重回帰分析におけるダミー変数の解釈について

Some Remarks of dummy variables in multiple regression analysis

石村　貞夫・石村　友二郎

Sadao ISHIMURA and Yujirou ISHIMURA

「鶴見大学紀要」第49号　第4部

人文・社会・自然科学編 （平成24年 3 月） 別刷

重回帰分析におけるダミー変数の解釈について

Some Remarks of dummy variables in multiple regression analysis

石村　貞夫・石村　友二郎

Sadao ISHIMURA and Yujirou ISHIMURA

1. 序文

社会調査、心理学、医学など、いろいろな分野の学術論文において、最も多く使用されている統計処理は重回帰分析である。その理由としては、独立変数を原因、従属変数を結果とみなすことにより、重回帰式や重回帰モデルを使って研究対象の因果関係をわかりやすく表現できるという点にある。

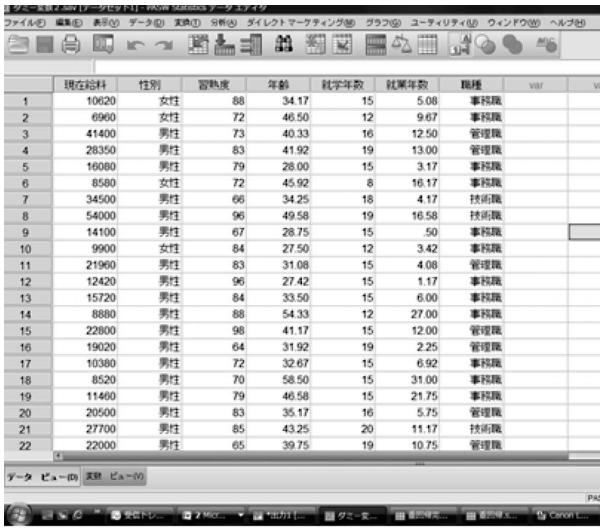
しかしながら、どのようなデータに対しても重回帰分析を適用できるわけではなく、例えば、重回帰モデルの検定を分散分析表で行う場合には、従属変数に対して正規性の仮定が必要であったり、重回帰式の計算においても、独立変数が名義データの場合には、各カテゴリをダミー変数に置き換えるなど、そのデータの取り扱い方については、細心の注意が必要である。

この掌編では、独立変数が名義変数の場合、その偏回帰係数の検定をどのように解釈すればよいかについて、1元配置の分散分析と比較しながら、その解説を行う。

2. 名義データ含んだ重回帰分析用データの型

次のデータは、重回帰分析でよく利用されている典型的な例である。従属変数は現在給料で、残りの変数が独立変数に対応している。

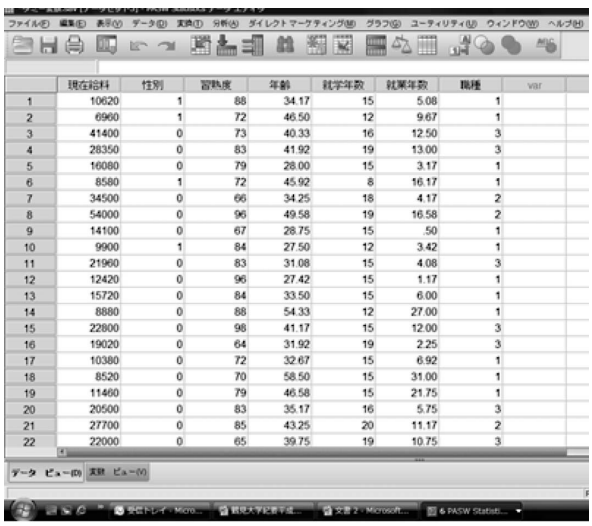
研究目的としては、従属変数に影響を与えている独立変数はどれかなど、給料とその要因といった因果関係が中心となっている。



