Non-linear Least-Squares Fitting

2015-16 MSc Ecology, Evolution and Conservation, Silwood Park

Samraat Pawar

Department of Life Sciences

(Silwood Park)

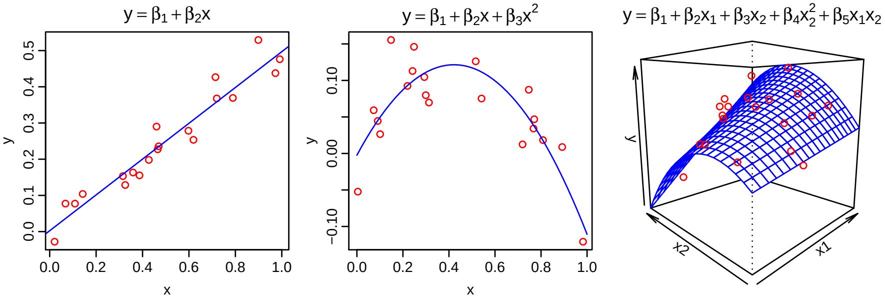
Imperial College London

February 15, 2016

OUTLINE

- Why Non-Linear Least Squares regression / fitting?
- The NLLS fitting method
- NLLS in `R`
- Afternoon practicals overview (two examples)

Linear models are great



- These are *all* good linear models (huh?!)
- The data can be modelled as *linear combination* of variables and coefficients
- Easily fitted using Ordinary Least Squares (OLS) regression
- Linear models can include curved relationships (e.g. polynomials)
- *OK, so then why Non-Linear Least Squares (NLLS) fitting?*

WHY NLLS? – FIRST, WHAT MAKES A MODEL NON-LINEAR?

- OLS regression can be used to fit both linear and nonlinear *equations* that *intrinsically linear*
  e.g., Straight line: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
  e.g., Polynomial: $y_i = \exp(\beta_0) + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$
- Indeed, for OLS to work, we need *intrinsic linearity* — i.e., the equation to be fitted (model) should be *linear in the parameters*
- Are these models linear in their parameters?
  - $y_i = \beta_0 + \beta_1 x_i^{\beta_2} + \varepsilon_i$
  - $y_i = \beta_0 e^{\beta_2 x_i} + \varepsilon_i$

  NO!

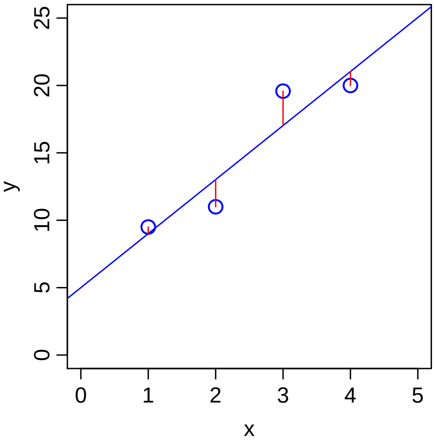
So what — why is intrinsic non-linearity a problem?

Recall what the Least Squares method does:

- Consider a predictor x, data y, n observations, and a model that we want to fit to the data:
  $f(x_i, \boldsymbol{\beta}) + \varepsilon_i$
  where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)$ are the model's k parameters

- The objective is to find estimates of values of the k parameters ($\hat{\beta}_j$) that minimize the sum (S) of squared residuals (r_i) (AKA RSS):
  $S = \sum_{i=1}^{n} [y_i - f(x_i, \boldsymbol{\beta})]^2 = \sum_{i=1}^{n} r_i^2$

So what — why is intrinsic non-linearity a problem?

