LIFEREGプロシジャを用いた打ち切りデータを含むワイブル回帰

2015年8月6日

BioStat研究所（株） 高橋 行雄
takahashi.stat@nifty.com

統計モデルの違い

- Lifregプロシジャの統計モデル

$$y = X\beta + \underline{\sigma\varepsilon}$$

- 通常の回帰モデル

$$y = X\beta + \varepsilon$$

- <u>基準化残差 ε</u> の前についている <u>σ</u> は，未知の尺度（scale）パラメータである.
- この違いに焦点をあて，ワイブル回帰の本質にせまる.

打ち切りのある人工データ

No	x	t（日）	$y = \ln t$	打ち切り
1	1	100	4.61	1
2	1	100	4.61	1
3	2	100	4.61	1
4	2	90	4.50	0
5	3	60	4.09	0
6	3	45	3.81	0
7	4	35	3.56	0
8	4	30	3.40	0

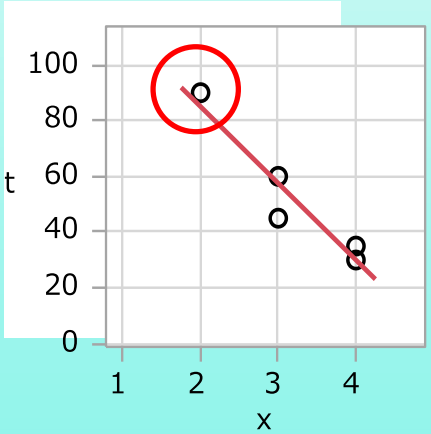
Excel による線形回帰分析（n=5）

t		X			β ^		t^		ε^		Se
90		1	2		140	=	85.0		5.0		212.50
60		1	3		-27.5		57.5		2.5		Ve
45		1	3				57.5		-12.5		70.83
35		1	4				30.0		5.0		n
30		1	4				30.0		0.0		5

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{t}$$

$$t_1 = 140 x_{0,1} - 27.5 x_{1,1} + \hat{\varepsilon}_1$$

$$\hat{\varepsilon}_1 = 90 - 85 = 5.0$$



Lifereg による最尤法

```
data d01 ;
    input x0  x1  t @@ ;
datalines;
1  2  90
1  3  60   1  3  45
1  4  35   1  4  30
;
proc lifereg data=d01 ;
    model t = x0 x1
            / d = Normal noint ;
run ;                /* 正規分布 */
```

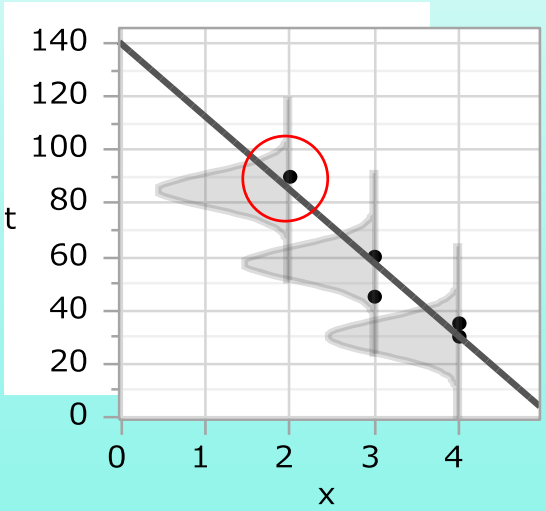
最尤法による回帰分析（*n*=5）

Parameter	DF	Estimate	Standard Error	95% Confidence Limits	
x0	1	140.00	12.80	114.91	165.09
x1	1	-27.50	3.90	-35.14	-19.86
Scale	1	6.52	2.06	3.51	12.12

$$t_1 = 140 - 27.5x_1 + \hat{\sigma}\hat{\varepsilon}_1$$

$$\hat{\sigma} = 6.52$$

$$\hat{\varepsilon}_1 = \frac{5.0}{6.52} = 0.77$$



Excel による対数尤度の計算

t		X		パラメータ			t^	ε^	σ ^ε^	ϕ_{nor}	Li	ln Li
90		1	2	β^$_0$=	140	=	85.0	0.77	5.00	0.2973	0.0456	-3.0878
60		1	3	β^$_1$=	-27.5		57.5	0.38	2.50	0.3707	0.0569	-2.8672
45		1	3	σ^=	6.5191		57.5	-1.92	-12.50	0.0635	0.0097	-4.6319
35		1	4				30.0	0.77	5.00	0.2973	0.0456	-3.0878
30		1	4				30.0	0.00	0.00	0.3989	0.0612	-2.7937
											L	ln L
											7.0E-08	-16.4685

$$L_1 = \frac{1}{6.52}\phi_{\mathrm{nor}}\left(\frac{90-85.0}{6.52}\right) = 0.0456$$

ϕ_{nor}：標準正規分布

$$\ln(L_1) = \ln(0.0456) = -3.0878$$

$$\ln(L) = \sum_{i=1}^{5} \ln(L_i) = -16.4685$$

打ち切りデータを含む回帰分析

- 最小2乗法による線形回帰では，打ち切りデータを含めることは，不可能である.
- 正規分布を誤差に仮定した最尤法
- Lifereg プロシジャは，最尤法が標準

```
proc Lifereg data=d02 ;
    model t * censor(1) = x0 x1
                  / d = Normal noint ;
                /* 打ち切りがあっても 正規分布 */
```

