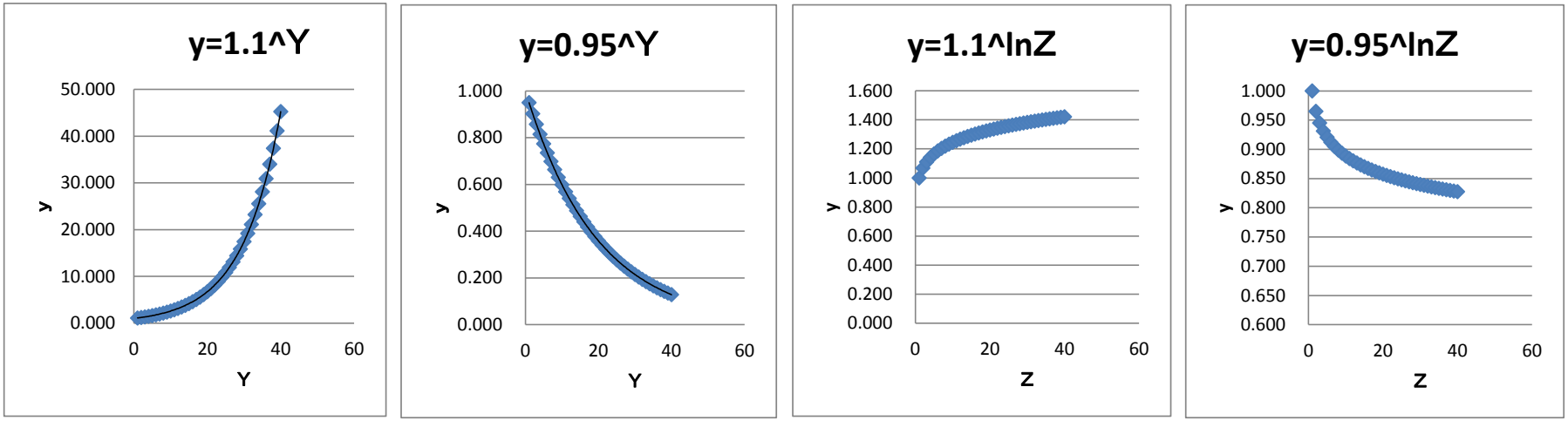
統計解析の基礎その2

目的変数と独立変数の関係は、線形、非線形、屈折点を持つ線形等さまざまなものを想定できるが、統計解析に入る前に標本モデルとしてどのようなモデルを想定するかが重要であると思われる。
即ち、標本そのものが線形である時に、非線形のモデルを想定して回帰式を求めても、線形モデルを基に求めた回帰式より精度が良くなるとは思えない。
それでは、実際に統計解析を行う場合に、標本モデルの想定はどのように行えばよいのか、または、その方法が解らなければ、各種のモデルを想定して解析を行い、その中で一番相関係数の良い方法を採用するしかないのか、シミュレーションにより検討してみた。

回帰式のモデル　　y＝Π aiXi　×　Π bj^(Yj)　×Π ck^(lnZk)　×　誤差項　・・・・・・・①
i=j=k＝1の場合、両辺の対数をとって

lny　＝　lna　+　lnX　+　(lnb)Y　+　(lnc)lnZ　＋　誤差項　・・・・・・・②

①式右辺第一項は、直線回帰である。第二項と第三項の関係を下図に示す。



標本モデルを、　　y = 0.5X　×　1.3^Y　×　2.0^(lnZ)　×　誤差項　とする。

X　:　　正規分布の乱数(平均μ ＝15、σ ＝5)
Y　:　　正規分布の乱数(平均μ ＝5、σ ＝1.6)
Z　:　　正規分布の乱数(平均μ ＝15、σ ＝5)
誤差項　:　正規分布の乱数(平均μ ＝1、σ ＝0　or　0.13)

