Nonlinear Regression: Some Ideas and a Little Guts

Nonlinear regression models are models of the form

$$\begin{aligned} y_i &= f(\underline{x}_i;\ \underline{\beta}) + \epsilon_i; \\ \epsilon_i &\overset{iid}{\sim} N(0, \sigma^2),\ i = 1, \ldots, n \end{aligned}$$

where $\underline{x}_i = (x_{i1}, \ldots x_{iJ})$ as in linear regression, and $\underline{\beta} = (beta_1, \beta_2, \ldots, \beta_K)$ are parameters that describe the function $f(\cdot; \underline{\beta})$.

Examples (all with $J = 1$):

- Exponential model: $f(x; \alpha, \beta) = \alpha e^{\beta x}$
- Logistic model: $f(x; \alpha.\beta, \gamma) = \dfrac{\alpha}{1 + \gamma e^{-\beta x}}$
- Gompertz growth model: $f(x; \alpha, \beta, \gamma) = \alpha e^{-\gamma e^{-\beta x}}$

etc. etc.

What is nonlinear regression doing in a linear regression course?

- The nonlinear model
$$y_i = f(x_{i1}, \ldots, x_{iJ});\ \underline{\beta}) + \epsilon_i,\ i = 1, \ldots, n$$
is a natural generalization of the linear model
$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{K-1} x_{i(K-1)} + \epsilon_i,\ i = 1, \ldots, n$$
- It is really linear regression in disguise!

Some Guts: How to estimate the $\underline{\beta}$'s.

Want to find the values $\hat{\underline{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_K)$ that minimize

$$RSS(\hat{\underline{\beta}}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - f(\underline{x}_i;\ \hat{\underline{\beta}}))^2$$

This is a differential calculus problem again.. For each β_j,

$$\frac{\partial RSS(\underline{\beta})}{\partial \beta_j} = -2 \sum_{i=1}^{n} (y_i - f(\underline{x}_i;\ \underline{\beta})) \frac{\partial f(\underline{x}_i;\ \underline{\beta})}{\partial \beta_j}$$

Setting these equal to zero leads to the system of equations

$$\begin{aligned} \sum_{i=1}^{n} \left(y_i - f(\underline{x}_i;\ \underline{\beta}) \right) \frac{\partial f(\underline{x}_i;\ \underline{\beta})}{\partial \beta_1} &= 0 \\ \sum_{i=1}^{n} \left(y_i - f(\underline{x}_i;\ \underline{\beta}) \right) \frac{\partial f(\underline{x}_i;\ \underline{\beta})}{\partial \beta_2} &= 0 \\ &\vdots \\ \sum_{i=1}^{n} \left(y_i - f(\underline{x}_i;\ \underline{\beta}) \right) \frac{\partial f(\underline{x}_i;\ \underline{\beta})}{\partial \beta_K} &= 0 \end{aligned}$$

We can arrange this in the form

$$F(X;\underline{\beta})^T \left[\underline{y} - \underline{f}(X;\underline{\beta})\right] \;=\; \underline{0} \qquad (*)$$

where

$$F(X;\underline{\beta}) = \begin{bmatrix} \frac{\partial f(x_1;\underline{\beta})}{\partial\beta_1} & \frac{\partial f(x_1;\underline{\beta})}{\partial\beta_2} & \cdots & \frac{\partial f(x_1;\underline{\beta})}{\partial\beta_J} \\ \frac{\partial f(x_2;\underline{\beta})}{\partial\beta_1} & \frac{\partial f(x_2;\underline{\beta})}{\partial\beta_2} & \cdots & \frac{\partial f(x_2;\underline{\beta})}{\partial\beta_J} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(x_n;\underline{\beta})}{\partial\beta_1} & \frac{\partial f(x_n;\underline{\beta})}{\partial\beta_2} & \cdots & \frac{\partial f(x_n;\underline{\beta})}{\partial\beta_J} \end{bmatrix}; \; \underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}; \text{ and } \underline{f}(X;\underline{\beta}) = \begin{bmatrix} f(\underline{x}_1;\underline{\beta}) \\ \vdots \\ f(\underline{x}_n;\underline{\beta}) \end{bmatrix}$$

Equation $(*)$ gives us the "normal equations", analogous to

$$X^T(\underline{y} - X\underline{\beta}) \;=\; \underline{0} \qquad (**)$$

in multiple regression. However we can't solve $(*)$ in one operation by pre-multiplying by $(X^TX)^{-1}$ to get $\hat{\beta} = (X^TX)^{-1}X^T\underline{y}$ as in multiple regression.

