

ポアソン回帰分析入門
—細胞数をカウントしたデータの解析—

第 17 回 高橋セミナー

2004 年 4 月 10 日

高橋 行雄

10/31/2005 7:37 PM

表紙裏

2004年1月19日, 新規作成,

ファイル名 : G:¥R3_高橋_2004_から¥R_ANZ_高橋 seminar_17_ポアソン¥Semi_17_ポアソン回帰.doc

非線形回帰分析入門

目 次

1. はじめに	1
2. データ例示	2
2.1. 遺伝毒性試験データ（15回セミナーより）	2
2.2. 光細胞毒性試験のデータ（16回セミナー）	3
2.3. PKパラメータの例	4
2.4. 建築面積と建築費の関係（芳賀のSASによる回帰分析より）	5
3. 反応の増加に伴う誤差の増加との関係	7
3.1. 細胞の数のカウントデータはポアソン分布	7
3.2. 細胞数を吸光度計で測定した場合	8
3.3. PKパラメータの例	10
3.4. 芳賀本の建築面積と建築費の例	11
4. ポアソン分布	13
4.1. ポアソン分布の基礎	13
4.2. JMPによるポアソン分布の作図	16
5. ポアソン回帰, Armitageの本, 4 th editon	17
6. ポアソン回帰の細胞数のカウントデータへの適用	25
7. 重み付き回帰分析の適用	25
7.1. 統計ソフトでの重み	25
7.2. JMPでの解析	26
7.3. 非線形回帰分析を用いる場合	30
8. 細胞毒性での応用	34

9. 各種の効力比、	34
9.1. 勾配比の推定、重み付き非線形最小2乗法の適用	35
9.2. 反応 y が対数変換されている場合	35
9.3. 投与量が対数変換されている場合	37

図 表 目 次

図 2.1 変異コロニー数の用量反応関係	3
図 2.2 光細胞毒性試験の吸光度の用量反応関係	4
図 2.3 投与量の増加と共に AUC が増加	5
図 3.1 吸光度データでの反応の変動との関連	10
図 3.2 原点を通る回帰直線	11
図 3.3 残差 e と予測値 y , e/y と y の関係	12
図 3.4 e^2 と推定値 y , e^2/y と y の関係	12
図 3.5 \sqrt{e} と推定値 y , \sqrt{e}/y と y の関係	12
図 4.1 λ が小さい場合の裾広がりのポアソン分布	16
図 4.2 λ が大きい場合のやや裾をひいているポアソン分布	16
図 5.1 Veterans vs. Non-Veterans	22
図 6.1 細胞毒性データに対する平方根変換と対数変換	25
図 7.1 0 次の結果	28
図 7.2 1 次の回帰分析の結果	29
図 7.3 2 次の回帰分析の結果	29
図 7.4 0 次結果	32
図 7.5 1 次の結果	32
図 7.6 2 次の結果	33
図 9.1 対数変換	36
表 2.1 変異コロニー数と比活性	2
表 2.2 化合物 A の光細胞毒性試験の吸光度と反応率	4
表 2.3 投与量と AUC の関係	5
表 3.1 用量の増加に伴う CV の変化	7
表 3.2 吸光度のばらつき	9
表 3.3 SD が平均に比例して増大	10
表 4.1 いくつかの λ (平均) に対するポアソン乱数の発生	16
表 5.1 退役軍人の群における癌の発生ケースの数とリスクの人年	21
表 5.2 GENMOD による推定	24
表 7.1 0 次の回帰分析データ	27
表 7.2 予測値の出力	28
表 7.3 初期値の設定	31
表 8.1 用量の増加に伴う CV の変化	34
表 8.2 重み付き非線形回帰	34
表 9.1 用量の増加に伴う CV の変化	35
表 9.2 重み付き非線形回帰	35
表 9.3 反応が対数の場合での切片の差の推定	36
表 9.4 S 薬の $y = 200$ および T 薬との差の逆推定	38

1. はじめに

化学物質の曝露量に関連してシャーレの中の死亡細胞をカウントするような実験では、死亡する細胞の数が増加すると共にばらつきも増大することが知られている。死亡数の増加とともに分散が比例的に増加するような場合に、死亡数はポアソン分布に従うのであるが、これまでのセミナーでは、これらの現象がうかがわれるような実験データであっても、分散が均一とみなせるものとして様々な解析方法を提示してきた。

しかしながら、これまで扱ってきた細胞を対象にした事例も、よく観察すれば平均値に関して分散も大きくなっている。今回は、この問題に焦点をあてる。さらに、平均値の増加に比例して SD が大きくなる、 $CV=SD/\text{平均}$ が一定となる場合についても、まじめに検討してみる。

2. データ例示

2.1. 遺伝毒性試験データ（15回セミナーより）

ある化学物質に復帰突然変異が陽性であるのかの判定は、2倍法、あるいは一般的な用量反応関係を検討するために、多くの統計手法が開発されてきた。さて、復帰突然変異が陽性である2つの化学物質の相対的な比較を行ないたい。どのような方法があるのだろうか。表 2.1 に復帰突然変異が陽性の物質Sと、代替物質Tの復帰突然変異の減弱の程度を比較したい。変異コロニー数の用量反応関係を 図 2.1 に示す。

表 2.1 変異コロニー数と比活性

溶液濃度 ($\mu\text{g}/\text{plate}$)	変異コロニー数		比活性	
	Test drug	Standard	Test drug	Standard
0	23	27	.	.
	26	33	平均 = 26.6667	
	26	25	.	.
	<hr/>			
50	68	68	0.827	0.827
	82	89	1.107	1.247
	72	81	0.907	1.087
75	99	131	0.964	1.391
	85	130	0.778	1.378
	115	117	1.178	1.204
100	137	144	1.103	1.173
	131	157	1.043	1.303
	134	159	1.073	1.323
125	189	199	1.299	1.379
	177	208	1.203	1.451
	168	198	1.131	1.371
150	197	260	1.136	1.556
	195	229	1.122	1.349
	220	228	1.289	1.342
*300	335	427	1.028	1.334
	332	407	1.018	1.268
	348	456	1.071	1.431

*2004年2月7日の安全研での事例研究ではこの用量は除いた

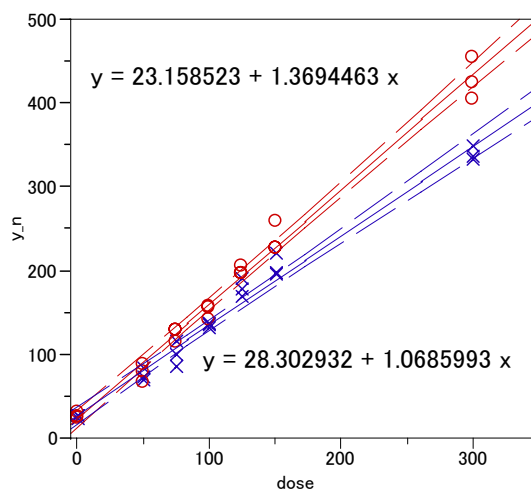


図 2.1 変異コロニー数の用量反応関係

2.2. 光細胞毒性試験のデータ（16回セミナー）

表 2.2 に光細胞毒性試験の実験データを示す。このデータは、ある物質Aを調整して光の照射をしないプレート上のウェル（穴）に 8 用量、 $2^3 \sim 2^{10}$ mg/mL（単位は要確認）、光を照射するプレートでも 8 用量とするが、 $2^0 \sim 2^7$ mg/mLと濃度を 3 段階低く設定し、それぞれ 2 回の繰り返しとする。陰性対照は、光の照射（具体的に光の波長、強さ、時間などを書く）によって影響がない細胞を含む溶液とする。陰性対照は、細胞を含まない溶液のみとする。

ある一定時間経過後に、プレート上のそれぞれのウェルで物質Aの濃度によって死亡細胞が変化する。生存した細胞に色のつく処理を施すと、死亡した細胞が多くなるウェル内の溶液の濁りが澄んで光が透過しやすくなる。微弱な測定用の光をそれぞれのウェルにあてて、ウェルから出てくる光の量と入れた光の量との差を測定する。溶液が濁っていると入れた光は拡散し出てくる光が見かけ上少なくなる。言い換えると光が吸収されてしまったようになり、この相対的な大きさが 0 から 2（要確認）ぐらいになるように設定し、それぞれのウェルの吸光度を測定する。これが、表 2.2 の吸光度の欄のデータである。

吸光度が実際の測定データであるが、一般的に陰性対照を 100%、陽性対照を 0%となるような反応率にして毒性の強さを基準化されている。反応率が高ければ毒性がなく、低ければ毒性が強いと判定する。反応率は相対的な毒性の強さなので、100%を超えることもあれば、0%以下になることもある。これは、光照射がある場合の反応率で 2^0 と 2^1 mg/mL 溶

液に 109.9%と 102.2%が 100%を超えるデータ例である.

表 2.2 化合物 A の光細胞毒性試験の吸光度と反応率

濃度	光：なし		濃度	光：あり			
	吸光度：y			吸光度：y			
陰性 対照	1.177 1.196	1.146 1.247	—	陰性 対照	1.194 1.270	1.330 1.170	—
2 ³	1.097	1.018	89.8 83.0	2 ⁰	1.331	1.084	109.9 88.7
2 ⁴	0.795	0.886	63.8 71.6	2 ¹	1.250	1.062	102.9 86.8
2 ⁵	0.731	0.701	58.3 55.8	2 ²	0.810	1.022	65.1 83.3
2 ⁶	0.409	0.426	30.7 32.1	2 ³	0.566	0.533	44.2 41.3
2 ⁷	0.217	0.220	14.2 14.4	2 ⁴	0.268	0.293	18.6 20.7
2 ⁸	0.178	0.156	10.8 8.9	2 ⁵	0.119	0.102	5.8 4.3
2 ⁹	0.122	0.133	6.0 7.0	2 ⁶	0.092	0.066	3.4 1.2
2 ¹⁰	0.135	0.111	7.1 5.1	2 ⁷	0.126	0.142	6.4 7.7
陽性 対照	0.041 0.016	0.054 0.075	—	陽性 対照	0.040 0.109	0.064 0.092	—

