第八章 参数模型

问题：如何选取合适的理赔额分布或理赔次数的分布。

分布拟合检验的一般步骤

（1）获得损失分布的经验分布信息，例如经验分布图、样本均值、样本方差、分位点等。

（2）选择一种概率分布作为损失的分布类型，估计所选择分布中

所包含的参数；

（3）对分布进行拟合检验，以确信所选择的分布类型和参数估计是否恰当；

（4）考虑是否还有其它适合的分布，如果有，重复第（1）—（3）步;

（5）在几种合适的分布中选取一个最优的分布作为损失额的分布。选择的标准有多种，常用的方法是比较 χ^2 统计量的值，比较最大似然函数的值等；

（6）模型的修正。选择模型后，要注意随时对模型修正，以反映未来发生的情况，如通货膨胀，免赔额变化等。

The Candidate will be able to:

a) Estimate the parameters for severity, frequency, and aggregate distributions using Maximum Likelihood Estimation for:

- Complete, individual data
- Complete, grouped data
- Truncated or censored data

b) Estimate the variance of the estimators and construct confidence intervals.

c) Use the delta method to estimate the variance of the maximum likelihood estimator of a function of the parameter(s).

d) Estimate the parameters for severity, frequency, and aggregate distributions using Bayesian Estimation.

e) Perform model selection using:

- Graphical procedures.
- Hypothesis tests, including Chi-square goodness-of-fit, Kolmogorov-Smirnov and Likelihood ratio (LRT) tests.
- Score-based approaches, including Schwarz Bayesian Criterion (SBC), Bayesian

Information Criterion (BIC) and Akaike Information Criterion (AIC).

一、常见数据类型

Data set A

下表是某保险公司在一年内小汽车发生事故次数的统计数据：

Number of accidence	Number of drivers
0	81,714
1	11,306
2	1618
3	250
4	40
5 or more	7

Data set B

下表是某劳工补偿险的部分原始损失数据

27	82	115	126	155	161	243	294	340	384
457	680	855	877	974	1193	1340	1884	2558	15743

Data set C

下表是某责任险的赔付数据：

Payment range	Number of payments
0-7500	99

7500-17500	42
17500-32500	29
32500-67500	28
67500-125000	17
125000-300000	9
Over 300000	3

Data setD1

寿险保单终止有三种状态：死亡，期满和退保（surrender）。下表是某寿险保单持有人在签订保单后 5 年内保单终止的时间记录。

policyholder	Time of death	Time of surrender
1	-	0.1

2	4.8	0.5
3	-	0.8
4	0.8	3.9
5	3.1	1.8
6	-	1.8
7	-	1.8
8	-	2.1
9	-	2.5
10	2.9	2.8
11	2.9	4.6
12	-	3.9
13	4.0	-

14	-	4.0
15	-	4.1
16	4.8	-
17	-	4.8
18	-	4.8
19-30	-	-

其中‘-’表示时间未知，最后 12 个保单持有人保单期满并退保。

Data set D2

下表表示寿险保单存活状态的两次观测值，其中 First observed 表示第一次观测的时间，若为 0 则表示保单签订后马上进行记录，Last observed 表示第二次观测的时间，Event 表示最后一次观测时保单持有人的状态，S 表示退保，D 表示死亡，E 表示保单期满。

Policy	First observed	Last observed	Event	Policy	First observed	Last observed	Event
1	0	0.1	S	16	0	4.8	D
2	0	0.5	S	17	0	4.8	S
3	0	0.8	S	18	0	5.0	S
4	0	0.8	D	19-30	0	5.0	E
5	0	1.8	S	31	0.3	5.0	E
6	0	1.8	S	32	0.7	5.0	E
7	0	2.1	S	33	1.0	4.1	D
8	0	2.5	S	34	1.8	3.1	D
9	0	2.8	S	35	2.1	3.9	S
10	0	2.9	D	36	2.9	5.0	E

11	0	2.9	D	37		4.8	S
12	0	3.9	S	38	3.2	4.0	D
13	0	4.0	D	39	3.4	5.0	E
14	0	4.0	S	40	3.9	5.0	E
15	0	4.1	s				

Data setE

这是一组责任险保单的赔付数据,这个数据中包含了不同的免赔额和限额。

年	免赔额	最大支付额	赔付额	年	免赔额	最大支付额	赔付额
90	0	1000000	2890	91	15000000	10000000	10000000
90	0	5000000	5851	92	0	1000000	1836
90	250000	10000000	15347	92	0	1000000	10705

90	0	1000000	15635	92	0	5000000	10973
90	0	3000000	20553	92	0	5000000	13408
90	0	10000000	34584	92	0	10000000	16339
90	0	10000000	79661	92	350000	5000000	95736
90	0	400000	132601	92	0	1000000	212313
90	1500000	5000000	1410989	92	0	5000000	439543
90	0	10000000	2784401	92	70000000	15000000	1098710
90	0	10000000	4894360	92	0	3000000	1211180
90	10000000	10000000	9316751	93	0	500000	10510
91	0	1000000	1891	93	0	3000000	14029
91	0	3000000	30893	93	0	10000000	15296
91	0	1000000	31392	93	50000	1000000	27516
91	500000	10000000	49488	93	0	10000000	53467
91	175000	1000000	67425	93	300000	5000000	87463
91	0	1000000	150310	93	100000	5000000	220995
91	45000000	33000000	1335735	93	150000	5000000	274086

91	0	10000000	3308199	93	0	5000000	1862304
91	12750000	10000000	10000000	93	0	5000000	5000000