Instead we solve "iteratively": Suppose after ℓ steps we our best guess for the parameters is $\underline{\beta}^{(\ell)}$, and we want to improve it to $\underline{\beta}^{(\ell+1)}$. We proceed as follows:

- Use Taylor's theorem to approximate $f(X;\underline{\beta}^{(\ell+1)})$ as

$$f(X;\underline{\beta}^{(\ell+1)}) \approx f(X;\underline{\beta}^{(\ell)}) + F(X;\underline{\beta}^{(\ell)})\left(\underline{\beta}^{(\ell+1)} - \underline{\beta}^{(\ell)}\right)$$

- Rewrite $(*)$ as

$$F(X;\underline{\beta}^{(\ell)})^T \left[\left(\underline{y} - f(X;\underline{\beta}^{(\ell)})\right) + F(X;\underline{\beta}^{(\ell)})\left(\underline{\beta}^{(\ell+1)} - \underline{\beta}^{(\ell)}\right)\right] \;\approx\; \underline{0}$$

  Note that this looks exactly like $(**)$, so that

$$\underline{\beta}^{(\ell+1)} - \underline{\beta}^{(\ell)} = \left(F(X;\underline{\beta}^{(\ell)})^T F(X;\underline{\beta}^{(\ell)})\right)^{-1} (F(X;\underline{\beta}^{(\ell)})^T \left(\underline{y} - f(X;\underline{\beta}^{(\ell)})\right) \qquad (***)$$

- Equivalently, we can obtain $\underline{\beta}^* = \underline{\beta}^{(\ell+1)} - \underline{\beta}^{(\ell)}$ by solving the least-squares regression problem

$$\underline{y}^* = X^*\underline{\beta}^* + \underline{\epsilon},$$

  where $\underline{y}^* = \underline{y} - f(X;\underline{\beta}^{(\ell)})$ and $X^* = F(X;\underline{\beta}^{(\ell)})$.

The ***Gauss-Newton method*** iterates $(***)$ until there is essentially no difference between $\underline{\beta}^{(\ell)}$ and $\underline{\beta}^{(\ell+1)}$.

The Gauss-Newton method requires

- Derivatives of $f(\underline{x}_i;\underline{\beta})$;
- Good starting values; and
- Some tuning of the step sizes in $(***)$

to work in practice, but most methods of estimating nonlinear regressions are based on variants of Gauss-Newton.

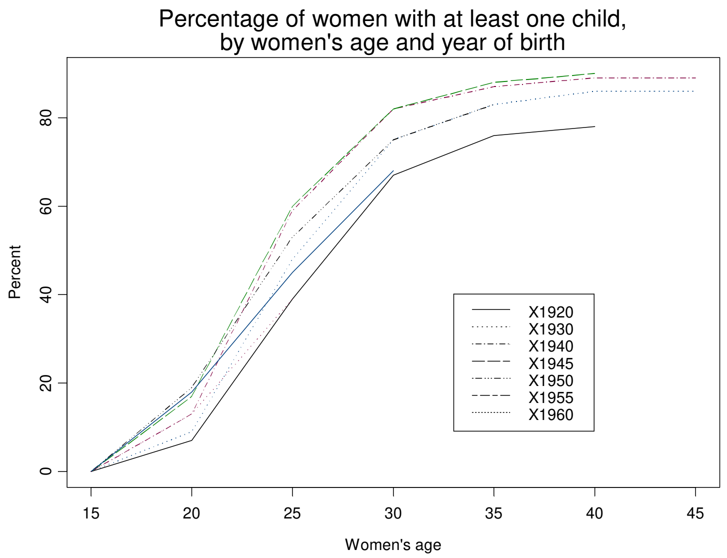
Comparison of Some Linear and Nonlinear Regression Facts.

