Chương 4
Hồi quy với biến giả

4.1. Bản chất của biến giả

- Trong nhiều mô hình hồi quy, chúng ta cần xét biến giải thích (thậm chí biến phụ thuộc) là biến chất lượng (biến định tính).
- Ví dụ biến về:
  - Vùng địa lý, tôn giáo, giới tính, loai hình đào tạo, loại hình công việc, mùa, …
- Loại thông tin này có tính chất tự nhiên như là biến chỉ dẫn.
- Trong kinh tế lượng, các biến như thế gọi là biến giả.

Ví dụ: Lương giáo viên phổ thông

- Chúng ta có số liệu về lương của giáo viên 51 địa điểm.
- Chia ra ba loại
  - Phía bắc (21 điểm)
  - Nam (17 điểm)
  - Trung (13 điểm)
- Làm thế nào để đặt các biến giả này?

Ví dụ: Lương giáo viên phổ thông (tiếp)

- Đặt 3 biến giả
  - D1 = 1 nếu là vùng miền Trung; =0 nếu ngược lại.
  - D2 = 1 nếu là vùng miền Bắc; =0 nếu ngược lại.
  - D3 = 1 nếu là vùng miền Nam; =0 nếu ngược lại.
- Câu hỏi: Lương trung bình của các giáo viên các miền có bằng nhau không?
- Mô hình: ANOVA

Mô hình là:

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i$$

Ta có:

$$E(Y_i \mid D_{2i} = 0, D_{3i} = 0) = \beta_1$$

$$E(Y_i \mid D_{2i} = 1, D_{3i} = 0) = \beta_1 + \beta_2$$

$$E(Y_i \mid D_{2i} = 0, D_{3i} = 1) = \beta_1 + \beta_3$$

Một biểu diễn thay thế

$$Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i$$

Chúng ta có:

$$E(Y_i \mid D_{1i} = 1, D_{2i} = 0, D_{3i} = 0) = \beta_1$$

$$E(Y_i \mid D_{1i} = 0, D_{2i} = 1, D_{3i} = 0) = \beta_2$$

$$E(Y_i \mid D_{1i} = 0, D_{2i} = 0, D_{3i} = 1) = \beta_3$$

$D_1+D_2+D_3=1$ nên có ĐCT.

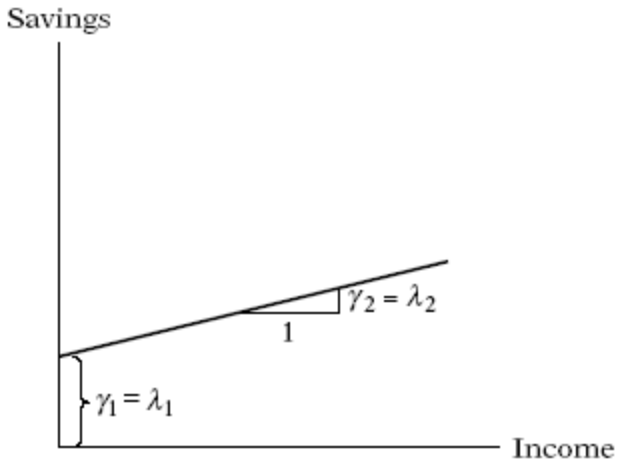
4.2. Hồi quy với một biến lượng và hai biến chất.

A dummy variable formulation of the Chow test

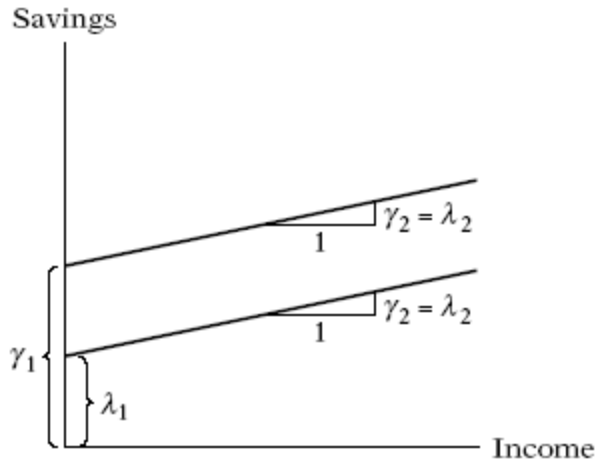
- If we have a simple grouping (say, two groups) we can ask if both the intercept and the slope changes across the groups
- This is another way of looking at the Chow test

$$Y_i = \alpha_1 + \alpha_2 D_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} D_i) + \beta_4 (X_{2i} D_i) + u_i$$