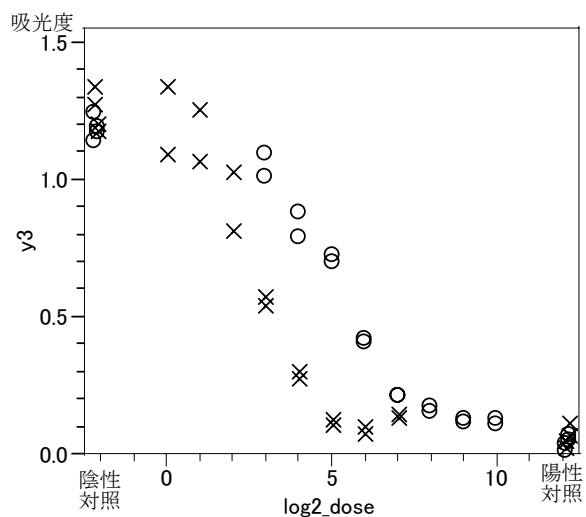


図 2.2 光細胞毒性試験の吸光度の用量反応関係

2.3. PK パラメータの例

表 2.3 投与量と AUC の関係

用量	AUC							
0.5	71.46	58.97	60.00	64.02	68.02	66.45	61.26	64.27
	67.37	68.02	79.21	61.72	82.18	60.61	71.99	70.68
1	147.73	118.22	144.89	140.64	146.15	121.84	106.58	109.62
	160.71	116.89	123.17	162.22	143.71	124.50	116.84	110.84
2	264.83	278.27	278.39	278.15	239.11	218.81	222.39	213.87
	237.92	278.47	249.27	238.62	266.57	191.40	198.34	270.00

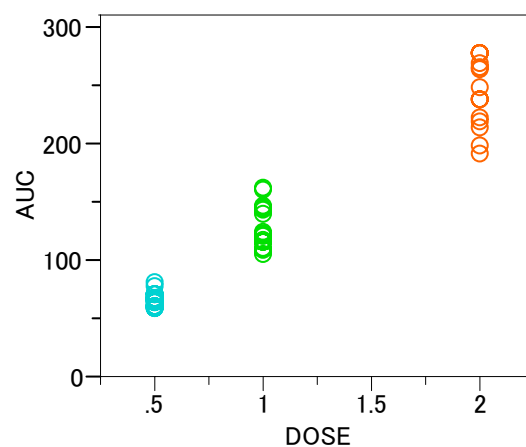


図 2.3 投与量の増加と共に AUC が増加

2.4. 建築面積と建築費の関係（芳賀の SAS による回帰分析より）

SASで学ぶ統計的データ解析 6

SASによる

回帰分析

竹内 啓 監修

芳賀敏郎・野澤昌弘・岸本淳司 著

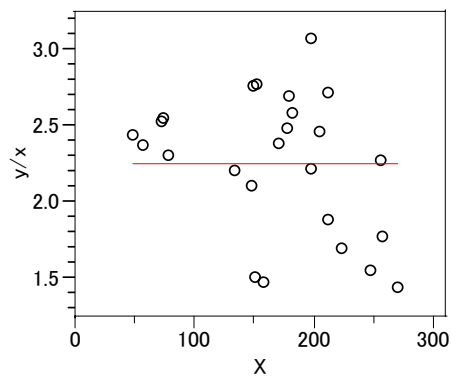
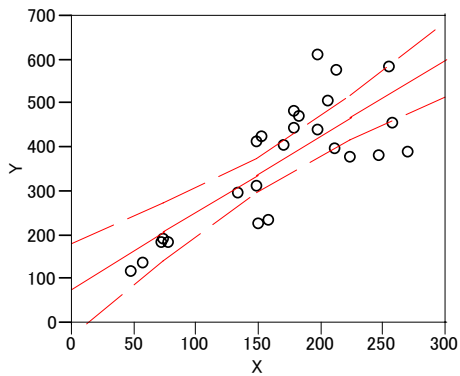
ある建設会社で、25の建物についてその面積 x_i と建築費 y_i の実績値を収集し、それから単位面積当たりの標準単価を求めようとしている。そのデータを図表9.1に示す。

図表 9.1 面積と建築費のデータ

	x	y	y/x		x	y	y/x		x	y	y/x
1	258	458	1.775	10	58	138	2.379	19	159	235	1.577
2	212	400	1.887	11	224	381	1.701	20	159	314	1.975
3	134	296	2.209	12	248	385	1.552	21	75	192	2.560
4	171	408	2.386	13	150	415	2.767	22	199	442	2.221
5	73	185	2.534	14	199	613	3.080	23	271	390	1.439
6	179	445	2.489	15	180	485	2.694	24	213	579	2.718
7	206	507	2.461	16	153	425	2.778	25	79	183	2.316
8	49	120	2.449	17	183	473	2.585	計	4229	9281	56.322
9	256	584	2.281	18	151	228	1.510	単価	2.195		2.253

単純に考えて、建物ごとに単価 $@_i = y_i/x_i$ を計算し、それを平均した単価 2.253 と、合計金額 9281 を合計面積 4229 で割った単価 2.195 とを求めたが、2つの値は一致しない。どちらの計算方法が良いのだろうか。また もっと良い方法はないのだろうか。

```
data TANKA;
  input X Y @@;
  cards;
    258 458 212 400 134 296 171 408 73 185
    179 445 206 507 49 120 256 584 58 138
    224 381 248 385 150 415 199 613 180 485
    153 425 183 473 151 228 159 235 149 314
    75 192 199 442 271 390 213 579 79 183
  ;
run;
```



パラメータ推定値				
項	推定値	標準誤差	t値	p値(Prob> t)
切片	74.96911	51.11789	1.47	0.1560
X	1.751424	0.283385	6.18	<.0001

3. 反応の増加に伴う誤差の増加との関係

3.1. 細胞の数のカウントデータはポアソン分布

用量の増加と共に変異コロニー数のSDが大きくなる場合に、重み付き最小 2 乗法の適用しなければならない。CVが一定の場合は、観測値 y_i の逆数 $w_i = 1/y_i$ を重みにするとよい。用量ごとの変異コロニー数のデータは各々3例しかないがCVを計算した結果を表 9.1 に示すがデータ数が各群 $n = 3$ と小さいので確かなことは言えないが、S薬についてみればSDがやや用量の増加に伴い大きくなっているが、CVが一定とはいえない。

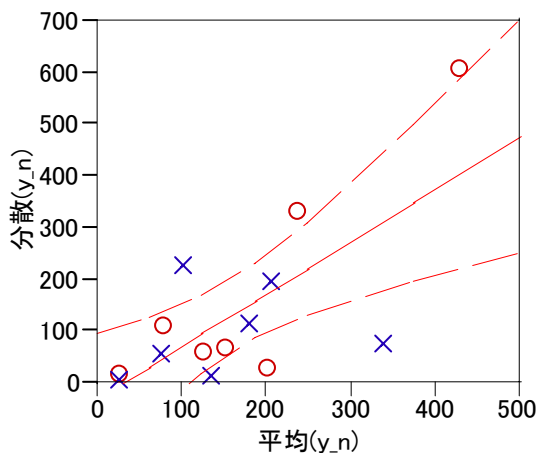
表 3.1 用量の増加に伴う CV の変化

dose	T 薬			S 薬		
	mean	SD	CV	mean	SD	CV
0	25.0	1.7	6.9	28.3	4.2	14.7
50	74.0	7.2	9.7	79.3	10.6	13.4
75	99.7	15.0	15.1	126.0	7.8	6.2
100	134.0	3.0	2.2	153.3	8.1	5.3
125	178.0	10.5	5.9	201.7	5.5	2.7
150	204.0	13.9	6.8	239.0	18.2	7.6
300	338.3	8.5	2.5	430.0	24.6	5.7

各群 $n = 3$

CV が用量の増加に伴って小さくなる場合は、分散が用量の増加に比例することが示唆される。また用量ごとの分散が平均値と同程度となるなれば、反応がポアソン分布に従っていると推測される。

Group	dose	N	平均 (y_n)	標準偏差 (y_n)	分散 (y_n)
S	0	3	28.33	4.16	17.33
S	50	3	79.33	10.60	112.33
S	75	3	126.00	7.81	61.00
S	100	3	153.33	8.14	66.33
S	125	3	201.67	5.51	30.33
S	150	3	239.00	18.19	331.00
S	300	3	430.00	24.64	607.00
T	0	3	25.00	1.73	3.00
T	50	3	74.00	7.21	52.00
T	75	3	99.67	15.01	225.33
T	100	3	134.00	3.00	9.00
T	125	3	178.00	10.54	111.00
T	150	3	204.00	13.89	193.00
T	300	3	338.33	8.50	72.33



パラメータ推定値

項	推定値	標準誤差	下側95%	上側95%
切片	-32.91	58.55	-160.49	94.66
平均(y _n)	1.02	0.30	0.37	1.66

用量ごとに計算した反応の平均と分散を、平均値を X 軸に、分散を Y 軸にして回帰分析を行なう。傾きが (95%ci) が 1.02 (0.37, 1.66)、切片もゼロに近いので、分散が平均と同じとみなせる。したがって、このデータはポアソン分布に従うものと推測される。

3.2. 細胞数を吸光度計で測定した場合

表 2.2 の吸光度のデータは、各用量に 2 個しかないが、SD、分散、範囲を計算し、平均値との関係を調べてみよう。図 2.2 に示すように、吸光度の平均とSDの関係は、吸光度が大きくなるにしたがって、2 次関数的に増加し、線形の関係にない。吸光度と範囲も同様である。吸光度と分散の関係は、SDの場合よりさらに線形が崩れている。SDの平方根との関係では、おおむね比例関係が見出される。

前節の例は、細胞数を直接カウントしていて、分散が平均値と同じと推定されたが、吸光度の場合は、このような関係は見出されなかった。

表 3.2 吸光度のばらつき

	photo	log2_dose	N	平均	SD	V	範囲	SD/平均	V/平均	root(SD)	root(SD)/平均	
○	1	1	-2	4	1.192	0.042	0.001794	0.101	0.036	0.001505	0.206	0.17
○	2	1	3	2	1.058	0.056	0.003121	0.079	0.053	0.002951	0.236	0.22
○	3	1	4	2	0.841	0.064	0.004140	0.091	0.077	0.004926	0.254	0.30
○	4	1	5	2	0.716	0.021	0.000450	0.030	0.030	0.000628	0.146	0.20
○	5	1	6	2	0.417	0.012	0.000145	0.017	0.029	0.000346	0.110	0.26
○	6	1	7	2	0.219	0.002	0.000005	0.003	0.010	0.000021	0.046	0.21
○	7	1	8	2	0.167	0.016	0.000242	0.022	0.093	0.001449	0.125	0.75
○	8	1	9	2	0.128	0.008	0.000061	0.011	0.061	0.000475	0.088	0.69
○	9	1	10	2	0.123	0.017	0.000288	0.024	0.138	0.002341	0.130	1.06
○	10	1	12	4	0.047	0.025	0.000610	0.059	0.531	0.013111	0.157	3.38
×	11	2	-2	4	1.241	0.073	0.005337	0.160	0.059	0.004301	0.270	0.22
×	12	2	0	2	1.208	0.175	0.030504	0.247	0.145	0.025263	0.418	0.35
×	13	2	1	2	1.156	0.133	0.017672	0.188	0.115	0.015287	0.365	0.32
×	14	2	2	2	0.916	0.150	0.022472	0.212	0.164	0.024533	0.387	0.42
×	15	2	3	2	0.549	0.023	0.000544	0.033	0.042	0.000991	0.153	0.28
×	16	2	4	2	0.280	0.018	0.000313	0.025	0.063	0.001114	0.133	0.47
×	17	2	5	2	0.110	0.012	0.000144	0.017	0.109	0.001308	0.110	0.99
×	18	2	6	2	0.079	0.018	0.000338	0.026	0.233	0.004278	0.136	1.72
×	19	2	7	2	0.134	0.011	0.000128	0.016	0.084	0.000955	0.106	0.79
×	20	2	12	4	0.058	0.039	0.001520	0.091	0.675	0.026325	0.197	3.42