Feature	Linear Regression	Nonlinear Regression
Model	$\underline{y} = X\underline{\beta} + \underline{\epsilon}$	$\underline{y} = \underline{f}(X; \underline{\beta}) + \underline{\epsilon}$
Normal equations	$X^T(\underline{y} - X\underline{\beta}) \; = \; \underline{0}$	$F(X; \underline{\beta})^T \; \left[\underline{y} - \underline{f}(X; \underline{\beta})\right] \; = \; \underline{0}$
Solution	$\hat{\underline{\beta}} = (X^T X)^{-1} X^T \underline{y}$	$\hat{\underline{\beta}} = \lim_{\ell\to\infty} \underline{\beta}^{(\ell)}$, where $\underline{\beta}^{(\ell+1)} \; =$ $\underline{\beta}^{(\ell)} \; + \; \left(F(X; \underline{\beta}^{(\ell)})^T F(X; \underline{\beta}^{(\ell)})\right)^{-1} (F(X; \underline{\beta}^{(\ell)})^T \left(\underline{y} - f(X; \underline{\beta}^{(\ell)})\right)$
Residual variance	$s^2 = \dfrac{\sum_i (y_i - \underline{x}_i \cdot \hat{\underline{\beta}})^2}{n - K}$	$s^2 = \dfrac{\sum_i (y_i - f(\underline{x}_i; \hat{\underline{\beta}}))^2}{n - K}$
$\widehat{\mathrm{Var}(\hat{\underline{\beta}})}$	$s^2 (X^T X)^{-1}$	$s^2 \left(F(X; \hat{\underline{\beta}})^T F(X; \hat{\underline{\beta}})\right)^{-1}$
R^2	$1 - \dfrac{\sum_i (y_i - \underline{x}_i \cdot \hat{\underline{\beta}})^2}{\sum_i (y_i - \overline{y})^2}$	$1 - \dfrac{\sum_i (y_i - f(\underline{x}_i; \hat{\underline{\beta}}))^2}{\sum_i (y_i - \overline{y})^2}$
t, F, SS's diagnostics	Based directly on X, $\underline{y}$.	Imitate linear regression with $X^* = F(X; \hat{\underline{\beta}})$, $\underline{y}$.

Example.

The following table lists British census data on the percentage of women with at least one child, arranged by women's age and birthyear. The graph shows curves for each birthyear, that is, each curve represents one column of the table (see RWG, pp. 168ff.). Women with the same birthyear are said to be in the same *birth cohort*.

Women's	Women's Birthyear						
Age	1920	1930	1940	1945	1950	1955	1960
15	0	0	0	0	0	0	0
20	7	9	13	17	19	18	13
25	39	48	59	60	53	45	39
30	67	75	82	82	75	68	NA
35	76	83	87	88	83	NA	NA
40	78	86	89	90	NA	NA	NA
45	NA	86	89	NA	NA	NA	NA



The SPLUS code that generated the plot above is given next. The asymmetric growth over time in each curve suggests we fit the curves using the Gompertz model.

```
> child _ read.table("child.dat",header=T)
> names(child)
[1] "Age"    "X1920" "X1930" "X1940" "X1945" "X1950" "X1955" "X1960"
> matplot(child$Age,as.matrix(child)[,-1],type="l")
> legend(locator(),legend=names(child)[-1],lty=1:7)
> title("Percentage of women with at least one child,\nby women’s age \
and year of birth")
> title(xlab="Women’s age",ylab="Percent")
```

The Gompertz model looks like

$$y_i = f(x_i; \alpha, \beta, \gamma) + \epsilon_i$$

where

$$f(x; \alpha, \beta, \gamma) = \alpha e^{-\gamma e^{-\beta x_i}}.$$

We can fit this to each birth cohort (that is, to the curve representing women from the same birth year), using the SPLUS function `nls`, for "nonlinear least squares".