请同学们观察上述几个数据集的特征

- 个体，完整数据
- 分组数据
- Truncated 和 Censored 数据

估计方法：

1、 矩估计法

2、 分位数估计法

3、 极大似然估计法

一、矩估计法（**了解**）

基本思想：

设未知参数为$\boldsymbol{\theta}=(\theta_1,\cdots,\theta_n)$，解方程

$$\hat{\mu}_k=\mu_k(\boldsymbol{\theta}),k=1,\cdots,n,$$

或$\mu'_k(\boldsymbol{\theta})=\hat{\mu}'_k$， $k=1,2,...,n$

得到$\boldsymbol{\theta}=(\theta_1,\cdots,\theta_n)$的估计值。

下面分三种情况来讨论：

1、个体完整数据的矩估计

略，具体步骤参看概率统计教材。

2、分组数据的样本矩，

$$\hat{\mu}'_k = \sum_{j=1}^{r} \int_{c_{j-1}}^{c_j} x^k \frac{n_j}{n(c_j - c_{j-1})} dx = \sum_{j=1}^{r} \frac{n_j (c_j^{k+1} - c_{j-1}^{k+1})}{n(c_j - c_{j-1})(k+1)}$$

3、非完整数据

二、分位数估计法（了解）

（1）设随机变量 X 的分布函数为 $F(x,\theta)$，称 $\pi_p(\theta)$ 为 $F(x,\theta)$ 的 **100%p 分位数**，如果 $\pi_p(\theta)$ 满足

$$F(\pi_p(\theta)\,|\,\theta)=p$$

(2)分位数的估计：令

$$p_j=F(\hat{\pi}_{p_i},\theta),\qquad \theta=(\theta_1,\cdots,\theta_n),\ j=1,\cdots,n\ ,$$

其中 $\hat{\pi}_{p_j}$ 为样本的 p_j 分位点。求解这 n 个方程就得到参数估计值 $\hat{\theta}=(\hat{\theta}_1,\cdots,\hat{\theta}_n)$。

个体数据的样本分位点：

将 $x_1,\cdots,x_n$ 按从小到大的顺序排列为 $x_{(1)},\cdots,x_{(n)}$。对于 $0<p<1$，

g=[(n+1)p]表示不超过(n+1)p 的最大整数，此时认为分位数应该在 $x_{(g)}$ 和 $x_{(g+1)}$之间。记 h=(n+1)p-g 表示(n+1)p 的小数部分，则样本的 100p％的分位数为

$$\pi_{(p)}=(1-h)\,x_{(g)}+h\,x_{(g+1)}$$

中位数

- 当 n 为奇数时,记 k=(n+1)/2, 中位数为 $x_{(k)}$,
- 当 n 为偶数时,记 k=n/2, 则中位数为

例 8.2：设下表中的理赔记录用 weibull 分布来拟合，用 25%和 75%分位数来估计参数的值。

0.1	0.5	2.2	4.1	28.1
0.2	0.7	2.6	5.9	30.0
0.2	0.9	2.9	6.2	49.2
0.3	1.3	3.2	12.1	63.8
0.4	1.8	3.3	13.65	118.0

解：weibull 的分布函数为

$$F_X(x) = 1 - e^{-x^\tau/\theta}$$

分别令

$$0.75 = 1 - e^{-x^\tau/\theta} \text{ 及 } 0.25 = 1 - e^{-x^\tau/\theta}$$

解得

$$x_1 = (\theta \log 4)^{\frac{1}{\tau}}, x_2 = (\theta \log(4/3))^{\frac{1}{\tau}}$$

由于 0.25×26＝6.5，因此，0.25 的分位点为

$$0.5\times 0.5+0.5\times 0.7=0.65$$

类似计算，0.75×26＝19.5，0.75 的分位点为

$$0.5\times 12.1+0.5\times 13.65=12.875$$

由分位数估计法

$$12.875 = (\theta \log 4)^{\frac{1}{\tau}}, \qquad 0.65 = (\theta \log(4/3))^{\frac{1}{\tau}}$$

解得 $\tau = 0.526$， θ ＝2.770。从而确定了损失分布。

分组数据的样本分位点：

对于$0<p<1$，分组数据的样本100p%的分位点$\hat{\pi}_p$定义为

$$p=\hat{F}_n(\hat{\pi}_p)$$

即

$$\hat{\pi}_p=\frac{p-F_n(c_{j-1})}{F_n(c_j)-F_n(c_{j-1})}\times c_j+\frac{F_n(c_j)-p}{F_n(c_j)-F_n(c_{j-1})}\times c_{j-1}$$

或者

$$\hat{\pi}_p=c_{j-1}+(np-\sum_{i=1}^{j-1}n_i)\frac{c_j-c_{j-1}}{n_j},$$

其中 c_{j-1} 和 c_j 满足 $F_n(c_{j-1}) \le p < F_n(c_j)$。

例 8.4 某责任险保单规定了保单限额为 300,000 元，表中的第一至三列给出了该险种 217 份保单的理赔额情况。假设理赔额服从对数正态分布。请用 30%和 70%分位数法估计参数。

表 2.2.5

理赔额	保单数	平均理赔额
0－2,500	41	1,389
2,500－7,500	48	4,661
7,500－12,500	24	9,991
12,500－17,500	18	15,482

17,500－22,500	15	20,232
22,500－32,500	14	26,616
32,500－47,500	16	40,278
47,500－67,500	12	56,414
67,500－87,500	6	74,985
87,500－125,000	11	106,851
125,000－225,000	5	184,735
225,000－300,000	4	264,025
300,000	3	300,000

解 经验分布 30％和 70％的分位点分别为

$$2{,}500+(65.1-41)5{,}000/48=5{,}010$$

$$22{,}500+(151.9-146)10{,}000/14=26{,}714$$

令 5,010 和 26,714 为对数正态分布的分位点，即

$$0.3 = \Phi[(\log 5,010 - \mu)/\sigma]$$

$$0.7 = \Phi[(\log 26,714 - \mu)/\sigma]$$

解方程组得 $\hat{\sigma} = 1.595871$， $\hat{\mu} = 9.356065$。

54. You are given:

(i) Losses follow an exponential distribution with mean θ.