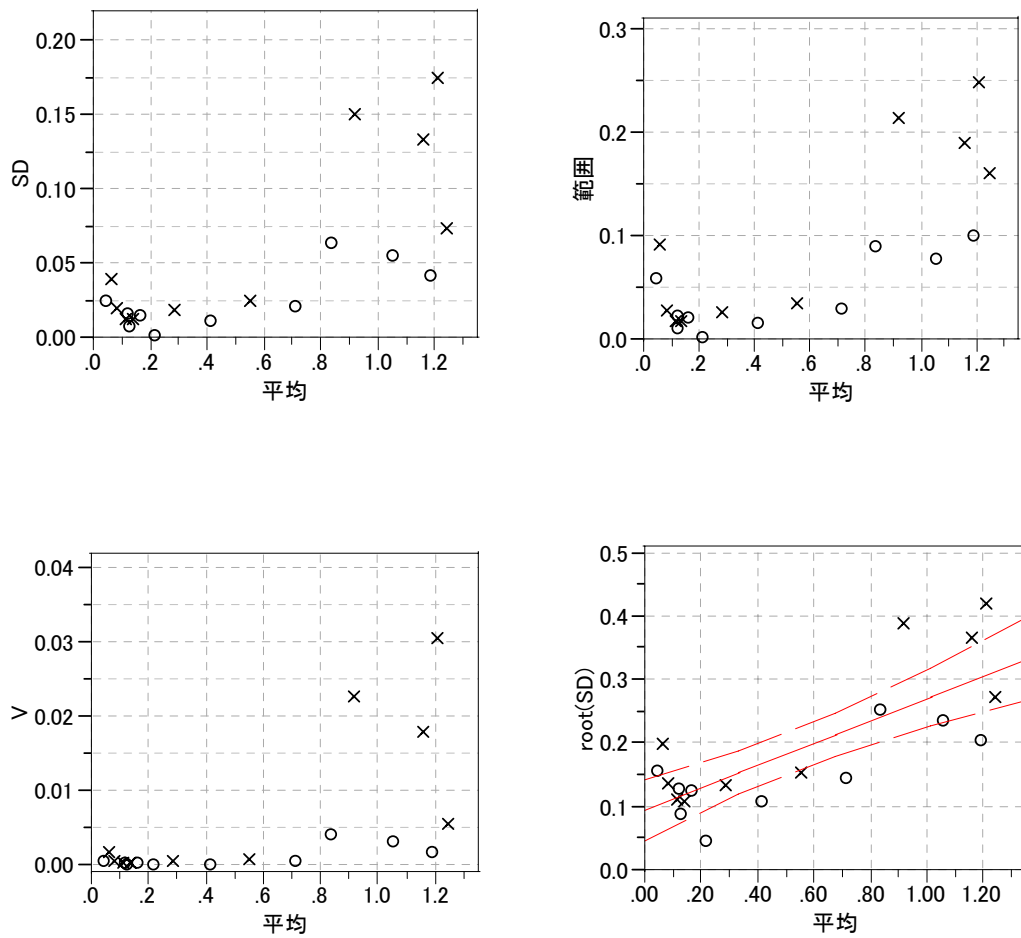


図 3.1 吸光度データでの反応の変動との関連

3.3. PK パラメータの例

反応の平均値の増加に伴い SD も比例的に増大する例である。CV=SD/平均値×100 は大よそ 12%と一定とみなせる。

表 3.3 SD が平均に比例して増大

	DOSE	N	平均	SD	V	SD/平均	V/平均	
○	1	0.5	16	67.26	6.69	44.71	0.10	0.66
●	2	1	16	130.91	18.30	334.95	0.14	2.56
○	3	2	16	245.28	29.76	885.55	0.12	3.61

3.4. 芳賀本の建築面積と建築費の例

芳賀の本では、建築面積 x と建築費 y のデータで、原点を通る回帰直線のあてはめに際して、推定誤差と推定値の関係について示されていない。重み付き最小 2 乗法を適用する最には、どのような関係なのかを見極める必要がある。図 3.2 から観察されるのは、建築面積 x の増加に伴い建築費 y は、ほぼ直線てきに増加するものの、直線からの残差の増大は、SDが平均値に比例する関係より大きいように思われる。

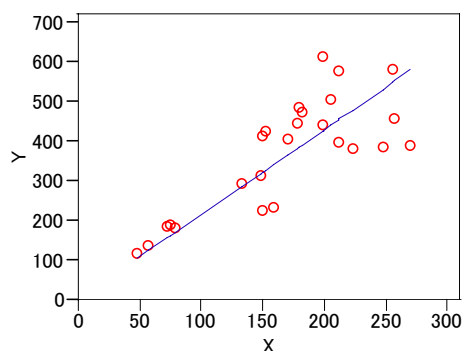


図 3.2 原点を通る回帰直線

同じ建築面積 x のデータがないので、SDおよび分散が推定できないので、原点を通る回帰直線からの残差をSDの代替変数とし、予測値を“平均値”とみなして散布図を作成した。結果を図 3.3 に示すが、 y の予測値が小さいほうから中程度までの残差が相対的に小さい。 e/y と y の関係でも定常とはみなせそうもない。分散との関連は、図 3.4 に示したが、さらに比例的な関連は見出し難くなっている。

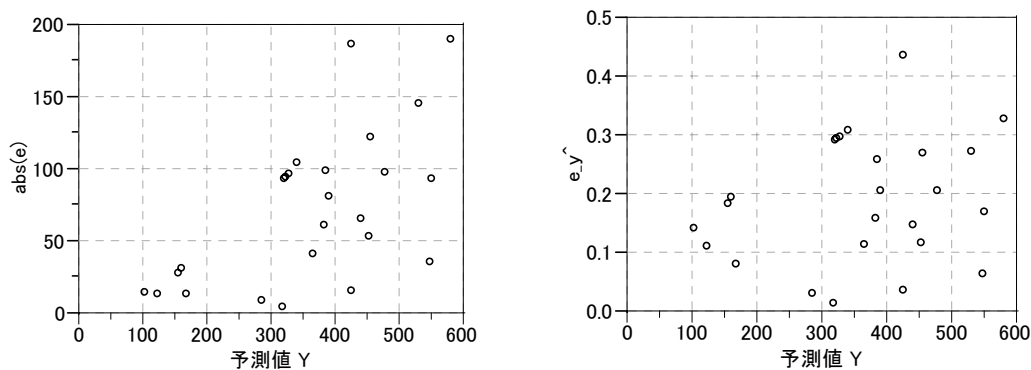


図 3.3 残差 e と予測値 y , e/y と y の関係

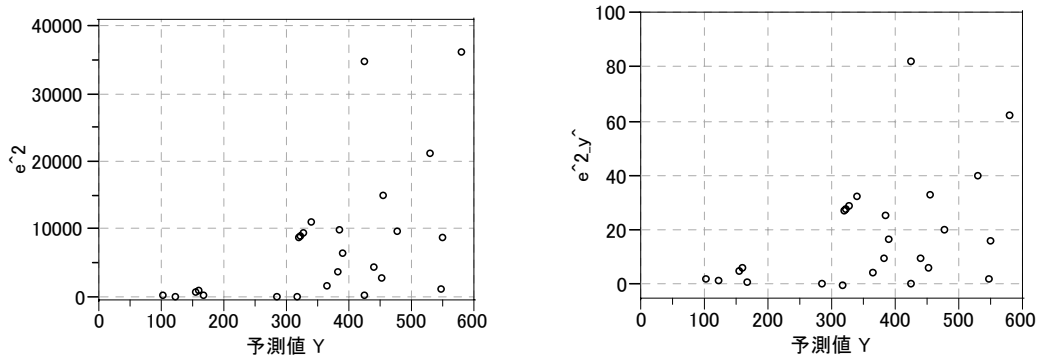


図 3.4 e^2 と推定値 y , e^2/y と y の関係

推定値 y の増大よりも残差 e の増加が大きいのであるから, \sqrt{e} と推定値 y の関係を
 図 3.5 に示した. \sqrt{e}/y と y の関係が, 定常とみなせる.

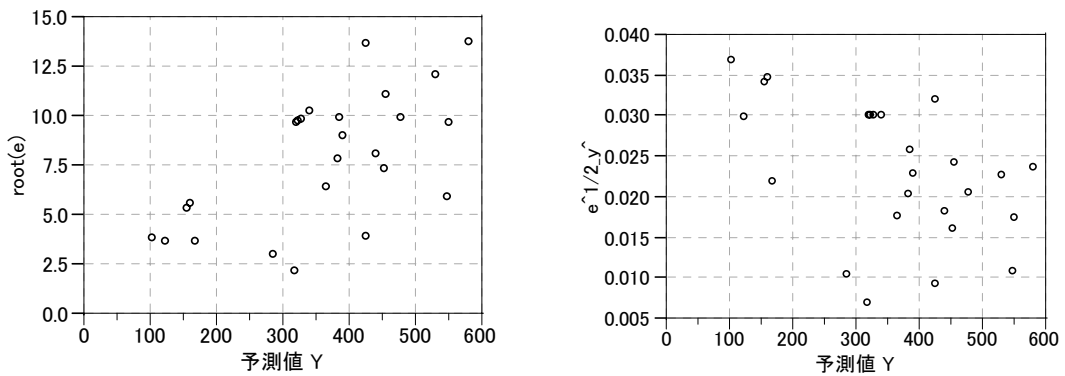


図 3.5 \sqrt{e} と推定値 y , \sqrt{e}/y と y の関係

4. ポアソン分布

4.1. ポアソン分布の基礎

医学研究のための統計的方法

P. Armitage, G. Berry 著
 橋 美智子・橋 広計 共訳
 Statistical Methods in Medical Research, 3rd ed.

2.6 ポアソン分布

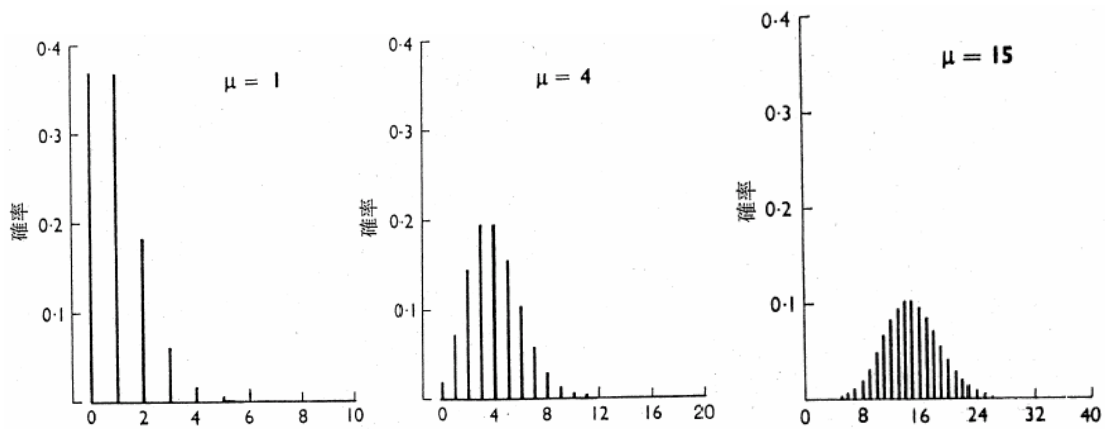


図 2.9 μ の様々な値に対するポアソン分布。各図の横軸は、 x の値を示している。

$\mu = 1, 4, 15$ に対する分布の形は、図 2.9 に示されている。 $\mu = 1$ の場合には、分布は非常に歪んでおり、 $\mu = 4$ の場合には歪みはずっと少なくなり、そして $\mu = 15$ の場合には、ほとんど歪みがないことに注目しよう。

