Solutions Chapter 4 - Likelihoods

Alexandra Hotti

October 2019

4.1 Blog blues

Suppose that visits to your newly launched blog occur sporadically. Imagine you are interested in the length of time between consecutive first-time visits to your homepage. You collect the time data for a random sample of 50 visits to your blog for a particular time period and day, and you decide to build a statistical model to fit the data.

Assignment 4.1.1:

What assumptions might you make about the first-time visits?

Answer:

- Each first time visit is independent of the other.
- The visit rate is constant such that the time between every visit is the same.

Assignment 4.1.2:

What might be an appropriate probability model for the time between visits?

Answer: The time between the visits should be exponentially distributed. First, since we have continuous, non-negative data. Second, since the visits are first-time visits they should be independent. However, it is hard to say whether the rate between the visits could be assumed to be constant. There are probably factors that impact the number of visitors at certain times.

Assignment 4.1.3:

Using your chosen probability distribution from the previous part, algebraically derive the maximum likelihood estimate (MLE) of the mean.

Answer: We start by defining the setting:

i = 1,..., T

$t_i \sim exp(\lambda)$

$p(t|\lambda) = \lambda \cdot e^{-\lambda x}$

This gives us the following likelihood (assuming iid) :

$$p(t_1, .., t_T|\lambda) = p(t_1|\lambda) \cdot p(t_2|\lambda) \cdot ... \cdot p(t_T|\lambda) = \prod_{i=1}^{T} \lambda \cdot e^{-\lambda t_i} \tag{1}$$

Since this is a monotonically increasing function we can differentiate the logged function instead.

$$log(p(t_1, .., t_N|\lambda)) = log(p(t_1|\lambda)) + \cdot ... \cdot + log(p(t_N|\lambda)) = \sum_{i=1}^{N} log(\lambda \cdot e^{-\lambda t_i}) = \sum_{i=1}^{N} (log(\lambda) - \lambda t_i) =$$

$$N \cdot log(\lambda) - N \cdot \lambda \bar{T} \tag{2}$$

Where $\bar{T}$ is the average time between first time visits.Now we differentiate (2) w.r.t. λ.

$$\frac{\partial log(p)}{\partial \lambda} = \frac{N}{\lambda} - N \cdot \lambda \bar{T} \tag{3}$$

We set (3) equal to zero to find the extreme point, i.e. the maximum likelihood estimate for λ.

$$\frac{N}{\hat{\lambda}} - N \cdot \bar{T} = 0$$

$$\hat{\lambda} = \frac{1}{\bar{\bar{T}}} \tag{4}$$

So, the mean first time vist rate is equivalent to the average time between first time visit.

Assignment 4.1.4

You collect data from Google Analytics that contains the time (in minutes) between each visit for a sample of 50 randomly chosen visits to your blog. The data set is called *likelihood_blogVisits.csv*. Derive an estimate for the mean number of visits per minute.

Answer: Using (4) we get $\hat{\lambda} = 1.6262145862718507 \approx 1.63$

Assignment 4.1.5

Graph the log-likelihood near the MLE. Why do we not plot the likelihood?

Answer: We do not plot the likelihoods since these values would be really small and have small discrepancy between each other.

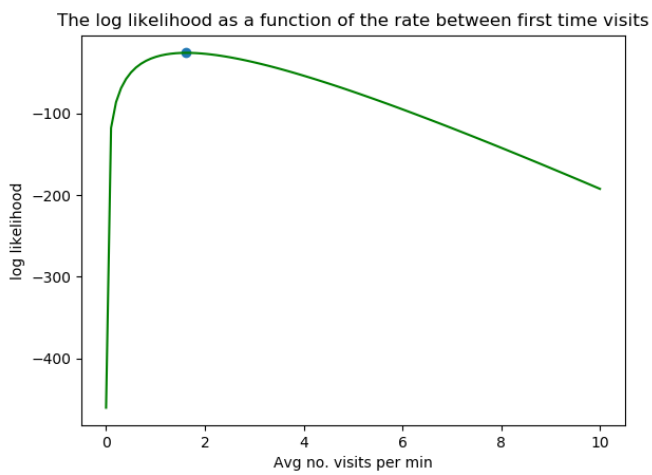

Figure 1: The log-likelihood near the MLE estimate.

