số hồi quy đứng trước các biến giải thích trong mô hình:

$$\ln(Y_i) = \alpha_1 + \alpha_2 \ln(N_i) + \alpha_3 \ln(P_i) + \alpha_4 \ln(I_i) + \hat{u}_i$$

Gõ lệnh:

Ls log(Y) c log(N) log(P) log(I)

Dependent Variable: LOG(Y)

Method: Least Squares

Date: 12/04/13 Time: 21:32

Sample: 1 33

Included observations: 33

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6,657823	0,772162	8,622317	0,0000
LOG(N)	-0,377950	0,064948	-5,819239	0,0000
LOG(P)	0,352044	0,055980	6,288779	0,0000
LOG(I)	0,159062	0,084535	1,881601	0,0700
R-squared	0,704924	Mean dependent var		11,72593
Adjusted R-squared	0,674399	S,D, dependent var		0,176593
S.E. of regression	0,100766	Akaike info criterion		-1,638809
Sum squared resid	0,294463	Schwarz criterion		-1,457414
Log likelihood	31,04035	Hannan-Quinn criter,		-1,577775
F-statistic	23,09324	Durbin-Watson stat		1,716252
Prob(F-statistic)	0,000000			

Hàm hồi quy mẫu như sau:

$$\ln(Y_i) = 6,6578 - 0,37795\ln(N_i) + 0,3520\ln(P_i) + 0,1591\ln(I_i) + \hat{u}_i$$

Các hệ số $\hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4$ chính là hệ số co giãn của Y theo từng biến giải thích tương ứng.

Hệ số co giãn của Y theo N là $\hat{\alpha}_2$: $\hat{\alpha}_2 = -0,37795$ cho biết trong điều kiện các yếu tố khác không đổi, khi I tăng 1% thì về mặt trung bình Y giảm -0,37795% và ngược lại.

Hệ số co giãn của Y theo P là $\hat{\alpha}_3$: $\hat{\alpha}_3 = 0,3520$ cho biết trong điều kiện các yếu tố khác không đổi, khi I tăng 1% thì về mặt trung bình Y tăng 0,3520% và ngược lại.

Hệ số co giãn của Y theo I là $\hat{\alpha}_4$: $\hat{\alpha}_4 = 0,1591$ cho biết trong điều kiện các yếu tố khác không đổi, khi I tăng 1% thì về mặt trung bình Y tăng 0,1591% và ngược lại.

Hệ số co giãn được tính theo cách này không thay đổi theo các giá trị của Y và biến giải thích, nhưng hệ số co giãn được tính theo cách ở câu g thì có thể thay đổi tại các giá trị của biến giải thích khác nhau.

i. Giả sử bạn chọn mô hình ở câu h để dự báo trung bình lượt khách đến một nhà hàng mà bạn quan tâm. Theo bạn giá trị dự báo điểm sẽ là bao nhiêu nếu biết nhà hàng này ở khu vực có N=5 đối thủ, P= 60000 người, I= 25000 USD (có thể tính gần đúng)

Thế các giá trị N=5, P=60000, I=25000 vào phương trình ta được ước lượng của $\ln(Y)$, sau đó lấy exp giá trị này sẽ được ước lượng gần đúng $\hat{Y}_i = 102060,2518 \approx 102060$ lượt khách.

j. Nếu chỉ sử dụng R^2 đã hiệu chỉnh để so sánh, theo bạn có thể chọn được mô hình ở câu e, và mô hình ở câu h mô hình nào tốt hơn hay không, và vì sao?

Mô hình ở câu e có Adjusted $R^2 = 0,57$; Mô hình ở câu h có Adjusted $R^2 = 0,67$. Không so sánh 2 giá trị này với nhau được bởi vì mô hình ở câu e có biến phụ thuộc là Y, còn mô hình ở câu h có biến phụ thuộc là $\ln(Y)$.