y　は、以上で発生させた乱数の積により求めた。

求める回帰式は、　(ln y)　＝　(ln 0.5)　+　(ln X)　+　(ln 1.3)*Y　+　(ln 2.0)*(ln Z)　　である。

誤差項 σ ＝0 の場合

ln y	y	X	Y	Z
4.53	92.63	16.38	2.37	13.50
4.44	84.92	8.51	5.72	8.61
6.10	443.91	23.63	6.46	16.22
5.42	225.62	16.40	4.54	21.38
4.59	98.84	19.57	0.77	20.99
5.78	325.06	16.08	5.74	23.67
4.20	66.55	22.17	3.11	4.08
4.65	104.41	6.44	6.32	13.83
5.00	148.77	16.80	2.98	20.48
5.01	149.71	13.78	5.77	9.57
5.07	158.42	17.63	4.55	11.55
4.92	137.13	19.37	5.14	6.55
3.85	47.18	17.82	1.72	5.77
4.09	59.98	9.21	3.67	10.11
3.92	50.65	15.92	0.69	11.13
4.09	59.76	15.41	3.89	4.41
3.98	53.38	10.82	2.13	12.16
5.00	148.65	16.81	4.18	12.98
4.94	139.75	10.97	5.07	15.67
5.20	181.91	16.81	4.91	13.17
3.88	48.56	5.10	4.38	13.37
4.30	73.46	8.16	4.21	13.15
4.95	141.11	7.29	5.81	21.71
5.11	166.11	19.05	3.82	14.57
5.46	234.67	17.77	5.49	14.07
5.76	318.31	21.11	6.32	12.43
5.49	241.40	11.77	5.67	24.86
5.04	154.09	18.88	2.82	19.33
6.00	404.38	18.64	5.67	26.88
4.39	80.53	12.58	3.22	11.73
5.00	148.69	10.41	4.46	23.31
4.69	108.36	15.02	5.06	6.94
5.99	400.04	24.27	5.73	17.69
4.96	143.12	10.78	4.65	19.51
6.06	429.07	12.60	7.63	24.59
5.42	226.04	22.59	4.34	14.58
5.71	300.95	14.73	7.49	12.38
6.43	619.04	11.19	10.25	18.38
4.89	132.35	16.59	3.76	13.09
6.24	510.87	18.28	7.59	18.79

基本統計量

ln y		y		X		Y		Z	
平均	5.01	平均	191.46	平均	15.18	平均	4.70	平均	14.93
標準誤差	0.11	標準誤差	22.24	標準誤差	0.76	標準誤差	0.30	標準誤差	0.92
中央値（メ	5.00	中央値（メ	148.67	中央値（メ	16.23	中央値（メ	4.60	中央値（メ	13.66
標準偏差	0.70	標準偏差	140.67	標準偏差	4.82	標準偏差	1.90	標準偏差	5.84
分散	0.49	分散	19788.57	分散	23.22	分散	3.60	分散	34.06
尖度	−0.77	尖度	1.26	尖度	−0.59	尖度	1.07	尖度	−0.60
歪度	0.15	歪度	1.35	歪度	−0.20	歪度	0.24	歪度	0.15
範囲	2.57	範囲	571.86	範囲	19.17	範囲	9.56	範囲	22.80
最小	3.85	最小	47.18	最小	5.10	最小	0.69	最小	4.08
最大	6.43	最大	619.04	最大	24.27	最大	10.25	最大	26.88
合計	200.55	合計	7658.40	合計	607.36	合計	188.08	合計	597.19
標本数	40	標本数	40	標本数	40	標本数	40	標本数	40

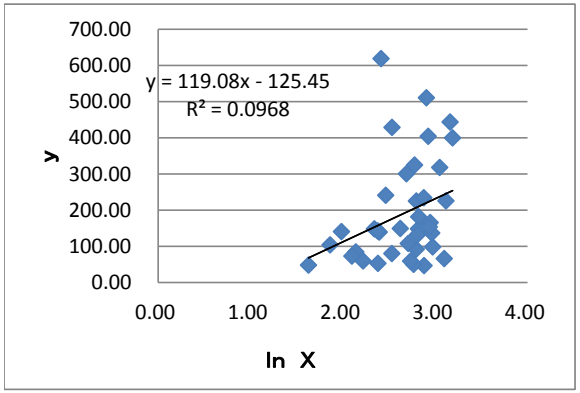
相関

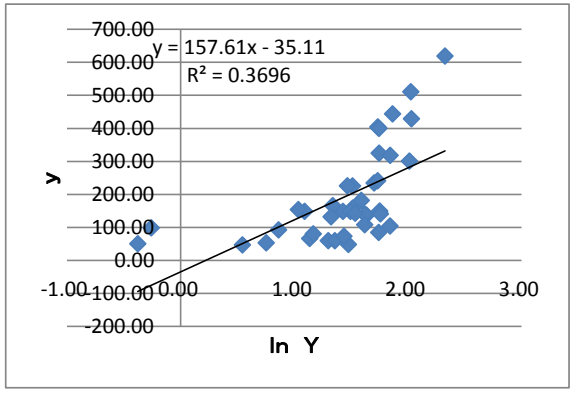
	y	X	Y	ln Z	Z	ln X	ln Y
y	1						
X	0.314651	1					
Y	0.774853	−0.13218	1				
ln Z	0.507811	−0.10362	0.27534	1			
Z	0.523166	−0.06693	0.272715	0.962834	1		
ln X	0.311105	0.978627	−0.14206	−0.09692	−0.05369	1	
ln Y	0.607956	−0.14361	0.921905	0.197964	0.199645	−0.16193	1