(ii) A random sample of losses is distributed as follows:

Loss Range	Number of Losses
(0 – 100]	32
(100 – 200]	21
(200 – 400]	27
(400 – 750]	16
(750 – 1000]	2
(1000 – 1500]	2
Total	100

Estimate θ by matching at the 80th percentile.

Question # 54
Key: A

Loss Range	Cum. Prob.
0 – 100	0.320
100 – 200	0.530
200 – 400	0.800
400 – 750	0.960
750 – 1000	0.980
1000 – 1500	1.000

At 400, $F(400) = 0.8 = 1 - e^{-400/\theta}$; solving gives $\theta = 248.53$.

三、极大似然估计法（掌握）

极大函数

$$L(\boldsymbol{\theta})=\prod_{j=1}^{n}P(X_j\in A_j\,|\,\boldsymbol{\theta})$$

1、个体完整数据：

$$L(\boldsymbol{\theta})=\prod_{j=1}^{n}f_{X_j}(x_j\,|\,\boldsymbol{\theta})\,,\quad l(\boldsymbol{\theta})=\sum_{j=1}^{n}\ln f_{X_j}(x_j\,|\,\boldsymbol{\theta})$$

137. You are given the following three observations:

0.74 0.81 0.95

You fit a distribution with the following density function to the data:

$$f(x) = (p+1)x^p, \quad 0 < x < 1, \; p > -1$$

Calculate the maximum likelihood estimate of p.

$$L(p)=f(0.74)f(0.81)f(0.95)=(p+1)0.74^{p}(p+1)0.81^{p}(p+1)0.95^{p}$$
$$=(p+1)^{3}(0.56943)^{p}$$
$$l(p)=\ln L(p)=3\ln(p+1)+p\ln(0.56943)$$
$$l'(p)=\frac{3}{p+1}-0.563119=0$$
$$p+1=\frac{3}{0.563119}=5.32747,\ p=4.32747.$$

2、分组完整数据：

$$L(\boldsymbol{\theta})=\prod_{j=1}^{r}[F(c_j\,|\,\boldsymbol{\theta})-F(c_{j-1}\,|\,\boldsymbol{\theta})]^{n_j}\ ,$$

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{r} n_j \ln[F(c_j \mid \boldsymbol{\theta}) - F(c_{j-1} \mid \boldsymbol{\theta})]$$

例 8.5：假设某险种的理赔数据如表

理赔额范围	理赔数
0~7,500	99
7,500~17,500	42
17,500~32,500	29
32,500~67,500	28
67,500~125,000	17
125,000~300,000	9
Over 300,000	3

若假设理赔额服从指数分布，则对数似然函数为

$$
\begin{aligned}
l(\theta) &= 99\ln[F(7,500)-F(0)]+42\ln[F(17,500)-F(7,500)]+\cdots \\
&\quad +3\ln[1-F(300,000)] \\
&= 99\ln[1-e^{-7,500/\theta}]+42\ln[e^{-7,500/\theta}-e^{-17,500/\theta}]+\cdots \\
&\quad +3\ln e^{-300,000/\theta}
\end{aligned}
$$

使用数值算法得到极大似然估计为 $\hat{\theta}=29,721$，似然函数值为－406.03。

```
rm(list=ls())
x<-c(0,7500,17500,32500,67500,125000,300000,10000000)#理赔额范围
n<-c(99,42,29,28,17,9,3)
f<-function(p,x,n){
  a=0
```

```r
  for(i in 1:length(n)){
    a=a+n[i]*log(exp(-x[i]/p)-exp(-x[i+1]/p))
  }
  return(a)
}
xmax <- optimize(f,c(10000,100000),tol = 0.0001, maximum = TR
UE,x,n)
xmax

## $maximum
## [1] 29720.77
##
## $objective
## [1] -406.0267
```

某医疗责任险的年理赔次数记录如表

年份	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
理赔次数	6	2	3	0	2	1	2	5	1	3

运用极大似然估计下列分布的参数:(1) 泊松分布(2)负二项分布。

解：首先按照理赔次数的值将 7-6 转化为频数表，见下表

理赔次数 k	观测样本数目 n_k
0	1
1	2
2	3
3	2
4	0
5	1
6	1
7+	0

n_k 是发生 k 次理赔的年份数。

假设 N 服从泊松分布，

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!}，\ \ln p_k = -\lambda + k\ln\lambda - \ln k!$$

$$l = \ln L = \sum_{k=0}^{\infty} n_k(-\lambda + k\ln\lambda - \ln k!)$$

$$= -\lambda n + \sum_{k=0}^{\infty} k n_k \ln\lambda + \sum_{k=0}^{\infty} n_k \ln k!$$

$$\frac{dl}{d\lambda} = -n + \sum_{k=0}^{\infty} k n_k \frac{1}{\lambda}$$

其中 $n = \sum_{k=0}^{\infty} n_k$ 是样本量。令上式等于 0 推出 $\hat{\lambda} = \frac{\sum_{k=0}^{\infty} k n_k}{n} = \overline{x}$ 。

代入本例数据得到 $\hat{\lambda}=2.5$。

注意，如果本例中 n_{7+} 的值不等于 0，则似然函数为

$$L=p_0^{n_0}p_1^{n_1}p_2^{n_2}p_3^{n_3}p_4^{n_4}p_5^{n_5}p_6^{n_6}(1-p_0-p_1-p_2-p_3-p_4-p_5-p_6)^{n_{7+}}$$