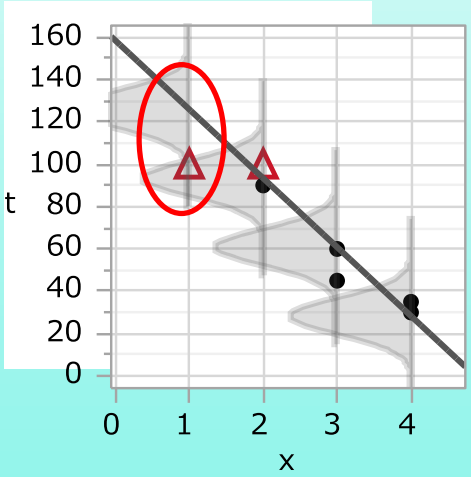
打ち切りデータを含んだ結果

Parameter	DF	Estimate	Standard Error	95% Confidence Limits	
x0	1	158.57	14.24	130.66	186.47
x1	1	-32.57	4.52	-41.42	-23.71
Scale	1	8.79	2.87	4.63	16.68

$$t_1 = 158x_{0,1} - 32.6x_{1,1} + \hat{\sigma}\hat{\varepsilon}_1$$

$$\hat{\sigma} = 8.79$$

$$\hat{\varepsilon}_1 = \frac{-26}{8.79} = 2.96(\sigma)$$



打ち切りデータの尤度

- 打ち切りがない場合：正規分布の確率密度

$$L_i = \frac{1}{\hat{\sigma}_{\mathrm{nor}}} \phi_{\mathrm{nor}} \left(\frac{t_i - \hat{t}_i}{\hat{\sigma}_{\mathrm{nor}}} \right) = \frac{1}{\hat{\sigma}_{\mathrm{nor}}} \phi_{\mathrm{nor}} \left(\hat{\varepsilon}_i \right)$$

- 打ち切りの場合： 生存関数 ***S* = 1 - *F***

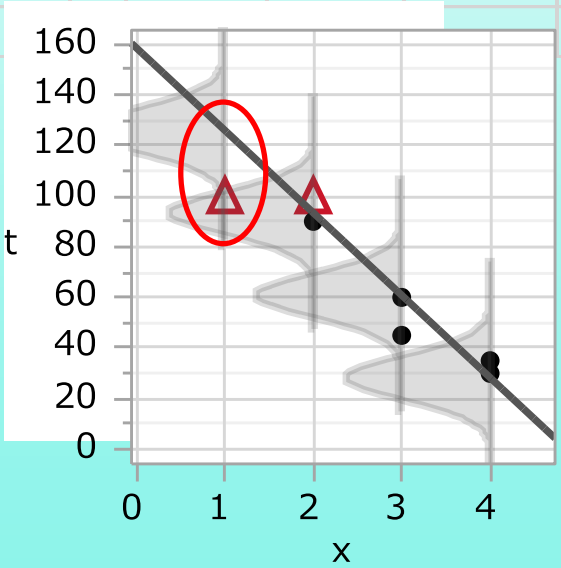
$$L_i = 1 - \Phi_{\mathrm{nor}} \left(\frac{t_i - \hat{t}_i}{\hat{\sigma}_{\mathrm{nor}}} \right) = 1 - \Phi_{\mathrm{nor}} \left(\hat{\varepsilon}_i \right)$$

打ち切りデータの尤度

- 右側の打ち切りがあったデータは，その時点までは，“生存”していたが，その後は，いつ“死亡”するかわからない．
- 早めに打ち切られた場合と，遅めに打ち切られた場合の，どちらの方が情報量が多いのだろうか．
- すぐ“死亡”した場合に，ほとんど情報がないので限りなく１に近ければ，対数尤度は，０に近づき，最尤法では，ほとんど無視される．
- これらを考えたときに，打ち切りデータには，生存関数を尤度とすることの妥当性となる．

Excel・JMP による可視化

t		X		censor	パラメータ			t^	ε^	σ^ε^	ϕ_{nor}	1-Φ_{nor}	Li	ln Li
100		1	1	1	β^$_0$=	158.57	=	126.0	-2.96	-26.00	0.0050	***0.9985***	***0.9985***	***-0.0016***
100		1	1	1	β^$_1$=	-32.57		126.0	-2.96	-26.00	0.0050	***0.9985***	***0.9985***	***-0.0016***
100		1	2	1	σ^=	8.79		93.4	0.75	6.57	0.3017	***0.2274***	***0.2274***	***-1.4810***
90		1	2	0				93.4	-0.39	-3.43	0.3697	0.6518	0.0421	-3.1687
60		1	3	0				60.9	-0.10	-0.86	0.3970	0.5391	0.0452	-3.0974
45		1	3	0				60.9	-1.80	-15.86	0.0783	0.9644	0.0089	-4.7211
35		1	4	0				28.3	0.76	6.70	0.2983	0.2228	0.0339	-3.3834
30		1	4	0				28.3	0.19	1.70	0.3915	0.4232	0.0445	-3.1113
													L	ln L
													5.8E-09	-18.9659