	現在給料	性別	習熟度	年齢	就学年数	就業年数	職種	var
1	10620	女性	88	34.17	15	5.08	事務職	
2	6960	女性	72	46.50	12	9.67	事務職	
3	41400	男性	73	40.33	16	12.50	管理職	
4	28350	男性	83	41.92	19	13.00	管理職	
5	16080	男性	79	28.00	15	3.17	事務職	
6	8580	女性	72	45.92	8	16.17	事務職	
7	34500	男性	66	34.25	18	4.17	技術職	
8	54000	男性	96	49.58	19	16.58	技術職	
9	14100	男性	67	28.75	15	.50	事務職	
10	9900	女性	84	27.50	12	3.42	事務職	
11	21960	男性	83	31.08	15	4.08	管理職	
12	12420	男性	96	27.42	15	1.17	事務職	
13	15720	男性	84	33.50	15	6.00	事務職	
14	8880	男性	88	54.33	12	27.00	事務職	
15	22800	男性	98	41.17	15	12.00	管理職	
16	19020	男性	64	31.92	19	2.25	管理職	
17	10380	男性	72	32.67	15	6.92	事務職	
18	8520	男性	70	58.50	15	31.00	事務職	
19	11460	男性	79	46.58	15	21.75	事務職	
20	20500	男性	83	35.17	16	5.75	管理職	
21	27700	男性	85	43.25	20	11.17	技術職	
22	22000	男性	65	39.75	19	10.75	管理職	

図1

重回帰分析を行うときには、性別や職種といった名義データはそのままでは計算が出来ないので、データの数値化を行わなければならない。このときに、よく行われる間違いは、職種が、事務職、技術職、管理職の3つのカテゴリの場合、事務職＝1、技術職＝2、管理職＝3　といったデータの数値化で、このような数値化を行うと、事務職は管理職の3倍といったことになり、この変換は無意味であることがすぐにわかる。



	現在給料	性別	習熟度	年齢	就学年数	就業年数	職種	var
1	10620	1	88	34.17	15	5.08	1	
2	6960	1	72	46.50	12	9.67	1	
3	41400	0	73	40.33	16	12.50	3	
4	28350	0	83	41.92	19	13.00	3	
5	16080	0	79	28.00	15	3.17	1	
6	8580	1	72	45.92	8	16.17	1	
7	34500	0	66	34.25	18	4.17	2	
8	54000	0	96	49.58	19	16.58	2	
9	14100	0	67	28.75	15	.50	1	
10	9900	1	84	27.50	12	3.42	1	
11	21960	0	83	31.08	15	4.08	3	
12	12420	0	96	27.42	15	1.17	1	
13	15720	0	84	33.50	15	6.00	1	
14	8880	0	88	54.33	12	27.00	1	
15	22800	0	98	41.17	15	12.00	3	
16	19020	0	64	31.92	19	2.25	3	
17	10380	0	72	32.67	15	6.92	1	
18	8520	0	70	58.50	15	31.00	1	
19	11460	0	79	46.58	15	21.75	1	
20	20500	0	83	35.17	16	5.75	3	
21	27700	0	85	43.25	20	11.17	2	
22	22000	0	65	39.75	19	10.75	3	

図2

このような名義データの場合には、カテゴリの数だけ変数を用意し、その変数に　0と1という2値データを適用する。このようにして変換された変数をダミー変数という。性別の場合は、カテゴリ数が2なので、変数は次のように女性、男性となる。職種の場合はカテゴリ数が3なので、変数は次のように事務職、技術職、管理職となる。