分布 (2.15) は、1つの母数 μ によって完全に決定される。その結果、興味ある分布の全ての特徴は、 μ のみの関数となる。特に、平均と分散は μ の関数でなければならない。平均は、

$$\begin{aligned} E(x) &= \sum_{x=0}^{\infty} xP_x, \\ &= \sum_{x=0}^{\infty} \frac{x\mu^x e^{-\mu}}{x!}, \\ &= \sum_{x=1}^{\infty} \frac{\mu^x e^{-\mu}}{(x-1)!}, \quad (\because \frac{x}{x!} = \frac{1}{(x-1)!}) \end{aligned}$$

となる。なお、総和の中で $x = 0$ に対応する項は 0 である。

$$\begin{aligned} E(x) &= \mu \sum_{x=1}^{\infty} \frac{\mu^{x-1} e^{-\mu}}{(x-1)!}, \\ &= \mu \sum_{i=0}^{\infty} \frac{\mu^i e^{-\mu}}{i!}, (i = x - 1) \\ &= \mu \end{aligned}$$

となる。何故ならば、合計はポアソン分布の全ての項を含んでおり、その和は 1 になるからである。

簡便公式 $E(x^2) - \mu^2$ によって分散を求めるには、まず最初に、 $E[x(x-1)] = E(x^2) - E(x)$ を確認する。

$$\begin{aligned} E[x(x-1)] &= \sum_{x=0}^{\infty} x(x-1)P_x \\ &= \mu^2 \sum_{x=2}^{\infty} \frac{\mu^{x-2} e^{-\mu}}{(x-2)!} \\ &= \mu^2. \end{aligned}$$

前述の議論と同様に、従って、

$$\begin{aligned} E(x^2) - E(x) &= \mu^2, \\ E(x^2) &= \mu^2 + E(x) \\ &= \mu^2 + \mu \end{aligned}$$

となる。故に、次式が成り立つ。

$$\begin{aligned} \text{var}(x) &= E(x^2) - \mu^2 \\ &= (\mu^2 + \mu) - \mu^2 \\ &= \mu. \end{aligned}$$

結局、 x の分散は平均と同様 μ に等しい。それゆえ、標準偏差は $\sqrt{\mu}$ である。

ポアソン分布は、細菌学でよく使われる。懸濁液中の生命体の密度を推定するために、細菌学者は一定倍、例えば 10^{-5} に懸濁液を希釈し、ピペットで例えば 1cm^3 の試料を採取し、バクテリア培養液の皿にピペットの内容物を滴下する。ある程度の時間の後、皿の上に滴下した各生命体はそれぞれ、コロニーを形成することになる。これらコロニーの数を数えることができる。元の懸濁液はよく混合され、試料の体積が正確に決定され、培養液が一様に成長維持にとって十分である場合は、一連の皿の中のコロニー数はポアソン分布に従うことが

期待できるであろう。皿 1 枚当たりの平均コロニー数 x は、元の懸濁液の 10^{-5} cm^3 当たりのバクテリアの平均数の推定値となる。ポアソン分布の理論的性質についての知識があれば、この推定値の精度を見積もることができる（4 章 10 節参照）。

同様に、生存している生命体および死んでいる生命体の総数は、一定の体積の試料を繰り返し取り、顕微鏡で調べ、生命体の数を直接数えることによって調べられるであろう。

例 2.5

例として、表 2.3 はペトロフハウゼム計算室における根粒細菌（トリホリウム症のリビヒウム）の数を観測した分布を示している。「期待」度数は、度数分布から単位正方形当たりの生命体の平均個数 \bar{x} を計算し $\bar{x} = 2.50$ となる。 μ を \bar{x} で置き換えたポアソン分布の確率 P_x を計算することによって得られる。期待度数は、 $400P_x$ によって与えられる。観察度数と期待度数は、非常によく一致している。通常、この生命体は沈殿物を生じ、容易に凝集する。このような状況のもとでは、ポアソン分布に従うとは期待できないだろう。表 2.3 のデータは凝集を弱める方法の有効性を示すために集められたものである。

表 2.3 ペトロフハウゼム計算室における根粒細菌（トリホリウム症のリビヒウム）の計数の分布（Wilson and Kullman(1931) からのデータ）

単位正方形当たりの 細菌の数	正方形の個数	
	観測値	期待値
0	34	32.8
1	68	82.1
2	112	102.6
3	94	85.5
4	55	53.4
5	21	26.7
6	12	11.1
7	4	5.7
	400	399.9

細菌の数	観測値	期待値	Poisson	JMP期待値
0	34	32.8	0.0821	32.83
1	68	82.1	0.2052	82.08
2	112	102.6	0.2565	102.61
3	94	85.5	0.2138	85.51
4	55	53.4	0.1336	53.44
5	21	26.7	0.0668	26.72
6	12	11.1	0.0278	11.13
7	4	5.7	0.0099	3.98

5.7 は本のミスプリ、3.98 が正しい。

4.2. JMP によるポアソン分布の作図

表 4.1 いくつかの λ (平均) に対するポアソン乱数の発生

	i	lambda_02	lambda_05	lambda_1	lambda_2	lambda_5	lambda_10	lambda_20	lambda_50
1	1	0	0	0	3	6	6	20	43
2	2	0	1	1	3	6	5	21	44
3	3	0	0	0	1	6	9	25	60
4	4	0	2	0	2	3	9	15	57
5	5	1	1	0	3	5	11	25	50
6	6	1	0	0	2	9	7	23	62
7	7	0	0	2	2	5	5	20	47
8	8	0	1	1	0	4	13	14	42
9	9	0	0	2	6	4	8	15	45
10	10	0	0	0	8	6	15	20	40

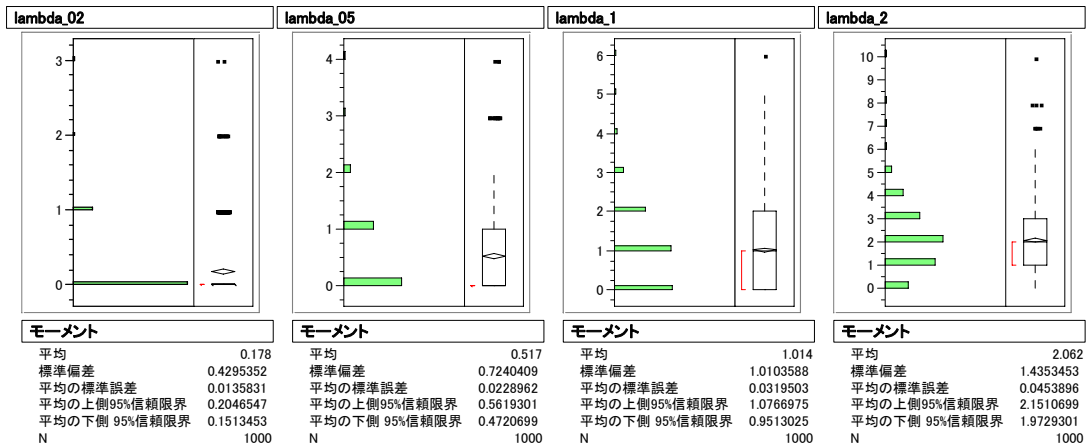


図 4.1 λ が小さい場合の裾広がりポアソン分布

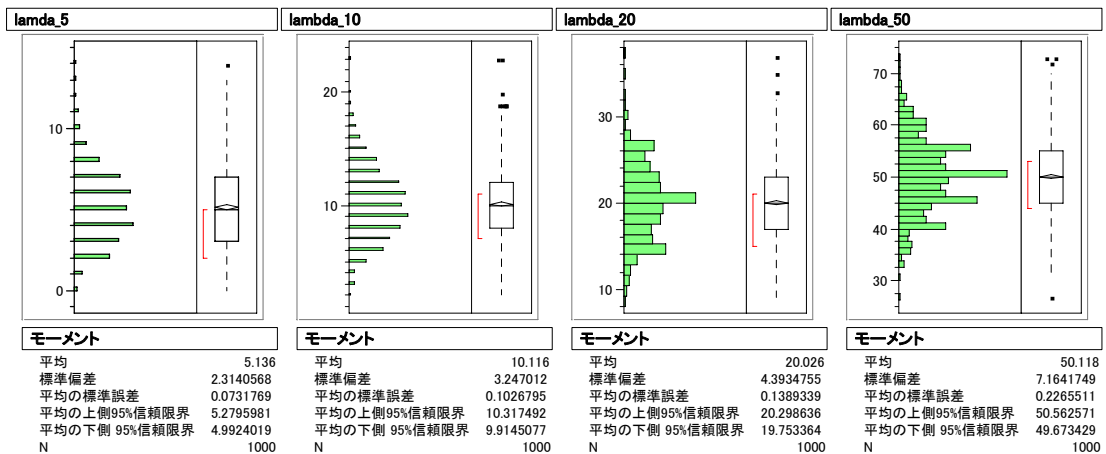


図 4.2 λ が大きい場合のやや裾をひいているポアソン分布

5. ポアソン回帰, Armitageの本, 4th editon

ここに引用したのは, 原著の4版である. 3版は椿らによって翻訳されているが, ポアソン回帰は独立した節にはなっていなかったが, 4版では, 14.4節としてまとめられている. 内容については同様なので, 原文と訳を対比して示す.