打ち切りデータの尤度

生存関数

$$L_1 = 1 - \Phi_{\mathrm{nor}}\left(\frac{t_1 - \hat{t}_1}{\hat{\sigma}_{\mathrm{nor}}}\right) = 1 - \Phi_{\mathrm{nor}}\left(\frac{100 - 126.0}{8.79}\right) = 0.9985$$

対数尤度

$$\ln(L_1) = \ln(0.9985) = -0.0016$$

対数正規分布を誤差分布

```
proc lifereg data=d02 ;
    model ln_t * censor(1) = x0 x1
                  / d = Normal noint ;
proc lifereg data=d02 ;
    model t * censor(1) = x0 x1
                  / d = Lnormal noint ;
                    /* 対数正規分布 */
```

どちらでも対数正規分布のあてはめ

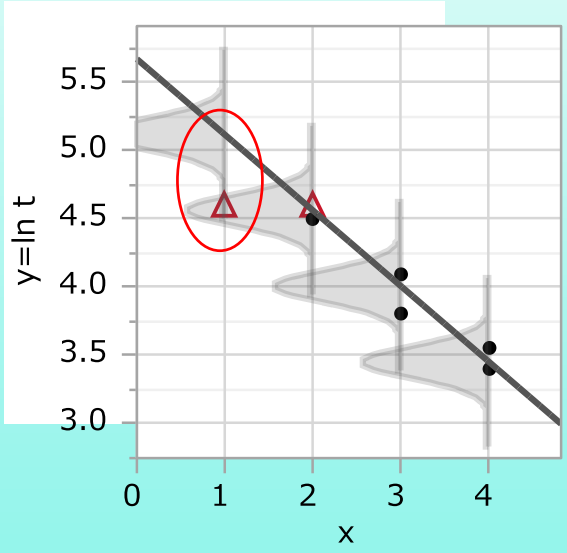
対数正規分布を仮定

Parameter	DF	Estimate	Standard Error	95% Confidence Limits	
x0	1	5.6798	0.1976	5.2924	6.0671
x1	1	-0.5575	0.0624	-0.6798	-0.4352
Scale	1	0.1187	0.0386	0.0627	0.2247

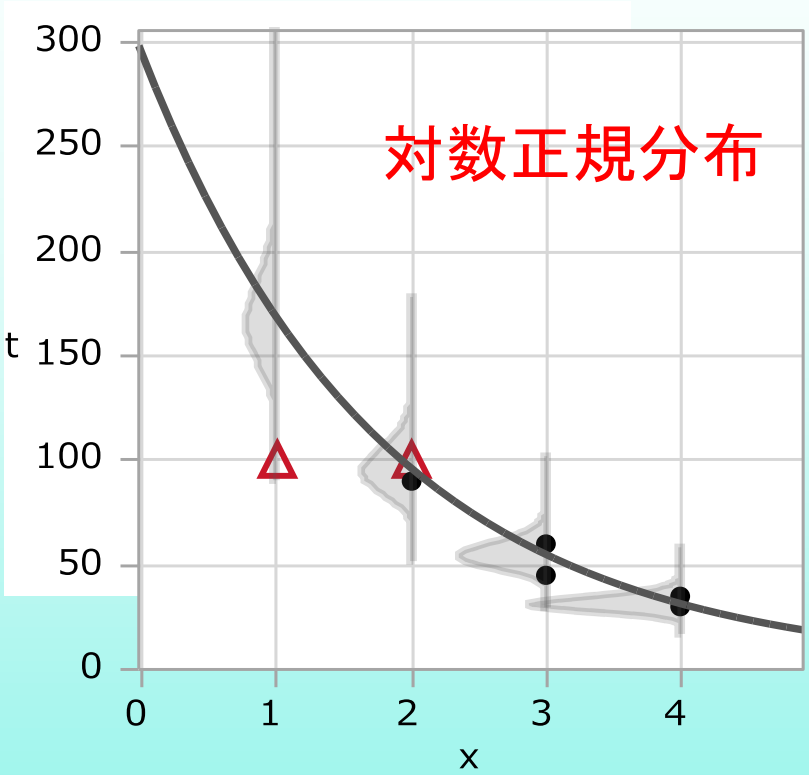
$$y_1 = 5.68x_{0,1} - 0.56x_{1,1} + \hat{\sigma}\hat{\varepsilon}_1$$