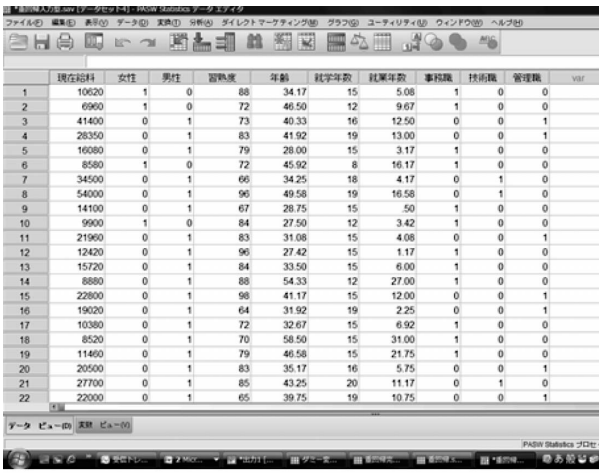


図3

3. カテゴリ数が2の場合の分散分析と単回帰分析

3. 1　カテゴリ数が2の場合の分散分析

独立変数の中で性別はカテゴリ数が2となっている。このとき、女性＝1と男性＝0の間で、給料に差があるかどうかを調べるときには、性別を因子、現在給料を従属変数として、1元配置の分散分析を行う。ＳＰＳＳでは、次のような手順となる。

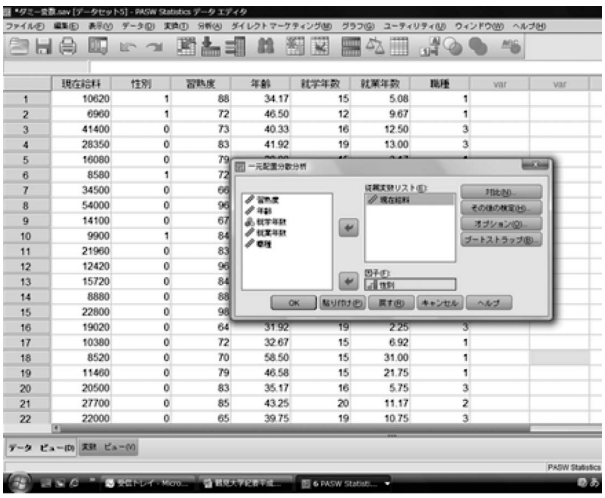


図3

この結果は、以下のように出力される。

表1　　1元配置の分散分析

現在の給与

	平方和	自由度	平均平方	F 値	有意確率
グループ間	2.781E9	1	2.781E9	98.224	.000
グループ内	1.093E10	283	4.169E7		
合計	1.372E10	284			

3. 2　カテゴリ数が2の場合の単回帰分析

性別を独立変数、現在給料を従属変数として、単回帰分析をしてみよう。ＳＰＳＳでは、次のような手順となる。

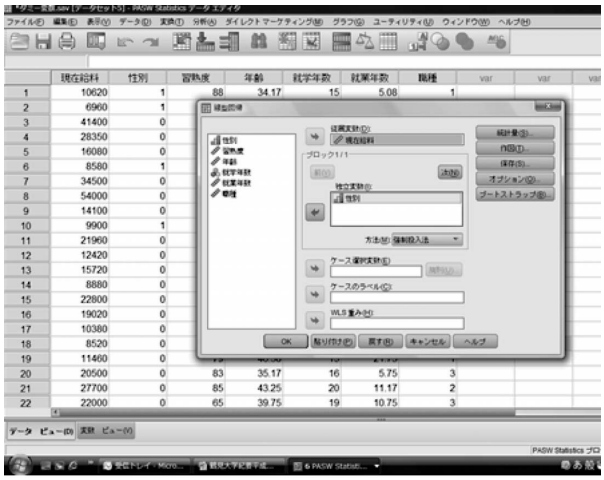


図5

この結果は、以下のように出力される。分散分析と単回帰分析とでは、分析の名前が異なるので、2つは別の分析と思えるのだが、分散分析のＦ値と単回帰分析のｔ値を比較すれば、この2つの検定は同じ検定であることが分かる。

