Nonlinear Regression

Let $\underline{y}$ be the vector of response variables, X the matrix of predictor variables, and $\underline{\beta}$ be the vector of parameters. Here are some rough definitions (no one will disagree with them as far as they go, but they could go farther):

A statistical model for $(X, \underline{y})$ is

- *linear* if
$$E[\underline{y}] = X\underline{\beta}$$

- *intrinsically linear* or *linearizable* if there are transformations $(\underline{y}^*, \underline{X}^*, \underline{\beta}^*)$ of $(\underline{y}, X, \underline{\beta})$ such that for some
$$E[\underline{y}^*] = X^*\underline{\beta}^*$$

- *nonlinear* or *intrinsically nonlinear* if no such transformations exist.

Note: these are essentially definitions of *linear in the parameters* or *nonlinear in the parameters*. A linearizable model, for example, can be very nonlinear in X!

<u>*Examples of linearizable models:*</u>

- *ANOVA models* (the "predictor variables" are factors with discrete categorical values, the X^* matrix consists of columns of dummy variables coding levels of the factors).

- Models with *interactions*

- Models with *polynomial terms*

- Models in which $\underline{y}$ and/or columns of X have been transformed to improve the appearance of residual plots (to better match the conventional assumptions of linear regression, to reduce high-leverage cases, etc.).

- Models that are nonlinear in their "natural" form but can be transformed to linearity. For example:

	"Natural form"	Transformation	"Linearized form"
Exponential	$y = \alpha e^{-\beta x}$	$y' = \log(y)$	$y' = \alpha' - \beta x$
Logistic	$y = 1/(1 + e^{-}\beta x)$	$y' = \log[y/(1-y)]$	$y' = \beta x$
Inverse	$y = 1/(\beta_0 + \beta_1 x)$	$y' = 1/y$	$y' = \beta_0 + \beta_1 x$

Aside:

A class of models closely related to the intrinsically linear models are *generalized linear models (glm's)*. A model is a *generalized linear model* if there is a function $L(u)$ such that

$$L(E[\underline{y}]) = X\underline{\beta}$$

Examples include logistic regression, poisson regression, log-linear models, etc. We will talk about these models at least briefly later on in the course.

End of aside.

Polynomial regression

Sometimes we want to fit a response surface about which we have little prior theory. In these cases we may want to try to fit a *polynomial regression model*, of the form:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} \\ &\quad + \beta_{11} X_{i1}^2 + \ldots + \beta_{kk} X_{ik}^2 \\ &\quad + \beta_{12} X_{i1} X_{i2} + \ldots + \beta_{(k-1)k} X_{i(k-1)} X_{ik} \\ &\quad + \beta_{111} X_{i1}^3 + \cdots \\ &\quad + \beta_{123} X_{i1} X_{i2} X_{i3} + \cdots + \text{etc.} \end{aligned}$$

Reasons:

- By Taylor's theorem, any sufficiently smooth surface is "locally" approximated by a polynomial.
- By a standard result in advanced calculus, a continuous function can be approximated uniformly well on a compact set with a polynomial.

Possible strategy:

- Start with approximately third-degree terms (that is, terms like X_{ij}^3 or $X_{i4}^2 X_{i6}$ whose powers add up to three). *(Nothing magic about 3^{rd} degree...)*
- "Step down" through the degrees of the terms, first eliminating any "insignificant" terms of 3^{rd} degree, then any "insignificant" terms of degree 2, etc.
- In every model considered, retain all "lower order terms" that are contained in the higher-order terms still in the model *(principle of marginality)*.
- The "essential structure" of the final model is often clarified by
  - Graphing "slices" in which all but one X are fixed;
  - Considering partial derivatives of Y with respect to each X

  Example:
  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \epsilon$$

Comments:

- Why retain lower order terms?
  - Interpretation of coefficients changes when lower-order terms are removed.
  - Locations of critical points, zeros of the polynomial, etc. are affected by taking out lower order terms.