Statistical Methods in Medical Research

P. Armitage
MA, PhD

*Emeritus Professor of Applied Statistics
University of Oxford*

G. Berry
MA, PhD

*Professor in Epidemiology and Biostatistics
University of Sydney*

J.N.S. Matthews
MA, PhD

*Professor of Medical Statistics
University of Newcastle upon Tyne*

14 Modelling categorical data, 485

- 14.1 Introduction, 485
- 14.2 Logistic regression, 488
- 14.3 Polytomous regression, 496
- 14.4 Poisson regression, 499

14.4 Poisson regression

Poisson distribution

The expectation of a Poisson variable is positive and so limited to the range 0 to ∞ . A link function is required to transform this to the unlimited range $-\infty$ to ∞ . The usual transformation is the logarithmic transformation

$$g(\mu) = \ln \mu,$$

leading to the log-linear model

$$\ln \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p. \quad (14.12)$$

ポアソン分布

ポアソン変数の期待値は正で, 範囲 $0 \sim \infty$ に制限されている. リンク関数は, これを制限のない範囲 $-\infty \sim \infty$ に変換することが必要とされる. 普通の変換は対数変換

$$g(\mu) = \ln \mu$$

であり, これは次の対数線形モデル (log-linear model) を導く.

$$\ln \mu = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p. \quad (12.26)$$

Example 14.4

Table 14.3 gives data on the number of incident cases of cancer in a large group of ex-servicemen, who had been followed up over a 20-year period. The servicemen are in two groups according to whether they served in a combat zone (veterans) or not, and the experience of each serviceman is classified into subject-years at risk in 5-year age groups. The study is described in Australian Institute of Health and Welfare (1992), where the analysis also controlled for calendar year. Each serviceman passed through several of these groups during the period of follow-up. The study was carried out in order to assess if there was a difference in cancer risk between veterans and non-veterans. The model used was a variant on (14.12). If y_{ij} is the number of cases of cancer in group i and age group j , and N_{ij} is the corresponding number of subject-years, then y_{ij}/N_{ij} is the incidence rate.

Table 14.3 Number of incident cases of cancer and subject-years at risk in a group of ex-servicemen (reproduced by permission of the Australian Institute of Health and Welfare).

Age	Veterans		Non-veterans	
	Number of cancers	Subject-years	Number of cancers	Subject-years
–24	6	60 840	18	208 487
25–29	21	157 175	60	303 832
30–34	54	176 134	122	325 421
35–39	118	186 514	191	312 242
40–44	97	135 475	108	165 597
45–49	58	42 620	74	54 396
50–54	56	25 001	88	40 716
55–59	54	13 710	120	33 801
60–64	34	6 163	141	26 618
65–69	9	1 575	108	17 404
70–	2	273	99	14 146
Total	509	805 480	1129	1 502 660

The log-linear model states that the logarithm of incidence will follow a linear model on variables representing the group and age. Thus if μ_{ij} is the expectation of y_{ij} , then

$$\ln \mu_{ij} = \ln N_{ij} + \alpha + \beta_i x_i + \gamma_j z_j, \quad (14.13)$$

where x_i and z_j are dummy variables representing the veteran groups and the age groups, respectively (the dummy variables were defined as in §11.7, with $x_1 = 1$ for the veterans group, and $z_1, z_2, \dots, z_{10} = 1$ for age groups 25–29, 30–34, ..., 70–; no dummy variable was required for the non-veterans or the youngest age group as their effects are included within the coefficient α). This model differs from (14.12) in the inclusion of the first term on the right-hand side, which ensures that the number of years at risk is taken into account (see (19.38)).

The model was fitted by maximum likelihood using GLIM with $\ln N_{ij}$ included as an OFFSET. The estimates of the regression coefficients were:

	Estimate	SE
a	-9.324	
b_1 (veterans)	-0.0035	0.0555
c_1 (25–29)	0.679	0.232
c_2 (30–34)	1.371	0.218
c_3 (35–39)	1.940	0.212
c_4 (40–44)	2.034	0.216
c_5 (45–49)	2.727	0.222
c_6 (50–54)	3.203	0.221
c_7 (55–59)	3.716	0.218
c_8 (60–64)	4.093	0.218
c_9 (65–69)	4.236	0.224
c_{10} (70–)	4.364	0.227

The estimate of the veterans effect is not significant, Wald $z = -0.0035/0.0555 = 0.06$ ($P = 0.95$). Converting back from the log scale, the estimate of the relative risk of cancer in veterans compared with non-veterans, after controlling for age, is $\exp(-0.0035) = 1.00$. The 95% confidence limits are $\exp(-0.0035 \pm 1.96 \times 0.0555) = 0.89$ and 1.11.

一般化線形モデル (12.24) の議論の最後に、ポアソン計数に対数線形モデルを適用した例を示す。

例 12.12

表 12.10 は、退役軍人の癌の発生数に関するデータを与えている。彼らは 20 年の期間にわたって追跡された。退役軍人は、彼らが実戦経験があるか否かで、2 群に分かれており、そして、5 歳毎の年齢層毎に各軍務経験をリスクとして人年が記録されている。研究は Australian Institute of Health and Welfare(1992) に記述されており、ここで解析は暦年に対しても調整されていた。各軍人は、追跡期間中に、これらの群のいくつかを通過した。研究は実戦経験のある群と実戦経験のない群間の発癌リスクに差が存在するか否かを評価するために行われた。使われたモデルは (12.26) の変形であった。 y_{ij} が群 i で年齢群 j の癌のケース数であり、 N_{ij} が対応する人年とすれば、 y_{ij}/N_{ij} が発生率である。対数線形モデルは、発生の対数が群と年齢を表す変数の線形モデルに従うというものである。従って、 μ_{ij} が y_{ij} の期待値ならば、

$$\ln \mu_{ij} = \ln N_{ij} + \alpha + \beta_i x_i + \gamma_j z_j$$

である。ここで、 x_i と z_j はそれぞれベテラン群と年齢群を表しているダミー変数である (ダミー変数は 10 章 2 節のように定義され、実戦経験のある群に対して $x_1 = 1$, 年齢群 25-29, 30-34, ..., 70-

表 12.10 退役軍人の群における癌の発生ケースの数とリスクの人年

年齢	実戦経験あり		実戦経験なし	
	癌の数	人年	癌の数	人年
-24	6	60840	18	208487
25-29	21	157175	60	303832
30-34	54	176134	122	325421
35-39	118	186514	191	312242
40-44	97	135475	108	165597
45-49	58	42620	74	54396
50-54	56	25001	88	40716
55-59	54	13710	120	33801
60-64	34	6163	141	26618
65-69	9	1575	108	17404
70-	2	273	99	14146
合計	509	805480	1129	1502660

に対しては $z_1, z_2, \dots, z_{10} = 1$ をとる)。実戦経験がない群あるいは最年少群に対しては、その効果は係数 α に含まれるので、ダミー変数は必要ではない。このモデルは (12.26) とは、リスクの年数の考慮を保証している右辺の第一項が含まれているという点が違っている。

モデルは、オフセット項として $\ln N_{ij}$ を含む GLIM を用いた最尤法であてはめられた。回帰係数の推定値は、

		推定値	SE
Y	a	-9.324	
	b_1 (実戦経験)	-0.0035	0.0555
	c_1 (25-29)	0.679	0.232
	c_2 (30-34)	1.371	0.218
	c_3 (35-39)	1.940	0.212
	c_4 (40-44)	2.034	0.216
	c_5 (45-49)	2.727	0.222
	c_6 (50-54)	3.203	0.221
	c_7 (55-59)	3.716	0.218
	c_8 (60-64)	4.093	0.218
	c_9 (65-69)	4.236	0.224
	c_{10} (70-)	4.364	0.227

となった。

実戦経験の効果の推定値は有意ではない (Wald $z = -0.0035/0.0555 = 0.06 (P = 0.95)$)。対数スケールから逆変換して、年齢に対してコントロールした後に、実戦経験のない群と比較した実戦経験のある群の癌の相対リスクの推定値は $\exp(-0.0035) = 1.00$ 。95%信頼限界は $\exp(-0.0035 \pm 1.96 \times 0.0555) = 0.89$ と 1.11 である。

文献には、癌の発生率が示されていないので、表 5.1 に計算結果を示した。“実戦経験あり”群は、年齢の増加に伴って癌の発生率は高まるが、対照の“実戦経験なし”群に比べて高いわけではない。

表 5.1 退役軍人の群における癌の発生ケースの数とリスクの年人

年齢	実戦経験あり			実戦経験なし		
	癌の数	人年	発生率	癌の数	人年	発生率
-24	6	60840	0.000099	18	208487	0.000086
25-29	21	157175	0.000134	60	303832	0.000197
30-34	54	176134	0.000307	122	325421	0.000375
35-39	118	186514	0.000633	191	312242	0.000612
40-44	97	135475	0.000716	108	165597	0.000652
45-49	58	42620	0.001361	88	54396	0.001618
50-54	56	25001	0.002240	74	40716	0.001817
55-59	54	13710	0.003939	120	33801	0.003550
60-64	34	6163	0.005517	141	26618	0.005297
65-69	9	1575	0.005714	108	17404	0.006205
70-	2	273	0.007326	99	14146	0.006998
total	509	805480	0.000632	1129	1502660	0.000751

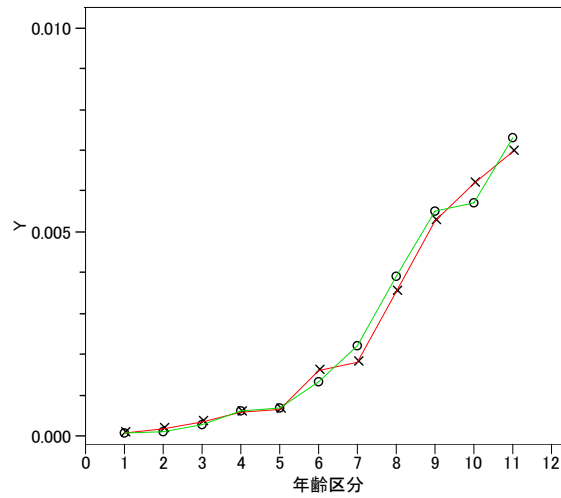


図 5.1 Veterans vs. Non-Veterans

○ : Veterans, × : Non-Veterans

SAS による追試を次に示す.

```
Title 'Poisson1.sas 2004-04-08 Y.Takahashi' ;
```

```
data d01;
```

```
input men $ age_no age_class $ y n ;
```

```
log_n = log(n) ;
```

```
x1 = (men='Veterans') ;
```

```
c1 = (age_class='25-29') ; c2 = (age_class='30-34') ;
```

```
c3 = (age_class='35-39') ; c4 = (age_class='40-44') ;
```

```
c5 = (age_class='45-49') ; c6 = (age_class='50-54') ;
```

```
c7 = (age_class='55-59') ; c8 = (age_class='60-64') ;
```

```
c9 = (age_class='65-69') ; c10= (age_class='70- ') ;
```

```
datalines ;
```

```
Veterans 1 20-24 6 60840 0.000099
```

```
Veterans 2 25-29 21 157175 0.000134
```

```
Veterans 3 30-34 54 176134 0.000307
```

```
Veterans 4 35-39 118 186514 0.000633
```

```
Veterans 5 40-44 97 135475 0.000716
```

```
Veterans 6 45-49 58 42620 0.001361
```

```
Veterans 7 50-54 56 25001 0.00224
```

```
Veterans 8 55-59 54 13710 0.003939
```

```
Veterans 9 60-64 34 6163 0.005517
```

```
Veterans 10 65-69 9 1575 0.005714
```

```
Veterans 11 70- 2 273 0.007326
```

```
Non_vete 1 20-24 18 208487 0.000086
```

```
Non_vete 2 25-29 60 303832 0.000197
```

```
Non_vete 3 30-34 122 325421 0.000375
```

```
Non_vete 4 35-39 191 312242 0.000612
```

```
Non_vete 5 40-44 108 165597 0.000652
```

```

Non_vete 6 45-49 88 54396 0.001618
Non_vete 7 50-54 74 40716 0.001817
Non_vete 8 55-59 120 33801 0.00355
Non_vete 9 60-64 141 26618 0.005297
Non_vete 10 65-69 108 17404 0.006205
Non_vete 11 70- 99 14146 0.006998
;
proc print data=d01 ;
run ;
proc genmod data=d01 ;
class men age_class ;
model y = x1 c1 c2 c3 c4 c5 c6 c7 c8 c9 c10
      / dist=poisson link=log offset=log_n ;
run ;
proc genmod data=d01 ;
class men age_class ;
model y = men age_class
      / dist=poisson link=log offset=log_n ;
lsmeans men age_class ;
run ;