Let us start with the 1940 birth cohort. First we need decent starting values. Taking logs, rearranging, and taking logs again, we get

$$\begin{aligned} \log f(x) &= \log \alpha - \gamma e^{-\beta x} \\ \beta x &= \log \gamma - \log(\log(\alpha / f(x)) \end{aligned}$$

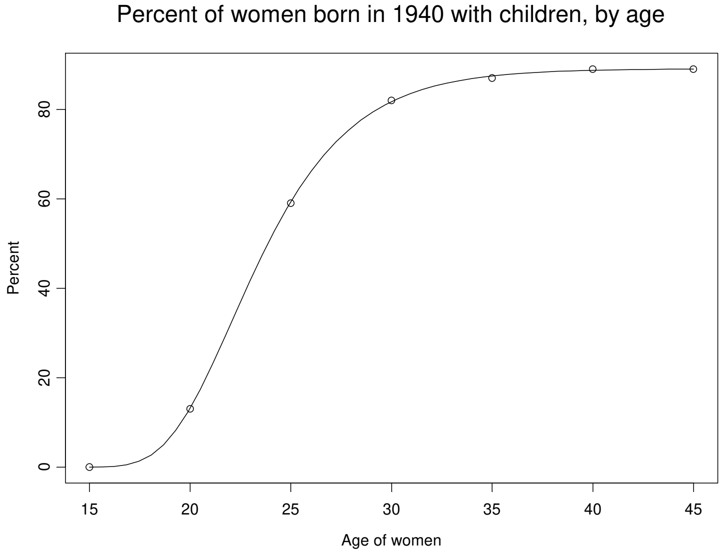
Letting $x \to \infty$ in the first equation, we see that α is the maximum value for $f(x)$, $\alpha \approx 90$ for the 1940 data. If we take $x = 0$ in the second equation to get rid of β, we discover γ should be quite large (since $f(0) \approx 0$ in the graphs); taking a cue from the text[1], we will take $\gamma = 1000$. Now put $x = 30$, so $f(30) = 82$ according to the table, and look at the second equation again. We get $\beta = (\log(1000) - \log\log(90/82))/30 \approx 0.31$.

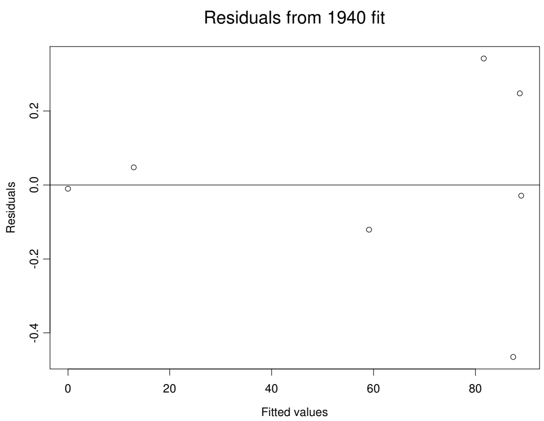
Now we are ready to estimate the first curve using SPLUS:

