多重共線性の意味について

重回帰分析などの多重共線性の目安として、説明変数の相関係数が 0.95 とか、VIF（Variance Inflation Factor）の値が 10 以上ということが言われているが、多重共線性は数学的な問題だけでなく、実用上どこが問題となるのか考えてみる。ここでの分析は、メニュー［分析－多変量解析他－予測手法－リッジ回帰分析他］を利用する。

VIF は、１つの説明変数 x_i を目的変数とした他の説明変数による重回帰分析での重相関係数 r_i を用いて以下のように定義される。

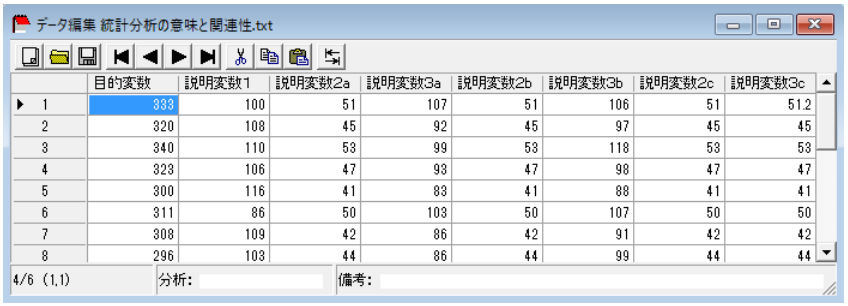
$$VIF_i = \frac{1}{1-r_i^2}$$

これによると、VIF の値が 10 程度というのは、重相関係数が約 0.95 ということになる。

$$VIF_i \simeq 10 \Leftrightarrow r_i \simeq 0.95$$

これから、説明変数同士の相関係数が 0.95 というのは納得の行く数値である。しかし、多重共線性は単に２つの変数同士の相関の問題ではなく、複数の変数間の関係性（束縛条件）の問題であるので、VIF は重要である。

図１のデータを元に多重共線性について調べてみる。



	目的変数	説明変数1	説明変数2a	説明変数3a	説明変数2b	説明変数3b	説明変数2c	説明変数3c
1	333	100	51	107	51	106	51	51.2
2	320	108	45	92	45	97	45	45
3	340	110	53	99	53	118	53	53
4	323	106	47	93	47	98	47	47
5	300	116	41	83	41	88	41	41
6	311	86	50	103	50	107	50	50
7	308	109	42	86	42	91	42	42
8	296	103	44	86	44	99	44	44

図１　データ

目的変数と説明変数１は共通で取り入れ、残りの変数は、説明変数２a と説明変数３a、説明変数２b と説明変数３b、説明変数２c と説明変数３c という順番に選ぶ。説明変数２と説明変数３については、a, b, c となるに連れて、相関が大きくなる。c については、最初のレコードだけが 0.2 違っている。

メニュー［分析－多変量解析他－予測手法－リッジ回帰分析他］を選択すると図 2 のような分析メニューが表示される。

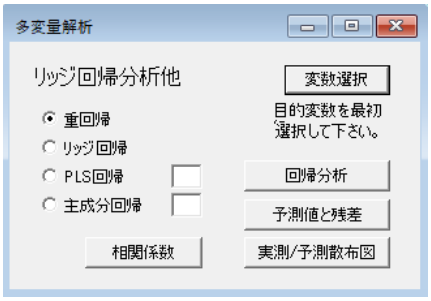


図２　分析実行メニュー

これは、多重共線性を見極め、それを回避する工学的に開発された手法を示すものである。

「重回帰分析」ラジオボックスを選択し、説明変数を a, b, c と選んで、「回帰分析」ボタンをクリックする。結果を図 3～図 5 に示す。

統計量（重回帰分析）

	偏回帰係数	標準偏差	標準化係数	VIF	残差分散	重相関R	寄与率R^2
▶ 説明変数1	1.3758	0.2679	0.7862	2.0106	51.7750	0.9019	0.8134
説明変数2a	0.6926	0.7616	0.2138	4.7396			
説明変数3a	1.5991	0.3787	1.0747	5.5556			
切片	-14.8325	46.2159	0.0000				

図 3　ほぼ問題のない結果

統計量（重回帰分析）

	偏回帰係数	標準偏差	標準化係数	VIF	残差分散	重相関R	寄与率R^2
▶ 説明変数1	0.9422	0.3204	0.5385	1.7112	86.9744	0.8286	0.6866
説明変数2b	6.1250	1.5267	1.8905	11.3370			
説明変数3b	-1.4149	0.6955	-0.9270	10.6017			
切片	67.8303	55.3885	0.0000				

図 4　問題のある結果

統計量（重回帰分析）

	偏回帰係数	標準偏差	標準化係数	VIF	残差分散	重相関R	寄与率R^2
▶ 説明変数1	0.9056	0.3472	0.5176	1.7239	101.4158	0.7966	0.6346
説明変数2c	-62.6317	58.4628	-19.3318	14257.2171			
説明変数3c	65.7902	58.3663	20.3274	14239.1764			
切片	68.1959	60.0788	0.0000				

図 5　完全に問題のある結果

図 5 をみると、寄与率は高くなっているが、偏回帰係数の値が大きくなって正と負で相殺している。またそれに伴い偏回帰係数の標準偏差も大きくなっている。これは、このデータでは予測値が当たっているが、新しいデータで少し値が異なると予測が大きくずれる可能性があることを意味している。これが多重共線性の問題点である。

ここでは c のデータについて、多重共線性を回避するためのリッジ回帰分析、PLS 回帰分析、主成分回帰分析の結果を図 6～図 8 に与えておく。特に後者 2 つについては独立と思われる説明変数の数を入力している。いずれも安定した解が得られている。

統計量（リッジ回帰分析）

	偏回帰係数	標準化係数	残差分散	重相関R	寄与率R^2	交差検証R	最良η
▶ 説明変数1	0.8760	0.5006	116.3666	0.7783	0.6057	0.7470	24.7000
説明変数2c	1.5081	0.4655					
説明変数3c	1.6077	0.4967					
切片	73.8804	0.0000					

図 6　リッジ回帰分析結果

統計量（PLS回帰分析）

	偏回帰係数	標準化係数	r-VIF	残差分散	重相関R	寄与率R^2	交差検証R	自由度
▶ 説明変数1	0.9397	0.5370	1.5028	109.0734	0.7791	0.6070	0.7458	2
説明変数2c	1.6260	0.5019	1.5028					
説明変数3c	1.6386	0.5063						
切片	60.2815	0.0000						

図 7　PLS 回帰分析結果

統計量（主成分回帰分析）

	偏回帰係数	標準化係数	r-VIF	残差分散	重相関R	寄与率R^2	交差検証R	自由度
▶ 説明変数1	0.9397	0.5370	1.0000	109.0748	0.7791	0.6070	0.7458	2
説明変数2c	1.6317	0.5036	1.0000					
説明変数3c	1.6328	0.5045						
切片	60.2825	0.0000						

図 8　主成分回帰分析結果