Statistics 520 - Assignment 6

Sam Olson

The Model

Define random variables $Y_1, \ldots, Y_n$ connected with the growth of some organism, observed at times $t_1, \ldots, t_n$. Assume the response variables are independent and follow the model:

$$Y_i = \mu_i + \sigma\varepsilon_i,$$

$$\mu_i = B + A\exp\Big[-\exp\{-k(t_i - T)\}\Big],$$

$$\varepsilon_i \overset{\text{iid}}{\sim} N(0,1).$$

This model describes a sigmoidal curve with both upper and lower asymptotes, but is not constrained to be symmetric about its inflection point in the way a logistic curve is.

Parameters

- A: distance between upper and lower asymptotes
- B: lower asymptote
- k: growth rate (slope of the linear portion of the curve)
- T: the time at which the inflection point occurs

Data

On the course web page in the Data module is a file `growthdat.txt` that contains data simulated from the Gompertz regression model. The variable names in this file are `x` and `y`, where:

- `x` = time of observation
- `y` = growth in appropriate units

Q1

Find generalized least squares estimates of the parameters A, B, k, T, and the associated moment-based estimate of σ^2. Describe how you determined effective starting values for the estimation procedure. Present a scatterplot along with the fitted expectation function.

Answer

As given, the form of four-parameter Gompertz model is:

$$\mu_i = B + A \exp\Big[- \exp\{-k(t_i - T)\}\Big]$$

Again, as given, the parameters are:

- (A): distance between the upper and lower asymptotes,
- (B): lower asymptote,
- (k): growth rate (slope at the inflection point),
- (T): time of inflection,

To determine "effective" starting values, we check the underlying data visually and descriptively, i.e., we intuited the values inductively. Also, since there is no reasonable basis to suppose numbers must be discrete, the observed values were largely used as starting values (treated observations as realizations of continuous R.V.s).

The values used (and how they were determined) are as follows:

- Lower asymptote (B): The min observed y-value is 0.948, so set ($B \approx 0.95$).
- Distance between the upper and lower asymptotes (A): The max observed y-value is 233.37. So, to get A we take the (observed) range ($A \approx 232.42$).
- Time of Inflection (T): From visual inspection, the curve rises steeply from t = 10 to t = 20; taking the midpoint of this period as an initial starting point for "inflection" ($T \approx 15$), knowing that the optimization method, particularly a Newton-type optimizer, will more accurately determine the "direction" to proceed for maximization.
- Growth rate (k): A more involved calculation The steepest rise in the data occurs between t=12 and t=18. Case in point, between t=13 ($y \approx 64$) and t=17 ($y \approx 140$), the increase is about 76 units over 4 time units, or roughly 19 units per time unit.

For the Gompertz model, the slope at the inflection point ($t = T$) is obtained by differentiation:

$$\left.\frac{d\mu}{dt}\right|_{t=T} = \frac{Ak}{e}$$

(Where "e" denotes Euler's number, not Euler's constant)

Equating this theoretical slope with the observed slope (≈ 19) and using $A \approx 232$ from prior determination, we obtain the starting value

$$k \approx \frac{19e}{232} \approx 0.22$$

Taken together, the starting parameter vector is

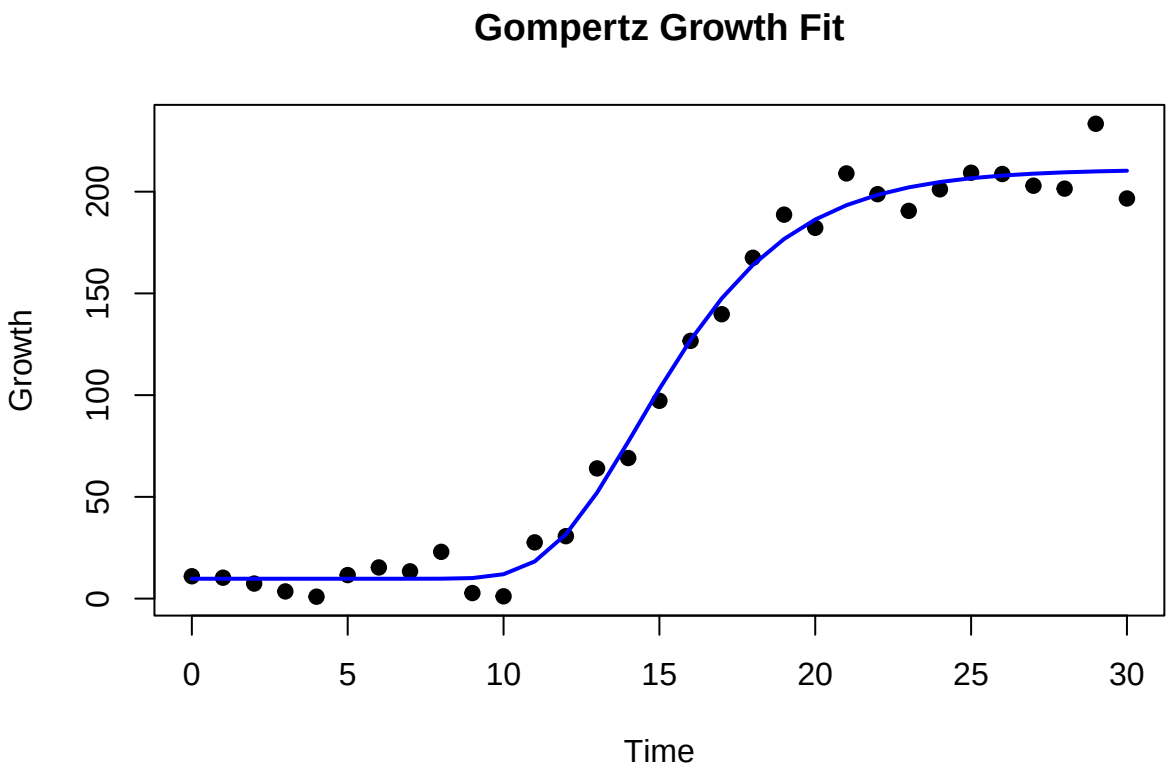
$$\theta^{(0)} = (A = 232.42, B = 0.95, k = 0.22, T = 15)$$