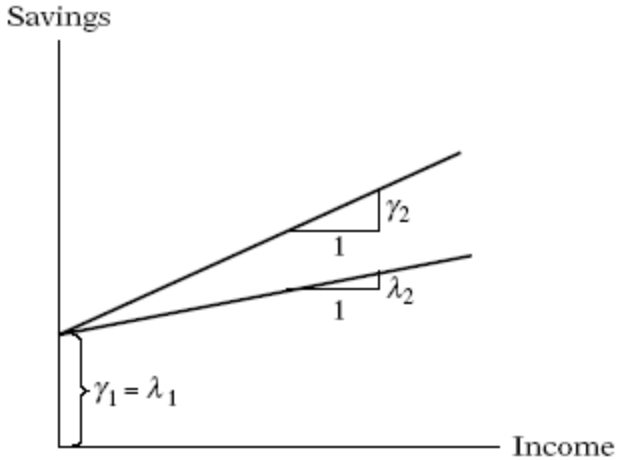
Interpretation of the possible regressions



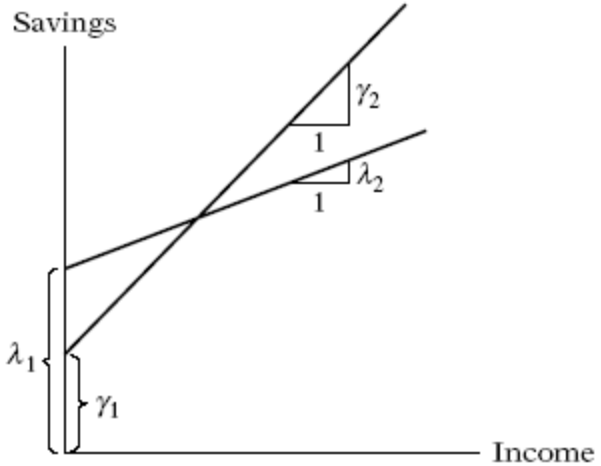
(*a*) Coincident regressions



(*b*) Parallel regressions



(*c*) Concurrent regressions



(*d*) Dissimilar regressions

Two Chow tests

```
. test  hhsize_D pelderly_D
  pchild_D pfemale_D urban98

( 1)  hhsize_D = 0
( 2)  pelderly_D = 0
( 3)  pchild_D = 0
( 4)  pfemale_D = 0
( 5)  urban98 = 0

      F(  5,  5989) =  415.48
           Prob > F =    0.0000
```

```
. test  hhsize_D pelderly_D
  pchild_D pfemale_D

( 1)  hhsize_D = 0
( 2)  pelderly_D = 0
( 3)  pchild_D = 0
( 4)  pfemale_D = 0

      F(  4,  5989) =    3.02
           Prob > F =    0.0167
```

A real, real life example
Some household characteristics in VLSS 1998

- Household size
- Proportion of elderly
- Proportion of children
- Proportion of females
- Spouse, 2 cat.
- Ethnic minority, 2 cat.
- Education (hhh), 6 cat.
- Education (hhs), 6 cat.
- Occupation (hhh), 7 cat.
- Type of house, 3 cat.
- Electricity, 2 cat.
- Source of drinking water, 3 cat.
- Type of toilet, 3 cat.
- TV ownership, 2 cat.
- Radio ownership, 2 cat.
- Region, 7 cat.

- How many regressors do we have in this model?

The expenditure regression for the rural area

Source	SS	df	MS	Number of obs =	4269
Model	567.404167	37	15.3352477	F(37, 4231) =	129.48
Residual	501.114754	4231	.118438845	Prob > F =	0.0000
Total	1068.51892	4268	.250355886	R-squared =	0.5310
				Adj R-squared =	0.5269
				Root MSE =	.34415

lnrpce	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hhsize	-.0752959	.0034388	-21.90	0.000	-.0820377	-.068554
pelderly	-.0206064	.0269051	-0.77	0.444	-.0733545	.0321417
pchild	-.3198785	.0282833	-11.31	0.000	-.3753286	-.2644283
pfemale	-.070512	.0280055	-2.52	0.012	-.1254175	-.0156065
ethnic	-.1090882	.0177687	-6.14	0.000	-.1439243	-.0742522
Iedcsp_1	.0284833	.0166373	1.71	0.087	-.0041345	.0611011
Iedchd_2	.0648492	.0150414	4.31	0.000	.0353602	.0943382
Iedchd_3	.1062095	.0173294	6.13	0.000	.0722347	.1401842
Iedchd_4	.1053885	.031159	3.38	0.001	.0443004	.1664765
Iedchd_5	.1522256	.0217644	6.99	0.000	.109556	.1948952
Iedchd_6	.2694558	.0549895	4.90	0.000	.1616475	.3772641