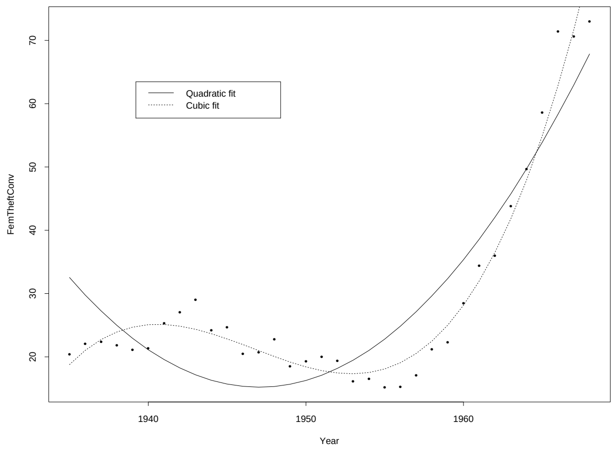
  Example:
  $x^2 - 4x + 3 = (x-1)(x-3) = (x-2)^2 - 1$ has roots at $x = 1, 3$ and a critical point at $x = 2$, and minimum $y = -1$. But $x^2 + 3$ has no real roots, a critical point at $x = 0$ and minimum $y = 3$.
- Possible problem: X, X^2, X^3, ..., can be very highly correlated—collinearity problems! Can be solved using "orthogonal polynomials".
  - Apply Gram-Schmidt process to the X matrix of the "natural" polynomials for your model
  - Use a standard set of orthogonal polynomials (cf. Abramowitz and Stegun, etc.); `poly(x,d)` generates a set of orthogonal polynomials of degree d for term x in `lm()`.

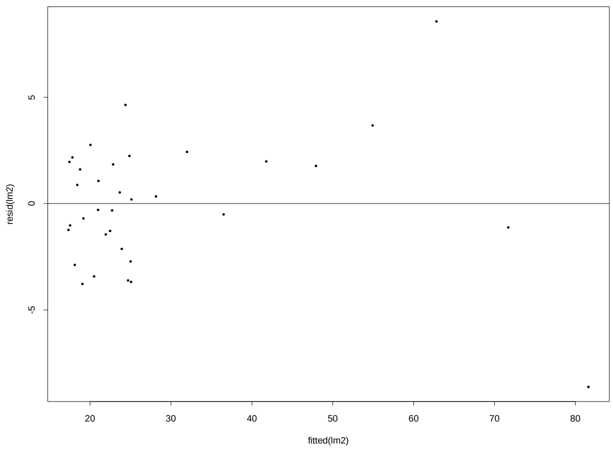
Example:

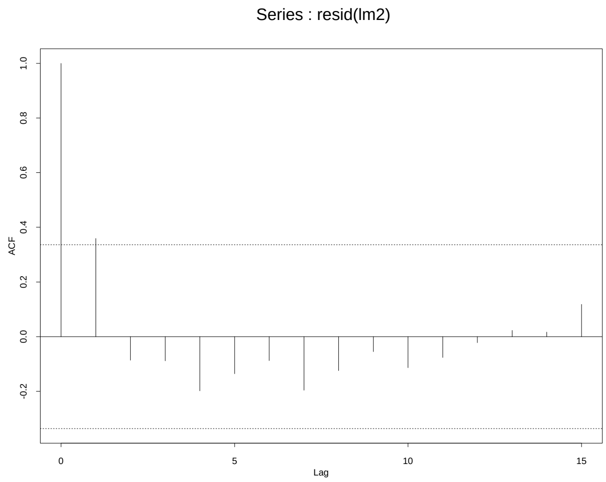
Using the data on Female theft conviction rates in Canada that we studied using GLS models, we try to fit just a polynomial predicting theft convictions from year:

```
> hart _ read.table("Hartnagl.dat",na.strings="?")
> names(hart) _ scan(what="")
1: Year
2: Fert
3: Partic
4: Coll
5: FemAllConv
6: FemTheftConv
7: MalAllConv
8: MalTheftConv
9:
> attach(hart)
> lm1 _ lm(FemTheftConv ~ Year + Year^2 + Year^3,na.action=na.omit)
Problem in lm.fit.qr(unset(x), unset(y)): computed fit is singular, rank 3
Use traceback() to see the call stack
> lm1 _ lm(FemTheftConv ~ Year + Year^2 ,na.action=na.omit)
> summary(lm1)

Call: lm(formula = FemTheftConv ~ Year + Year^2, na.action = na.omit)
Residuals:
    Min     1Q Median    3Q  Max
 -12.14 -5.748 0.1385 5.405 13.1

Coefficients:
                   Value   Std. Error      t value     Pr(>|t|)
(Intercept)  453895.1867   54829.4485       8.2783       0.0000
       Year    -466.2280      56.1931      -8.2969       0.0000
  I(Year^2)       0.1197       0.0144       8.3160       0.0000

Residual standard error: 7.218 on 31 degrees of freedom
Multiple R-Squared: 0.8199
F-statistic: 70.56 on 2 and 31 degrees of freedom, the p-value is 2.888e-12
4 observations deleted due to missing values

Correlation of Coefficients:
          (Intercept) Year
     Year -1
I(Year^2)  1          -1
> lm2 _ lm(FemTheftConv ~ poly(Year,3) ,na.action=na.omit)
> summary(lm2)

Call: lm(formula = FemTheftConv ~ poly(Year, 3), na.action = na.omit)
Residuals:
    Min   1Q  Median    3Q   Max
 -8.618 -1.4 -0.0499 1.929 8.568

Coefficients:
                  Value Std. Error t value Pr(>|t|)
   (Intercept) 27.1191  0.6523    41.5731  0.0000
poly(Year, 3)1 71.3393  4.9893    14.2985  0.0000
poly(Year, 3)2 45.9030  5.1494     8.9143  0.0000
poly(Year, 3)3 53.4591  4.7213    11.3229  0.0000
```

```
Residual standard error: 3.195 on 30 degrees of freedom
Multiple R-Squared: 0.9658
F-statistic: 282.8 on 3 and 30 degrees of freedom, the p-value is 0
4 observations deleted due to missing values

Correlation of Coefficients:
               (Intercept) poly(Year, 3)1 poly(Year, 3)2
poly(Year, 3)1 -0.4662
poly(Year, 3)2  0.4741     -0.6010
poly(Year, 3)3 -0.4362      0.5551        -0.5728
> plot(Year,FemTheftConv)
> lines(Year[!is.na(FemTheftConv)],fitted(lm1),lty=1)
> lines(Year[!is.na(FemTheftConv)],fitted(lm2),lty=2)
> legend(locator(),legend=c("Quadratic fit","Cubic fit"),lty=1:2)
> plot(fitted(lm2),resid(lm2))
```