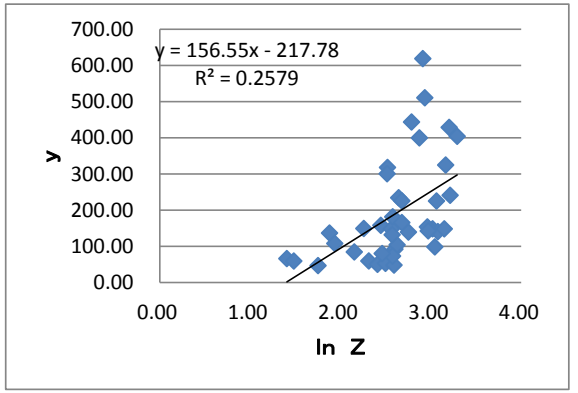
	ln y	X	Y	ln Z	Z	ln X	ln Y
ln y	1						
X	0.371138	1					
Y	0.756953	−0.13218	1				
ln Z	0.593706	−0.10362	0.27534	1			
Z	0.597728	−0.06693	0.272715	0.962834	1		
ln X	0.378325	0.978627	−0.14206	−0.09692	−0.05369	1	
ln Y	0.656545	−0.14361	0.921905	0.197964	0.199645	−0.16193	1

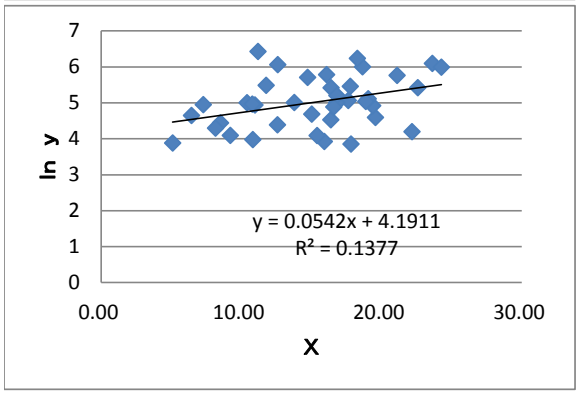
目的変数 y または ln yと相関係数の高いのは、ln X、Y、ln Zである。
そこで、独立変数の形体としては、ln X、Y、ln Zを採用する。

