Modeling web browsing at the user level.

How well does the Inverse Gaussian model fit?

Computer Scientists are being called more and more frequently to provide computer log data that can be used to find out how users interact with the Internet (see, for example, Sen and Hansen (2003), Cadez et al. (forthcoming), Huberman et al (1998)). In addition to that, the number of Internet user surveys is growing at an amazing speed (see, for example, Wellman and Haythornthwaite (2002)). The following problem uses a data set published in the UCI KDD Archive (Kederman, 2003) to obtain data on the number of different pages visited by users who entered the msnbc.com page on September 28, 1999 and other information. The variables we analyze in this activity are usually investigated in the references provided above.

1 Data Description

The data comes from Internet Information Server (IIS) logs for msnbc.com and news-related portions of msn.com for the entire day of September 26, 1999 (Pacific Standard Time). Each sequence in the dataset corresponds to page views of a user during that twenty-four hours period. Each event in the sequence corresponds to a user's request for a page. Requests are not recorded at the finest level of detail–that is, at the level of URL, but rather, they are recorded at the levels of page category (as determined by a site administrator). The categories are "frontpage", "news", "tech", "local", "opinion", "on-air", "misc", "weather", "health", "living", "business", "sports", "summary", "bbs (bulletin board service)", "travel", "msn-news", and "msn-sports". The variable "depth" shows how many different sites are visited by users, and the variable "length" represents the actual total number of pages visited by each user (total number of clicks).

Data URL

`http://kdd.ics.uci.edu/databases/msnbc/msnbc.data.html`

Number of users: $n = 989818$

2 Activity

Huberman et al.(1998) derive the probability $P(L)$ of the number of links L that a user follows in a web site. They determined that this distribution is given asymptotically by the two-parameter inverse Gaussian Distribution

$$P(L) = \sqrt{\frac{\lambda}{2\pi L^3}} \exp\left[\frac{-\lambda(L-\mu)^2}{2\mu^2 L}\right] \tag{1}$$

with mean $E(L) = \mu$ and variance $Var(L) = \mu^3/\lambda$, where λ is a scale parameter. This distribution "has a very long tail, which extends much farther than that of a normal distribution with comparable mean and variance. This implies a finite probability for events that would be unlikely if described by a normal distribution. Consequently, large deviations from the average number of user-clicks computed at a site will be observed." Another property is that "because of the asymmetry of the distribution function, the typical behavior of users will not be the same as their average behavior. Thus, because the mode is lower than the mean, care must be exercised with available data on the average number of clicks, as this average overestimates the typical depth being surfed."

It can be shown that the cumulative distribution function of the inverse Gaussian distribution is

$$F(L,\mu,\lambda) = \Phi\left[\sqrt{\frac{\lambda}{L}}\left(\frac{L}{\mu}-1\right)\right] + e^{2\lambda/\mu}\Phi\left[\sqrt{\frac{\lambda}{L}}\left(\frac{L}{\mu}+1\right)\right] \tag{2}$$

where $\Phi[\,]$ is the standard normal distribution function.

Question 1. Find theoretically the equations for the maximum likelihood estimator (MLE) of λ and μ in the inverse Gaussian distribution given above. Show all the steps of your computations.

Question 2. Open R in the directory where the data are and then read the data into R using the following command. Do a histogram to get a feeling for the distribution of the length of visits. What is the message in the histogram? What do the summary statistics indicate? Find the mle and determine the goodness of fit of the inverse gaussian to this data by looking at the goodness of fit statistic.

```
library(dse)       # We need special packages
library(SuppDists)  # We need special packages

msnbc = read.table("http://www.stat.ucla.edu/~jsanchez/teaching/stat258
/reading/msnbclength.txt",nrow=100000,header=T)
length = msnbc[,2]  # the length variable is the second column
n=length(msnbc[,2])  #sample size
hist(length)   # get a feeling for the distribution and summaries
summary(length)
```

Remember that if your picture gets masked by very large values, you might want to look at the picture only for smaller values, for instance, length smaller than 500. You can do that by putting that condition. For example, hist(length[length< 100)

Find the empirical MLEs derived from these data. You will need the formulas you derived in question 1 to present your results and know what the code is doing. You may find these R commands useful.

```
mu=mean(length)                   ## mu
temp=(((length/mu)-1)^2)/length
lambda=n/sum(temp)               ##lambda
lengthsum=sum(length)
lengthtable=table(length)         ## frequency of length
lengthfrequency=table(length)/n   ## relative frequency of length, empirical pdf
cumlength=cumsum(lengthfrequency)   ##cumulative distribution of the data
```

Find the estimates directly by asking the computer to do MLE estimation for you with a numerical routine..

```
library(stats4)   # need this library to do mle

fn=function(mu,lambda)
{
 -(   (n/2) *log(lambda)-(n/2)*log(2*pi) -(3/2)*sum(log(length))-((lambda)/2)*sum((((length/mu)
 }

est=mle(minuslogl=fn,start=list(mu=0.1,lambda=0.1))
summary(est)
```

To see how good is this model, Huberman et al.(1998) and Sen and Hansen(2003) compare the cumulative distribution function implied by the model to the empirical cumulative distribution function derived from the data. Then they use a quantile-quantile against the fitted distribution. We will do this here.