There is still substantial autocorrelation in the residuals, but at least some of the "structure" (V-shape) in the residuals from earlier years when we fitted the GLS model has been accounted for by the cubic fit.

Nonlinear regression

Linear regression models are models of the form

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}$$

where $\underline{\epsilon}$ has mean zero and (hopefully) equal variances and zero covariances. Usually of course $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

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 $f(\cdot; \underline{\beta})$.

When $\underline{\epsilon}$ is normal, the likeihood is again

$$\begin{aligned} L(\underline{y}; \underline{\beta}, \sigma_\epsilon^2) &= \prod_1^n \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left[-\frac{1}{2}(y_i - f(\underline{x}_i;\ \underline{\beta}))^2/\sigma_\epsilon^2\right] \\ &= \frac{1}{(2\pi)^{n/2}|\sigma^2 I|^{1/2}} \exp\left[-\frac{1}{2}(\underline{y} - \underline{f}(X;\ \underline{\beta}))^T(\sigma_\epsilon^2 I)^{-1}(\underline{y} - \underline{f}(X;\ \underline{\beta}))\right] \end{aligned}$$

(where $\underline{f}(X;\ \underline{\beta}) = [f(\underline{x}_1;\ \underline{\beta}), f(\underline{x}_2;\ \underline{\beta}), \ldots, f(\underline{x}_n;\ \underline{\beta})]^T$); and least squares and maximum likelihood yield the same estimation results.

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.

Linearizable regression or nonlinear regresion??

$$y_i = \alpha e^{-}\beta x_i \cdot \epsilon_i \qquad \text{vs.} \qquad y_i = \alpha e^{-}\beta x_i + \epsilon_i$$

What does the residual pattern look like??

Estimation ideas…

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

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

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

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

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

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

- 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.

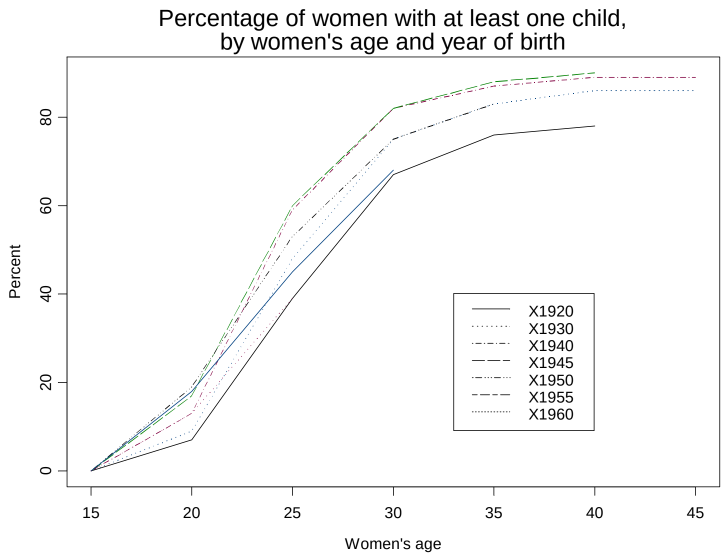
Comparison of Some Linear and Nonlinear Regression Facts.