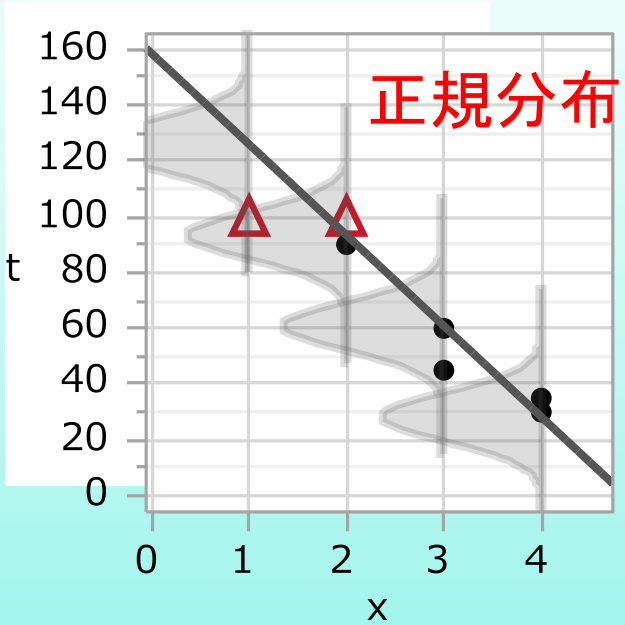
$$\hat{\sigma} = 0.12$$

$$\hat{\varepsilon}_1 = \frac{-0.52}{0.12} = -4.36$$



実時間 *t* での表示





$x = 1$ での推定値が 167.7 と増加

Weibull分布

- 分布関数

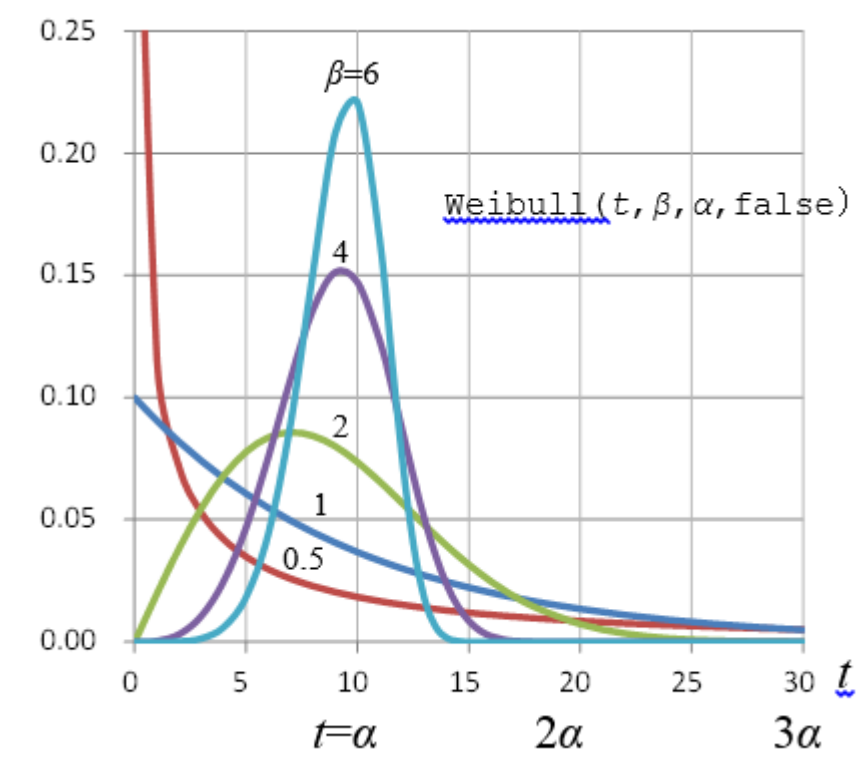
$$F(t) = 1 - \exp\left\{ -\left(\frac{t}{\alpha} \right)^{\beta} \right\}$$

- 確率密度関数

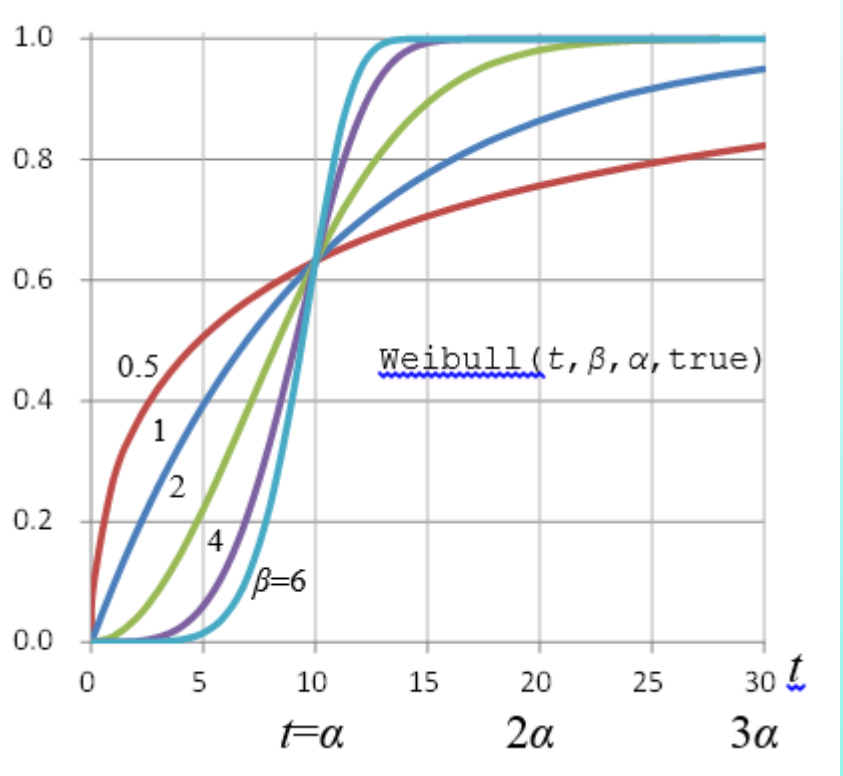
$$f(t) = \frac{\beta}{\alpha^{\beta}} t^{\beta - 1} \exp\left\{ -\left(\frac{t}{\alpha} \right)^{\beta} \right\}$$

Weibull分布の形状

確率密度関数



分布関数



Weibull分布を誤差に仮定？

- 1群の生存時間にWeibull分布をあてはめ，推定されたパラメータ β から“故障”の発現メカニズムを解釈できることは有用である.
- 正規分布のパラメータ
  - 平均 μ_{nor}，標準偏差 σ_{nor}
- Weibull分布のパラメータ
  - 平均と標準偏差は別途計算が必要
  - 回帰分析での誤差分布として用い難い.