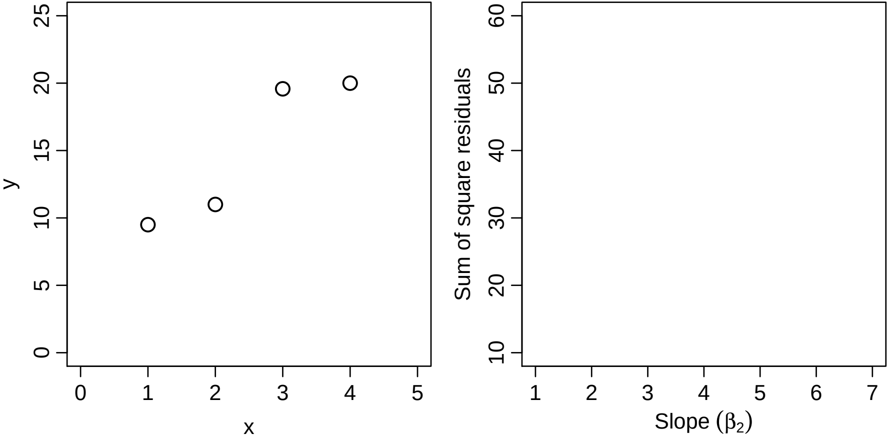


$$y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_i$$

$$9.50 = 5 + 4 \times 1 + 0.50$$
$$11.00 = 5 + 4 \times 2 - 2.00$$
$$19.58 = 5 + 4 \times 3 + 2.58$$
$$20.00 = 5 + 4 \times 4 - 1.00$$

$$\beta_0 = 5; \beta_1 = 4$$

OLS minimizes the *sum* of the *squared* residuals



So what — why is intrinsic non-linearity a problem?

That's all well and good, but we can use maths instead of brute-force computation!

- Our model is $f(x_i, \boldsymbol{\beta}) + \varepsilon_i$
- We want to find estimates of values of the parameters ($\hat{\beta}_j$) that *minimize* the sum (S) of squared residuals (r_i) (AKA "RSS")
  $S = \sum_{i=1}^{n} [y_i - f(x_i, \boldsymbol{\beta})]^2 = \sum_{i=1}^{n} r_i^2$
- For this we can solve $\frac{\partial S}{\partial \beta_j} = 0, j = 0, 1, 2, \ldots, k$ to find the *minimum*
- That is, we need to solve $\frac{\partial \sum_{i=1}^{n} r_i^2}{\partial \beta_j} = 0$
- Or, $2 \sum_{i=1}^{n} r_i \frac{\partial r_i}{\partial \beta_j} = 0$

So what — why is intrinsic non-linearity a problem?

- Thus, solving $2\sum_{i=1}^{n} r_i \frac{\partial r_i}{\partial \beta_j} = 0$
  boils down to finding the "gradient" $\frac{\partial r_i}{\partial \beta_j}$
- This is not a problem in linear models, because this gradient is fully solvable as the equation is *intrinsically linear*
- That is, the solution of $\frac{\partial r_i}{\partial \beta_j}$ is simple (enough)

So what — why is intrinsic non-linearity a problem?

- For example, if $f(x_i, \boldsymbol{\beta}) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- That is, our model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (Linear Regression)
- Then we want to solve
  $\frac{\partial S}{\partial \beta_0} = \sum_{i=1}^{n} \frac{\partial [y_i - (\beta_0 + \beta_1 x_i)]^2}{\partial \beta_0} = 0$
  $\frac{\partial S}{\partial \beta_1} = \sum_{i=1}^{n} \frac{\partial [y_i - (\beta_0 + \beta_1 x_i)]^2}{\partial \beta_1} = 0$

So what — why is intrinsic non-linearity a problem?

