Principles of Statistical Data Analysis

Homework 1

Academic Year 2024-2025

Introduction

You must prepare this homework in groups of 2 students. Your report must be submitted as a pdf file via Ufora. The R file must also be submitted. Please use the following format for the file name

HW1_LastNameStudent1_LastNameStudent2....pdf

and use a similar format for the R file.

We anticipate that you can finish the assignment in about 8 hours. The report should not be longer than 3 pages (excluding tables and graphs). The submission deadline is October 28 (please submit via Ufora).

Assignment

In ecology, the abundance of a particular species in a fixed area or volume is often described by the negative binomial distribution, with densitiy function

$$p(y) = \frac{\Gamma(y + 1/\phi)}{\Gamma(1/\phi)y!} \left(\frac{1}{1+\mu\phi}\right)^{1/\phi} \left(\frac{\mu\phi}{1+\mu\phi}\right)^{y},$$

where $\Gamma(\cdot)$ is the gamma function ($\Gamma(n) = (n-1)!$) and with $\mu = \mathrm{E}(Y) > 0$ the mean of Y and $\phi > 0$ the overdispersion parameter. The interpretation of the latter comes from $\mathrm{Var}(Y) = \mu + \phi\mu^2$, i.e. the variance of the negative binomial distribution is overdispersed as compared to the Poisson distribution.

You are given data of the abundances of a particular species of ants (Hymenoptera: Formicidae). The data are collected from 104 randomly sampled soil samples in Belgium during the summer months. For each soil sample, the

soil moisture is also measured (expressed in percentage). To investigate the relationship between soil moisture and the average abundance of ants, the ecologist suggest the following model for the mean,

$$\log(\mathrm{E}\,(Y \mid x)) = \log(\mu) = \beta_0 + \beta_1 x.$$

The unknown parameters are thus ϕ, β_0 and β_1.

The mean of the abundance depends on the moisture x. If μ is replaced by $\exp(\beta_0 + \beta_1 x)$ in the density function, then it is more appropriate to use the notation $p(y \mid x)$ instead of $p(y)$; this stresses that $p(y \mid x)$ is the density function of the conditional distribution of Y given x. For the distribution of X, you may choose an arbitrary distribution; use $f(x)$ to denote its density function.

The data set (ants) is available on Ufora.

1. Explore the data, both with numerical summary statistics and with graph(s). Include an exploration of the relationship between the mean and the variance.
2. Construct the likelihood and the log-likelihood functions. You may assume that all observations are independently distributed.
3. Give the estimating equation for the β_1 parameter (assuming that the other parameters are known). The estimating equation is the equation that must be solved for β_1; the solution is the maximum likelihood estimate.
4. Make a plot of the log-likelihood function as a function of β_1. For the other two parameters you may use $\hat{\beta}_0 = 2.509067$ and $\hat{\phi} = 2.289377$ (these are the maximum likelihood estimates).
5. Calculate the maximum likelihood estimate of β_1 via numerical optimisation in R (e.g. with the optim function).
6. Plug-in the maximum likelihood estimate of β_1 (result of the previous question) in the estimating equation of question 3. What is the result? Is this what you expect?
7. With the estimated parameters, visualise the mean / variance relationship. Does this agree with your data exploration (question 1)?