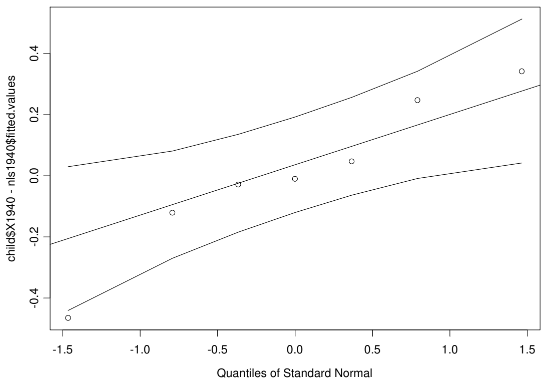
```
> nls1940 _ nls(X1940 ~ alpha * exp(-gamma*exp(-beta*Age)),
+ data=child,start=list(alpha=90,beta=0.31,gamma=1000),trace=T)
4.50236 : 90 0.31 1000
0.4183 : 89.1071 0.309252 933.509
0.412138 : 89.1039 0.30956 941.892
0.412135 : 89.1039 0.309562 941.97
> summary(nls1940)

Formula: X1940 ~ alpha * exp( - gamma * exp( - beta * Age))

Parameters:
           Value  Std. Error  t value
alpha  89.103900  0.19578900 455.1010
 beta   0.309562  0.00358636  86.3165
gamma 941.970000 75.35320000  12.5007

Residual standard error: 0.320989 on 4 degrees of freedom

Correlation of Parameter Estimates:
       alpha   beta
 beta -0.508
gamma -0.464  0.993
> 1 - ( (0.320989)^2 * 4 ) / sum((child$X1940-mean(child$X1940))^2)
[1] 0.9999527  # = R^2 --- wow, that's pretty good!
>
```

[1]Since I don't have much experience with the Gompertz model, it would have been hard for me to see how large γ should have been, and it really seems to matter for this problem!

```
> # Make some plots of the fit...
>
> plot(child$Age,child$X1940,xlab="",ylab="")
> x _ seq(15,45,length=50)
> lines(x,89.1*exp(-942*exp(-.31*x)))
> title("Percent of women born in 1940 with children, by age",
+ xlab="Age of women", ylab="Percent")
> plot(nls1940$fitted.values,child$X1940-nls1940$fitted.values,xlab="",ylab="")
> abline(h=0)
> title("Residuals from 1940 fit",xlab="Fitted values",ylab="Residuals")
> qqn(child$X1940-nls1940$fitted.values)
```







Breakout session

Here is the code for fitting the 1945 birth cohort data, using the 1940 parameter estimates as starting values.

```
> child45 _ na.omit(child[,c("Age","X1945")])
> nls1945 _ nls(X1945 ~ alpha * exp(-gamma*exp(-beta*Age)),
+ data=child45, start=list(alpha=89.1,beta=0.31,gamma=942), trace=T)
16.9186 : 89.1 0.31 942
5.01028 : 89.7436 0.294916 590.214
1.98967 : 90.398 0.280863 443.121
0.121371 : 90.4283 0.281552 465.866
0.118423 : 90.4253 0.281703 468.054
0.118423 : 90.4253 0.281703 468.057
> summary(nls1945)

Formula: X1945 ~ alpha * exp( - gamma * exp( - beta * Age))

Parameters:
           Value  Std. Error  t value
alpha  90.425300  0.16066900 562.8050
 beta   0.281703  0.00221843 126.9830
gamma 468.057000 22.54640000  20.7597

Residual standard error: 0.198681 on 3 degrees of freedom

Correlation of Parameter Estimates:
       alpha   beta
 beta -0.634
gamma -0.587  0.993

> 1 - ( (??????)^2 * ?? ) / sum(child45$X1945-mean(child45$X1945)^2)
[1] ????????  # = R^2
>
> # Make some plots of the fit...
>
> plot(child45$Age,child45$X1945,xlab="",ylab="")
> x _ seq(15,45,length=50)
> lines(x,??????)
> title("Percent of women born in 1945 with children, by age",
+ xlab="Age of women", ylab="Percent")
> plot(nls1945$fitted.values,child45$X1945-nls1945$fitted.values,
+ xlab="",ylab="")
> abline(h=0)
> title("Residuals from 1945 fit",xlab="Fitted values",ylab="Residuals")
> qqn(child45$X1945-nls1945$fitted.values)
```