Bài 2 (25 điểm)

Một nghiên cứu của tác giả Khương Ninh đăng trên tạp chí Công Nghệ Ngân Hàng số tháng 8/2010 có kết quả hồi quy sau:

Mô hình 1: $INV = 0.9762 + 0.0768PRO + \hat{u}_i$
(thống kê t) (13.0593) (6.7066)

n=810, $\overline{R}_2 = 0.074$, P-value(F)=0.000

Mô hình 2: $INV = 0.9426 + 0.4941PRO - 0.0667PRO \times ASS + \hat{u}_i$
(thống kê t) (13.0248) (7.5750) (-6.4683)

n=810, $\overline{R}_2 = 0.138$, P-value(F)=0.000

Mô hình 3:

$$INV = 0.66 + 0.50^{***} PRO - 0.07^{***} PRO \times ASS + 0.0033^{**} DASL + 0.04^{*} EDU - 0.0095^{*} AGE + 0.0159VOCA - 0.04INP - 0.10ENL + 0.05COM + \hat{u}_i$$

n=810, $\overline{R}_2 = 0.314$, P-value(F)=0.000

Ghi chú: dấu *, **, *** sau các hệ số hồi quy cho biết hệ số hồi quy có ý nghĩa thống kê tương ứng ở mức ý nghĩa 10%, 5%, và 1%

Trong đó, thông tin của một số biến như sau: INV là tỷ số giữa giá trị đầu tư vào máy móc thiết bị, nhà xưởng ... và tổng giá trị tài sản cố định của doanh nghiệp; PRO là tỷ số giữa lợi nhuận của năm trước và giá trị tài sản cố định của doanh nghiệp, PROxASS là tích số giữa PRO và

ln(giá trị tài sản cố định), DASL là tốc độ tăng trưởng của doanh số (%), EDU cho biết trình độ học vấn của người quản lý doanh nghiệp, AGE là tuổi của doanh nghiệp tính từ khi thành lập đến năm điều tra, VOCA là số lần dự các lớp tập huấn chuyên môn của người quản lý doanh nghiệp.

Hãy tạm thời bỏ qua những sai sót của tác giả (nếu có) trong việc xây dựng và ước lượng mô hình, cũng như giả sử rằng mô hình này không vi phạm các giả định của phương pháp OLS, từ kết quả trên, bạn hãy trả lời các câu hỏi sau:

a. Tính hệ số tương quan của biến INV và PRO

$$\text{Ta có } \bar{R}^2 = 1 - (1 - R^2) \frac{n-1}{n-k}$$

$$\Rightarrow R^2 = 1 - (1 - \bar{R}^2) \frac{n-k}{n-1} = 1 - (1 - 0,074)(810-2)/(810-1) = 0,075145$$

$$r_{(\text{INV}, \text{PRO})} = \sqrt{0,075145} = 0,274125$$

b. Trong mô hình 1, theo bạn biến PRO có ý nghĩa thống kê ở độ tin cậy 95% hay không, hãy cho biết ý nghĩa kinh tế của hệ số này?

Biến PRO có ý nghĩa thống kê ở độ tin cậy 95% do $|T\text{-stat}(\text{PRO})| = 6,7066 > 2$

Hệ số này cho biết: trong điều kiện các yếu tố khác không đổi, khi PRO tăng (hoặc giảm) 1 đơn vị thì trung bình INV tăng (hoặc giảm) 0,0768 đơn vị.

c. Theo bạn thống kê F của mô hình 1 sẽ bằng bao nhiêu?

Trong hàm hồi quy đơn ta có thể tính $F = T\text{-stat}(\text{PRO})^2 = 6,7066^2 = 44,97848$

d. Từ các mô hình trên, theo bạn tác động biên của PRO lên INV có phụ thuộc vào biến ASS hay không?

