36-401 Class Notes and Exercise for 10/26/00

Non-linear Regression (RwG pp. 163-173)

Howard Seltman

- Curvilinear regression vs. non-linear regression

- Non-linear exponential models

$$Y_i = \alpha e^{\beta X_i} + \epsilon_i$$

- Negative exponential models (rise to asymptote with $\alpha, \beta > 0$)

$$Y_i = \alpha(1 - e^{-\beta X_i}) + \epsilon_i$$

- Two term exponential models (rise and fall with $0 < \theta_2 < \theta_1$)

$$Y_i = \frac{\theta_1}{\theta_1 - \theta_2}\left(e^{-\theta_2 X_i} - e^{-\theta_1 X_i}\right) + \epsilon_i$$

- (Three parameter) Logistic models (symmetric S shape)

$$Y_i = \frac{\alpha}{(1 + \gamma e^{-\beta X_i})} + \epsilon_i$$

- Gompertz models (asymmetric S shape)

$$Y_i = \alpha e^{-\gamma e^{-\beta X_i}} + \epsilon_i$$

- Minimizing residual sum of squares requires iterative techniques. Need starting values. May not converge. May converge to local minimum.

- R^2 defined as $1 - \frac{\text{RSS}}{\text{cTSS}}$. May be negative.

- Under normal i.i.d. errors, estimates and tests are *asymptotically* correct. Use the usual graphs to check model assumptions.

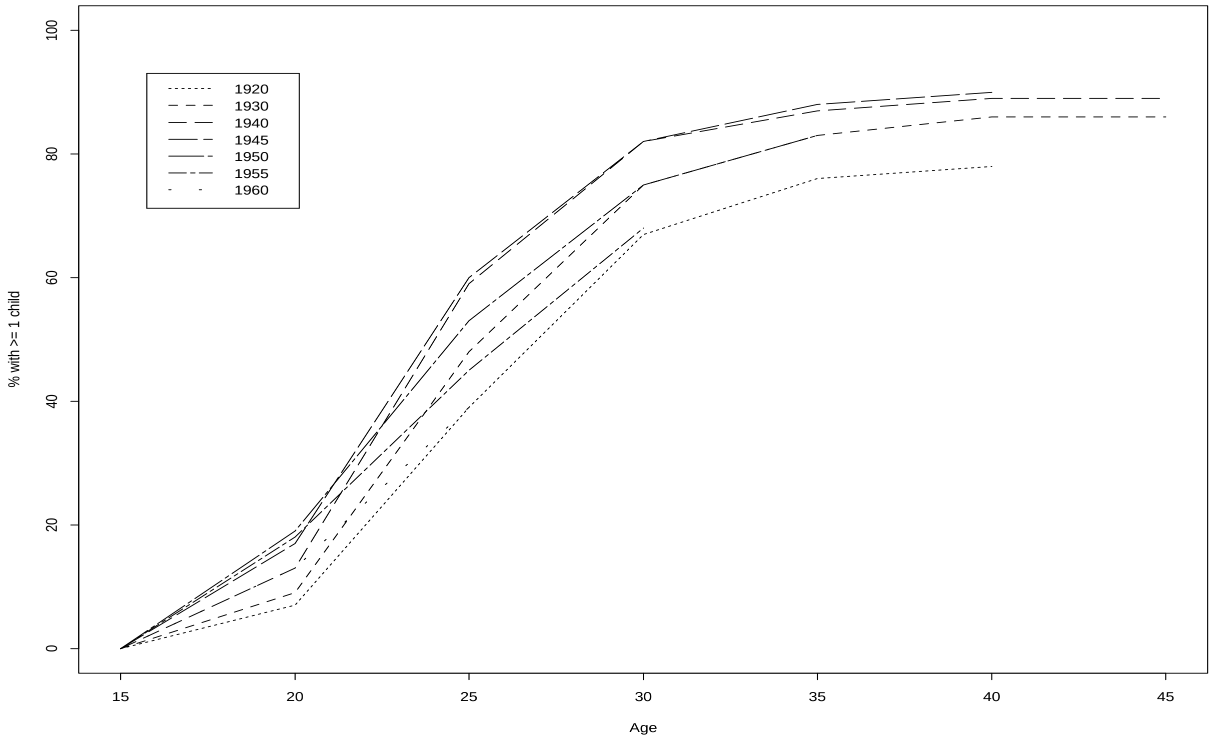
- Splus: 1) turn a data.frame into a "parameterized data.frame" using parameters(df)=list(parm1=c1, parm2=c2) to name the parameters and set starting values; 2) get estimates of the parameters using nls(formula, pdf) where formula is the non-linear formula in Splus syntax and pdf is the parameterized data.frame containing the data and the parameter starting values; 3) use summary() to get asymptotic standard errors and tests. Note that R^2 is not calculated and that we have $param instead of $coef.