这时求极大似然函数值较为困难，需采用数值算法来计算。

（2）若采用负二项分布拟合表 7-7 的数据，则对数似然函数可写为

$$l=\sum_{k=0}^{\infty}n_k\ln p_k$$

$$=\sum_{k=0}^{\infty}n_k\ln\left[\ln\binom{r+k-1}{k}-r\ln(1+\beta)+k\ln\beta-k\ln(1+\beta)\right]$$

$$\frac{\partial l}{\partial\beta}=\sum_{k=0}^{\infty}n_k\left(\frac{k}{\beta}-\frac{r+k}{1+\beta}\right)\quad\text{（7-1-11）}$$

$$\begin{aligned}\frac{\partial l}{\partial r} &= -\sum_{k=0}^{\infty} n_k \ln(1+\beta) + \sum_{k=0}^{\infty} n_k \frac{\partial}{\partial r} \ln \frac{(r+k-1)\cdots r}{k!} \\ &= -n\ln(1+\beta) + \sum_{k=0}^{\infty} n_k \frac{\partial}{\partial r} \ln \prod_{m=0}^{k-1}(r+m) \\ &= -n\ln(1+\beta) + \sum_{k=0}^{\infty} n_k \frac{\partial}{\partial r} \sum_{m=0}^{k-1} \ln(r+m) \\ &= -n\ln(1+\beta) + \sum_{k=1}^{\infty} n_k \sum_{m=0}^{k-1} \frac{1}{r+m}.\end{aligned} \quad (7\text{-}1\text{-}12)$$

令上述两式等于 0 得到

$$\hat{\mu} = \hat{r}\hat{\beta} = \frac{\sum_{k=0}^{\infty} k n_k}{n} = \bar{x} \quad (7\text{-}1\text{-}13)$$

和

$$n\ln(1+\hat{\beta}) = \sum_{k=1}^{\infty} n_k \left(\sum_{m=0}^{k-1} \frac{1}{\hat{r}+m} \right) \quad (7\text{-}1\text{-}14)$$

从式（7-1-12）可看出，均值的极大似然估计值等于样本均值。将 $\hat{\beta}=\frac{\hat{\mu}}{\hat{r}}$ 代入式（7-1-4）得到方程

$$H(\hat{r})=n\ln(1+\frac{\overline{x}}{\hat{r}})-\sum_{k=1}^{\infty}n_k\left(\sum_{m=0}^{k-1}\frac{1}{\hat{r}+m}\right)=0$$ 。（7-1-15）

可采用 Newton-Raphson 法求解方程（7-1-15），迭代公式为

$$r_k=r_{k-1}-\frac{H(r_{k-1})}{H'(r_{k-1})}$$

一般选取迭代初始值 r_0 等于 r 的矩估计值。使用本例的数据可得

$$\hat{r}=10.965,\hat{\beta}=0.2280$$

```
#加载 MASS、survival、fitdistrplus 包
library(MASS)
library(survival)
library(fitdistrplus)
#理赔次数观察值
claim = c(0, 1, 2, 3, 4, 5, 6)
#样本数的观察值
freq = c(1, 2, 3, 2, 0, 1, 1)
#把理赔次数观察值整理为一个向量
num = rep(claim, freq)
#用泊松分布拟合数据，应用极大似然法估计参数
fit1 = fitdist(num, "pois", method = "mle")
#得到所估计的参数
coef(fit1)
##    lambda
##       2.5
```

```
#用负二项分布拟合数据，应用极大似然法估计参数
fit2 = fitdist(num, "nbinom", method = "mle")
#得到所估计的参数
coef(fit2)
#其中 size 表示的是参数 r，mu 表示的是均值，即 rβ
##       size         mu
##   10.96929    2.49980
```

EXAMPLE 14.9

Determine the maximum likelihood estimate of the Poisson parameter for the data Table 14.8.

Table 14.8 Automobile claims by year.

Year	Exposure	Claims
1986	2,145	207
1987	2,452	227
1988	3,112	341
1989	3,458	335
1990	3,698	362
1991	3,872	359

$$L = \prod_{k=1}^{6} \frac{e^{-\lambda e_k} (\lambda e_k)^{n_k}}{n_k!}.$$

The maximum likelihood estimate is found by

$$
\begin{aligned}
l &= \ln L = \sum_{k=1}^{6} [-\lambda e_k + n_k \ln(\lambda e_k) - \ln(n_k!)], \\
\frac{\partial l}{\partial \lambda} &= \sum_{k=1}^{6} \left(-e_k + n_k \lambda^{-1}\right) = 0, \\
\hat{\lambda} &= \frac{\sum_{k=1}^{6} n_k}{\sum_{k=1}^{6} e_k} = \frac{1{,}831}{18{,}737} = 0.09772.
\end{aligned}
$$

276. For a group of policies, you are given:

(i) Losses follow the distribution function

$$F(x) = 1 - \theta / x, \quad x > 0.$$

(ii) A sample of 20 losses resulted in the following:

Interval	Number of Losses
(0,10]	9
(10, 25]	6
(25,∞)	5

Calculate the maximum likelihood estimate of θ.

(A) 5.00

(B) 5.50

(C) 5.75

(D) 6.00

(E) 6.25

Question #276
Key: B

$$L(\theta)=\left(1-\frac{\theta}{10}\right)^{9}\left(\frac{\theta}{10}-\frac{\theta}{25}\right)^{6}\left(\frac{\theta}{25}\right)^{5}\propto(10-\theta)^{9}\theta^{11}$$

$$l(\theta)=9\ln(10-\theta)+11\ln(\theta)$$

$$l'(\theta)=-\frac{9}{10-\theta}+\frac{11}{\theta}=0$$

$$11(10-\theta)=9\theta$$

$$110=20\theta$$

$$\theta=110/20=5.5.$$

256. You are given:

(i) The distribution of the number of claims per policy during a one-year period for 10,000 insurance policies is:

Number of Claims per Policy	Number of Policies
0	5000
1	5000
2 or more	0

(ii) You fit a binomial model with parameters m and q using the method of maximum likelihood.