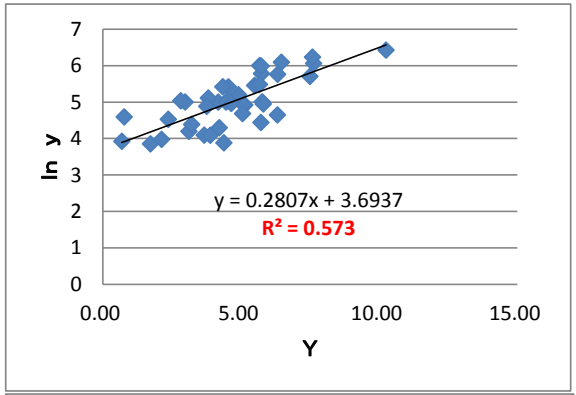
散布図　及び　相関係数

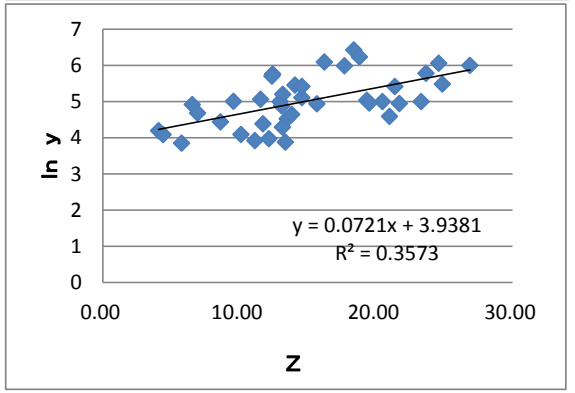

y = 119.08x - 125.45
R² = 0.0968
y
ln X

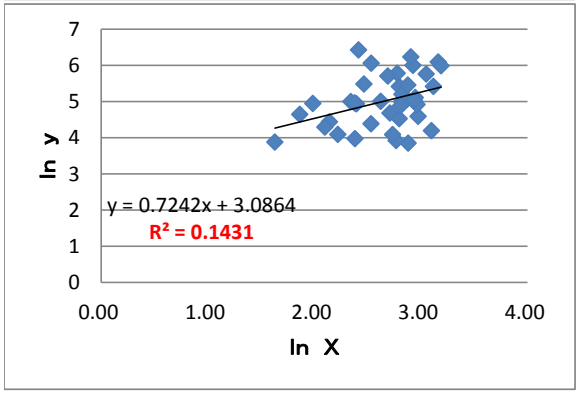

y = 157.61x - 35.11
R² = 0.3696
y
ln Y

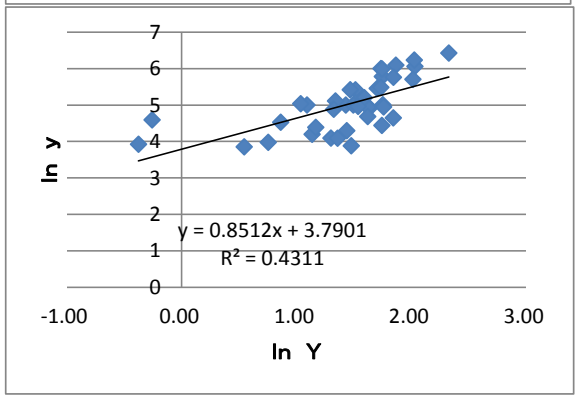

y = 156.55x - 217.78
R² = 0.2579
y
ln Z

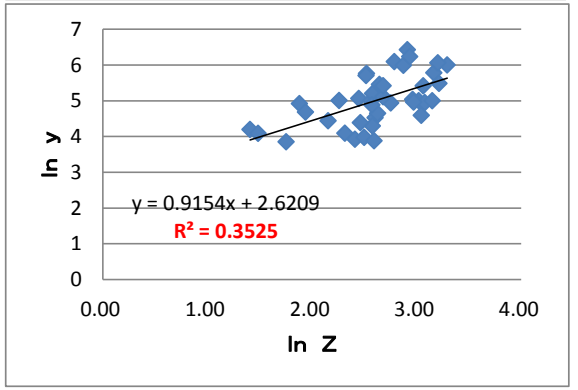

y = 0.0542x + 4.1911
R² = 0.1377
ln y
X


y = 0.2807x + 3.6937
R² = 0.573
ln y
Y


y = 0.0721x + 3.9381
R² = 0.3573
ln y
Z


y = 0.7242x + 3.0864
R² = 0.1431
ln y
ln X


y = 0.8512x + 3.7901
R² = 0.4311
ln y
ln Y


y = 0.9154x + 2.6209
R² = 0.3525
ln y
ln Z

標準化 及び 重回帰分析

ln y	ln X	Y	ln Z
−0.69	0.37	−1.23	−0.03
−0.81	−1.41	0.54	−1.01
1.54	1.36	0.93	0.38
0.58	0.37	−0.08	0.98
−0.60	0.85	−2.07	0.94
1.09	0.32	0.55	1.21
−1.16	1.19	−0.84	−2.65
−0.52	−2.17	0.85	0.03
−0.02	0.44	−0.91	0.89
−0.01	−0.10	0.56	−0.78
0.07	0.57	−0.08	−0.37
−0.13	0.82	0.23	−1.61
−1.65	0.60	−1.57	−1.89
−1.31	−1.20	−0.54	−0.66
−1.55	0.29	−2.12	−0.45
−1.31	0.20	−0.43	−2.48
−1.47	−0.76	−1.36	−0.25
−0.02	0.44	−0.28	−0.11
−0.11	−0.72	0.19	0.30
0.27	0.44	0.11	−0.08
−1.61	−2.81	−0.17	−0.05
−1.02	−1.53	−0.26	−0.08
−0.09	−1.84	0.58	1.02
0.14	0.78	−0.47	0.14
0.63	0.59	0.42	0.07
1.07	1.06	0.86	−0.21
0.67	−0.53	0.51	1.31
0.03	0.75	−0.99	0.76
1.41	0.72	0.51	1.48
−0.89	−0.35	−0.78	−0.33
−0.02	−0.87	−0.13	1.17
−0.47	0.13	0.19	−1.48
1.39	1.44	0.54	0.57
−0.07	−0.77	−0.03	0.78
1.49	−0.35	1.54	1.29
0.58	1.24	−0.19	0.14
0.99	0.08	1.47	−0.21
2.01	−0.67	2.92	0.65
−0.18	0.40	−0.50	−0.09
1.74	0.67	1.52	0.70