```

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-9.3243	0.2045	-9.7251	-8.9234	2078.84	<.0001
x1	1	<u>-0.0067</u>	0.0555	-0.1148	0.1027	0.01	0.9130
c1	1	0.6796	0.2325	0.2239	1.1353	8.54	0.0035
c2	1	1.3714	0.2177	0.9447	1.7981	39.68	<.0001
c3	1	1.9400	0.2121	1.5244	2.3556	83.69	<.0001
c4	1	2.0349	0.2161	1.6113	2.4584	88.67	<.0001
c5	1	2.8279	0.2206	2.3956	3.2602	164.36	<.0001
c6	1	3.1010	0.2223	2.6652	3.5368	194.53	<.0001
c7	1	3.7163	0.2178	3.2895	4.1432	291.22	<.0001
c8	1	4.0926	0.2177	3.6659	4.5192	353.47	<.0001
c9	1	4.2358	0.2242	3.7964	4.6753	356.88	<.0001
c10	1	4.3632	0.2274	3.9176	4.8088	368.24	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

文献は、0.0035となっている。SEが文献と同じなので、ダミー変数を (1, -1) としている。

SASのGENMODでダミー変数を生成して実行した結果の推定値は、基準となる水準からの自然対数の差になっている。元の発生率を水準ごとに求めるために、classステートメントを用い、lsmeansステートメントを加えて、推定値を求めた。この推定値も自然対数であるので、EXCELで発生率に直した結果を表 5.2 に示した。

表 5.2 GENMOD による推定

年齢	実践経験あり			実戦経験なし			推定値	
	癌の数	人年	発生率	癌の数	人年	発生率	log_e()	発生率 [^]
-24	6	60840	0.000099	18	208487	0.000086	-9.3273	0.000089
25-29	21	157175	0.000134	60	303832	0.000197	-8.6477	0.000176
30-34	54	176134	0.000307	122	325421	0.000375	-7.9559	0.000351
35-39	118	186514	0.000633	191	312242	0.000612	-7.3873	0.000619
40-44	97	135475	0.000716	108	165597	0.000652	-7.2924	0.000681
45-49	58	42620	0.001361	88	54396	0.001618	-6.4994	0.001504
50-54	56	25001	0.002240	74	40716	0.001817	-6.2263	0.001977
55-59	54	13710	0.003939	120	33801	0.003550	-5.6109	0.003658
60-64	34	6163	0.005517	141	26618	0.005297	-5.2347	0.005328
65-69	9	1575	0.005714	108	17404	0.006205	-5.0914	0.006149
70-	2	273	0.007326	99	14146	0.006998	-4.9641	0.006984
total	509	805480	0.000632	1129	1502660	0.000751		
	推定値:exp(-6.7458)=		0.001176	exp(-6.7519)=		0.001169		

6. ポアソン回帰の細胞数のカウントデータへの適用

平方根を取ると分散は安定するが，用量に対して曲線となり，扱いにくくなる．対数線形モデルによる線形化も，投与量に関して曲線となり，扱い難い．

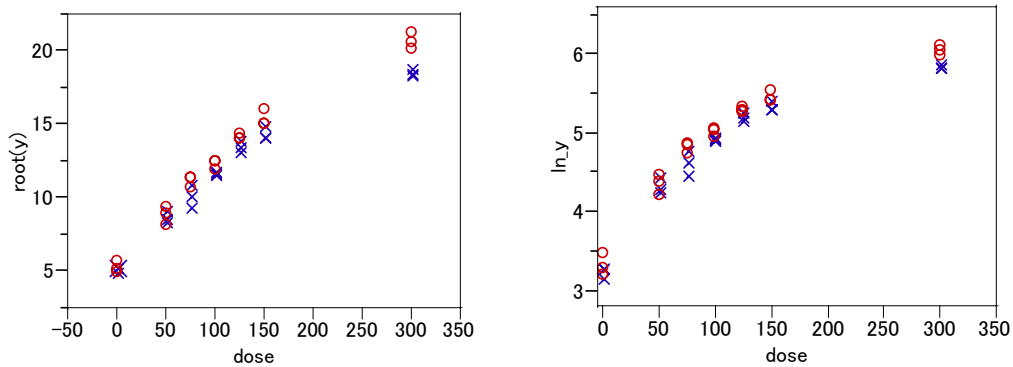


図 6.1 細胞毒性データに対する平方根変換と対数変換

7. 重み付き回帰分析の適用

7.1. 統計ソフトでの重み

重み付き回帰分析の計算方法について JMP では見当たらないので，SAS の GLM プロシジャのマニュアルを参照すると，測定値 y_i と予測値 \hat{y}_i との偏差の平方和に重み w_i を掛けている．

$$\sum_i w_i (y_i - \hat{y}_i)^2$$

重み w_i を分散の逆数とすると，偏差を標準偏差で割っていることになり，分散の期待値が 1 となるように変数変換したと理解される．

WEIGHT Statement

WEIGHT *variable* ;

When a WEIGHT statement is used, a weighted residual sum of squares

$$\sum_i w_i (y_i - \hat{y}_i)^2$$

is minimized, where w_i is the value of the variable specified in the WEIGHT statement, y_i is the observed value of the response variable, and \hat{y}_i is the predicted value of the response variable.

If you specify the WEIGHT statement, it must appear before the first RUN statement or it is ignored.

An observation is used in the analysis only if the value of the WEIGHT statement variable is nonmissing and greater than zero.

The WEIGHT statement has no effect on degrees of freedom or number of observations, but it is used by the MEANS statement when calculating means and performing multiple comparison tests (as described in the “MEANS Statement” section beginning on page 1497). The normal equations used when a WEIGHT statement is present are

$$X'WX\beta = X'WY$$

where W is a diagonal matrix consisting of the values of the variable specified in the WEIGHT statement.

If the weights for the observations are proportional to the reciprocals of the error variances, then the weighted least-squares estimates are best linear unbiased estimators (BLUE).

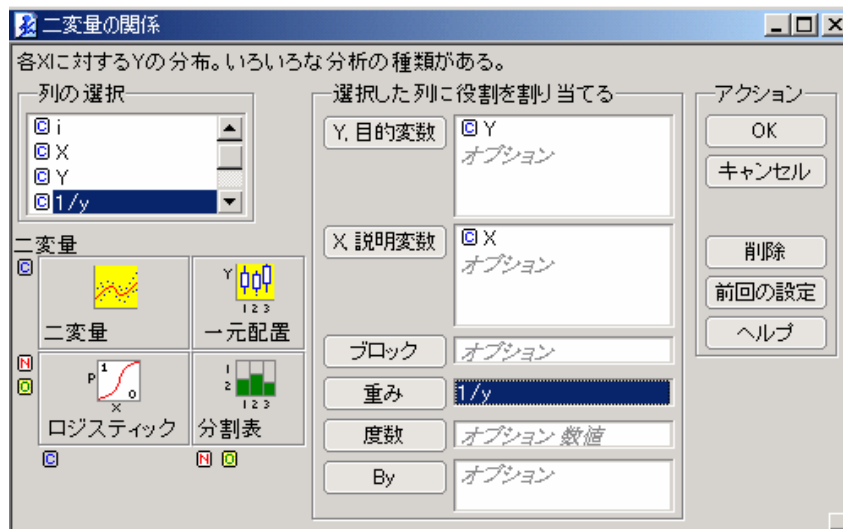
7.2. JMP での解析

芳賀本の建築面積と建築費用のデータを用いて、重み付き回帰分析を試みる。データは、建築面積 x の小さい順に並べ直し、重みの初期値として $1/y$ を計算した結果を、

表 7.1 0 次の回帰分析データ

	i	X	Y	1/y
1	8	49	120	0.0083
2	10	58	138	0.0072
3	5	73	185	0.0054
4	21	75	192	0.0052
5	25	79	183	0.0055
6	3	134	296	0.0034
7	20	149	314	0.0032
8	13	150	415	0.0024
9	18	151	228	0.0044
10	16	153	425	0.0024
11	19	159	235	0.0043
12	4	171	408	0.0025
13	6	179	445	0.0022
14	15	180	485	0.0021
15	17	183	473	0.0021
16	14	199	613	0.0016
17	22	199	442	0.0023
18	7	206	507	0.0020
19	2	212	400	0.0025
20	24	213	579	0.0017
21	11	224	381	0.0026
22	12	248	385	0.0026
23	9	256	584	0.0017
24	1	258	458	0.0022
25	23	271	390	0.0026