Calculate the maximum value of the loglikelihood function when $m = 2$.

$$L(q) = \left[\binom{2}{0}(1-q)^2\right]^{5000}\left[\binom{2}{1}q(1-q)\right]^{5000} = 2^{5000}q^{5000}(1-q)^{15000}$$

$$l(q) = 5000\ln(2) + 5000\ln(q) + 15000\ln(1-q)$$

$$l'(q) = 5000q^{-1} - 15000(1-q)^{-1} = 0$$

$$\hat{q} = 0.25$$

$$l(0.25) = 5000\ln(2) + 5000\ln(0.25) + 15000\ln(0.75) = -7780.97.$$

3、非完整数据

（1）只存在一个右删失censored数据

设 $x_1, x_2, ..., x_n$ 是随机变量 X 的 n 个观测，将它们从小到大排列 $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$。假设样本数据在 $x = u$ 处存在右删失，且存在 $1 \le k < n$ 使得 $x_{(k)} < u \le x_{(k+1)}$，则样本数据将变为 $x_{(1)}, x_{(2)}, \cdots, x_{(k)}, u, \cdots u$。似然函数可以写为

$$L(\theta) = [\prod_{i=1}^{k} f(x_{(i)}; \theta)]\mathrm{S}(u; \theta)^{n-k}$$

例8.7 You are given:

(i) Losses follow a Single-parameter Pareto distribution with density function:

$$f(x)=\frac{\alpha}{x^{(\alpha+1)}}, x>1, 0<\alpha<\infty$$

(ii) A random sample of size five produced three losses with values 3, 6 and 14, and two losses exceeding 25.

Determine the maximum likelihood estimate of α .

(A) 0.25
(B) 0.30
(C) 0.34
(D) 0.38
(E) 0.42

Key:A

解：

The distribution function is $F(x) = \int_1^x \alpha t^{-\alpha-1} dt = -t^{-\alpha}\Big|_1^x = 1 - x^{-\alpha}$. The likelihood function is

$$L = f(3)f(6)f(14)[1 - F(25)]^2$$
$$= \alpha 3^{-\alpha-1} \alpha 6^{-\alpha-1} \alpha 14^{-\alpha-1} (25^{-\alpha})^2$$
$$\propto \alpha^3 [3(6)(14)(625)]^{-\alpha}.$$

Taking logs, differentiating, setting equal to zero, and solving:

$$\ln L = 3\ln\alpha - \alpha \ln 157{,}500 \text{ plus a constant}$$
$$(\ln L)' = 3\alpha^{-1} - \ln 157{,}500 = 0$$
$$\hat{\alpha} = 3/\ln 157{,}500 = .2507.$$

例 8.8：假设 data set B（某劳工补偿险的原始损失数据）中的数据 censored

at 250, 假设总体分布为指数分布，求θ的极大似然估计。

27	82	115	126	155	161	243	294	340	384
457	680	855	877	974	1193	1340	1884	2558	15743

解：由于有 13 个数据大于 250，因此似然函数为

$$
\begin{aligned}
&f(27)f(82)\cdots f(243)P(X>250)^{13}\\
&=\theta^{-1}e^{-27/\theta}\theta^{-1}e^{-82/\theta}\cdots\theta^{-1}e^{-243/\theta}(e^{-250/\theta})^{13}\\
&=\theta^{-7}e^{-4159/\theta}
\end{aligned}
$$

$$
l(\theta)=-7\ln\theta-4159\theta^{-1}
$$

$$
l(\theta)=-7\theta^{-1}-4159\theta^{-2}=0
$$

$$
\theta=4159/7=594.14
$$

```
rm(list=ls())
censored=250 #删失值
```

```r
x<-c(27,82,115,126,155,161,243,294,340,384,457,680,855,877,9
74,1193,1340,1844,2558,15743) #原始损失数据
xx<-c(x[x<=censored])
n1<-length(x)
n2<-length(xx)
x_x<-c(xx,rep(250,n1-n2))

f<-function(x,x_x,n1,n2){
  a=0
  for(i in 1:n2){
    a=a-log(x)-x_x[i]/x
  }
  a=a-(n1-n2)*censored/x
  return(a)
}
xmax <- optimize(f,c(100,1000),tol = 0.0001, maximum = TRUE,x
```

```
_x,n1,n2)
xmax

## $maximum
## [1] 594.1429
##
## $objective
## [1] -51.70984
```

练习

218. The random variable X has survival function:

$$S_X(x) = \frac{\theta^4}{(\theta^2 + x^2)^2}$$

Two values of X are observed to be 2 and 4. One other value exceeds 4.

Calculate the maximum likelihood estimate of θ.

$$f(x) = -S'(x) = \frac{4x\theta^4}{(\theta^2 + x^2)^3}$$

$$L(\theta) = f(2)f(4)S(4) = \frac{4(2)\theta^4}{(\theta^2+2^2)^3}\frac{4(4)\theta^4}{(\theta^2+4^2)^3}\frac{\theta^4}{(\theta^2+4^2)^2} = \frac{128\theta^{12}}{(\theta^2+4)^3(\theta^2+16)^5}$$

$$l(\theta) = \ln 128 + 12\ln\theta - 3\ln(\theta^2+4) - 5\ln(\theta^2+16)$$

$$l'(\theta) = \frac{12}{\theta} - \frac{6\theta}{\theta^2+4} - \frac{10\theta}{\theta^2+16} = 0; 12(\theta^4+20\theta^2+64) - 6(\theta^4+16\theta^2) - 10(\theta^4+4\theta^2) = 0$$