概要

回帰統計	
重相関 R	1
重決定 R2	1
補正 R2	1
標準誤差	4.48E−16
観測数	40

分散分析表

	自由度	変動	分散	測された分散	有意 F
回帰	3	39.00266	13.00089	6.47E+31	0
残差	36	7.24E−30	2.01E−31		
合計	39	39.00266			

	係数	標準誤差	t	P−値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	−0.0005	7.09E−17	−7.1E+12	0	−0.0005	−0.0005	−0.0005	−0.0005
ln X	0.522388	7.27E−17	7.19E+15	0	0.522388	0.522388	0.522388	0.522388
Y	0.707395	7.52E−17	9.4E+15	0	0.707395	0.707395	0.707395	0.707395
ln Z	0.449585	7.48E−17	6.01E+15	0	0.449585	0.449585	0.449585	0.449585

回帰式 : **(ln y) ＝ (ln 0.5) + (ln X) + (ln 1.3)*Y + (ln 2.0)*(ln Z)**

以上により、当初想定した標本モデルの回帰式を求めることができた。

σ ＝0.13の場合

lny	y	X	Y	Z
4.52	91.96	16.38	2.37	13.50
4.44	84.62	8.51	5.72	8.61
5.79	325.45	23.63	6.46	16.22
5.42	226.51	16.40	4.54	21.38
4.78	119.35	19.57	0.77	20.99
5.96	387.39	16.08	5.74	23.67
4.23	68.67	22.17	3.11	4.08
4.76	116.49	6.44	6.32	13.83
5.09	162.58	16.80	2.98	20.48
5.15	171.74	13.78	5.77	9.57
5.23	187.63	17.63	4.55	11.55
4.89	132.64	19.37	5.14	6.55
3.78	43.63	17.82	1.72	5.77
4.11	61.09	9.21	3.67	10.11
4.06	58.12	15.92	0.69	11.13
4.03	56.17	15.41	3.89	4.41
4.09	59.99	10.82	2.13	12.16
4.81	122.85	16.81	4.18	12.98
4.90	133.92	10.97	5.07	15.67
5.23	186.23	16.81	4.91	13.17
3.91	50.01	5.10	4.38	13.37
4.47	87.11	8.16	4.21	13.15
4.98	144.75	7.29	5.81	21.71
5.18	177.86	19.05	3.82	14.57
5.42	225.74	17.77	5.49	14.07
5.81	332.38	21.11	6.32	12.43
5.33	206.07	11.77	5.67	24.86
4.89	132.98	18.88	2.82	19.33
5.96	388.43	18.64	5.67	26.88
4.55	95.00	12.58	3.22	11.73
5.12	166.94	10.41	4.46	23.31
4.63	102.05	15.02	5.06	6.94
5.91	368.34	24.27	5.73	17.69
4.84	126.72	10.78	4.65	19.51
6.04	421.63	12.60	7.63	24.59
5.37	215.05	22.59	4.34	14.58
5.89	362.11	14.73	7.49	12.38
6.49	661.66	11.19	10.25	18.38
4.55	95.05	16.59	3.76	13.09
6.17	477.76	18.28	7.59	18.79

ln y		y		X		Y		Z	
平均	5.02	平均	190.87	平均	15.18	平均	4.70	平均	14.93
標準誤差	0.11	標準誤差	22.00	標準誤差	0.76	標準誤差	0.30	標準誤差	0.92
中央値（メ	4.94	中央値（メ	139.34	中央値（メ	16.23	中央値（メ	4.60	中央値（メ	13.66
標準偏差	0.69	標準偏差	139.12	標準偏差	4.82	標準偏差	1.90	標準偏差	5.84
分散	0.47	分散	19354.86	分散	23.22	分散	3.60	分散	34.06
尖度	−0.70	尖度	2.13	尖度	−0.59	尖度	1.07	尖度	−0.60
歪度	0.17	歪度	1.47	歪度	−0.20	歪度	0.24	歪度	0.15
範囲	2.72	範囲	618.03	範囲	19.17	範囲	9.56	範囲	22.80
最小	3.78	最小	43.63	最小	5.10	最小	0.69	最小	4.08
最大	6.49	最大	661.66	最大	24.27	最大	10.25	最大	26.88
合計	200.78	合計	7634.65	合計	607.36	合計	188.08	合計	597.19
標本数	40	標本数	40	標本数	40	標本数	40	標本数	40