JMP の「二変量の関係」で $1/y$ を「重み」に指定して実行する。結果を



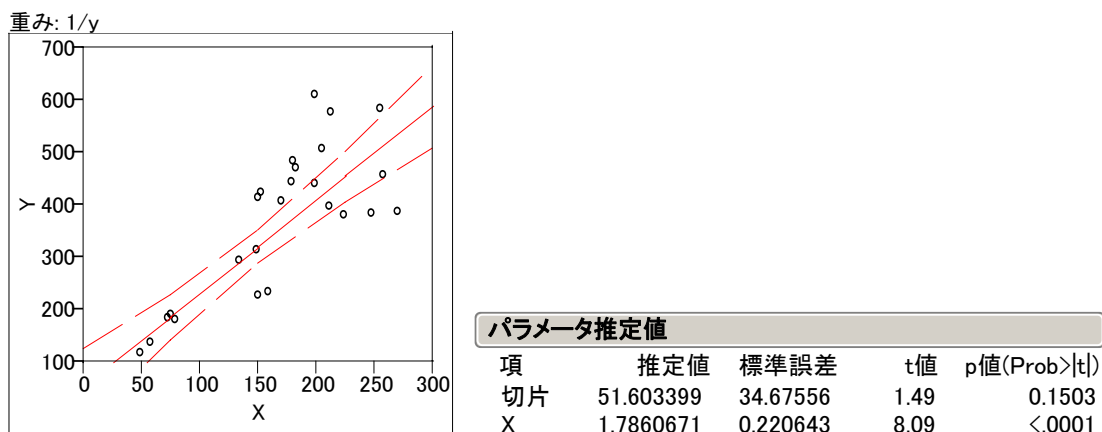


図 7.1 0 次の結果

推定値 $\hat{y}^{(1)}$ をデータテーブルに出力し、その逆数 $1/\hat{y}^{(1)}$ を再度計算する。結果を表 7.2 に示す。

表 7.2 予測値の出力

	i	X	Y	1/y	予測値 Y	1/y^	
○	1	8	49	120	0.0083	139.12	0.0072
○	2	10	58	138	0.0072	155.20	0.0064
○	3	5	73	185	0.0054	181.99	0.0055
○	4	21	75	192	0.0052	185.56	0.0054
○	5	25	79	183	0.0055	192.70	0.0052
○	6	3	134	296	0.0034	290.94	0.0034
○	7	20	149	314	0.0032	317.73	0.0031
○	8	13	150	415	0.0024	319.51	0.0031
○	9	18	151	228	0.0044	321.30	0.0031
○	10	16	153	425	0.0024	324.87	0.0031
○	11	19	159	235	0.0043	335.59	0.0030
○	12	4	171	408	0.0025	357.02	0.0028

推定値 \hat{y} の逆数 $1/\hat{y}^{(1)}$ を新たに重みとして重み付き回帰分析を行なう。結果を図 7.2 に示す。同様に 2 次の解析を行なう。この繰り返しが収束するまで続ける。

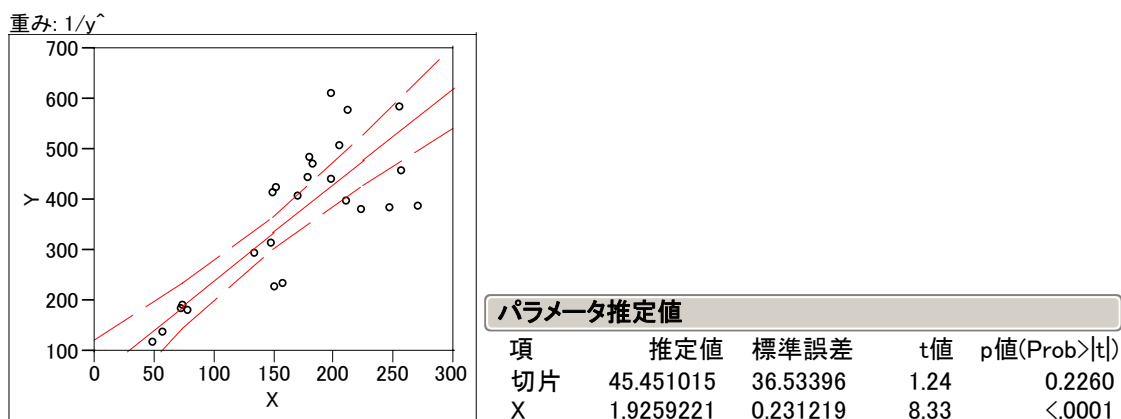


図 7.2 1 次の回帰分析の結果

パラメータ推定値				
項	推定値	標準誤差	t値	p値(Prob> t)
切片	44.703672	36.01575	1.24	0.2270
X	1.9303401	0.229303	8.42	<.0001

図 7.3 2 次の回帰分析の結果

参考のため、芳賀本の結果を示す。

9.4 誤差分散が η の関数である場合

ここで、回帰直線が原点を通らない一般の場合

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{9.15}$$

で、 ε の標準偏差 σ_i または分散 σ_i^2 が y_i の期待値 η_i に比例していると考えられる場合に、重みつき回帰分析によって、 β_0, β_1 を推定する方法を考える。

原点を通る回帰式では「 η_i に比例する」を「 x_i に比例する」と置き換えることができたが、式(9.15)のモデルではそれができない。 β_0 が小さいときは「誤差の標準偏差または分散が x_i に比例する」と近似的に置き換えることができるが、一般的ではない。

誤差 ε_i の標準偏差 σ_i が y_i の期待値 η_i に比例するというモデルを

$$\begin{aligned} y_i &= \eta_i + \varepsilon_i \\ &= \beta_0 + \beta_1 x_i + \varepsilon_i \quad \varepsilon_i \sim N(0, (\eta_i \sigma)^2) \end{aligned} \quad (9.16)$$

と表す。ここで、 η_i は面積が x_i のときの金額 y_i の期待値である。実際のコスト y_i は η_i に誤差 ε_i を加えたものである。両辺を η_i で割ると等分散のモデル式

$$\frac{y_i}{\eta_i} = \frac{\beta_0}{\eta_i} + \beta_1 \frac{x_i}{\eta_i} + \frac{\varepsilon_i}{\eta_i} \quad (9.17)$$

が得られる。ここで η_i がわからないから、その近似値として y_i を使う。

$$y_i^* = \frac{y_i}{y_i} = 1 = \beta_0 \frac{1}{y_i} + \beta_1 \frac{x_i}{y_i} \quad (9.18)$$

これは、 $1/y_i$ と x_i/y_i の2つを説明変数とする常数項を含まない重回帰式となる。 y_i^* はすべて1である。回帰係数の近似推定値 b_0, b_1 を求める。それを使って、 y_i の予測値 \hat{y}_i を計算する。 \hat{y}_i を式(9.18)の η_i のより良い近似値として用い、

$$y_i^* = \frac{y_i}{\hat{y}_i} = \frac{b_0}{\hat{y}_i} + b_1 \frac{x_i}{\hat{y}_i} \quad (9.19)$$

とする。今度は y_i^* が一定ではない。これから回帰係数の改善された近似推定値 b_0, b_1 を求める。このような逐次近似を収束するまで数回繰り返すと、

$$\hat{y} = 27.334 + 2.052x$$

が得られる。

分散が η_i に比例する場合も同様な手順を繰り返すことによって解かれ、

$$\hat{y} = 43.510 + 1.937x$$

となる。

以上の計算を proc reg を繰り返し使って解く手順を示す。

JMP の結果が芳賀本と少し異なる。原因はこれから調査する。

7.3. 非線形回帰分析を用いる場合

JMPの「二変量の関係」で再計算するのは、煩雑であるので、再計算が容易にできる非線形回帰分析を“重み付き回帰分析”として用いる。パラメータの初期値の設定を表7.3に示す。

表 7.3 初期値の設定

The screenshot shows the Tanka_NLIN software window. The main window displays a data table with columns for 'i', 'X', 'Y', 'y-hat', and '1/y-hat'. A dialog box titled 'y-hat' is open, showing parameter settings for a linear regression model. The parameters are b0 = 45 and b1 = 2. The function selected is 'b0 + b1 * X'.

	i	X	Y	y-hat	1/y-hat	
○	1	8	49	120	143	0.006993
○	2	10	58	138	161	0.006211
○	3	5	73	185	191	0.005236
○	4	21	75	192	195	0.005128
○	5	25	79	183	203	0.004926
○	6	3	134	296	313	0.003195
				4	343	0.002915
				5	345	0.002899
				8	347	0.002882
				15	351	0.002849
				15	363	0.002755
				18	387	0.002584
				5	403	0.002481
				15	405	0.002469
				13	411	0.002433
				3	443	0.002257
				3	443	0.002257
				2	443	0.002257
				17	457	0.002188
				10	469	0.002132
				9	471	0.002123
				11	493	0.002028
				15	541	0.001848
				14	557	0.001795
				18	561	0.001783
				30	587	0.001704

The screenshot shows the '非線形回帰' (Non-linear Regression) dialog box. It is used to configure the regression model. The '列の選択' (Column Selection) section shows 'i', 'X', 'Y', 'y-hat', and '1/y-hat' as available columns. The 'Y, 応答変数' (Y, Response Variable) is set to 'Y', and the 'X, 予測式列' (X, Prediction Formula Column) is set to 'y-hat'. The '重み' (Weight) is set to '1/y-hat'. The '計算式' (Formula) section shows the prediction formula 'Parameter (b0 = 45, b1 = 2); b0 + b1 * :X'. The '損失' (Loss) section is empty. The '2次微分した式' (Second-order differentiated formula) checkbox is unchecked.

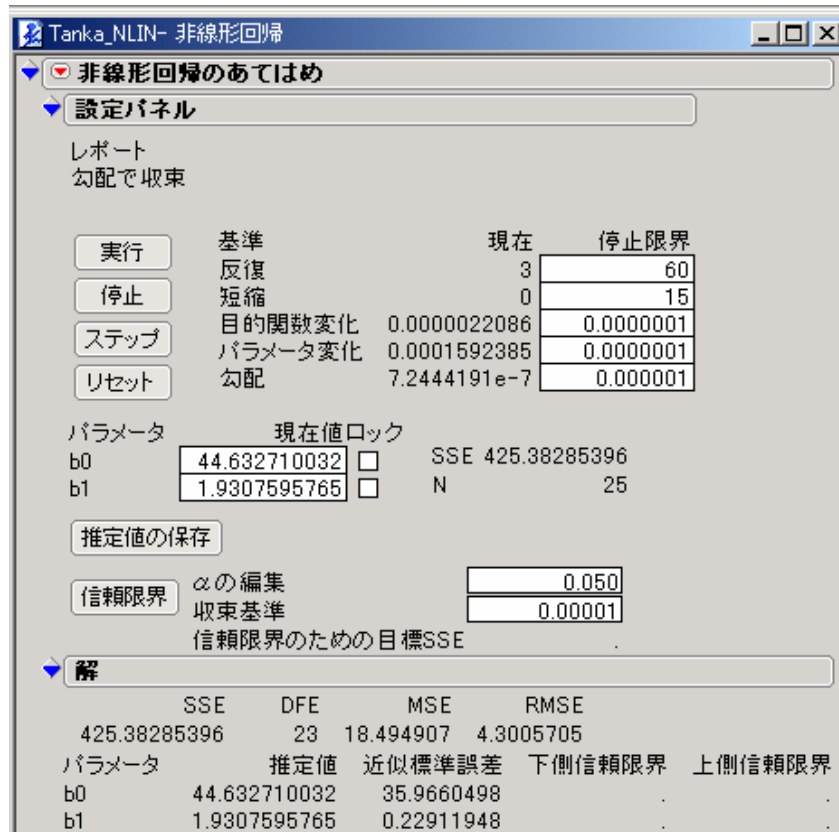


図 7.4 0 次結果

図 7.4 の 0 次の結果の「推定値の保存」ボタンをクリックすると、設定したパラメータが書き換えられる。そこで、「リセット」、「実行」をして、新たに推定された重みを使って計算をする。

	SSE	DFE	MSE	RMSE
	425.3829348	23	18.49491	4.3005709
パラメータ	推定値	近似標準誤差	下側信頼限界	上側信頼限界
b0	44.633321145	35.9664422	.	.
b1	1.9307559639	0.22912083	.	.