対数“Weibull”分布

- 対数“Weibull”分布➔最小極値分布

$$F(t)=1-\exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right]=1-\exp\left[-\left(\frac{t}{\underline{\exp(\mu_{\mathrm{sev}})}}\right)^{\underline{1/\sigma_{\mathrm{sev}}}}\right]$$

$$=1-\exp\left\{\underline{-\exp}\left[\underline{\ln}\left(\frac{t}{\exp(\mu_{\mathrm{sev}})}\right)^{1/\sigma_{\mathrm{sev}}}\right]\right\}=1-\exp\left[-\exp\left(\frac{\ln(t)-\mu_{\mathrm{sev}}}{\sigma_{\mathrm{sev}}}\right)\right]$$

/* 二重指数分布 */

/* 最小極値分布 */

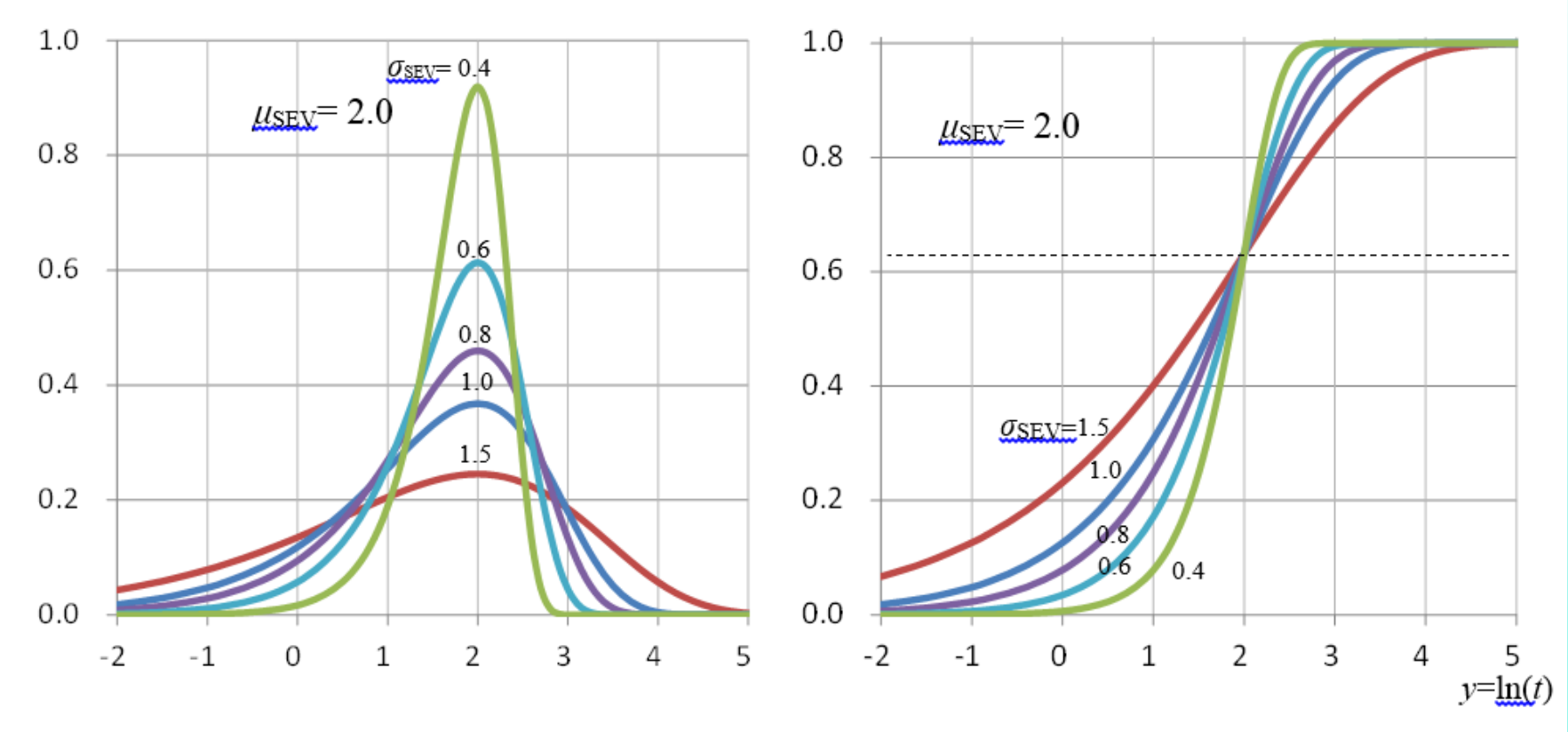
$$\alpha=\underline{\exp(\mu_{\mathrm{sev}})}$$

$$\beta=\underline{1/\sigma_{\mathrm{sev}}}$$

/* SEV(Smallest Extreme Value Distribution */

最小極値分布の形状

確率密度関数　　分布関数



最小極値分布を誤差分布

```
proc lifereg data=d02 ;
    model  t * censor(1) = x0 x1
                / d = Weibull noint ;
      /* 内部では Weibull としても最小極値分布 */
proc lifereg data=d02 ;
    model  ln_t * censor(1) = x0 x1
              / d = Weibull noint nolog;
```