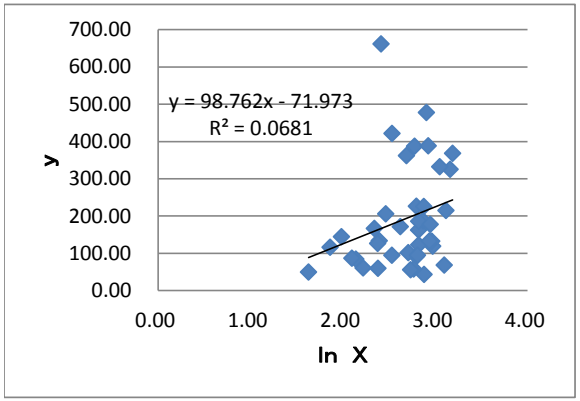
相関

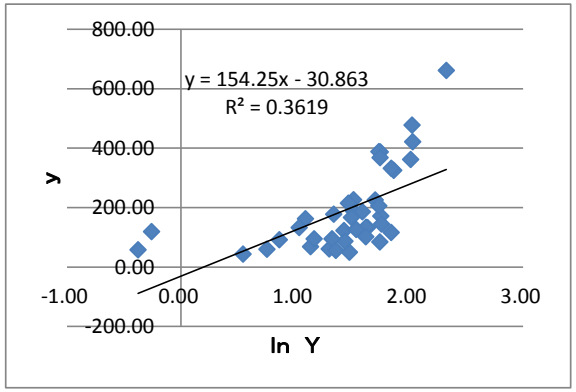
	y	X	Y	ln Z	Z	ln X	ln Y
y	1						
X	0.251192	1					
Y	0.785948	−0.13218	1				
ln Z	0.502626	−0.10362	0.27534	1			
Z	0.517369	−0.06693	0.272715	0.962834	1		
ln X	0.260899	0.978627	−0.14206	−0.09692	−0.05369	1	
ln Y	0.601599	−0.14361	0.921905	0.197964	0.199645	−0.16193	1

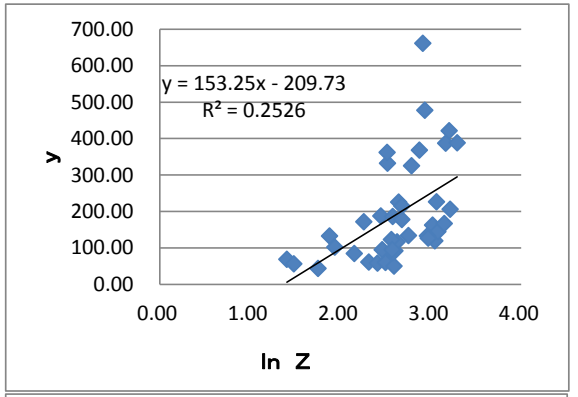
	ln y	X	Y	ln Z	Z	ln X	ln Y
ln y	1						
X	0.325937	1					
Y	0.755863	−0.13218	1				
ln Z	0.606228	−0.10362	0.27534	1			
Z	0.605818	−0.06693	0.272715	0.962834	1		
ln X	0.338015	0.978627	−0.14206	−0.09692	−0.05369	1	
ln Y	0.635106	−0.14361	0.921905	0.197964	0.199645	−0.16193	1