Assignment 4.1.6

Estimate 95% confidence intervals around your estimate of the mean visit rate.

Answer:

The MLE estimate $\hat{\lambda}$ is approximately normally distributed about λ in large samples. This property is called the "asymptotic normality of the MLE," and the technique of forming confidence intervals is called the "asymptotic normal approximation." This method works for a wide variety of statistical models. [**?**].

Thus, the asymptotic 95% confidence interval for a parameter λ has the following form:

$$\hat{\lambda} \pm 1.96 \frac{1}{\sqrt{-l''(\hat{\lambda}; x)}} \tag{5}$$

where $-l''(\hat{\lambda}; x)$ is the second derivative of the log likelihood function with respect to λ, evaluated at $\hat{\lambda}$.

$-l''(\hat{\lambda}; x)$ is called the "observed information, " and $\frac{1}{\sqrt{-l''(\hat{\lambda};x)}}$ is an approximate standard error for $\hat{\lambda}$. As the log-likelihood becomes more sharply peaked about the MLE, the second derivative drops and the standard error goes down When calculating asymptotic confidence intervals, statisticians often replace the second derivative of the loglikelihood by its expectation; such that:

$$I(\lambda) = -E[-l''(\hat{\lambda}; x)] \tag{6}$$

which is called the expected information or the Fisher information. Now, we apply this to our problem. First we calculate:

$$I(\lambda) = -E[\frac{\partial^2 log(p)}{\partial \lambda^2}] = \frac{T}{\lambda^2} \tag{7}$$

To convert this to a confidence interval asymptotically we get due to Cramer-Rao lower bound assuming that the MLE parameter is unbiased. The central limit theorem gives us (for the error) :

$$\sqrt{T}(\hat{\lambda} - \lambda) \xrightarrow{D} N(0, I(\hat{\lambda})^{\frac{-1}{2}})$$

where $Var(\hat{\lambda}) = I^{-1/2}(\lambda) = E[-l''(\hat{\lambda}; x)]^{-1/2}$ since we are more certain of the MLE estimate when the second derivative is higher, i.e. we get a smaller variance. This gives us that:

$$\hat{\lambda} \xrightarrow{approx} N(\lambda, \frac{I(\hat{\lambda})^{\frac{-1}{2}}}{\sqrt{T}}) \tag{8}$$

Using (7) gives us:

$$\frac{I(\hat{\lambda})^{\frac{-1}{2}}}{\sqrt{T}}) = \frac{\hat{\lambda}}{T} \approx \frac{1.63}{50} = 0.0326 \tag{9}$$

Now, using the above information we calculate the 95% confidence interval as:

$$\hat{\lambda} \pm 1.96\sigma = 1.6262145862718507 \pm 1.96 \cdot 0.0326 \tag{10}$$

Which gives us:

$$1.56231858627 < \hat{\lambda} < 1.69011058627 \tag{11}$$

This is quite a small interval. This is good since a tight interval implies that we are quite certain about where the MLE estimate is located!

Assignment 4.1.7

What does this interval mean?

Answer: The interval means that the confidence intervals for 95 % of our samples will cover the true λ parameter, while the other 5 % will miss the true value. Since we have a small interval above, one can deduce that we are quite certain about where the true λ parameter lies.

Assignment 4.1.8

Using your maximum likelihood estimate, what is the probability you will wait: (a) 1 minute or more, (b) 5 minutes or more, (c) half an hour or more before your next visit?

