First homework assignment

Professional Master in Economics
Instructor: Hedibert Freitas Lopes
Insper Institute of Education and Research

Advanced Econometrics
TA: Guilherme Piantino
October 2024

Count and time-to-event datasets

Dataset A – count data: The data below represents the yearly counts of coal mining disasters (y_i) in Great Britain from 1851 to 1962, i.e. for $i = 1, \ldots, n$ and $n = 112$ observations.

```
y = c(4,5,4,1,0,4,3,4,0,6,3,3,4,0,2,6,3,3,5,4,5,3,1,4,4,1,5,5,3,4,2,5,2,2,3,4,2,1,3,2,2,
      1,1,1,1,3,0,0,1,0,1,1,0,0,3,1,0,3,2,2,0,1,1,1,0,1,0,1,0,0,0,2,1,0,0,0,1,1,0,2,3,3,
      1,1,2,1,1,1,1,2,4,2,0,0,0,1,4,0,0,0,1,0,0,0,0,0,1,0,0,1,0,1)
years = 1851:1962
n = length(y)
plot(years,y,ylab="count of disasters",type="b")
```

This data was analyzed, for example, by Carlin, Gelfand and Smith (1992) in their paper entitled *Hierarchical Bayesian Analysis of Changepoint Problems* that appeared in the prestigious statistical journal *Applied Statistics*, volume 41, number 2 and pages 389-405 (`https://www.jstor.org/stable/2347570`). They make the following statement: "A much analysed data set of intervals between British coal-mining disasters during the 112-year period 1851-1962 was gathered by Maguire et al. (1952), extended and corrected by Jarrett (1979). Frequentist changepoint investigations appear in Worsley (1986) and in Siegmund (1988) while Raftery and Akman (1986) apply their Bayesian model." For completion, all additional references appear at the end of this document.

Dataset B – time-to-event data:

The data below represent the time intervals in days between explosions in mines, involving 10 or more men killed, from 15 March 1851 to 22 March 1962.

```
days = c(157,123,2,124,12,4,10,216,80,12,33,66,232,826,40,12,29,190,97,65,186,23,92,197,
431,16,154,95,25,19,78,202,36,110,276,16,88,225,53,17,538,187,34,101,41,139,42,1,250,80,
3,324,56,31,96,70,41,93,24,91,143,16,27,144,45,6,208,29,112,43,193,134,420,95,125,34,127,
218,2,378,36,15,31,215,11,137,4,15,72,96,124,50,120,203,176,55,93,59,315,59,61,1,13,189,
345,20,81,286,114,108,188,233,28,22,61,78,99,326,275,54,217,113,32,388,151,361,312,354,
307,275,78,17,1205,644,467,871,48,123,456,498,49,131,182,255,194,224,566,462,228,806,
517,1643,54,326,1312,348,745,217,120,275,20,66,292,4,368,307,336,19,329,330,312,536,
145,76,364,37,19,156,47,129,1630,29,217,7,18,1358,2366,952,632)
```

Working questions to discuss with TA during the office hours

Assuming the sample from dataset A follows a Poisson model, i.e. $y_1, \ldots, y_n$ are conditionally independent and identically distribution Poisson(λ), denoted by

$$y_1, \ldots, y_n | \lambda \sim Poi(\lambda).$$

a) Obtain $\widehat{\lambda}_{mle}$, the maximum likelihood estimator (MLE) of λ. Recall that

$$\begin{aligned} p(y_i|\lambda) &= \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}, \\ \log p(y_i|\lambda) &= y_1 \log(\lambda) - \lambda - \log(y_i!), \\ l(\lambda|y_1, \ldots, y_n) &= Log\left\{\prod_{i=1}^{n} p(y_i|\lambda)\right\} = \sum_{i=1}^{n} \log p(y_i|\lambda) \\ &= \log(\lambda) n\bar{y} - n\lambda - \sum_{i=1}^{n} \log y_i! \end{aligned}$$

for $\lambda > 0$, $y_i \in \{0, 1, 2, \ldots\}$, $i = 1, \ldots, n$ and $\bar{y} = \sum_{i=1}^{n} y_i/n$.

