StAN Exercise Sheet 5

Richard D. Gill

Mathematical Institute, University of Leiden, Netherlands

http://www.math.leidenuniv.nl/~gill

1 November, 2012

1 Least squares, weighted least squares, minimum chi-square

1.1 Pareto distribution (Salpeter function)

Cf. Feigelson and Babu, Chapter 4.

Suppose $X_1, \ldots, X_n$ is a random sample from the probability density $\alpha^{-1}x^{-\alpha-1}$, $x > 1$, where $\alpha > 0$ is an unknown positive constant.

What is the maximum likelihood estimate of α?

A histogram of data from this distribution shows numbers of observations in a bin of width h around a midpoint x of about $nh\alpha^{-1}x^{-\alpha-1}$. This is both (approximately) the expected number of observations in this bin, and the variance of the number of observations in this bin.

This suggests estimating α by minimum chi-squared: minimize

$$\sum_{\text{bins}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

where the expected number per bin, "Expected", is given by $Cx^{-\alpha-1}$, with x being the bin mid-point, and we minimise by choice of C and α. Of course, "Observed" stands for the observed number per bin.

It might be clever to transform the counts first by taking logarithms. The log counts have mean values approximately const $-$ $(\alpha + 1)\log x$ and variances (by propagation of error) 1/Expected count. This suggests estimating α by minimising

$$\sum_{\text{bins}} \text{Expected} \cdot (\text{Log Observed} - \text{Expected Log Observed})^2.$$

Two-stage least squares is probably the most effective way to solve this last problem: in step 1 just minimise $\sum_{\text{bins}}(\text{Log Observed} - \text{Expected Log Observed})^2$, in step 2 minimize the "correct" target function $\sum_{\text{bins}} \text{Expected} \cdot (\text{Log Observed} - \text{Expected Log Observed})^2$ where the weights, Expected, are taken from the first stage. Minimization is done over the term α (and constant) appearing in Expected Log Observed.

Investigate this procedure through a small simulation experiment of your own, and compare with maximum likelihood. Use the `R` function "`lm`" for estimating unweighted and weighted linear least squares.

It might be sensible to take logarithms of the data before beginning this whole procedure. Of course the expected value of counts of binned log observations have to be corrected for the transformation "log". Can you figure out how to do this?