したがって、ダミー変数を利用した回帰分析の場合、回帰係数の検定の解釈は、1元配置の分散分析を同じように、女性と男性の間の差を調べているという事になる。

表2　　単回帰係数a

モデル		標準化されていない係数		標準化係数	t 値	有意確率
		B	標準誤差	ベータ		
1	(定数)	16373.058	534.363		30.642	.000
	性別	-6489.120	797.403	-.448	-8.138	.000

a. 従属変数 現在の給与

次に、性別はカテゴリ数が2となっているので、ダミー変数を使って、単回帰分析をしてみよう。女性を独立変数、現在給料を従属変数として、単回帰分析を行う。独立変数として、女性と男性の2つの変数を利用しようとすると、共線性が起こるので、どちらか1つの変数しか分析に使えない。ＳＰＳＳでは、次のような手順となる。

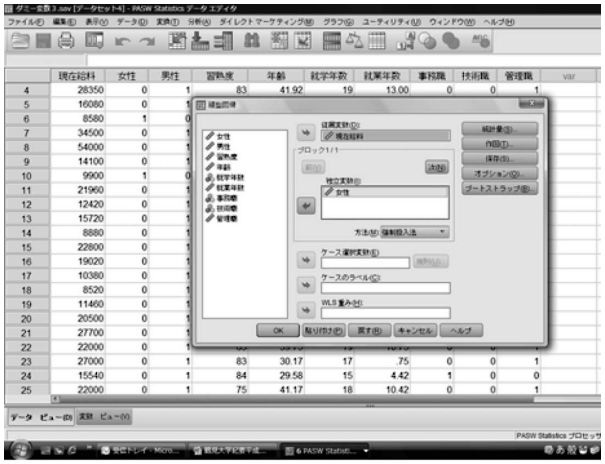


図6

この結果は、以下のように出力される。表1の分散分析のＦ値、表2の単回帰分析のｔ値、表3の単回帰分析のｔ値の3つを比較すれば、これらの検定はすべて同じ検定で、要するに2つのグループ間の差の検定ということが分かる。

表 3　　単回帰係数a

モデル		標準化されていない係数		標準化係数		
		B	標準誤差	ベータ	t 値	有意確率
1	(定数)	16373.658	534.363		30.642	.000
	女性	-6489.120	797.403	-.448	-8.138	.000

a. 従属変数 現在の給与

男性を独立変数として、単回帰分析を行ってみよう。

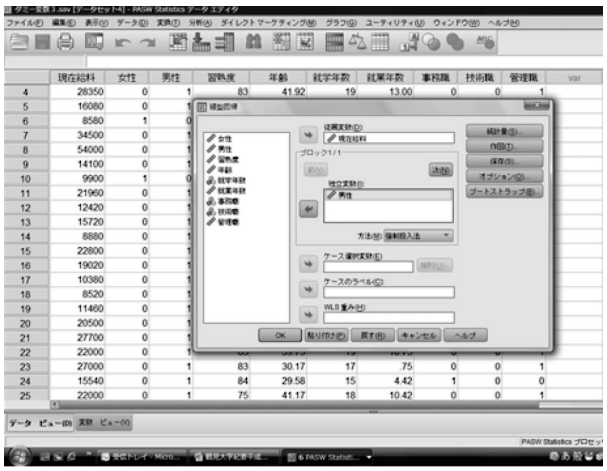


図7

その出力は、次のようになり、表3と表4との違いは、回帰係数の符号だけであることが分かる。

表4　　単回帰係数a

モデル		標準化されていない係数		標準化係数		
		B	標準誤差	ベータ	t 値	有意確率
1	(定数)	9884.538	591.877		16.700	.000
	男性	6489.120	797.403	.448	8.138	.000

a. 従属変数 現在の給与

4.　カテゴリ数が3の場合の重回帰分析

4. 1　技術職と管理職の場合

次に職種に注目してみよう。職種は、事務職、技術職、管理職の3つのカテゴリに分かれている。この3つの変数を用いて重回帰分析をしようとすると、多重共線性が起こるので、どれか1つの変数を分析からのぞかなければならない。始めに、事務職を除いてみよう。