Breakout Session for 10/26/00

Here we fit a Gompertz model to the 1940 column of the following dataframe, named "child" (RwG p. 168). The table tells us that e.g. by age 30, 82% of women born in 1940 had at least one child.

	Age	X1920	X1930	X1940	X1945	X1950	X1955	X1960
1	15	0	0	0	0	0	0	0
2	20	7	9	13	17	19	18	13
3	25	39	48	59	60	53	45	39
4	30	67	75	82	82	75	68	NA
5	35	76	83	87	88	83	NA	NA
6	40	78	86	89	90	NA	NA	NA
7	45	NA	86	89	NA	NA	NA	NA

1. Here is a plot of all 7 birth cohorts:



Why is Gompertz a reasonable model? What mathematic problem prevents a perfect fit to these data?

2. We change the data.frame into a parameterized data.frame with
```
parameters(child)_list(alpha=c1, beta=c2, gamma=c3)
```
What are reasonable choices for c1, c3, and c2, i.e. what are your best guesses for values that approximately fit the data?

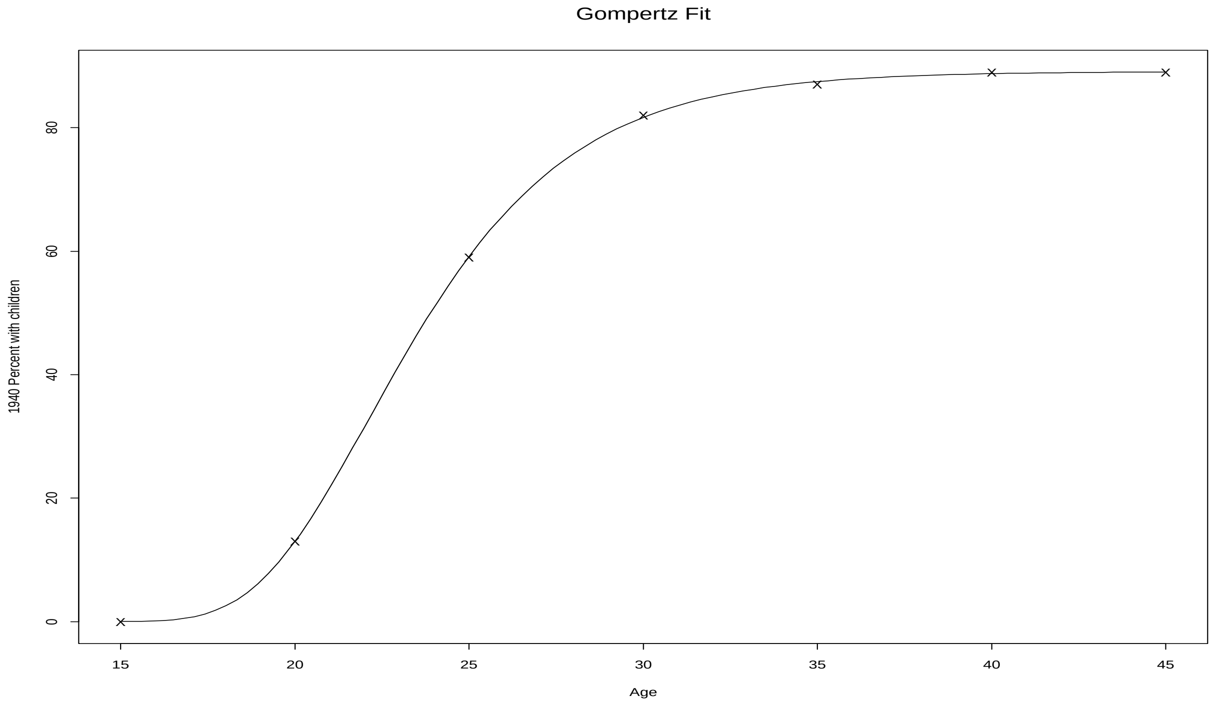
3. chnls_nls(X1940~alpha*exp(-gamma*exp(-beta*Age)), child) is the command to fit a Gompertz model to the 1940 data. What command would fit a logistic curve to the 1945 data?

4. "chnls" is as follows. Are the parameters as expected? How can we interpret "Residual sum of squares"?
```
Residual sum of squares : 0.4121352
parameters:
    alpha   gamma      beta
 89.10387 941.971 0.3095621
formula: X1940 ~ alpha * exp( - gamma * exp( - beta * Age))
7 observations
```

5. Starting values of list(alpha=1, beta=1, gamma=1) give an error (singular gradient matrix). If you know something about finding minima/maxima, take a guess as to what this means.

6. A fitted vs actual plot is drawn using

```
x_seq(15,45,len=100)
cf_chnls$param
y_???
plot(child$Age, child$X1940, xlab="Age", ylab="1940 Percent with children",
  main="Gompertz Fit", pch="X")
lines(x,y)
```



Fill in the missing code. Does the plot look good? What does this tell you about local minima?

7. What diagnostic plots would you do?

8. If you have time, write the complete steps to try a logistic fit.