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	$\underline{\hat{\beta}} = (X^TX)^{-1}X^T\underline{y}$	$\underline{\hat{\beta}} = \lim_{\ell\to\infty} \underline{\beta}^{(\ell)}$, 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	$s^2 = \dfrac{\sum_i (y_i - \underline{x}_i \cdot \underline{\hat{\beta}})^2}{n-K}$	$s^2 = \dfrac{\sum_i (y_i - f(\underline{x}_i; \underline{\hat{\beta}}))^2}{n-K}$
$\widehat{\mathrm{Var}}(\underline{\hat{\beta}})$	$s^2(X^TX)^{-1}$	$s^2 \left(F(X; \underline{\hat{\beta}})^T F(X; \underline{\hat{\beta}})\right)^{-1}$
R^2	$1 - \dfrac{\sum_i (y_i - \underline{x}_i \cdot \underline{\hat{\beta}})^2}{\sum_i (y_i - \overline{y})^2}$	$1 - \dfrac{\sum_i (y_i - f(\underline{x}_i; \underline{\hat{\beta}}))^2}{\sum_i (y_i - \overline{y})^2}$
t, F, SS's diagnostics	Based directly on X, $\underline{y}$.	Imitate linear regression with $X^* = F(X; \underline{\hat{\beta}})$, $\underline{y}$.

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. 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

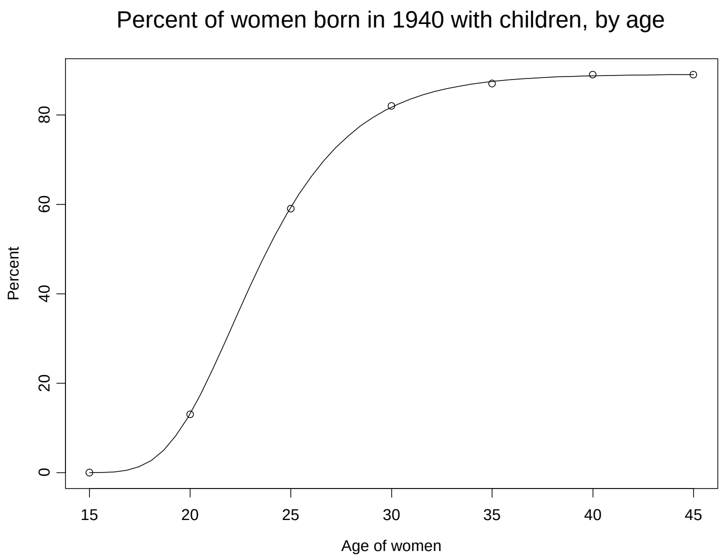
$$\begin{aligned} \log f(x) &= \log \alpha - \gamma e^{-\beta x} \\ \beta x &= \log \gamma - \log(\log(\alpha / f(x)) \end{aligned}$$

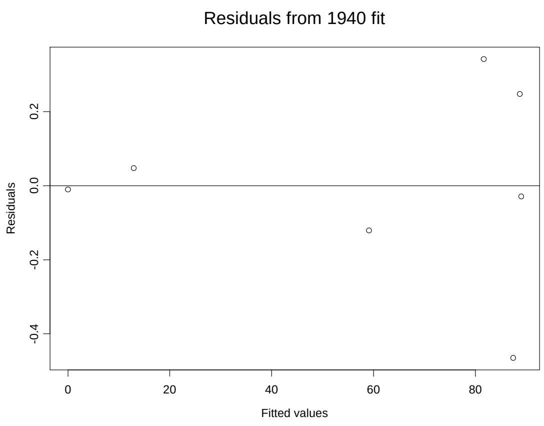
Letting $x \to \infty$ in the first equation, we see that α is the maximum value for $f(x)$, $\alpha \approx 90$ for the 1940 data. If we take $x = 0$ in the second equation to get rid of β, we discover γ should be quite large (since $f(0) \approx 0$ in the graphs); after some trial and error, we take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