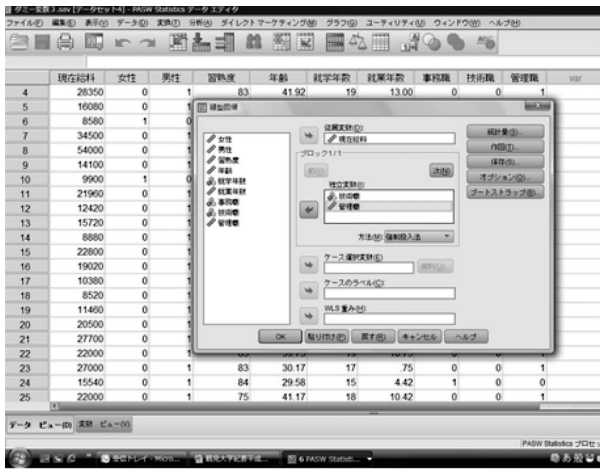


図8

この出力は、次のような結果となる。

表5　　単回帰係数a

モデル		標準化されていない係数		標準化係数		
		B	標準誤差	ベータ	t 値	有意確率
1	(定数)	11134.819	276.483		40.273	.000
	技術職	26956.847	1722.944	.528	14.833	.000
	管理職	14480.806	786.582	.695	18.384	.000

a. 従属変数 現在の給与

4. 2　事務職と管理職の場合

次に技術職を除いて、重回帰分析をしてみよう。

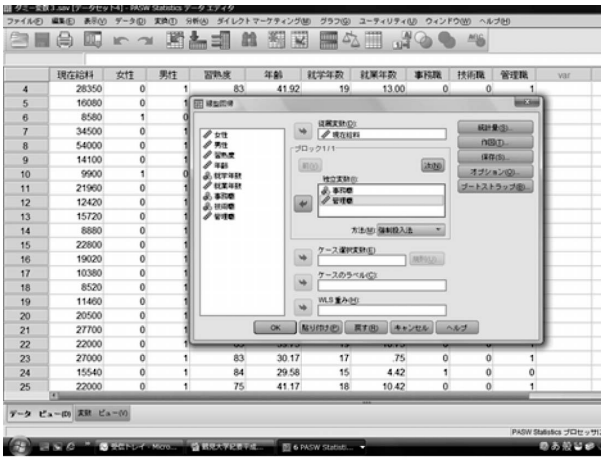


図9

この出力は、次のような結果となる。

表6　係数[a]

モデル		標準化されていない係数		標準化係数	t 値	有意確率
		B	標準偏差誤差	ベータ		
1	(定数)	38891.667	1700.618		21.576	.000
	事務職	-25956.847	1722.944	-1.245	-14.833	.000
	管理職	-11096.042	1853.203	-.502	-5.987	.000

a. 従属変数 現在の給与

4. 3　事務職と管理職の場合

最後に、管理職を除いてみよう

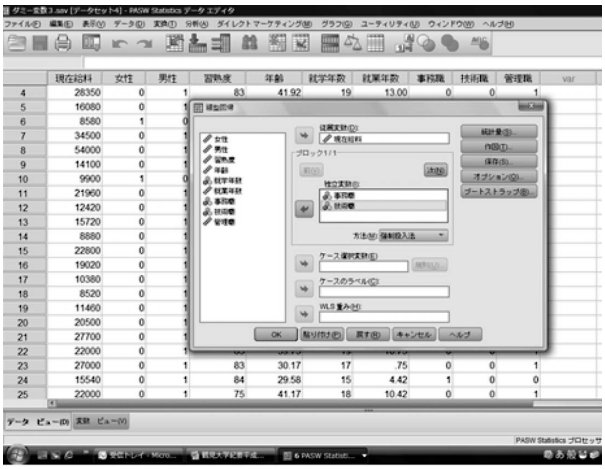


図10

この出力は、次のような結果となる。

表7　係数[a]