$$0 = -4\theta^4 + 104\theta^2 + 768 = \theta^4 - 26\theta^2 - 192$$

$$\theta^2 = \frac{26 \pm \sqrt{26^2 + 4(192)}}{2} = 32; \theta = 5.657$$

（2）只存在一个 Truncated 点的数据

当数据存在左截断时，截断点 d 下方数据将不会被观测。例如假设

保单具有免赔额d，那么意味着小于d的实际损失额将不会被报告。截断数据有两种记录方式。如果数据只被截断，则截断后的观测值为

$$Y = X \mid X > d$$

称Y为在d点截断数据。若X表示实际损失额（死亡时间），则Y即为高于d的损失（死亡时间）。

另一种记录方式是记录超过截断点d的部分损失，这时截断后的观测值为

$$W = (X - d) \mid X > d$$

称W为截断且被平移数据。若d表示免赔额，X表示损失事件的实际损失额，则W就是每次理赔事件的理赔额。当实际损失小于免赔额d时，

被保险人没有获得理赔，理赔额不存在，因而没有定义。

设总体分布为$F(x,\theta)$， $x_1,\cdots,x_n$为没有被截断的观测值，假设样本数据在$x=d$处存在左截断，且存在$1\le k<n$使得$x_{(k-1)}\le d<x_{(k)}$。如果只记录d点截断数据，则样本数据将变为$x_{(k)},x_{(k+1)},\cdots,x_{(n)}$。如果记录的是截断且被平移的数据，则样本数据为$x_{(k)}-d,x_{(k+1)}-d,\cdots,x_{(n)}-d$，用$w_1,\cdots,w_{n-k}$来表示。

对于只存在 truncated 数据，有几种方法来估计。

（一） 平移法（*The shifted model*）

对于被截断且被平移的观测值，由于这些观测值都是从 0 开始被记录，从数据本身来看，就像没有被截断的数据。假设被截断且被平移的观测值 $(w_1, w_2 \ldots w_k)$ 具有与没有被截断的样本相同的分布类型，只是参数不同。那么就可以不考虑截断点对数据的影响直接使用分布进行估计。比如，若 d 表示免赔额，假设实际损失 X 的分布为 $F(x,\theta)$，若每次理赔事件的理赔额 W 的分布为 $F(x,\theta')$，则可以直接使用理赔额观测值对参数 θ' 进行估计。

对于只被截断的观测值，可以先对数据平移，即将每个观测点减去截断点，然后使用平移后分布进行估计。这时，似然函数为

$$L(\theta)=\prod_{i=k}^{n} f(x_i-d;\theta)=\prod_{i=1}^{n-k} f(w_i;\theta)$$

例 8.10 已知一组的工伤险赔付数据如下：35，72，120，135，165，144，243，302，332，378，465，664，854，858，986，1185，1332，1892，2567，15730；假设它在 200 处从下方被截断。使用平移法估计已知θ=800 的帕累托分布的参数α，并计算当免赔额为 0、200、400 时理赔事件理赔额的期望值。

解：因为数据在 200 处截断，所以被记录下来的样本数据只有 14 个。 使用数据平移法，得到截断且被平移的数据为

43　102　132　178　265　464　654　658　786　985　1132　1692　2367　15530

对于帕累托分布，$f(x)=\dfrac{\alpha\theta^{\alpha}}{(x+\theta)^{\alpha+1}}$，因此似然函数为

$$L(\alpha)=\prod_{j=1}^{14}\frac{\alpha(800^{\alpha})}{(800+w_j)^{\alpha+1}}$$

对数似然函数：

$$\begin{aligned}l(\alpha)&=\sum_{j=1}^{14}[\ln\alpha+\alpha\ln 800-(\alpha+1)\ln(w_j+800)]\\&=14\ln\alpha-10.353\alpha-103.938\end{aligned}$$

令$l'(\alpha)=14/\alpha-10.353=0$，得$\hat{\alpha}=1.352$.

下面计算不同免赔额对应的理赔额期望值.

（1）免赔额等于 0 因为数据被移动了，所以无法估计无免赔时的成本。

（2）当免赔额为 200 时，这些在 200 处被截断且被平移的数据可看作是每次理赔事件理赔额的样本数据，因此每次理赔事件的理赔额期望值为 $E(W)=\frac{\theta}{\alpha-1}$，即 800/0.352=2272.7。

（3） 当免赔额为 400 时，等价于在这些截断的数据上再次附加免赔额为 200 的条款，因此，每次理赔事件的理赔额的期望为

$$\frac{E(W)-E(W\wedge 200)}{1-F(200)}=\frac{\frac{800}{0.352}\left(\frac{800}{200+800}\right)^{0.352}}{\left(\frac{800}{200+800}\right)^{1.352}}=2840.85$$

例 8.11：假设 data set B（某劳工补偿险的原始损失数据）中的数据

存在免赔额 200，总体分布为 Paerto 分布，参数α未知，$\theta=800$，使用 shift 模型极大似然估计α，并计算当免赔额为 200、400 是理赔事件理赔额的期望值。

某劳工补偿险的原始损失数据

27	82	115	126	155	161	243	294	340	384
457	680	855	877	974	1193	1340	1884	2558	15743

解：但免赔额 200 存在时，原始数据将变为

43 94 140 184 257 480 655 677 774 993 1140 1684 2358 15543

似然函数为

$$L(\alpha)=\prod_{j=11}^{14}\frac{\alpha(800)^{\alpha}}{(800+x_j)^{\alpha+1}}$$

$$\begin{aligned}l(\alpha)&=\sum_{j=1}^{n}[\ln\alpha+\alpha\ln 800-(\alpha+1)\ln(x_j+800)]\\&=14\ln\alpha+93.5846\alpha-103.969(\alpha+1)\\&=14\ln\alpha-103.969-10.384\alpha\\l'(\alpha)&=14\alpha^{-1}-10.384\\\hat{\alpha}&=\frac{14}{10.384}=1.3482\end{aligned}$$