Actual Computation yields:

Convergence met with estimates:

$\hat{A} = 201.30$ $\hat{B} = 9.76$ $\hat{k} = 0.35$ $\hat{T} = 14.26$ $\hat{\sigma}^2 = 85.57$

Plotting the observed vs. estimated values:



Q2

Compute 95% approximate confidence intervals for the parameters of the expectation function (A, B, k, T).

Answer

The `nonlin` function returns the parameter estimates ($\hat{\boldsymbol{\beta}} = (\hat{A}, \hat{B}, \hat{k}, \hat{T})$) and an estimated covariance matrix ($\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \mathbf{V}$)

Then, following the standard convention, the formula for the 95% (Wald-based approximate) confidence intervals for each parameter ($\beta_j \in A, B, k, T$) are:

$$\hat{\beta}_j \ \pm \ z_{0.975}, \mathrm{SE}(\hat{\beta}_j)$$

param	estimate	SE	CI_L	CI_U
A	201.296	4.974	191.546	211.045
B	9.763	2.809	4.258	15.268
k	0.354	0.032	0.290	0.417
T	14.257	0.195	13.875	14.639

Corresponding to 95% approximate confidence intervals of:

- (A: $201.296 \pm 1.96(4.974) \ \Rightarrow \ (191.546, \ 211.045)$)
- (B: $9.763 \pm 1.96(2.809) \ \Rightarrow \ \ (4.258, \ \ 15.268)$)
- (k: $0.354 \pm 1.96(0.032) \ \Rightarrow \ \ (0.290, \ \ 0.417)$)
- (T: $14.257 \pm 1.96(0.195) \ \Rightarrow \ (13.875, \ 14.639)$)

Q3

Compute pairwise correlations between $\hat{A}, \hat{B}, \hat{k}, \hat{T}$.

Answer

The `nonlin()` output includes the estimated covariance matrix of the parameter estimates (`out$covb`). From this matrix we can derive the correlations:

$$\text{Corr}(\hat{\theta}_i, \hat{\theta}_j) = \frac{\widehat{\text{Cov}}(\hat{\theta}_i, \hat{\theta}_j)}{\sqrt{\widehat{\text{Var}}(\hat{\theta}_i)}; \sqrt{\widehat{\text{Var}}(\hat{\theta}_j)}}$$

for parameters ($\theta_i, \theta_j \in \{A, B, k, T\}$).

Computing these quantities, using the `cov2cor` function from the base `stats` package:

	A	B	k	T
A	1.000	-0.639	-0.632	-0.029
B	-0.639	1.000	0.243	0.439
k	-0.632	0.243	1.000	0.104
T	-0.029	0.439	0.104	1.000

Q4

Two quantities of interest to scientists are called the maximum relative growth rate and the maximum absolute growth rate. These quantities are related to the slope of the growth curve at the inflection point and give the per time unit increase in growth relative to the upper asymptote and in absolute scale at that time point. The maximum absolute growth rate is defined as,

$$k_{\text{abs}} = \frac{k(A+B)}{\exp(1)}$$

A plug-in estimate of k_{abs} is then

$$\hat{k}_{\text{abs}} = \frac{\hat{k}(\hat{A}+\hat{B})}{\exp(1)}$$

and note that k_{abs} is an absolute function of its components. Given this, compute a 90% approximate confidence interval for k_{abs} (note a 90% interval – I get tired of using 95% all the time). Outline the procedure you used to calculate the quantities needed.

Answer

As defined,

$$k_{\text{abs}} \;=\; \frac{k(A+B)}{e}$$

Given $(\hat{A}, \hat{B}, \hat{k}, \hat{T})$ from Q1 and the estimated covariance matrix from Q3, we can (again, like Q2) use a Wald (delta-method) interval.

Let $g(A, B, k, T) = k(A+B)/e$. Its gradient at $\hat{\theta}$ is

$$\nabla g(\hat{\theta}) = \left(\frac{\partial g}{\partial A}, \frac{\partial g}{\partial B}, \frac{\partial g}{\partial k}, \frac{\partial g}{\partial T}\right)^{\top} = \left(\frac{\hat{k}}{e},\; \frac{\hat{k}}{e},\; \frac{\hat{A}+\hat{B}}{e},\; 0\right)^{\top}$$

Then

$$\widehat{\text{Var}}(\hat{k}_{\text{abs}}) = \nabla g(\hat{\theta})^{\top}\; \widehat{\text{Cov}}(\hat{\theta})\; \nabla g(\hat{\theta})$$

And

$$\text{SE}(\hat{k}_{\text{abs}}) = \sqrt{\widehat{\text{Var}}(\hat{k}_{\text{abs}})}$$

For a (presumably two-sided) 90% CI, using $z_{0.95} = 1.645$:

$$\hat{k}_{\text{abs}} \;\pm\; 1.645 \cdot \text{SE}(\hat{k}_{\text{abs}})$$

Computing:

```
## [1] "A 90% approximate confidence interval for k_abs is (23.811, 31.115)"
```