Tác động biên của PRO lên INV chính là đạo hàm của INV theo PRO. Nếu xét mô hình 3, tác động này bằng $0,5 - 0,07\text{ASS}$. Hệ số đứng trước biến tương tác $\text{PRO} \times \text{ASS}$ bằng $-0,07$, có ý nghĩa thống kê ở mức ý nghĩa 1% nên tác động biên của PRO lên INV phụ thuộc vào ASS.²

e. Theo bạn, hệ số 0.0033 đứng trước biến DASL (trong mô hình 3) cho bạn biết điều gì? đây có phải là hệ số co giãn của INV theo DASL hay không, và vì sao?

Hệ số này cho biết DASL tăng 1 đơn vị (hay tăng 1 điểm %) [ví dụ từ 7% lên 8% chẳng hạn] thì về mặt trung bình INV tăng lên 0,0033, trong điều kiện các yếu tố khác không đổi.

Con số này không phải là hệ số co giãn, vì hệ số co giãn phải cho biết rằng khi DASL tăng 1% [ví dụ từ DASL tăng từ 7% lên 7,07%] thì trung bình INV tăng lên bao nhiêu %, và ngược lại.

Bài 3 (25 điểm)

Từ dữ liệu VHLSS2008, sử dụng phần mềm Stata;

(a) bạn hãy ước lượng hàm hồi quy tuyến tính bội thể hiện mối quan hệ giữa chi tiêu của hộ và các biến giải thích sau đây: thu nhập của hộ, tuổi của chủ hộ, số người trong hộ (có sử dụng trọng số trong hồi quy). Sau đó, bạn hãy cho biết những nhận xét của bạn từ kết quả tính toán.

Trong bài này, bạn cần nối các file dữ liệu cần thiết lại với nhau, dữ liệu liên quan đến các file: ho15.dta, muc123a.dta, hhxpe08.dta. Chú ý rằng file muc123a.dta chỉ giữ lại các quan sát là chủ hộ. Sau khi có file dữ liệu, thực hiện hàm hồi quy theo cấu trúc lệnh

Mở các file và sắp xếp thứ tự

² Làm tương tự với mô hình 2

```
. use "E:\z.VHLSS2008\Bai 8 - MPP6\ muc123a.dta", clear
. keep if matv==1
(29064 observations deleted)
. sort tinh huyen xa diaban hoso matv
. save "E:\z.VHLSS2008\Bai 8 - MPP6\ muc123a_cut_sorted.dta"
file E:\z.VHLSS2008\Bai 8 - MPP6\ muc123a_cut_sorted.dta saved
. use "E:\z.VHLSS2008\Bai 8 - MPP6\ ho15.dta", clear
. sort tinh huyen xa diaban hoso
. save "E:\z.VHLSS2008\Bai 8 - MPP6\ ho15_sorted.dta"
file E:\z.VHLSS2008\Bai 8 - MPP6\ ho15_sorted.dta saved
. use "E:\z.VHLSS2008\Bai 8 - MPP6\ hhexpe08.dta", clear
(Household expenditures: 2008 VHLSS)
. sort tinh huyen xa diaban hoso
. save "E:\z.VHLSS2008\Bai 8 - MPP6\ hhexpe08_sorted.dta"
file E:\z.VHLSS2008\Bai 8 - MPP6\ hhexpe08_sorted.dta saved
.
```

Nối các file với nhau

```
. use "E:\z.VHLSS2008\Bai 8 - MPP6\ ho15_sorted.dta", clear
. merge 1:1 tinh huyen xa diaban hoso using "E:\z.VHLSS2008\Bai 8 -
MPP6\ muc123a_cut_sorted.dta"
Result # of obs.
-----
not matched 0
matched 9,189 (_merge==3)
-----

. drop _merge
. . merge 1:1 tinh huyen xa diaban hoso using "E:\z.VHLSS2008\Bai 8 -
MPP6\ hhexpe08_sorted.dta"
Result # of obs.
-----
not matched 0
matched 9,189 (_merge==3)
-----

. save "E:\z.VHLSS2008\Bai 8 - MPP6\ All_data_combine.dta"
file E:\z.VHLSS2008\Bai 8 - MPP6\ All_data_combine.dta saved
```