Iedcsp_3	.0145598	.01664	0.87	0.382	-.0180632	.0471829
Iedcsp_4	-.0011391	.0190424	-0.06	0.952	-.0384722	.0361941
Iedcsp_5	-.001556	.0348009	-0.04	0.964	-.0697839	.066672
Iedcsp_6	.1026943	.0270398	3.80	0.000	.0496822	.1557064
Iedcsp_7	.158244	.0755996	2.09	0.036	.010029	.306459
Ioccup_1	.1712163	.0477333	3.59	0.000	.077634	.2647985
Ioccup_2	.1256021	.0393633	3.19	0.001	.0484294	.2027748
Ioccup_3	.1250224	.0318177	3.93	0.000	.062643	.1874018
Ioccup_4	.0058642	.0209146	0.28	0.779	-.0351394	.0468678
Ioccup_5	.0738985	.0295435	2.50	0.012	.0159777	.1318193
Ioccup_6	-.0461344	.0299045	-1.54	0.123	-.1047629	.0124942
Ihouse_1	.2592556	.0249945	10.37	0.000	.2102531	.308258
Ihouse_2	.1603518	.0147686	10.86	0.000	.1313976	.1893061
electric	.0952981	.0142953	6.67	0.000	.0672718	.1233244
Inwate_1	.0838858	.0431485	1.94	0.052	-.0007079	.1684795
Inwate_2	.1218378	.0160064	7.61	0.000	.0904569	.1532186
Itoile_1	.3282756	.0302283	10.86	0.000	.2690122	.387539
Itoile_2	.056766	.0139646	4.07	0.000	.0293881	.084144
tv	.2310695	.0121537	19.01	0.000	.2072418	.2548972
radio	.1009113	.0111825	9.02	0.000	.0789876	.1228349
reg7_2	-.0179041	.0202812	-0.88	0.377	-.0576659	.0218576
reg7_3	.0147107	.0211819	0.69	0.487	-.026817	.0562383
reg7_4	.1320784	.0223846	5.90	0.000	.0881929	.1759639
reg7_5	.1523564	.0229629	6.63	0.000	.107337	.1973758
reg7_6	.4852656	.0229432	21.15	0.000	.4402849	.5302463
reg7_7	.3444246	.0227196	15.16	0.000	.2998823	.3889669
cons	7.419036	.0411953	180.09	0.000	7.338271	7.4998

Testing the different categories

- In models with many categories you should not test the individual regressors but the groups, say education of household head

```
. test Iedchd_2  Iedchd_3  Iedchd_4  Iedchd_5
  Iedchd_6

 ( 1)  Iedchd_2 = 0
 ( 2)  Iedchd_3 = 0
 ( 3)  Iedchd_4 = 0
 ( 4)  Iedchd_5 = 0
 ( 5)  Iedchd_6 = 0

       F(  5,  4231) =   14.38
            Prob > F =    0.0000
```

Next time

- Introduction to extensions of the classical linear regression model
- Multicollinearity (Chapter 10)

Equation: UNTITLED Workfile: CH4BT1

View | Procs | Objects | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Q
Method: Least Squares
Date: 02/01/07 Time: 09:39
Sample: 97M1 to 99M3
Included observations: 27

Variable	Coefficient	Std. Error	t-Statistic	Prob.
PG	-7.0673	.20832	33.9252	.000
D	106.0104	98.5409	1.0758	.293
DPG	.278299	.078845	2.6307	.012
C	2403.548	564.049	4.2609	.000

R-squared	.99252	Mean dependent var	1831.4
Adjusted R-squared	.99154	S.D. dependent var	451.9370
S.E. of regression	41.5680	Akaike info criterion	
Sum squared resid	39741.7	Schwarz criterion	
Log likelihood	-136.78	F-statistic F(3,23)	1016.8
Durbin-Watson stat	1.9506	Prob(F-statistic)	0.000