因此，当免赔额为 200 时，每次理赔事件的理赔额期望值为 800/0.3482=2298。

当免赔额为 400 时，等价于在这些 truncated 数据上再次附加免赔额为 200 的条款，因此，每次理赔事件的理赔额的期望为

$$\frac{E(X)-E(X\wedge 200)}{1-F(200)}=\frac{\frac{800}{0.3482}\left(\frac{800}{200+800}\right)^{0.3482}}{\left(\frac{800}{200+800}\right)^{1.3482}}=\frac{1000}{0.3482}=2872$$

```
rm(list = ls())
truncated=200
x<-c(27,82,115,126,155,161,243,294,340,384,457,680,855,877,9
74,1193,1340,1844,2558,15743) #原始损失数据
xx<-c(x[x>=truncated]-truncated)
n1=length(xx)
theta = 800

f<-function(x,x_x,n1,theta){
  a=0
```

```
  for(i in 1:n1){
    a=a+log(x)+x*log(theta)-(x+1)*log(xx[i]+theta)
  }
  return(a)
}

xmax <- optimize(f,c(0,20),tol = 0.0001, maximum = TRUE,x_x,n
1,theta)
xmax

## $maximum
## [1] 1.350301
##
## $objective
## [1] -113.748
```

```r
Ex=theta/(xmax$maximum-1)
Ex
```

```
## [1] 2283.75
```

（二）非平移法

如果不对数据进行平移，建立似然函数的一个问题是对小于截断点 d 的那些观测值设置概率。设 d 为截断点，Y 是被截断后 X 的值，。设 X 的分布函数为 $F(x)$， Y 的分布为

$$F_Y(x)=\begin{cases}0 & x\le d\\ P(X\le x\mid X>d) & x>d\end{cases}=\begin{cases}0 & x\le d\\ \dfrac{F(x)-F(d)}{1-F(d)} & x>d\end{cases}$$

若 X 是连续随机变量，则 Y 的密度函数为

$$f_Y(x)=\begin{cases}0, & x\le d\\ \dfrac{f(x)}{1-F(d)}, & x>d\end{cases}。$$

因此被截断后的观测值 $y_1,...,y_k$ 似然函数值为

$$\prod_{i=1}^{k}\frac{f(y_i;\theta)}{1-F(d;\theta)}.$$

设 W 为被截断且被平移后的值，类似的可推出 W 的分布为

$$F_W(x)=\begin{cases}0, & x\le 0\\ \dfrac{F(x+d)-F(d)}{1-F(d)}, & x>0\end{cases}$$

$$f_W(x)=\begin{cases}0 & x\le 0\\ \dfrac{f(x+d)}{1-F(d)} & x>0\end{cases} \qquad (0.0.6)$$

因此被截断且被平移后的观测值 $w_1,...,w_k$ 似然函数值为

$$L(\theta)=\prod_{i=1}^{k}\frac{f(w_i+d;\theta)}{1-F(d;\theta)}$$

例 8.11（续）：假设 data set B（某劳工补偿险的原始损失数据）中的数据存在免赔额 200，总体分布为 Paerto 分布，参数α未知，$\theta=800$，使用 shift 模型极大似然估计α，使用 unshift 模型来计算当免赔额为 200、400 是理赔事件理赔额的期望值。

解：假设总体分布为 Paerto 分布，当免赔额 200 存在时，没有 truncated 的数据的条件分布为

$$f_X(x \mid X > 200) = \frac{f(x)}{1 - F(200)}$$

似然函数为

$$L(\alpha) = \prod_{j=1}^{14} \frac{f(x_j \mid \alpha)}{1 - F(200 \mid \alpha)} = \prod_{j=1}^{14} \left[\frac{\alpha(800^\alpha)}{(800 + x_j)^{\alpha+1}} \Big/ \left(\frac{800}{800 + 200} \right)^\alpha \right]$$

$$= \prod_{j=1}^{14} \frac{\alpha(1000^\alpha)}{(800 + x_j)^{\alpha+1}}$$

$$l(\alpha) = 14 \ln \alpha + 14\alpha \ln 1000 - (\alpha + 1) \sum_{j=1}^{14} \ln(800 + x_j)$$

$$= 14 \ln \alpha + 96.709\alpha - (\alpha + 1)105.810$$

$$l'(\alpha) = 14\alpha^{-1} - 9.101 \Rightarrow \hat{\alpha} = 1.5383$$

例8.12 You observe the following five ground-up claims from a data set that is truncated from below at 100:

125 150 165 175 250

You fit a ground-up exponential distribution using maximum likelihood estimation.

Determine the mean of the fitted distribution.

(A) 73

(B) 100

(C) 125

(D) 156

(E) 173

Key:A

Because the exponential distribution is memoryless, the excess over the deductible is also exponential with the same parameter. So subtracting 100 from each observation yields data from an exponential distribution and noting that the maximum likelihood estimate is the sample mean gives the answer of 73.

Working from first principles,

$$L(\theta) = \frac{f(x_1)f(x_2)f(x_3)f(x_4)f(x_5)}{[1-F(100)]^5} = \frac{\theta^{-1}e^{-125/\theta}\theta^{-1}e^{-150/\theta}\theta^{-1}e^{-165/\theta}\theta^{-1}e^{-175/\theta}\theta^{-1}e^{-250/\theta}}{(e^{-100/\theta})^5}$$
$$= \theta^{-5}e^{-365/\theta}.$$

Taking logarithms and then a derivative gives

$$l(\theta) = -5\ln(\theta) - 365/\theta, l'(\theta) = -5/\theta + 365/\theta^2 = 0.$$

The solution is $\hat{\theta} = 365/5 = 73$.