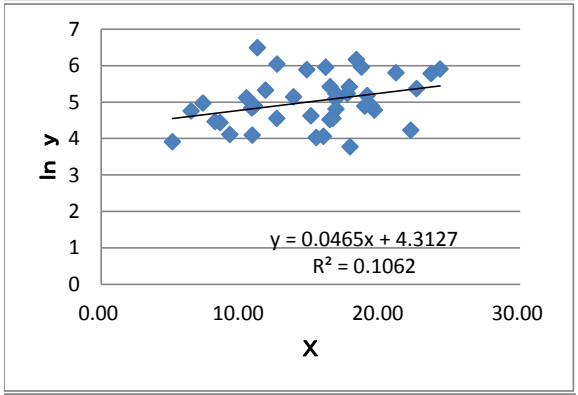
最も相関係数の高い、ln X、Y、ln Zを採用する。

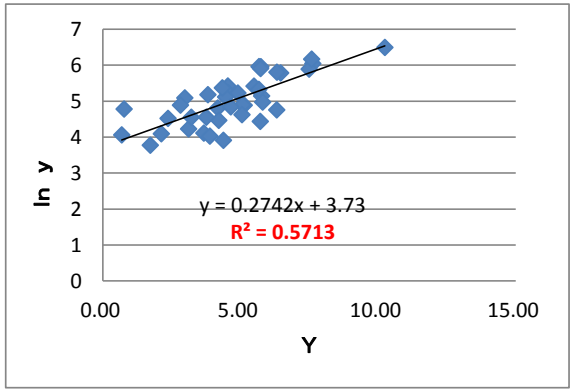
散布図　及び　相関係数

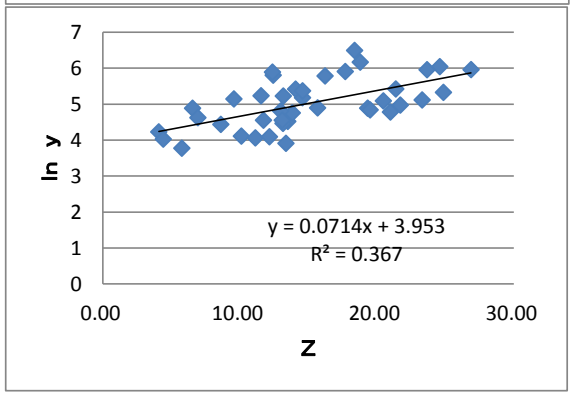

y = 98.762x - 71.973
R² = 0.0681
y
ln X

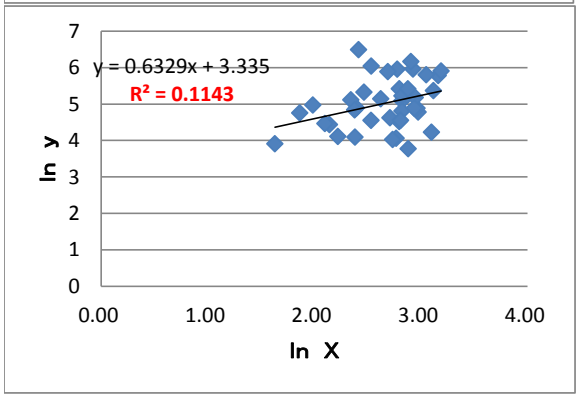

y = 154.25x - 30.863
R² = 0.3619
y
ln Y

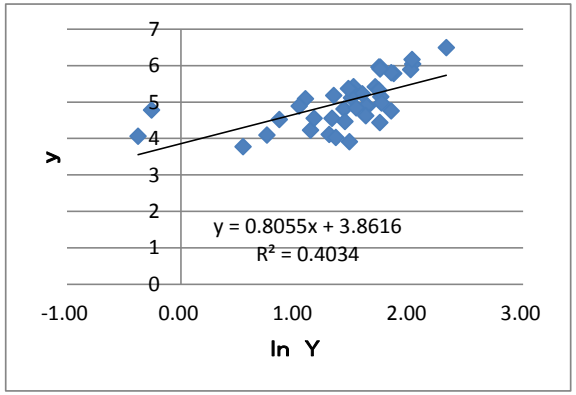

y = 153.25x - 209.73
R² = 0.2526
y
ln Z

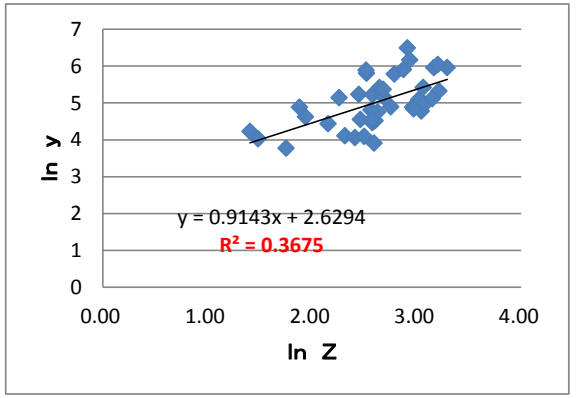

y = 0.0465x + 4.3127
R² = 0.1062
ln y
X


y = 0.2742x + 3.73
R² = 0.5713
ln y
Y


y = 0.0714x + 3.953
R² = 0.367
ln y
Z


y = 0.6329x + 3.335
R² = 0.1143
ln y
ln X


y = 0.8055x + 3.8616
R² = 0.4034
y
ln Y


y = 0.9143x + 2.6294
R² = 0.3675
ln y
ln Z

標準化 及び 重回帰分析

標準化

lny	ln X	Y	ln Z
−0.72	0.37	−1.23	−0.03
−0.84	−1.41	0.54	−1.01
1.11	1.36	0.93	0.38
0.59	0.37	−0.08	0.98
−0.34	0.85	−2.07	0.94
1.37	0.32	0.55	1.21
−1.15	1.19	−0.84	−2.65
−0.38	−2.17	0.85	0.03
0.10	0.44	−0.91	0.89
0.18	−0.10	0.56	−0.78
0.31	0.57	−0.08	−0.37
−0.19	0.82	0.23	−1.61
−1.81	0.60	−1.57	−1.89
−1.32	−1.20	−0.54	−0.66
−1.39	0.29	−2.12	−0.45
−1.44	0.20	−0.43	−2.48
−1.34	−0.76	−1.36	−0.25
−0.30	0.44	−0.28	−0.11
−0.18	−0.72	0.19	0.30
0.30	0.44	0.11	−0.08
−1.61	−2.81	−0.17	−0.05
−0.80	−1.53	−0.26	−0.08
−0.06	−1.84	0.58	1.02
0.23	0.78	−0.47	0.14
0.58	0.59	0.42	0.07
1.14	1.06	0.86	−0.21
0.45	−0.53	0.51	1.31
−0.19	0.75	−0.99	0.76
1.37	0.72	0.51	1.48
−0.68	−0.35	−0.78	−0.33
0.14	−0.87	−0.13	1.17
−0.57	0.13	0.19	−1.48
1.29	1.44	0.54	0.57
−0.26	−0.77	−0.03	0.78
1.49	−0.35	1.54	1.29
0.51	1.24	−0.19	0.14
1.27	0.08	1.47	−0.21
2.14	−0.67	2.92	0.65
−0.68	0.40	−0.50	−0.09
1.67	0.66	1.52	0.70

概要

回帰統計	
重相関 R	0.984578
重決定 R2	0.969394
補正 R2	0.966843