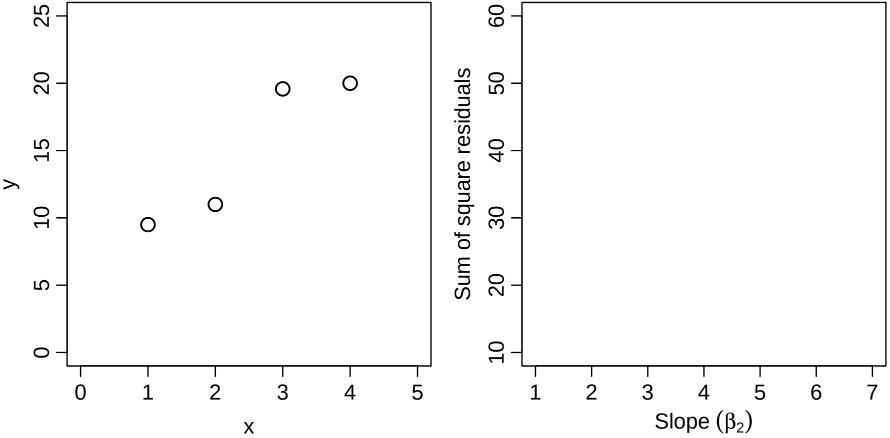
- And, solving
  $\frac{\partial S}{\partial \beta_0} = \sum_{i=1}^{n} \frac{\partial [y_i - (\beta_0 + \beta_1 x_i)]^2}{\partial \beta_0} = 0$
  $\frac{\partial S}{\partial \beta_1} = \sum_{i=1}^{n} \frac{\partial [y_i - (\beta_0 + \beta_1 x_i)]^2}{\partial \beta_1} = 0$
  just boils down to solving two simultaneous equations because $\frac{\partial r_i}{\partial \beta_j}$ is simple *because* the model is intrinsically linear:

  $-n\beta_0 + \sum_{i=1}^{n} y_i + \beta_1 \sum_{i=1}^{n} x_i = 0$
  $\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = 0$

- That is, we need to solve $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ (*this is what R solves when you use* `lm()`)

The linear least-squares solution

There is only one unique solution, which gives the OLS fitted parameter values for $\boldsymbol{\beta} = (\beta_0, \beta_1)$:



So what — why is intrinsic non-linearity a problem?

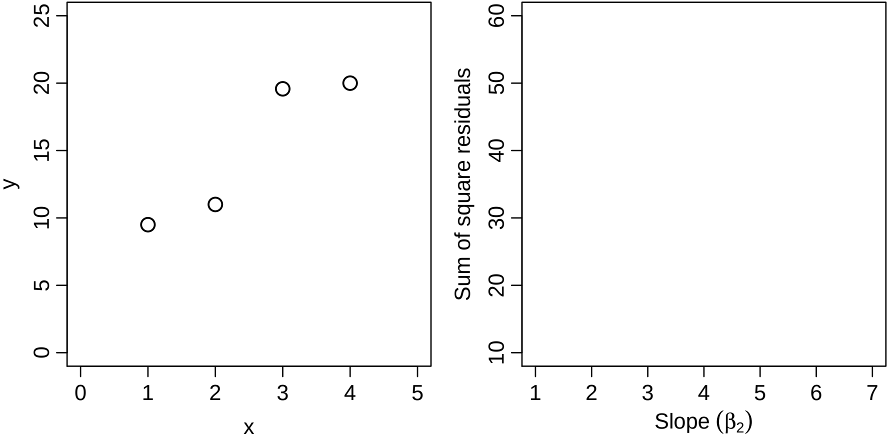
- So, then, in an intrinsically non-linear model such as $y_i = \beta_0 e^{\beta_2 x_i} + \varepsilon_i$ the derivatives $\frac{\partial r_i}{\partial \beta_j}$ are naughty
- That is, they are functions of both x and the parameters β_j, so the gradient equations do not have a solution like the OLS case
- So the nice trick of solving $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is impossible *mathematically*

So — Enter NLLS!

But we can use a computer!

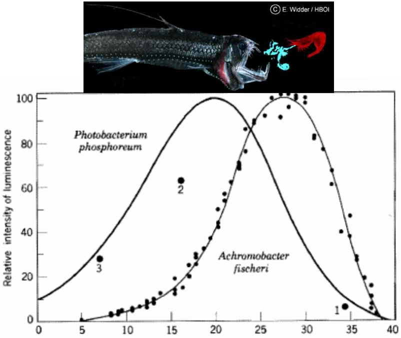
- Choose initial values for the β_j's
- Then, "refine" the parameters *iterately* by calculating $\frac{\partial r_i}{\partial \beta_j}$ *approximately* — this approximation is the *Jacobian* (the gradient), which is a matrix of the $\frac{\partial r_i}{\partial \beta_j}$'s
- Whether a refinement has taken place in any step of the iteration is determined by re-calculating the residuals at that step
- Eventually, if it all goes well, we find a combination of β_j's that is *very close* to the desired solution $\frac{\partial S}{\partial \beta_j} = 0, j = 0, 1, 2, \ldots, k$

The end result is (approximately) the same:



OK, FINE, WHY WOULD *I* EVER NEED TO WORRY ABOUT NLLS?