（3）同时存在多个右删失点和左截断点的数据

在大多数情况下，右删失和左截断是同时存在的，而且删失点和截断点不一致。例如保险责任包括免赔和赔偿限额的保单理赔额。又或，在特定观察期内，对某个观察对象的观测会分很多种不同情况；从初期开始观测，从期中开始观测，对象生存到观察期末，对象在观察期内死亡。对各种情况下的似然函数，需要分情况来考虑：

（1） 如果观测值 x_i 没有被截断或删失，则似然函数值为 $f(x_i)$；

（2） 如果观测值 x_i 是在 d_i 点被截断的，则似然函数值为 $f(x_i)/(1-F(d_i))$；

（3） 如果观测值 x_i 是在 d_i 点被截断且被平移的，则似然函数值为

$f(x_i+d_i)/(1-F(d_i))$；

（4） 如果观测值 x_i 是在 u_i 点被删失的，则似然函数值为 $1-F(u_i)$；

（5） 如果观测值 x_i 是在 d_i 点被截断且在 u_i 点删失的，则似然函数值为 $\dfrac{1-F(u_i)}{1-F(d_i)}$；

注意到 $S(x)=1-F(x)$， $f(x)=h(x)S(x)$， $h(x)$ 是危险率函数。因此对于没有被平移的数据，可以用一个公式来表示似然函数：

$$L=\prod_{i=1}^{n}\frac{S(x_i)\left[h(x_i)\right]^{\delta_i}}{S(d_i)}，\quad \delta_i=\begin{cases}1 & \text{未删失}\\ 0 & \text{在}x_i\text{删失}\end{cases} \qquad (*)$$

其中，若观测值 x_i 是在 u_i 被删失，则记 $x_i = u_i$。

例 8.13 已知观测到如下 20 例损失额（原始值）如下表：

基本数据

损失额	观测数	免赔额	保单限额
750	3	200	∞
200	3	0	10000
300	4	0	20000

>10000	6	0	10000
400	4	300	∞

过去经验显示，损失额服从参数为α和θ的帕累托分布，且已知参数θ=10000，用极大似然法估计α。

解: 帕累托分布的密度函数为 $f(x)=\dfrac{\alpha\theta^{\alpha}}{(\theta+x)^{\alpha+1}}$

所以可以写出似然函数：

$$L=[\frac{f(750)}{1-F(200)}]^3 f(200)^3 f(300)^4[1-F(10000)]^6[\frac{f(400)}{1-F(300)}]^4$$
$$=[\frac{\alpha 10200^{\alpha}}{10750^{\alpha+1}}]^3[\frac{\alpha 10000^{\alpha}}{10200^{\alpha+1}}]^3[\frac{\alpha 10000^{\alpha}}{10300^{\alpha+1}}]^4[\frac{10000^{\alpha}}{20000^{\alpha}}]^6[\frac{\alpha 10300^{\alpha}}{10400^{\alpha+1}}]^4$$

$$\propto \alpha^{14}10000^{13\alpha}10750^{-3\alpha}20000^{-6\alpha}10400^{-4\alpha}.$$

于是得到对数似然函数

$$\ln L=14\ln\alpha+13\alpha\ln(10000)-3\alpha\ln(10750)-6\alpha\ln(20000)-4\alpha\ln(10400)=14$$

对 α 求导令导数等于零，解得 $\hat{\alpha}=3.089$。 □

例 8.14 考察以下 6 个接受人工心脏的病人的样本。其中 4 人在 2006 年 12 月 31 日之前死亡（见表），观察期为日历年 2006 年。假设人工移植心脏病人的存活时间服从指数分布，由所给样本数据估计指数分布参数。

接受人工心脏的病人的生存情况

病人	移植时期	死亡时间
1	2005.1.1	2006.4.1
2	2005.4.1	2006.4.1
3	2005.7.1	-
4	2005.10.1	2006.7.1

5	2006.1.1	-
6	2006.4.1	2006.10.1

解：由于观察期是日历年 2006，在此之前进行心脏移植手术的病人，在进入观察期时已经存活了一段时间，因此数据是被截断数据。而观察期结束还存活的病人，其死亡时间未知，因此是删失数据。根据上述分析，这 6 个样本的截断点和真实值（删失点）为：

$d_1=1, d_2=0.75, d_3=0.5, d_4=0.25, d_5=0, d_6=0$ ，

$x_1=1.25, x_2=1, x_3=1.5+, x_4=0.75, x_5=1+, x_6=0.5$，其中 x_3, x_5 是删失点。

则根据式（*）

$$L=\prod_{i=1}^{6}\frac{e^{-x_i/\theta}(\frac{1}{\theta})^{\delta_i}}{e^{-d_i/\theta}}$$

$$\ln L=\sum_{i=1}^{6}[-x_i/\theta+d_i/\theta+\delta_i\ln(1/\theta)]$$

令$\frac{d\ln L}{d\theta}=0$，得到$\hat{\theta}=\sum_{i=1}^{6}(x_i-d_i)/\sum_{i=1}^{6}\delta_i=3.5/4=0.875$

262. You are given:

(i) At time 4 hours, there are 5 working light bulbs.

(ii) The 5 bulbs are observed for p more hours.

(iii) Three light bulbs burn out at times 5, 9, and 13 hours, while the remaining light bulbs are still working at time $4 + p$ hours.

(iv) The distribution of failure times is uniform on $(0, \omega)$.

(v) The maximum likelihood estimate of ω is 29.

Calculate p.

(A) Less than 10

(B) At least 10, but less than 12

(C) At least 12, but less than 14

(D) At least 14, but less than 16

(E) At least 16

Question #262
Key: D

$$L(\omega)=\frac{\frac{1}{\omega}\frac{1}{\omega}\frac{1}{\omega}\left(\frac{\omega-4-p}{\omega}\right)^2}{\left(\frac{\omega-4}{\omega}\right)^5}=\frac{(\omega-4-p)^2}{(\omega-4)^5}$$

$$l(\omega)=2\ln(\omega-4-p)-5\ln(\omega-4)$$

$$l'(\omega)=\frac{2}{\omega-4-p}-\frac{5}{\omega-4}=0$$

$$0=l'(29)=\frac{2}{25-p}-\frac{5}{25}$$

$p=15.$

The denominator in the likelihood function is $S(4)$ to the power of five to reflect the fact that it is known that each observation is greater than 4.