データ解析

http://coconut.sys.eng.shizuoka.ac.jp/data/

静岡大学工学部
安藤和敏

2005.11.09

質的変数の取り扱い

No	広さ (平米)	住所 u	価格 (千万円)
1	51	田町	3.0
2	38	鍛冶町	3.2
3	57	鍛冶町	3.3
4	51	肴町	3.9
5	53	板屋町	4.4
6	77	田町	4.5
7	63	板屋町	4.5
8	69	鍛冶町	5.4
9	72	板屋町	5.4
10	73	肴町	6.0

ダミー変数

住所を表す変数 u は，田町，鍛冶町，肴町，板屋町のいずれかの値をとる．

住所を表す変数 u の取りうる値のそれぞれに対して，0か1の値をとる変数を導入する．

$$u_{田}, u_{鍛冶}, u_{肴}, u_{板屋}$$

そして，例えば，$u_{鍛冶}$は以下のように定義する．

$$u_{鍛冶} = \begin{cases} 1 & (u = 鍛冶町のとき) \\ 0 & (u \neq 鍛冶町のとき) \end{cases}$$

質的変数の取り扱い

No	面積 (平米)	住所				価格 (千万円)
		田町	鍛冶町	肴町	板屋町	
1	51	1	0	0	0	3.0
2	38	0	1	0	0	3.2
3	57	0	1	0	0	3.3
4	51	0	0	1	0	3.9
5	53	0	0	0	1	4.4
6	77	1	0	0	0	4.5
7	63	0	0	0	1	4.5
8	69	0	1	0	0	5.4
9	72	0	0	0	1	5.4
10	73	0	0	1	0	6.0

ダミー変数の問題点とその解決

この多変量データに対して重回帰分析を行うと，偏回帰係数は一意に定まらない．

なぜならば，こうして導入されたダミー変数 $u_{田}$, $u_{鍛冶}$, $u_{肴}$, $u_{板屋}$は以下の式を必ず満たしている．

$$u_{田} + u_{鍛冶} + u_{肴} + u_{板屋} = 1$$

このように変数の間に線型従属性が存在するときは，偏回帰係数は一意に定まらない．

ダミー変数を1つ削除してしまえばよい．

質的変数の取り扱い

No	面積 (平米)	住所			価格 (千万円)
		鍛冶町	肴町	板屋町	
1	51	0	0	0	3.0
2	38	1	0	0	3.2
3	57	1	0	0	3.3
4	51	0	1	0	3.9
5	53	0	0	1	4.4
6	77	0	0	0	4.5
7	63	0	0	1	4.5
8	69	1	0	0	5.4
9	72	0	0	1	5.4
10	73	0	1	0	6.0

質的変数の取り扱い

一般にある質的変数uが，R個の値をとるときには，$R-1$個のダミー変数

$$u_1, u_2, \cdots, u_{R-1}$$

を

$$u_j = \begin{cases} 1 & (u = j \quad \text{のとき}) \\ 0 & (u \neq j \quad \text{のとき}) \end{cases}$$

で定義してuの代わりに置き換えればよい．

質的変数の取り扱い

このようにして，質的変数uを$R-1$個のダミー変数で置き換えたデータに対して前回までと同様にして，重回帰分析を行えばよい．

2個以上の質的変数が存在する場合は，それぞれの質的変数に対して，上で述べたようにダミー変数を導入すればよい．