どちらでも同じ結果となる

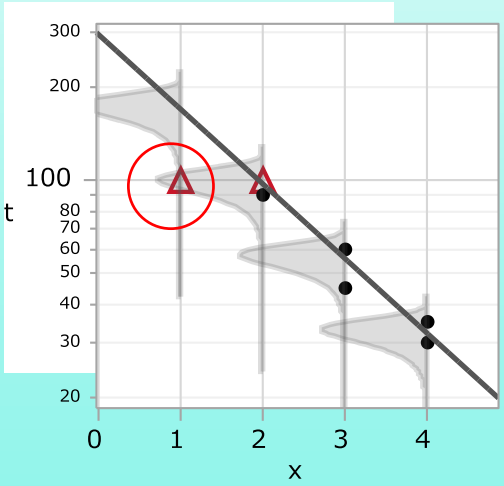
最小極値分布を誤差とする回帰

Parameter	DF	Estimate	Standard Error	95% Confidence Limits	
x0	1	5.7259	0.1660	5.4006	6.0512
x1	1	-0.5565	0.0507	-0.6560	-0.4571
Scale	1	0.0885	0.0335	0.0421	0.1859
Weibull Shape	1	11.2981	4.2793	5.3778	23.7360

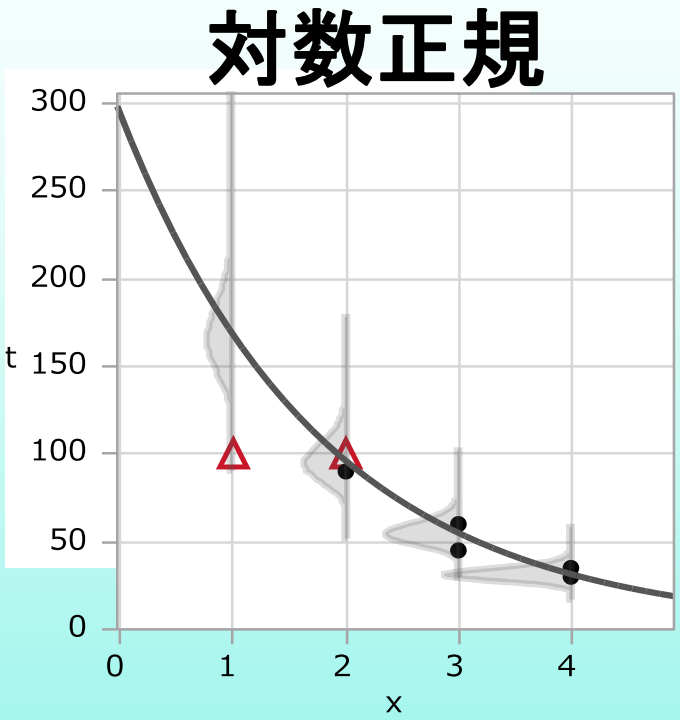
$$y_1 = 5.73x_{0,1} - 0.56x_{1,1} + \hat{\sigma}\hat{\varepsilon}_1$$