- Many observations in biology are just not well-fitted by a linear model
- That is, the underlying biological phenomena are not well-described by a linear model
- Examples:
  - Logistic growth model
  - Michaelis-Menten biochemical kinetics (two parameters $V_{\max}$ and K_m: $v = \frac{V_{\max}[S]}{K_m+[S]}$
  - Responses of metabolic rates to changing temperature (practical 1)
  - Consumer-Resource (e.g., predator-prey) functional responses (practical 2)
  - Time-series data (e.g., fitting sinusoidal function)



Johnson et al 1974

$$B = B_0 \boxed{e^{-\frac{E}{kT}}} f(T, T_{pk}, E_D)$$

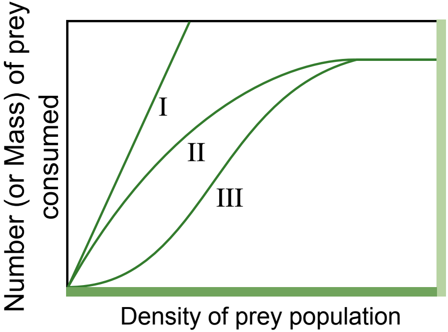
T = temperature (K)
k = Boltzmann constant (eV K^{-1})
E = Activation energy (eV)
T_{pk} = Temperature of peak performance
E_D = Deactivation energy (eV)
(J H vant Hoff 1884, S Arrhenius 1889)

- Surely there is more to thermal responses?
  - Oxygen limitation
  - Complexity of metabolic network
  - Hormonal regulation
- *What about alternative models?*

Example 2: Functional responses

$$f(x_R) = \frac{a x_R^{q+1}}{1 + h a x_R^{q+1}} \text{ (Holling, 1959)}$$

x_R = Resource density (Mass / Area or Volume)
a = Search rate (Area or Volume / Time)
h = Handling time
q = Shape parameter (dimensionless)



Note that:

- NLLS fitting can yield $h < 0$, $q < 0$, or both
- $h < 0$ is biologically impossible but indicates an upward curving response
- $q < 0$ is biologically unlikely as it indicates a decline in search rate with resource density (but is useful as a measure of deviation away from a type III response)

So the general procedure is:

1. Start with an initial value for each parameter in the model
2. Generate the curve defined by the initial values
3. Calculate the residual sum-of-squares (RSS)
4. Adjust the parameters to make the curve come closer to the data points *This the tricky part*
5. Adjust the parameters again so that the curve comes even closer to the points (RSS decreases)
6. Repeat 4–5
7. Stop simulations when the adjustments make virtually no difference to the RSS

NLLS FITTING IN R

The *tricky part — adjust parameters to make curve come closer to the data points* (step 4) has at least two algorithms:

- The Gauss-Newton algorithm is the default in the `nls` package (part of the `stats` base package) — good in many cases, but doesn't work very well if the model is mathematically weird (the optimization landscape is difficult) and the starting values for parameters are far-off-optimal

- The Levenberg-Marquardt (LM) switches switches between Gauss-Newton and "gradient descent" and is more robust against starting values that are far-off-optimal — available in `R` through the the `minpack.lm` package
  `http://cran.r-project.org/web/packages/minpack.lm`

- The command is `nlsLM`

NLLS FITTING IN R

- Once the algorithm as converged (hopefully – but you may be surprised how well it usually works), you need to get the goodness of fit measures
- First, of course, examine the fits visually
- Also, report the best-fit results, including:
  - Sums of deviations of the data points from the final model fit (final RSS)
  - R^2
  - Estimated coefficients
  - For each coefficient, standard error (can be used for CI's), t-statistic and corresponding (two-sided) p-value
- The function `summary.nls` will give you all these measures
- Remember, the precise parameter values you obtain will depend in part on the initial values chosen and the convergence criteria
- You may also want to compare multiple models...

