Assignment Nonlinear Regression

A report containing solutions to the problems below is to be uploaded in the Blackboard as a single pdf-file, due no later than the deadline for this assignment. Submitted report should be well-organized and include step-by-step explanations, relevant plots, tables, R code and conclusions.

I. Computational problems

1. The dataset `muscle` is contained in the add-on `R`-package `MASS`. To load the package, type `library(MASS)` at the `R`-prompt. These data concern an experiment on muscle contraction for 21 animals. The observed variables are `Strip` (muscle identifier, a factor), `Conc` (CaCl concentration, a numeric vector) and `Length` (resulting length of muscle section, a numeric vector). There are 60 observations on each variable. The nonlinear model we are considering is

$$\texttt{Length}_i = \theta_1 + \theta_2 \exp(-\texttt{Conc}_i/\theta_3) + \varepsilon_i, \quad i = 1, \ldots, 60,$$

   where $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. We wish to estimate $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$.

   (i) Use `R` to construct a scatterplot of the data, according to the proposed model. State your observations about the plot.

   (ii) Obtain an estimate of $\boldsymbol{\theta}$, using nonlinear least squares estimation. Plot the fitted curve along with the data and comment on the quality of the fit.

   (iii) Possibly different values for θ_1 and θ_2 should be used (why?) for the different animals. The parameter vector is then $\boldsymbol{\theta}^* = (\theta_{1,1}, \ldots, \theta_{1,21}, \theta_{2,1}, \ldots, \theta_{2,21}, \theta_3)$ and the nonlinear model becomes

$$Y_{jk} = \theta_{1,k} + \theta_{2,k} \exp(-X_{jk}/\theta_3) + \varepsilon_{jk}, \qquad j = 1, \ldots, n_k, \quad k = 1, \ldots, 21,$$

   where Y_{jk} and X_{jk} denote respectively `Length` and `CaCl` for the jth observation from the kth animal. Use nonlinear least squares to obtain an estimate of $\boldsymbol{\theta}^*$.

   Hints: you can use the estimates from (ii) as the starting values for estimating θ^*; use `Length~th1[Strip]+th2[Strip]*exp(-Conc/th3)` to design your formula for the `nls` function to make sure that only the data corresponding to each level of the variables `Strip` are used to estimate the corresponding components of $\boldsymbol{\theta}^*$. Here `th1`, `th2` and `th3` are the names given to the three components of the original parameter vector $\boldsymbol{\theta}$.

2. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ and $f(x, \boldsymbol{\theta}) = \theta_1 x + \frac{\theta_2}{\theta_3 + x^2}$. Suppose

$$Y_i = f(x_i, \boldsymbol{\theta}) + \varepsilon_i, \quad \varepsilon_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n.$$

   Take $n = 100$, choose reasonable values for $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$, $\sigma^2 > 0$ and $x_i \in [0, 3]$, $i = 1, \ldots, n$. Generate independent $\varepsilon_1, \ldots, \varepsilon_{100} \sim \mathcal{N}(0, \sigma^2)$ distribution and use these values to form $Y_1, \ldots, Y_{100}$.

(i) Make a scatterplot of your simulated data, along with a plot of the true curve $f(x, \boldsymbol{\theta})$. Comment. Then use nonlinear regression to estimate $\boldsymbol{\theta}$ (take reasonable starting values). Also report your results for $\hat{\sigma}^2$ and $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\theta}})$. Add a plot of the curve $f(x, \hat{\boldsymbol{\theta}})$ to your scatterplot and comment on the quality of the fit.

(ii) Plot the residuals against the fitted values, generate a normal quantile plot, and comment on the validity of the model assumptions.

(iii) Obtain 98% confidence intervals for θ_1, θ_2 and θ_3, first using asymptotic normality, and then using the bootstrap. Comment on the agreement of the confidence intervals between the two methods, and their accuracy.

(iv) Obtain a 98% confidence interval for the expected value of Y when $x = 1$.

3. The impact of weeds on the productivity of a crop is investigated. The file `Weeds.txt` (available on Blackboard) contains a column for the number N of healthy crops produced in each of 56 plots of land, along with the rate r of weeds found in each plot (in kg/ha). A proposed approach is to model the relationship between N and R using one of the following two functions: $f_1(r) = \delta + \frac{\alpha-\delta}{1+\exp[\beta \log(\gamma r)]}$ (the log-logistic function), where $\beta, \gamma > 0$ and $\alpha, \delta \in \mathbb{R}$; or $f_2(r) = \gamma + \alpha \exp(-\beta r)$, where $\alpha, \beta > 0$ and $\gamma \in \mathbb{R}$.

   Apply the nonlinear least squares method to fit model (2.1) from the lecture notes to the dataset `Weeds` using both f_1 and f_2. Use appropriate methods to determine which function produces the best fit.

II. Theoretical problems

1. Suppose $Y_i = f(x_i, \theta) + \varepsilon_i$, $i = 1, \ldots, n$, where $\theta \in \mathbb{R}$, f is everywhere differentiable and $\varepsilon_1, \ldots, \varepsilon_n$ are independent $\mathcal{N}(0, \sigma^2)$. Derive the expression for the log-likelihood function $\ell(\theta; \mathbf{x})$. Relate the MLE to the LSE and derive the MLE for σ^2.

2. Suppose we have a data set which we want to describe by the nonlinear model

$$Y_i = \theta_1 x + 2\theta_3 x^2 + \frac{\theta_2 x^3}{3 + \exp(-\theta_4 x + x^2)} + \theta_5 + \varepsilon_i, \quad \varepsilon_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n,$$

   Formulate the hypotheses that the model is in fact linear. Describe how you would test this hypothesis.

3. Suppose we have a dataset $\{(Y_1, x_1), \ldots, (Y_n, x_n)\}$ which is modeled as follows:

$$Y_i = \sin(\theta_1 x_i) + \theta_1 \exp\{-\theta_2 x_i\} + \varepsilon_i, \quad i = 1, \ldots, n,$$

   where (θ_1, θ_2) is to be estimated and such that $\theta_1 \neq 0$, $\varepsilon_1, \ldots \varepsilon_n$ are independent random errors such that $\mathrm{E}\varepsilon_i = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $i = 1, \ldots, n$.

   Give the normal equations for the above model.

   Suppose $n = 200$, $x_1 = x_2 = \ldots = x_{100} = 0$ and $x_{101} = x_{102} = \ldots = x_{200} = 1$. Propose a starting value for the LSE $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2)$ in the Gauss-Newton method and explain your choice.