Question 3. Find the cumulative distribution of the data and the cumulative distribution implied by the model just estimated and superimpose them. Do they match each other? Do also a q-q plot and determine how good is the fit based on that. The following R commands may be useful. Be selective in choosing them and in choosing the output to answer this question.

```
##CDF plot for the model using estimated inverse Gaussian Model

l=seq(1, max(length))
```

```
density=dinvGauss(l, mu, lambda)

prob=pinvGauss(l, mu, lambda)   ##calculate the cdf  of model

plot(l, prob, type="o", xlab="length", ylab="CDF(model)")

dev.print(dev=pdf, file="CDF(data).pdf")   ##plot of CDF for model



##plot of CDF for model for length<=100
l=seq(1, 100)

density=dinvGauss(l, mu, lambda)     ##density

prob=pinvGauss(l, mu, lambda)        ##calculate the density of length

 data implied by the estimated  model

plot(l, prob, type="o", xlab="length", ylab="CDF(model)")

dev.print(dev=pdf, file="CDF(model).pdf")   ##plot of CDF for model


##plot of empirical CDF for the data
plot(l, cumlength[1:100], type="o", xlab="length", ylab="CDF(empirical)")

dev.print(dev=pdf, file="CDF(empirical).pdf")  ##plot of empirical CDF


##put CDF on same graph to check the fit

click=l

CDF=cumlength[1:100]

 plot(click, CDF, type="p")                ##point for empirical data

lines(click, prob)                               ##lines for inverse Gaussian

dev.print(dev=pdf, file="CDF plot.pdf")


##overlapping the data(Q-Q plot)
```

```
plot(prob, cumlength[1:100], xlab="CDF(model)", ylab="CDF(empirical)")

lines(1/100, 1/100)

dev.print(dev=pdf, file="CDF(overlapping).pdf")

qqplot(prob, cumlength[1:100])   ##same as above


fit=lm(cumlength[1:100]~prob)
```

If you take logs in both sides of the inverse gaussian formula you obtain an expression such that a plot of $\log(L)$ vs $\log(L)$ shows a straigt line whose slope approximates 3/2 for small values of L and large values of the variance. This could be another way of determining whether the model is good for these data.

With the data that they used, Sen and Hansen did not find much support for the inverse Gaussian model. They used the tools you used in question 3, and those you will use in question 4. It could be that the model is good for the msnbc data set, so it is worth trying further and see what conclusions we reach.

Question 4. Do a plot of the frequency distribution of surfing clicks on the log-log scale and fit a least squares regression line to these data. Do the results confirm the theoretical implication of a slope of 3/2? What do you conclude about the model? You may want to use the following commands.

To answer this question, it may help if you read the data set “frequency”, which contains the frequency table that you obtained earlier (the one you called ”lengthfrequency” in question 2. Type the data address

```
freqtable = read.table("http://www.stat.ucla.edu/~jsanchez/oid03/datasets/
frequencytable.txt",header=TRUE)
length=freqtable$length
frequency=freqtable$frequency
reg=lm(log(frequency)~log(length))  #regression to see if we get -3/2 slope
reg
plot(log(length),log(frequency))  #will do a plot on the log-log scale.
abline(reg)
```

Question 5. Based on your results above, would you recommend the inverse Gaussian model for the length of visits (or number of links that a user visits) in the msnbc.com data set? Why?

References

(1) Cadez, I., Heckerman, D., Meek, C., Smyth, P., White, S. (2003). Model-based clustering and visualization of navigation patterns on a Web site. Accepted for publication on Journal of Data Mining and Knowledge Discovery, 7(4).

(2) Hansen, M.H. and Sen, R.(2003). Predicting Web User's next access based on log data. Journal of Computational and Graphical Statistics, Vol 12, No. 1, March, p. 143-155.

(3) Heckerman, D. The UCI KDD Archive

`http://kdd.ics.uci.edu`

Irvine, CA: University of California, Department of Information and Computer Science.

(4) Huberman, B.A., Pirolli, P.L.T., Pitkow, J.E., Lukose, R.M. (1998). Strong Regularities in World Wide Web Surfing. Science, Vol. 280, 3 April.

(5) Wellman, B. and Haythornthwaite edts.(2002). The Internet in Everyday Life. Blackwell Publishing.