標準誤差	0.18209
観測数	40

分散分析表

	自由度	変動	分散	測された分散	有意 F
回帰	3	37.8064	12.60213	380.0755	2.67E−27
残差	36	1.193649	0.033157		
合計	39	39.00005			

	係数	標準誤差	t	P−値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	6.95E−07	0.028791	2.41E−05	0.999981	−0.05839	0.058392	−0.05839	0.058392
ln X	0.481764	0.029511	16.32484	3.21E−18	0.421912	0.541615	0.421912	0.541615
Y	0.6974	0.030553	22.82585	5.28E−23	0.635436	0.759365	0.635436	0.759365
ln Z	0.460901	0.030386	15.16807	3.27E−17	0.399275	0.522527	0.399275	0.522527

回帰式 : **ln y** ＝ 0.9021*(ln X) +0.253019*Y + 0.695117*(ln Z)

＝ 0.90*(ln X) ＋ (ln 1.29)*Y + (ln 2.004)*(ln Z)

誤差項を含んだ分精度が悪くなっている。

まとめ

目的変数y または ln yに対して、各独立変数の二乗、、平方根、対数等との相関係数を比較し、各独立変数毎に相関係数の大きなものを選べば、
ほぼ、最適な重回帰が行えそうである。
過去の知見等により、二乗が良いのか平方根が良いのかが解っていれば、より精度の高い関数を選ぶことができることになる。

念のために他の組み合わせを採用した時の回帰統計等を比較してみる。
σ ＝0.13の場合、
ケース Ⅰ 目的変数 ln y
独立変数 ln X、 Y、 ln Z
ケース Ⅱ 目的変数 ln y
独立変数 X、 ln Y、 Z

	ケース Ⅰ					ケース Ⅱ				
回帰統計	重相関 R	0.98458				重相関 R	0.91485			
	重決定 R2	0.96939				重決定 R2	0.83696			
	補正 R2	0.96684				補正 R2	0.82337			
	標準誤差	0.18209				標準誤差	0.42027			
	観測数	40				観測数	40			
分散分析表		自由度	変動	分散	測された分散		自由度	変動	分散	測された分散	有意 F
	回帰	3	37.8064	12.60213	380.0755	2.67E-27	3	32.64137	10.88046	61.60022	2.98E-14
	残差	36	1.193649	0.033157			36	6.358686	0.17663		
	合計	39	39.00005				39	39.00005			
		係数	標準誤差	t	P-値		係数	標準誤差	t	P-値	
	切片	6.95E-07	0.028791	2.41E-05	0.999981		6.33E-07	0.066451	9.53E-06	0.999992	
	ln X	0.481764	0.029511	16.32484	3.21E-18		0.446115	0.068056	6.555151	1.27E-07	
	Y	0.6974	0.030553	22.82585	5.28E-23		ln Y	0.596022	0.069298	8.600837	2.96E-10
	ln Z	0.460901	0.030386	15.16807	3.27E-17		Z	0.516683	0.068734	7.517153	6.97E-09