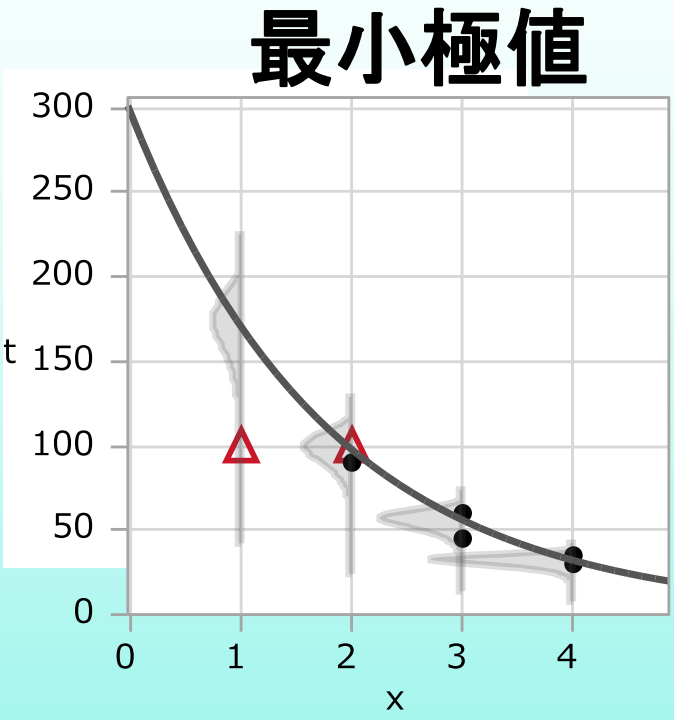
$$\hat{\sigma} = 0.0885$$

$$\hat{\varepsilon}_1 = \frac{-0.56}{0.0885} = -6.73$$



対数正規 vs 最小極値





打ち切りデータに関して，左に裾を引く最小極値分布の仮定が望ましい．

モータの絶縁組織の寿命試験

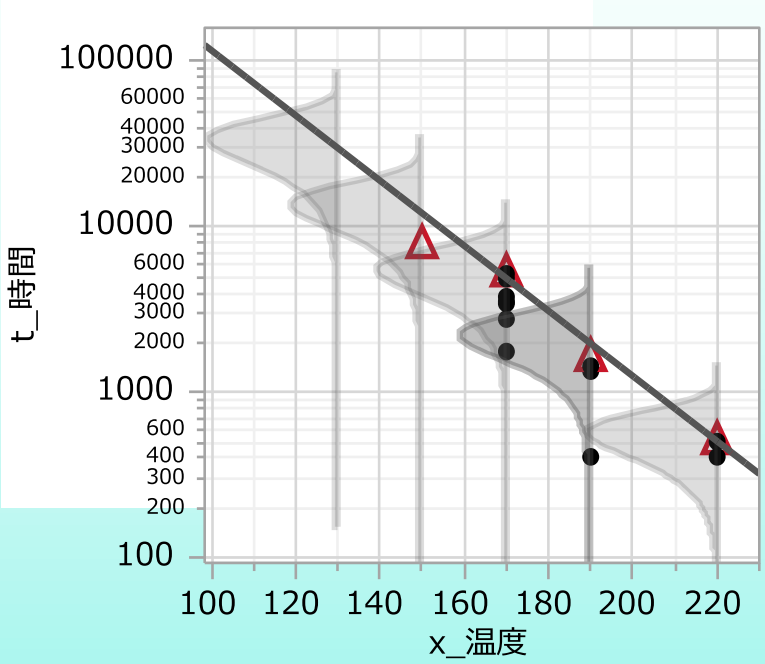
Kalbfleisch and Prentice（2002） Table 1.3

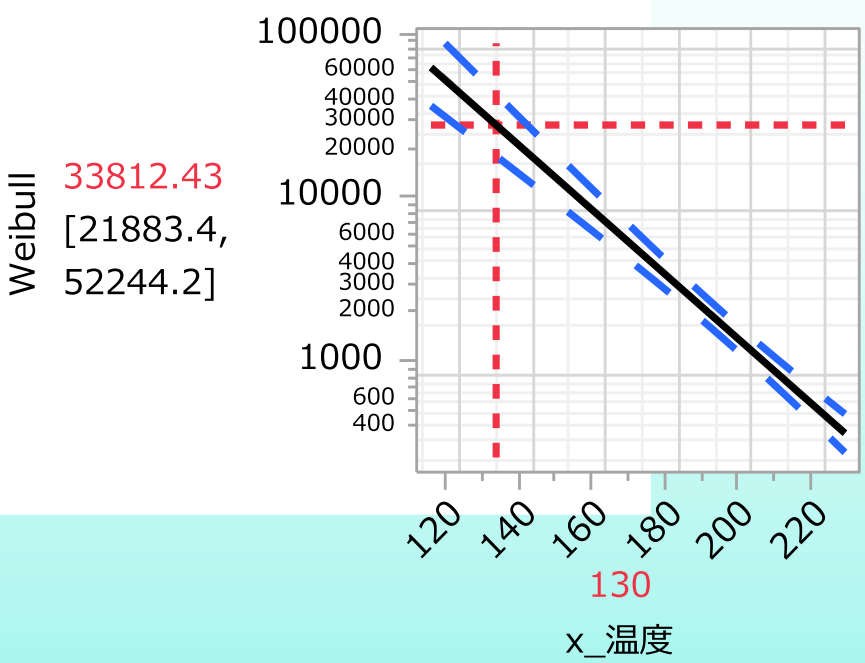
150℃	All 10 motorettes without failure at 8064 hours
170℃	1764，2772，3444，3542，3780，4860，5196
	3 motorettes without failure at 5448 hours
190℃	408，408，1344，1344，1440
	5 motorettes without failure at 1680 hours
220℃	408，408，504，504，504
	5 motorettes without failure at 528 hours

40件分のデータリスト

OBS	x	c	are_x	t	ln_t
1	150	1	26.149	8064	8.995
2	150	1	26.149	8064	8.995
:					
10	150	1	26.149	8064	8.995
11	170	0	24.969	1764	7.475
12	170	0	24.969	2772	7.927
:					
20	170	1	24.969	5448	8.603
21	190	0	23.891	408	6.011
22	190	0	23.891	408	6.011
:					
30	190	1	23.891	1680	7.427
31	220	0	22.437	408	6.011
32	220	0	22.437	408	6.011
:					
40	220	1	22.437	528	6.269

“最小極値” 回帰





130℃での推定値（下側 63.21%点）

$$t_{130} = \exp(16.318519 - 0.045307 \times 130) = 33,812$$

原著（p70）およびSAS9.3（Ex 50.1）

- 3.7 ILLUSTRATIONS IN MORE COMPLEX DATA SETS
- 3.7.1 Accelerated Life Testing
  - 150℃のデータを除く
  - アレニウス温度 = 1,000/（temp + 273.2）
  - Weibull（最小極値）モデルでの結果