Answer:

$$p(t \geq 1|\hat{\lambda} = 1.63) = \int_1^\infty 1.63e^{-1.63t}dt = [-e^{-1.63t}]_1^\infty = 0 + \frac{1}{e^{1.63}} = 0.19592957412 \approx 19.6\% \tag{12}$$

$$p(t \geq 5|\hat{\lambda} = 1.63) = \int_5^\infty 1.63e^{-1.63t}dt = [-e^{-1.63t}]_5^\infty = 0 + \frac{1}{e^{1.63 \cdot 5}} = 0.00028873535 \approx 0.0289\% \tag{13}$$

$$p(t \geq 30|\hat{\lambda} = 1.63) = \int_{30}^\infty 1.63e^{-1.63t}dt = [-e^{-1.63t}]_{30}^\infty = 0 + \frac{1}{e^{1.63 \cdot 5}} = 5.7942848e - 22 \approx 0\% \tag{14}$$

Assignment 4.1.9

Evaluate your model.

Answer: One way to evaluate the model is to generate data and then compare the generated data with the data used to train the model.

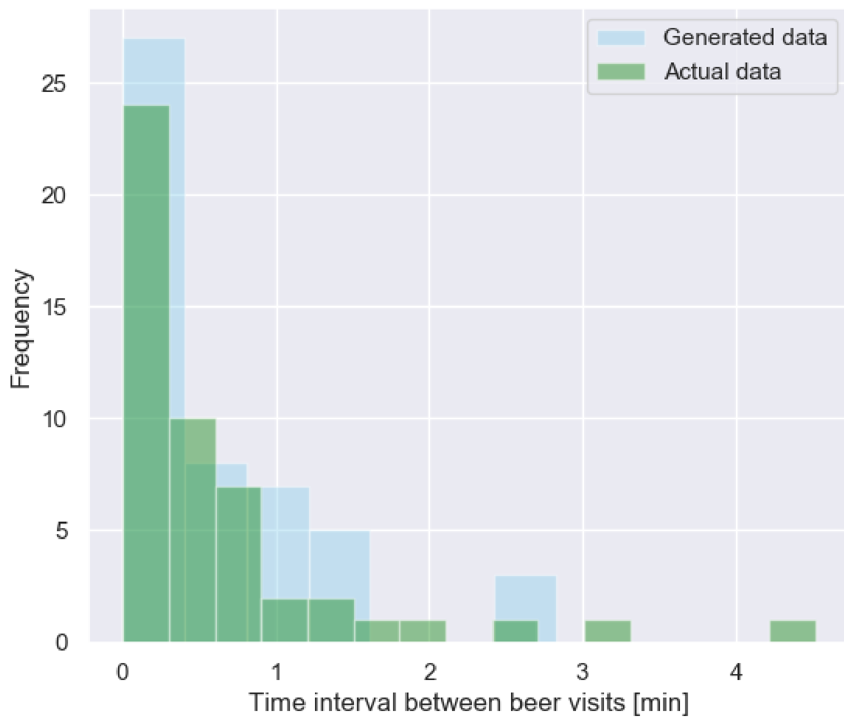

Figure 2: Comparison between real and simulated data from the exponential model at the maximum likelihood estimate of the rate parameter.

Assignment 4.1.10

Can you think of a better model to use? What assumptions are relaxed in this model?

Answer: Both of the assumptions seem to be relaxed. First, the visit rate should probably not be considered constant. For instance, people probably visit blogs more often during certain hours during the day. Also, first-time visits might not always be independent. For instance, if I tell my friends about a blog during a dinner then their visits would not be independent. Because if one of them visits the blog and then confirm what I told them, the other would be more likely to also visit the blog.

A better alternative would be to use the negative binomial distribution as events are not assumed to be independent. The events for the are no longer constrained to be independent.

Assignment 4.1.11

Estimate the parameters of your new model, and hence estimate the mean number of website visits per minute.

Answer: The waiting time distribution, i.e. the pdf for the negative binomial distribution is a member of the generalized Pareto distributions, type II [1].

The pdf has the form:

$$p(t_i|\alpha,\beta) = \frac{\alpha}{\beta}(1+\frac{t_i}{\beta})^{-(\alpha+1)} \tag{15}$$

4.2 Violent crime counts in New York counties

In data file *likelihood_NewYorkCrimeUnemployment.csv* is a data set of the population, violent crime count and unemployment across New York counties in 2014 (openly available from the New York Criminal Justice website).