NLLS Assumptions

NLLS-regression has all the assumptions of OLS-regression:

- No (in practice, minimal) measurement error in explanatory variable (x-axis variable)
- Data have constant normal variance — errors in the y-axis are homogeneously distributed over the x-axis range
- The measurement/observation error distribution is Gaussian — for example, what would the error distribution of this non-linear model be: $y_i = \beta_0 e^{\beta_2 x_i} + \varepsilon_i$
- What if the errors are not normal? — use maximum likelihood instead! (e.g., using `nlm` for optimizing/fitting)

Comparing models

You can use information theory (including AIC and BIC) to compare models. The lower the AIC or BIC, the better. This is how you can calculate these (using `R` syntax):

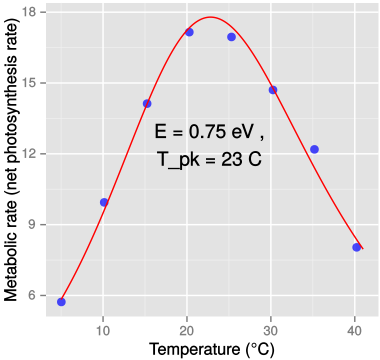
- residuals = Observations - Predictions
- `rss = sum(residuals ^ 2)`
- Then, AIC is `n * log((2 * pi) / n) + n + 2 + n * log(rss) + 2 * k`
  (*what is* `n` *and* `k`*?*)
- And BIC is `n + n * log(2 * pi) + n * log(rss / n) + (log(n)) * (k + 1)`
- For both AIC and BIC, If model **A** has AIC lower by 2-3 or more than model **B**, its better — Differences of less than 2-3 dont really matter

Also note that:

- R^2 = 1 - (rss/tss), where tss is total sum of squares:
  `tss = sum((Observations - mean(Predictions)) ^ 2)`

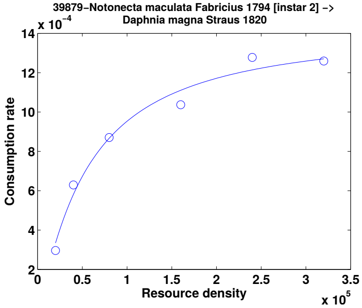
Practical 1: Fitting thermal responses

- Use `nls` (or `nlsLM`) to fit the `ThermRespData.csv` dataset to the model: $B = B_0 \boxed{e^{-\frac{E}{kT}}} f(T, T_{pk}, E_D)$
- Plot the data and output coefficient estimates and fit stats to a file
- Here's an example of how a fitted curve looks (only two fitted parameters are shown):



PRACTICAL 2: FITTING FUNCTIONAL RESPONSES

- Use `nls` (or `nlsLM`) to fit `CRat.csv` dataset to the model:
  $f(x_R) = c = \frac{a x_R^{q+1}}{1 + h a x_R^{q+1}}$
  (c is consumption rate)
- Plot the data and output coefficient estimates and fit stats to a file
- Here's an example of how a fitted curve looks :



PRACTICALS: GENERAL INSTRUCTIONS

- You can choose either practical, or do both
- Make sure you have a good look at the data first by plotting them up in a loop
- Keep workflow organized in `Code`, `Results`, `Data` !
- You may also find your own model to fit some data related to your interests or project
- Your demonstrators and I will help you get started in any case

You can also use mixed-effects models in NLLS

- You can use mixed-effects modelling with NLLS in R
- The package is `nlme` `https://stat.ethz.ch/R-manual/R-devel/library/nlme/html/nlme.html`
- You are probably stuck with the Gauss-Newton algorithm with `nlme` though...

Readings

- Motulsky, Harvey, and Arthur Christopoulos. Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting. OUP USA, 2004.