Reg Y X1 X2...Xk; sau đó nhận xét kết quả

Reg chitieu thunhap hhsiz m1ac5 [pw=wt9], beta

```
. reg chitieu thunhap hhsiz m1ac5 [pw=wt9], beta
(sum of wgt is 2.0960e+07)
```

Linear regression

```
Number of obs = 9189
F( 3, 9185) = 358.91
Prob > F = 0.0000
R-squared = 0.4783
Root MSE = 28357
```

chitieu	Coef.	Robust Std. Err.	t	P> t	Beta
thunhap	.3547346	.0696622	5.09	0.000	.6611453
hhsiz	2879.859	629.1372	4.58	0.000	.1204261
m1ac5	30.2907	21.93291	1.38	0.167	.0104938
_cons	7286.227	1516.938	4.80	0.000	.

Hàm hồi quy bội là:

$$\text{Chitieu}_i = 7286,227 + 0,3547346\text{thunhap}_i + 2879,859\text{hhsiz}_i + 30,2907\text{m1ac5}_i + u_i$$

Trong mô hình, biến m1ac5 (tuổi của chủ hộ) không có ý nghĩa thống kê ở mức ý nghĩa 10%, các biến giải thích khác như thunhap, hhsiz (quy mô hộ) có ý nghĩa thống kê ở mức ý nghĩa 1%.

Hhsiz, và thunhap ảnh hưởng thuận đến chitieu. Trong điều kiện cố định các yếu tố khác, khi thunhap tăng 1 ngàn đồng thì chitieu về mặt trung bình sẽ tăng 0.354 ngàn đồng và ngược lại. Trong điều kiện cố định các yếu tố khác, khi số người trong hộ tăng 1 người, thì trung bình chitieu sẽ tăng 2879 ngàn đồng.

Mô hình giải thích được 47.8% biến thiên của biến chitieu về mặt trung bình.

(b) Trong các biến giải thích trên, theo bạn biến nào ảnh hưởng mạnh nhất (quan trọng nhất) đến chi tiêu của hộ.

Để xem biến nào ảnh hưởng mạnh nhất, đến chi tiêu của hộ, học viên có thể sử dụng hệ số hồi quy đã chuẩn hóa. Trị tuyệt đối của hệ số hồi quy đã chuẩn hóa (hệ số Beta) của biến giải thích nào càng lớn thì biến giải thích ấy ảnh hưởng càng mạnh, hay càng quan trọng đến biến phụ thuộc.

|Hệ số Beta của biến thunhap|=0,661 lớn nhất trong các |hệ số Beta| nên biến này ảnh hưởng mạnh nhất lên chitieu

(c) Nếu sử dụng VIF và hệ số tương quan cặp giữa các biến giải thích để chẩn đoán hiện tượng đa cộng tuyến, theo bạn mô hình có dấu hiệu bị vi phạm đa cộng tuyến hay không?

Tính hệ số VIF, nếu có ít nhất một biến giải thích có $VIF > 10$, hoặc có ít nhất một cặp biến giải thích có trị tuyệt đối hệ số tương quan lớn (> 0.8) ta có thể xem mô hình đã bị vi phạm đa cộng tuyến.

. vif

variable	VIF	1/VIF
hhsz	1.05	0.955420
thunhap	1.03	0.970953
m1ac5	1.02	0.983746
Mean VIF	1.03	

. cor thunhap hhsz m1ac5
 (obs=9189)

	thunhap	hhsz	m1ac5
thunhap	1.0000		
hhsz	0.1672	1.0000	
m1ac5	-0.0155	-0.1478	1.0000

Từ kết quả trên cho thấy, mô hình không có dấu hiệu bị đa cộng tuyến.