図 7.5 1 次の結果

解					
	SSE	DFE	MSE	RMSE	
	425.38294175	23	18.494911	4.300571	
パラメータ	推定値	近似標準誤差	下側信頼限界	上側信頼限界	
b0	44.633373692	35.966476	.	.	
b1	1.9307556533	0.22912095	.	.	

図 7.6 2 次の結果

8. 細胞毒性での応用

用量の増加と共に変異コロニー数のSDが大きくなる場合に、重み付き最小 2 乗法の適用しなければならない。CVが一定の場合は、観測値 y_i の逆数 $w_i = 1/y_i$ を重みにするとよい。用量ごとの変異コロニー数のデータは各々3例しかないがCVを計算した結果を表 9.1 に示すがデータ数が各群 $n=3$ と小さいので確かなことは言えないが、S薬についてみればSDがやや用量の増加に伴い大きくなっているが、CVが一定とはいえない。

表 8.1 用量の増加に伴う CV の変化

dose	T 薬			S 薬		
	mean	SD	CV	mean	SD	CV
0	25.0	1.7	6.9	28.3	4.2	14.7
50	74.0	7.2	9.7	79.3	10.6	13.4
75	99.7	15.0	15.1	126.0	7.8	6.2
100	134.0	3.0	2.2	153.3	8.1	5.3
125	178.0	10.5	5.9	201.7	5.5	2.7
150	204.0	13.9	6.8	239.0	18.2	7.6
300	338.3	8.5	2.5	430.0	24.6	5.7

各群 $n = 3$

したがって、重みを考慮した解析は必要としないが、重み付き非線形回帰の練習問題として結果を表 9.2 に示す。

表 8.2 重み付き非線形回帰

Solution				
	SSE	DFE	MSE	RMSE
	46.338791699	39	1.1881741	1.090034
Parameter	Estimate	ApproxStdErr	Lower CL	Upper CL
b0	24.327837651	2.00098085	20.2804719	28.3752034
b1_2	1.3491164349	0.03116247	1.28608439	1.41214848
b2	0.8110285231	0.02493155	0.76182665	0.86282786

$ratio = 0.811$, $95\%cl = (0.762, 0.863)$ 、重み = $1/y$

この例では、効力比の SE が少し大きくなっていて、この面からも重みを考慮することは指示されない。

9. 各種の効力比、

9.1. 勾配比の推定、重み付き非線形最小 2 乗法の適用

用量の増加と共に変異コロニー数のSDが大きくなる場合に、重み付き最小 2 乗法の適用しなければならない。CVが一定の場合は、観測値 y_i の 逆数 $w_i = 1/y_i$ を重みにする とよい。用量ごとの変異コロニー数のデータは各々3例しかないがCVを計算した結果を表 9.1 に示すがデータ数が各群 $n = 3$ と小さいので確かなことは言えないが、S薬についてみればSDがやや用量の増加に伴い大きくなっているが、CVが一定とは言いがたい。

表 9.1 用量の増加に伴う CV の変化

dose	T 薬			S 薬		
	mean	SD	CV	mean	SD	CV
0	25.0	1.7	6.9	28.3	4.2	14.7
50	74.0	7.2	9.7	79.3	10.6	13.4
75	99.7	15.0	15.1	126.0	7.8	6.2
100	134.0	3.0	2.2	153.3	8.1	5.3
125	178.0	10.5	5.9	201.7	5.5	2.7
150	204.0	13.9	6.8	239.0	18.2	7.6
300	338.3	8.5	2.5	430.0	24.6	5.7

各群 $n = 3$

したがって、重みを考慮した解析は必要としないが、重み付き非線形回帰の練習問題として結果を表 9.2 に示す。

表 9.2 重み付き非線形回帰

Solution					
	SSE	DFE	MSE	RMSE	
	46.338791699	39	1.1881741	1.090034	
Parameter	Estimate	ApproxStdErr	Lower CL	Upper CL	
b0	24.327837651	2.00098085	20.2804719	28.3752034	
b1_2	1.3491164349	0.03116247	1.28608439	1.41214848	
b2	0.8110285231	0.02493155	0.76182665	0.86282786	

$$ratio = 0.811, \quad 95\%cl = (0.762, 0.863), \quad \text{重み} = 1/y$$

この例では、効力比の SE が少し大きくなっていて、この面からも重みを考慮することは指示されない。

9.2. 反応 y が対数変換されている場合

反応 y が対数変換されている場合で、S薬の反応とT薬の反応に切片は同じであるが傾きが異なる直線があてはまり、かつ用量が等比で設定されている場合は、反応 y および

用量 x について対数をとることにより図 9.1 に示すように平行な直線関係となる。

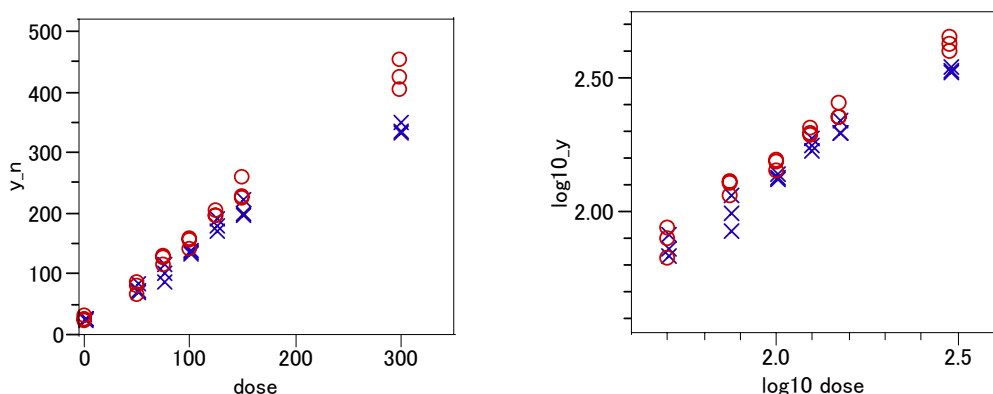


図 9.1 対数変換

反応を対数として S 薬と T 薬の用量反応に平行な直線があてはめられる場合には、その直線の差の逆対数が効力比となる。それぞれの薬剤の回帰式は、切片が異なるが傾きは β_1 と共通である回帰式を求め、切片の差を求める問題である。

$$\text{T 薬: } y^{(T)} = \beta_{0,1} + \beta_1 x$$

$$\text{S 薬: } y^{(S)} = \beta_{0,2} + \beta_1 x$$

この 2 つの回帰式を同時に推定するためにダミー変数を用いた

$$y_i = \beta_{0,2} + \beta_{0,(1-2)} z_1 + \beta_1 x_i + \varepsilon_i$$

を考える。ここで、 $\beta_{0,(1-2)} = \beta_{0,1} - \beta_{0,2}$ である。

表 9.3 反応が対数の場合での切片の差の推定

Parameter Estimates				
Term	Estimate	Std Error	Lower 95%	Upper 95%
Intercept	0.3812	0.0518	0.2758	0.4866
z1	-0.0698	0.0121	-0.0945	-0.0451
log10_dose	0.9096	0.0249	0.8589	0.9602

回帰直線

$$\text{S 薬の切片: } \hat{\beta}_{0,2} = 0.3812$$

$$\text{切片の差: } \hat{\beta}_{0,(1-2)} = -0.0698$$

$$\text{T 薬の切片: } \hat{\beta}_{0,1} = \hat{\beta}_{0,2} + \hat{\beta}_{0,(1-2)}$$

$$\text{T 薬: } y = (0.3812 - 0.0698) + 0.9096x$$

$$\text{S 薬: } y = 0.3812 + 0.9096x$$

ある用量における反応の差(対数)は一定であり、このことから、効力比: $10^{-0.0698} = 0.852$ 、95%信頼区間は (0.805, 0.9014) と推定される。

9.3. 投与量が対数変換されている場合

投与量が等比で設定されていて、S薬とT薬の用量反応が平行な直線である場合に、反応 y が対数変換されていない場合には、平行線の Y 軸方向の距離は、反応の差であり、これからは効力比を求めることができない。しかし、用量についての X 軸が対数となっていれば、こちらの平行線の距離の逆対数は、同じ反応が得られる T 薬の S 薬に対する効力比となる。この考え方は、平行線検定法として生物検定法の代表的な解析方法として知られている。

効力比の推定は、あるコロニー数、例えば 200 個としたときに、200 個となる両群の投与量を逆推定し、その用量の差と信頼区間を求める問題となる。この問題は、前節の線形回帰ではなく、非線形回帰が必要となる。

非線形回帰式は、次の線形回帰式

$$y - \log_{10}(200) = \beta_{0,2} + \beta_{0,(1-2)}z_1 + \beta_1 \log_{10}(x)$$

を考え、 β_1 を共通項として取り出すことにより、次式を得る。

$$y = \beta_1 \left(\frac{\beta_{0,2}}{\beta_1} + \frac{\beta_{0,(1-2)}}{\beta_1} z_1 + x \right) + \log_{10}(200)$$

S 薬の回帰式 $y^{(S)} - \log_{10}(200) = \beta_{0,2} + \beta_1 \log_{10}(x)$ で、 $y^{(S)}$ が $\log_{10}(200)$ となる x の逆推定値は、 $\log_{10}(x) = -\beta_{0,2} / \beta_1$ となるので、 $\beta_{0,2} / \beta_1$ を $-\beta_S$ と置き換えて S 薬の逆推定用量値を作り、 $\beta_{0,(1-2)} / \beta_1$ を β_{S-T} と置き換えて S 薬と T 薬の差を直接推定する。なお、この方法は、投与量が 0 であるような陰性対照を含む場合は、0 の対数はマイナス無限大となり、解析に含めることができない。

表 9.4 S 薬の $y = 200$ および T 薬との差の逆推定

Solution				
	SSE	DFE	MSE	RMSE
	0.0437366784	33	0.0013254	0.0364054
Parameter	Estimate	ApproxStdErr	Lower CL	Upper CL
b1	0.9095681726	0.02488295	0.85923765	0.9598987
b_s	2.1106937811	0.0095605	2.09149902	2.1302356
b_s_t	-0.076704423	0.01350566	-0.1042999	-0.0495787

$\hat{\beta}_{S-T} = -0.0767$ が対数用量の差の推定値である。この逆対数 $10^{-0.0767} = 0.838$ 倍となる。
 95%信頼区間は、(0.787, 0.892) となる。