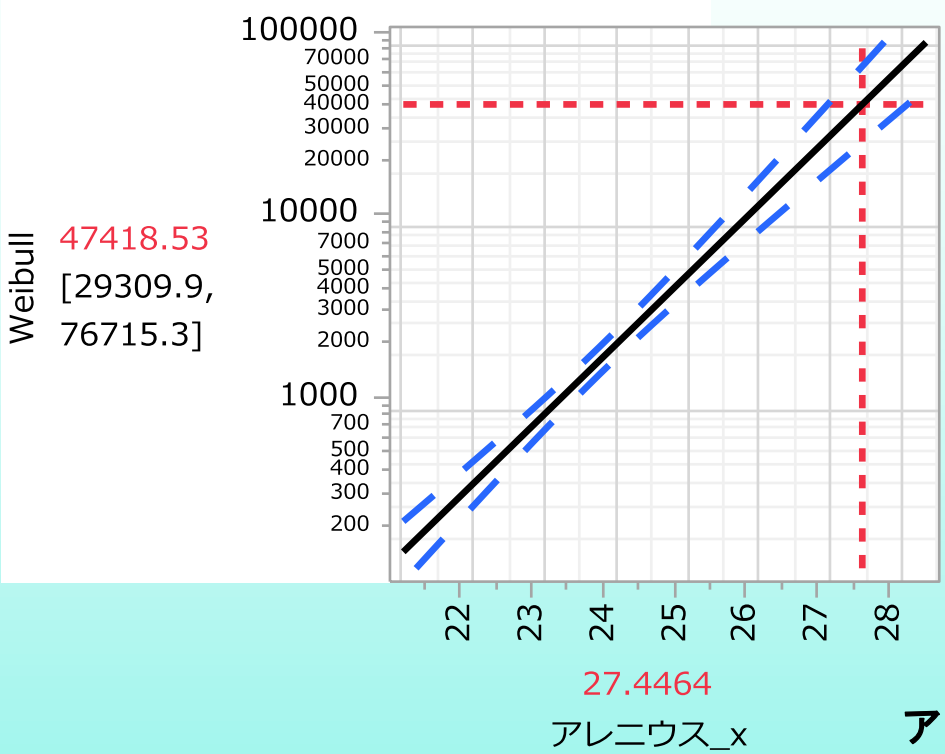
$$\ln(t_i) = -11.891 + 9.038\,x_{are} + 0.3613\varepsilon_i$$

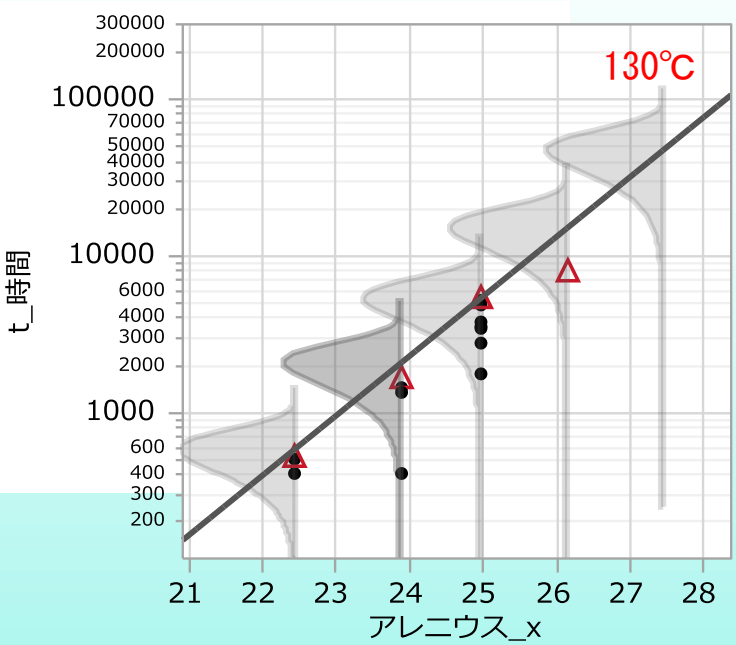
- 高橋の結果 アレニウス温度 = 11,065/（temp + 273.15）

$$\ln(t_i) = -13.353 + 0.8788\,x_{are} + 0.3343\varepsilon_i$$

ボルツマン定数：8.6171×10^{-5} 1/8.6171×10-5=11065

アレニウス温度での回帰





アレニウス $x = 11065/(x + 273.15)$

温度の大小関係は逆点し，推定は伸びる

まとめ

- パラメトリックな生存時間解析は，ノンパラメトリック，セミパラメトリック生存時間解析に引き続いて紹介される場合には，ハザードを用いた説明となりがちである.
- 最小2乗法による回帰分析から始めて，正規分布を誤差とする最尤法による回帰分析，さらに打ち切りデータを含めた回帰分析に拡張した.
- 引き続き“対数正規”回帰を導入し，それと対比しつつ，“対数ワイブル”回帰を示した.
- さらに，加速寿命試験データの解析結果を示した.