Problem 4.2.1.

Graph the violent crime count against population size across all the counties. What type of relationship does this suggest?

Answer: From the plot it looks like there is a correlation between the count of violent crime in a city and the population size.

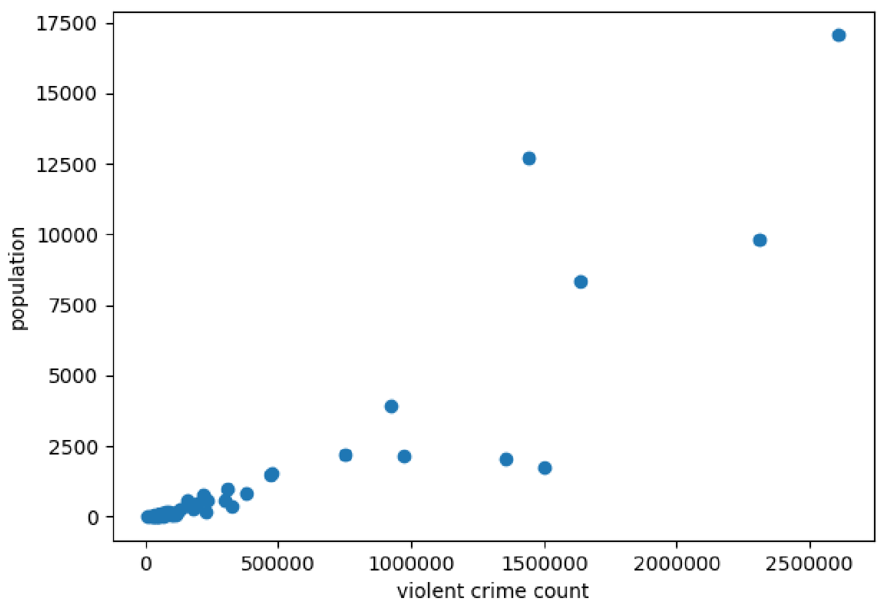

Figure 3: Relationship between crime and population.

Problem 4.2.2.

A simple model here might be to assume that the crime count in a particular county is related to the population size by a Poisson model:

$$crime_i \sim Poisson(n_i\theta) \tag{16}$$

where $crime_i$ and n_i are the crime count and population in county i. Write down an expression for the likelihood.

Answer:
We assume that the crime rate and population size in one city is conditionally independent of the other city and get the likelihood as:

$$l = p((crime_1, n_1), ..., (crime_N, n_N)|\theta) = p((crime_1, n_1) \cdot ... \cdot p((crime_N, n_N)|\theta) = \prod_{i=1}^{N} \frac{(n_i\theta)^{crime_i} \cdot e^{-n_i\theta}}{crime_i!} \tag{17}$$

$$log(l) = -N\bar{n}\theta + \sum_{i=1}^{N} crime_i \cdot log(n_i\theta) + const \tag{18}$$

$$\frac{\partial log(l)}{\partial\theta} = -N\bar{n} + \sum_{i=1}^{N} crime_i \cdot \frac{1}{n_i\theta} \cdot n_i = -N\bar{n} + N \cdot \overline{crime} \cdot \frac{1}{\theta} \tag{19}$$

Now we get the MLE:

$$-N\bar{n} + N \cdot \overline{crime} \cdot \frac{1}{\hat{\theta}} = 0$$

$$-\bar{n} + \cdot \overline{crime} \cdot \frac{1}{\hat{\theta}} = 0$$

$$\hat{\theta} = \frac{\overline{crime}}{\bar{n}} = 0.00366090 \approx 0.0037 \tag{20}$$

Thus, we get a 95% confidence interval in the range: $[-0.00853, 0.0159]$, i.e. the difference between the largest and the smallest value is 0.02443 and the upper range is 4.3 times larger than the MLE. Thus, it seems like the estimate is quite uncertain.

References

[1] probabilityislogic. Poisson is to exponential as gamma-poisson is to what?