モデル		標準化されていない係数		標準化係数	t 値	有意確率
		B	標準偏差誤差	ベータ		
1	(定数)	25995.625	736.388		34.758	.000
	事務職	-14460.806	786.582	-.704	-18.384	.000
	技術職	11096.042	1853.203	.229	5.987	.000

a. 従属変数 現在の給与

以上のことから､カテゴリが3つのダミー変数の場合、多重共線性を回避するため、重回帰分析からどれか一つのダミー変数を除くと、その出力結果の解釈は、除いた変数を基準にした2つのグループの差の検定を行っているの同じであることが分かる。

例えば、事務職を分析からのぞいた場合、技術職の偏回帰係数の検定は、事務職を基準としたときの技術職との差の検定と同じであり、管理職の偏回帰係数の検定は、事務職を基準としたときの管理職との差の検定と同じである。

4. ダミー変数とその他の独立変数が含まれている場合の重回帰分析

図1のように、重回帰分析では多くの変数を独立変数として利用するのが一般的である。そこで、職種だけではなく、それ以外の変数も独立変数として取り上げた場合の重回帰分析をおこない、ダミー変数の偏回帰係数の検定を比較してみよう。その結果は、次の3つの表のようになる。

表8　事務職を除いたときの係数[a]

モデル		標準化されていない係数		標準化係数	t 値	有意確率
		B	標準偏差誤差	ベータ		
1	(定数)	3896.044	2695.201		1.525	.129
	技術職	22633.302	1612.526	.468	14.038	.000
	管理職	11902.670	818.938	.539	14.570	.000
	仕事の熟練度	46.425	22.653	.065	2.049	.041
	年齢	-90.449	28.827	-.148	-3.138	.002
	就学年数	548.452	97.440	.226	5.629	.000
	就業年数	10.611	38.948	.012	.272	.786

a. 従属変数 現在の給与

表9　技術職を除いたときの回帰係数a

モデル		標準化されていない係数		標準化係数		
		B	標準誤差	ベータ	t 値	有意確率
1	(定数)	28529.346	3307.251		8.022	.000
	事務職	-22633.302	1612.526	-1.102	-14.036	.000
	管理職	-10730.632	1643.722	-.486	-6.528	.000
	仕事の熟練度	46.425	22.653	.065	2.049	.041
	年齢	-90.449	28.827	-.148	-3.138	.002
	就学年数	548.452	97.440	.226	5.629	.000
	就業年数	10.611	38.948	.012	.272	.785

a. 従属変数 現在の給与

表10　管理職を除いたときの回帰係数a

モデル		標準化されていない係数		標準化係数		
		B	標準誤差	ベータ	t 値	有意確率
1	(定数)	15798.714	2808.915		5.624	.000
	事務職	-11902.670	818.938	-.580	-14.570	.000
	技術職	10730.632	1643.722	.222	6.528	.000
	仕事の熟練度	46.425	22.653	.065	2.049	.041
	年齢	-90.449	28.827	-.148	-3.138	.002
	就学年数	548.452	97.440	.226	5.629	.000
	就業年数	10.611	38.948	.012	.272	.785

a. 従属変数 現在の給与

参考文献

1）M.G.Kendall　Kendall's Advanced Theory of Statistics, Volume 1,2,3 CHARLS&GRIFIN

2）石村貞夫　他　「入門はじめての多変量解析」　東京図書

3）石村貞夫　他　「ＳＰＳＳによる多変量データ解析」第4版　東京図書

重回帰分析におけるダミー変数の解釈について

Some Remarks of dummy variables in multiple regression analysis

歯学部　准教授　石村貞夫

早稲田大学大学院 基幹理工学研究科　応用数学科

石村友二郎