University of Leicester

MSc in Medical Statistics

Computer Intensive Methods: Likelihood

Coursework 2014

This coursework should take about 10 hours to complete, including the write-up. Allocate 1 hour for every 10 marks. Hand in whatever you have managed to do in that time.

Background

The data for this coursework come from the Women's Interagency HIV Study (Barkan et al 1998) from which 195 women, who had been on antiretroviral therapy, were selected to have their viral load measured in both plasma and saliva samples. The objective was to estimate the correlation between the two viral loads. However, some women had viral loads below the limit of detection of the assay, which at the time of the study, was 80 RNA copies per millilitre. So 65% had saliva viral loads that were below the limit of detection, 39% had plasma viral loads below the limit of detection and 34% of women had both measurements below the limit of detection.

Data

The data are available on blackboard in the Excel file viral.xlsx. Missing data in that file correspond to values below the limit of detection.

References

The Women's Interagency HIV Study is described in,

Barkan SE, Melnick SL et al
The Women's Interagency HIV Study
Epidemiology 1998;9:117-125

The general problem of estimating the correlation between two measurements of viral load was first considered by Lyles et al (2001) and then by Chu et al (2005). The method that you will use follows the approach of Lyles et al, although Chu et al used the Women's Interagency HIV data to illustrate an alternative approach.

Lyles RH, Williams JK, Chuachoowong R
Correlating two viral load assays with known detection limits
Biometrics 2001;57:1238-1244

Chu H, Moulton LH et al
Correlating two continuous variables subject to detection limits in the context of mixture distributions
Applied Statistics 2005;54:831-845

Preliminaries

Viral loads are usually log-transformed and then modelled by a normal distribution. Because there are two viral load measurements these data will be modelled by a bivariate normal distribution. If you are unfamiliar with the bivariate normal you might want to read about it. See for example
http://en.wikipedia.org/wiki/Multivariate_normal_distribution
or the video (it is a bit slow) http://www.youtube.com/watch?v=4BrMSVV7mBM
To program the log-likelihood for the bivariate normal you will need Stata's binormal function.

Questions

1. Discuss the use of viral load measurements in the context of HIV. Your answer might include some or all of; the definition of viral load, the way that it is measured, why it is monitored, why it is monitored together with CD4 count, the range of values that you would expect to find in the UK, the relationship between viral load and the risk of conversion to AIDS, or the relationship with the risk of transmission.

   Your answer must be less than 300 words in length and it should include a minimum of two references. Usually an answer that considers a few of the suggested issues in depth is better to one that considers them all but without any detail.

   [10%]

2. Although the ultimate aim of the analysis is to measure the correlation between plasma and saliva measurements, you will start with the simpler problem of estimating the mean and standard deviation of the plasma measurements alone.

   (a) A simple strategy that is sometimes used when analysing viral loads is to replace all of the values below the level of detection (LD) with the value LD/√2. Do this for the plasma measurements and then calculate the mean and standard deviation of the **$\log_{10}$** plasma measurements for the 195 women. Plot a histogram of the **$\log_{10}$**-plasma viral load measurements that were above the level of detection and superimpose over the histogram, a normal distribution with the mean and standard deviation that you calculated using LD/√2 for the missing values.

   [8%]

   Will the use of LD/√2 under- or over-estimate the true mean and standard deviation? Explain why.

   [4%]

   (b) If the true mean and standard deviation of the **$\log_{10}$**-plasma measurements in a population of people with HIV were 3 and 1.5 respectively, what percentage of measurements would you expect to fall below a LD of 80?

   [2%]

   In a sample of N people, n_0 have a measured viral load below the LD of λ and n_1 have viral load measurements, y_i , above λ, i=1,…,n_1 and N=n_0+n_1. The mean and standard deviation of the $\log_{10}$ measurements in the whole population are μ and σ respectively and λ is known. Derive an expression for the log-likelihood of the data.

   [6%]

Write a Stata program that evaluates the log-likelihood of the plasma measurements from the Women's Interagency HIV Study for specific values of μ and σ, setting λ=80. Use Stata's ml command to estimate the two parameters, their standard errors and 95% confidence intervals. Find the size of the correlation between the estimates of μ and σ for these data.

[10%]

Simulate a set of 195 values, z_i, from a normal distribution with mean 3 and standard deviation 1.5. Create viral loads by taking $y_i=10^{z_i}$. Replace any values below λ=10 with missing values and then estimate μ and σ. Repeat for λ=20(10)100. Describe the general relationship between the size of this correlation and the percentage of measurements that are below LD.

[10%]

3. Lyles et al. extend the method of Question 2 to the situation in which there are two measurements of viral load for each patient and the $\log_{10}$ measurements follow a bivariate normal distribution.

(a) Use the notation, y_{1i} and y_{2i} for the two viral load measurements on subject i, λ for the LD (same for both measurements), μ_1 and σ_1 for the mean and standard deviation of $\log_{10}(y_1)$ and μ_2 and σ_2 for the mean and standard deviation of $\log_{10}(y_2)$ and ρ for the correlation between $\log_{10}(y_1)$ and $\log_{10}(y_2)$.

Suppose that μ_1=4 and σ_1=2, μ_2=3 and σ_2=1.6, and ρ=0.5.

(i) If you know nothing about a person's plasma viral load (y_{1i}), what would be the probability that their saliva viral load (y_{2i}) is less that 80?

(ii) If you know that a person's plasma viral load (y_{1i}) is 50,000, what would be the probability that their saliva viral load (y_{2i}) is less that 80?

[5%]

Further assume that the subjects are sorted so that,
the first n_1 have both measurements above LD,
the next n_2 have y_1 above LD but y_2 below LD,
the next n_3 have y_1 below LD but y_2 above LD,
the last n_4 have both y_1 and y_2 below LD.
Write down an expression for the log-likelihood.

[15%]

Write a Stata program that evaluates the log-likelihood for the Women's Interagency HIV Study plasma and saliva measurements for specific values of μ_1, μ_2 ,σ_1, σ_2, ρ. Use Stata's ml command to estimate the five parameters, their standard errors and 95% confidence intervals.

[20%]

(b) Sort the plasma measurements that were above the level of detection into ascending order and based on the MLEs from Q3(b) and the bivariate normal model, calculate the expected quantiles corresponding to these observed values. Make a QQ plot of the data. Repeat for the saliva measurements. Your plots should resemble Figure 2(a) and (d) from Chu et al. although it is more common to put the expected quantile on the x-axis. What is the use of these plots? Describe your interpretation of them in this case.

[10%]