Supplementary article data

Projections of total hip replacement in Sweden from 2013 to 2030

Szilárd Nemes[1], Max Gordon[1,2], Cecilia Rogmark[1,3], and Ola Rolfson[1,4]

[1]The Swedish Hip Arthroplasty Register, Gothenburg; [2]Department of Clinical Sciences at Danderyd Hospital, Karolinska Institutet, Stockholm; [3]Department of Orthopedics, Lund University, Skåne University Hospital, Malmö; [4]Department of Orthopedics, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden.
Correspondence: szilard.nemes@registercentrum.se
Submitted 13-10-21. Accepted 14-02-20

DOI 10.3109/17453674.2014.913224

Technical details and implementation in R of the statistical methods used in '*Projections of total hip replacement in Sweden from 2013 to 2030'* by Nemes and collaborators.

1. Nonlinear Least Squares Regression

Nonlinear regression extends linear least squares regression to a more wide class of functions, given that the function can be written in closed form. These models are parametric as the they include a well-defined function with unknown parameters. Moreover they are efficient and the parameters generally have intuitive interpretations with a formulaic way of describing the relationship between the outcome and predictors.

The general form for the nonlinear regression model is

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i .$$

The linear regression is a special case of nonlinear regression models and some models become linear after appropriate transformation. Nonlinear regression has the following assumptions:

1.1. Functional form: $E[\mathrm{y}_i \mid \mathbf{x}_i] = f(\mathbf{x}_i, \boldsymbol{\beta})$

1.2. Zero conditional mean of ε : $E[\varepsilon_i \mid f(\mathbf{x}_i, \boldsymbol{\beta})] = 0$

1.3. Homoscedastic errors: $\mathrm{Var}(\varepsilon_i \mid \mathbf{x}_i) = \sigma^2$ for all *i* and $\sigma^2 < \infty$

1.4. Uncorrelated errors: $E[\varepsilon_i \varepsilon_j \mid f(\mathbf{x}_i, \boldsymbol{\beta}) f(\mathbf{x}_j, \boldsymbol{\beta})] = 0$

1.5. Identifiability of the model parameters

1.6. Underlying probability model: the errors ε follows a well-defined probability distribution function, generally the mean zero normal distribution.

Nonlinear least squares modeling assume that the function is smooth with respect to the unknown parameters, and the least squares criterion has a unique solution.

Due to the more or less complicated structural form closed form solutions such as the normal equations for the ordinary least squares regressions are hard or at times impossible to find. As a result different optimization algorithms are used for parameter estimation. A popular optimization algorithm in this context is the Levenberg–Marquardt algorithm. Briefly, the Levenberg–Marquardt algorithm is a blend of Gradient Descent and the Gauss-Newton iteration. The steps in the Gradient Descent are augmented with information about the curvature of the function, by using Newton's method for estimating the second derivatives. The resulting algorithm has a large step in the direction with low curvature and a small step in the direction with high curvature.

2. Asymptotic-models

A specific sub-family of the Nonlinear Least Squares Regression models assumes that the outcome cannot take any value but it is restricted by bounds. While this is implicitly assumed by most researchers, regression models ignore this assumption and at least theoretically the predicted values can take-up of any value imaginable. The upper or lower value that the outcome can attain is often a direct interest. For disease prognosis studies that try to estimate the number of patients who will have a specific illness in the future (near or distant) is of immediate interest is the upper asymptote denoting the maximum number of patients with a specific illness.

Different generic models are presented in the literature and additional study specific models can be specified by finding the integrated form of the differential equation that was derived to meet the study requirements.

Here we run three asymptotic regression models (Table 1). We chose models that allow estimation of an upper-asymptote without the requirement of a lower.

Table 1. Equations and parameters of the competing models used for projection.

Model	Equation
Asymptotic-regression	$y = A\left\{1 - e^{-e^{\beta_1}(x-\beta_0)}\right\}$
Logistic-regression	$y = \frac{A}{1 + e^{\left(\frac{\delta - x}{\gamma}\right)}}$
Gompertz-regression	$y = Ae^{(-\beta_0 \gamma^x)}$

Herein, y represents the outcome in this case the recorded incidence while x the predictor denotes the calendar years. The upper-asymptote is denoted by A and represents the maximum possible incidence of the total hip replacement. Nuisance parameters for the asymptotic regression are β_0 the value of the predictor at which the outcome is zero, i.e. the calendar year with no-operation; and β_1 which denotes the change in outcome with one unit change in the predictor, however without a direct interpretation. Nuisance parameters for the logistic regression inflection point δ at which point the maximum growth occurs and scale parameter γ denoting the change in outcome with one unit change in the predictor. The Gomperetz model is similar to the logistic one, but more asymmetric where β_0 is the intercept and γ is the growth rate.

3. Model Selection

Studies evaluating the feasibility of different model selection and comparison techniques for non-linear models are somewhat scarce.