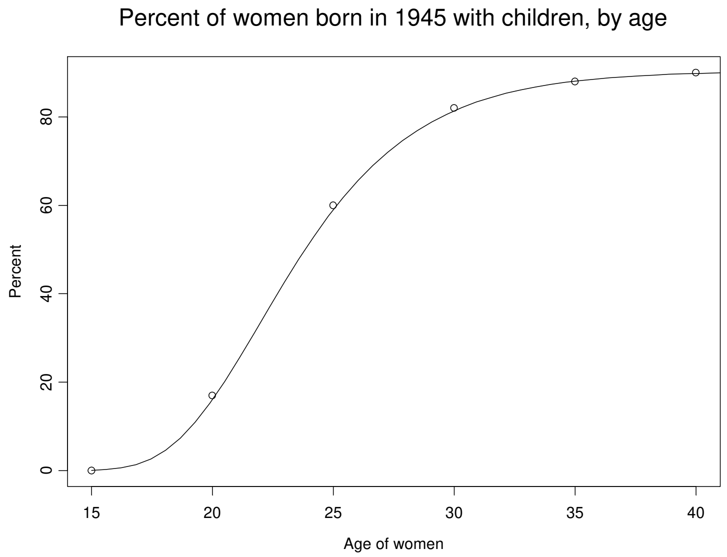
1. Explain *carefully* what `child45` is, what it contains, and why it had to be made.

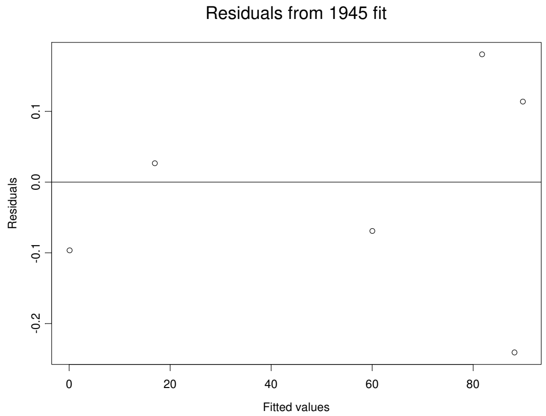
2. Comment on the fitted nonlinear regression: How many iterations were required? Are all the parameters estimated to be (significantly) different from zero? Which parameter is most poorly estimated? Are there any problems analogous to collinearity in this fit?

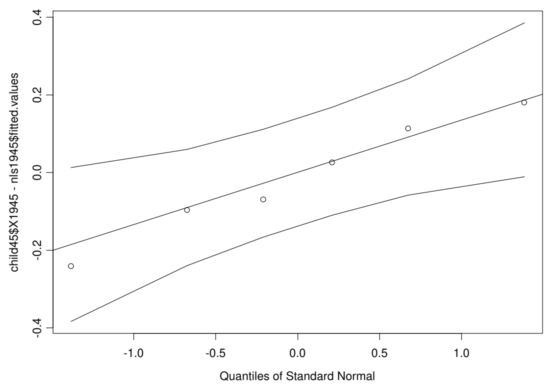
3. If we had used the starting values `start=list(alpha=1, beta=1, gamma=1)`, the `nls()` function would have stopped with the message `Problem in nls(X1945 ~ alpha * exp( - gamma * exp( -..:  singular gradient matrix`. What is going wrong?

4. The value for `sum(child45$X1945-mean(child45$X1945)^2)` is approximately 7529. Fill in the "?????" and the "??" in the line that calculates R^2, and then calculate R^2. Comment on this value.

5. Fill in the "??????" in the `lines()` command that makes the plot of the fitted Gompertz curve for the 145 birth cohort.







6. Do the plots look good? What kind of plot could you make, to explore whether there are other local minima besides the ones `nls()` found?

7. The logistic function for nonlinear regression is

$$f(x;\alpha.\beta,\gamma) = \frac{\alpha}{1+e^{-\beta(x-\gamma)}}$$

(a) Do you expect the logistic fit to be better or worse than the Gompertz fit? Why or why not?

(b) If you have time, write out the `nls()` command you would use to fit the logistic function to the 1945 data (including starting values, if you can). *Hint for starting values:*

$$\log\left(\frac{f(x)/\alpha}{1-f(x)/\alpha}\right) = \beta(x-\gamma)$$