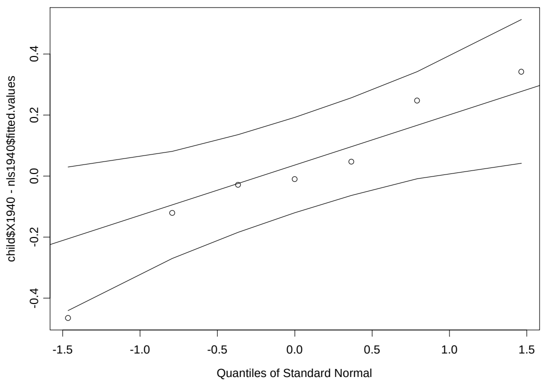
Now we are ready to estimate the first curve using SPLUS:

```
> 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!
>
> # 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)
```

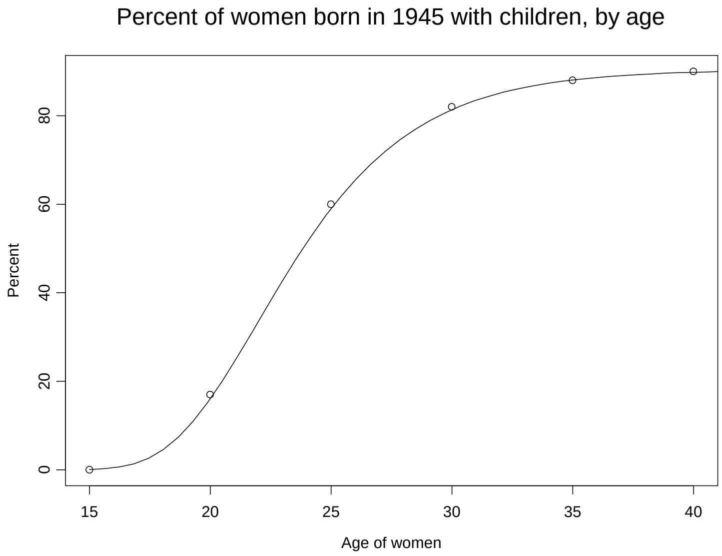


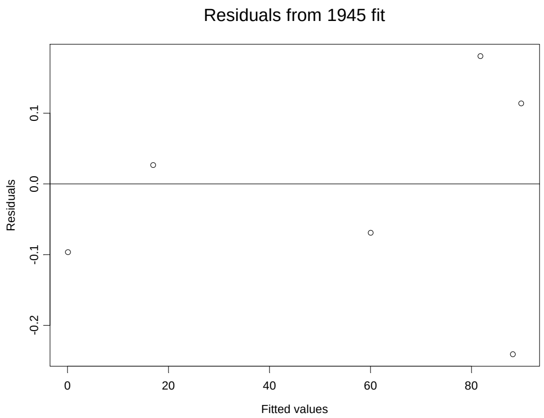


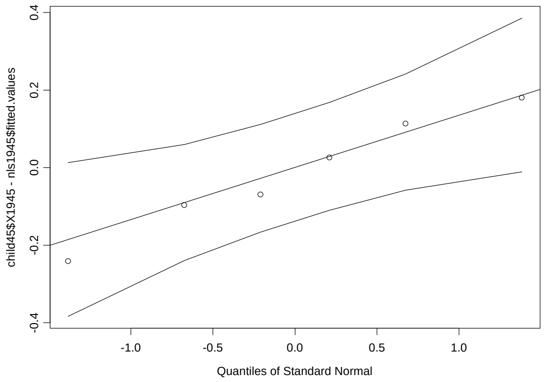


Here is the code for fitting the 1945 birth cohort data, using the 1940 parameter estimates as starting values.

```
> 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 - ( (0.198681)^2 * 3 ) / sum(child45$X1945-mean(child45$X1945)^2)
[1] 0.9999843 # = R^2
>
> # Make some plots of the fit...
>
> plot(child45$Age,child45$X1945,xlab="",ylab="")
> x _ seq(15,45,length=50)
> lines(x,90.4*exp(-468*exp(-.28*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)
```


Percent of women born in 1945 with children, by age
Percent
0
20
40
60
80
15
20
25
30
35
40
Age of women


Residuals from 1945 fit
Residuals
-0.2
-0.1
0.0
0.1
0
20
40
60
80
Fitted values


child45$X1945 - nls1945$fitted.values
-0.4
-0.2
0.0
0.2
0.4
-1.0
-0.5
0.0
0.5
1.0
Quantiles of Standard Normal