We chose the Akaike's Information Criterion as a model selection tool. AIC is estimated as $AIC = -2\log L(\hat{\theta}) + 2p$, where $L(\hat{\theta})$ is the likelihood and p the number of parameters in the model.

AIC has been reported to find the 'true' model more reliably than F-test. The model with lowest AIC value is considered to be closest to the unknown truth. The drawback of AIC and Information Criterions in general is the lack of a straightforward universal interpretation and a proper scale with easily interpretable values. The lack of scale makes hard to get insight how much statistical importance we can attach to a difference in AIC between two models. Raw comparison of AIC values does not provide a weight of evidence in favor of the chosen model. If the considered models have almost equal AIC values raw comparison becomes even more difficult. In this case it's attractive to calculate the Akaike weights that serve as estimates for the conditional probabilities for each model. First we estimate the differences in AIC between models

$$\Delta_i(AIC) = AIC_i - \min(AIC)$$

then we can estimate the relative likelihood of model i given the data as

$$L(M_i \mid \text{data}) \propto \exp\{-0.5\Delta_i(AIC)\}.$$

The normalized relative likelihoods function as Akaike weights (w_i)

$$w_i(AIC) = \frac{\exp\{-0.5\Delta_i(AIC)\}}{\sum_k \exp\{-0.5\Delta_k(AIC)\}}$$

so that $\sum_i w_i(AIC) = 1$. The interpretation is straightforward, the probability that the chosen model best describes the data given the considered candidate models. Dividing the Akaike weights of two competing models gives the strength of evidence of one model over the other. AIC will not give a degree of belief about the model's truthfulness; it merely gives us objective tool to compare the degree to which extent the data supports the various models we wish to consider

4. Implementation in R

The R programing language offers numerous routines and applications for nonlinear least stares regression. The base `stats` package has the `nls` function and the different `selfStart` (`SSasympOff` , `SSgompertz` , `SSlogis` among others) functions that allows optimal start value and straightforward programing syntax. The package `minpack.lm` contains the Levenberg–Marquardt algorithm for optimization, while the package `nls2` a 'brute-force' parameter estimation. The package `nlstools` offers different routines for model evaluation.

```
rm(list = ls())
library(minpack.lm)
library(nls2)
library(propagate)
library(qpcR)
library(nlstools)
timeprog <- NULL
timeprog$year <- 1968:2012
timeprog$inc40 <- c(4.999062, 11.532154, 25.020892, 38.026046, 51.174894,
59.811730, 67.297406, 75.145132, 94.269069, 114.677618, 128.503150, 131.793292,
```

```
139.219985, 155.891680, 165.090324, 177.761185, 193.135288, 200.524853, 194.714938,
240.277227, 213.238804, 214.424084, 225.686785, 274.522637, 264.173518, 223.267211,
221.964295, 216.989864, 252.034770, 240.635941, 249.304712, 243.026163, 258.968934,
276.876948, 285.678462, 283.828886, 296.176272, 305.714016, 304.469996, 306.303163,
306.442262, 330.565806, 332.027297, 328.944112, 326.419377)

timeprog <- as.data.frame(timeprog)
## Model fit
   # 1. Asymptotic Regression
fit.asymptOff <- nlsLM(inc40 ~ SSasympOff(year, Asym, Ro, co) , data = timeprog)
summary(fit.asymptOff) ## Summmary statitics
confint(fit.asymptOff) ## 95 % Confidence Intervals

   # 2. Logistic Regression
fit.logis <- nlsLM(inc40 ~ SSlogis(year, Asym, xmid, scal),data = timeprog)
summary(fit.logis)

   # 3. Gomperez Regression
fit.gomp <- nlsLM(inc40 ~ SSgompertz(I(year-1968), Asym, b2, b3) , data = timeprog)
summary(fit.gomp)

akaike.weights(c(AIC(fit.asymptOff), AIC(fit.gomp), AIC(fit.logis)))

## Jackknife for the asymptotic model
fitJack <- nls(inc40 ~ SSasympOff(year, Asym, Ro, co) , data = timeprog)
summary(fitJack)
jackEval <- nlsJack(fitJack)
plot(jackEval)
summary(jackEval)

## Examine the effect of the identified influential points

summary(nlsLM(inc40 ~ SSasympOff(year, Asym, Ro, co) , data = timeprog[-20,]))
summary(nlsLM(inc40 ~ SSasympOff(year, Asym, Ro, co) , data = timeprog[-24,]))
summary(nlsLM(inc40 ~ SSasympOff(year, Asym, Ro, co) , data = timeprog[-42,]))
summary(nlsLM(inc40 ~ SSasympOff(year, Asym, Ro, co) , data = timeprog[-43,]))




## Prediction intervals for the estimated incidences
fit2 <- nls(inc40 ~ Asym*(1 - exp(-exp(lrc)*(year - c0))) , data = timeprog,
      start = list(Asym = 396, lrc = -3, c0 = 1968))
summary(fit2)

newdata <- data.frame(year=seq(2013, 2030))

predCI <- predictNLS(fit2, newdata, interval = "prediction")
predCI$summary
```