b) We know that when $Y \sim Poi(\lambda)$, it follows that $E(Y) = V(Y) = \lambda$. Is $\widehat{\lambda}_{mle}$ close to the sample variance of the ys?

c) Looking at the data more closely (draw a time series plot and you will see), it seems that around 1897 (observation $i = 47$) there is a break in the count of disasters. If that is correct, then we obtain MLEs for λ before and after 1987. We could repeat this for, say, years 1890 to 1905, and compare the MLEs before and after in order to grasp more or less where this potential break may actually occur.

d) In c), instead of comparing MLEs, we can compare log-likelihoods. More precisely, we can compute

$$S_k = l(\widehat{\lambda}_{mle,1:k}|y_1, \ldots, y_k) + l(\widehat{\lambda}_{mle,(k+1):n}|y_{k+1}, \ldots, y_n),$$

where $\widehat{\lambda}_{mle,a:b}$ is the MLE of λ based on data $\{y_a, \ldots, y_b\}$. We can, say, make $k = 40, \ldots, 55$. One possible estimator of k could be

$$\widehat{k}_{mle} \equiv arg \max_{k \in \{40:55\}} S_k.$$

Plotting k against S_k also helps.

Solution

```
ks = 20:90
lambda.mle = NULL
S = NULL
for (k in ks){
  lambdas = c(mean(y[1:k]),mean(y[(k+1):n]))
  lambda.mle = rbind(lambda.mle,lambdas)
  S = c(S,sum(dpois(y[1:k],lambdas[1],log=TRUE)) + sum(dpois(y[(k+1):n],lambdas[2],log=TRUE)))
}
kmax = ks[S==max(S)]
yearmax = years[kmax]
lambda1 = lambda.mle[ks==kmax,1]
lambda2 = lambda.mle[ks==kmax,2]

par(mfrow=c(1,2))
plot(years,y,ylab="count of disasters",type="b")
abline(v=yearmax,col=2)
plot(years[ks],S,xlim=range(years))
title(paste("Year with highest likelihood = ",yearmax,sep=""))
abline(v=yearmax,col=2)
legend("topright",legend=c(paste("lambda (left) = ",round(lambda1,3),sep=""),
paste("lambda (right) = ",round(lambda2,3),sep="")),bty="n")
```

Homework questions

Repeat the above analysis for the time-to-event dataset (dataset B), which obviously cannot be modeled as Poisson. For time-to-event data, we usually assume the Exponential model, or the Gamma model, or the Weibul model, to name a few. To keep it simpler, let us assume that $y_1, \ldots, y_n$ are conditionally independent and identically distribution Exponential(λ), denoted by

$$y_1, \ldots, y_n | \lambda \sim Exp(\lambda),$$

where $p(y_i|\lambda) = \lambda e^{-\lambda y_i}$, with $E(y_i|\lambda) = 1/\lambda$ and $V(y_i|\lambda) = 1/\lambda^2$. It is very easy to see that

$$\begin{aligned} p(y_1, \ldots, y_n|\lambda) &= \lambda^n e^{-\lambda \sum_{i=1}^n y_i} \\ logp(y_1, \ldots, y_n|\lambda) &= nlog(\lambda) - \lambda \sum_{i=1}^n y_i, \end{aligned}$$

such that $\widehat{\lambda} = n/\sum_{i=1}^n y_i = 1/\bar{y}$.

Additional references

1. Jarrett, R. G. (1979) A note on the intervals between coal-mining disasters. *Biometrika*, 66, 191-193.

2. Maguire, B. A., Pearson, E. S. and Wynn, A. H. A. (1952) The time intervals between industrial accidents. *Biometrika*, 38, 168-180.

3. Raftery, A. E. and Akman, V. E. (1986) Bayesian analysis of a Poisson process with a change-point. *Biometrika*, 73, 85-89.

4. Siegmund, D. (1986) Boundary crossing probabilities and statistical applications. *Annals of Statistics*, 14, 361-404.

5. Worsley, K. J. (1986) Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika*, 73, 91-104.