Non-Linear Regression Analysis

By Chanaka Kaluarachchi

UNIVERSITY of OTAGO
Te Whare Wānanga o Otāgo
NEW ZEALAND

Presentation outline

- Linear regression
- Checking linear Assumptions
- Linear vs non-linear
- Non linear regression analysis

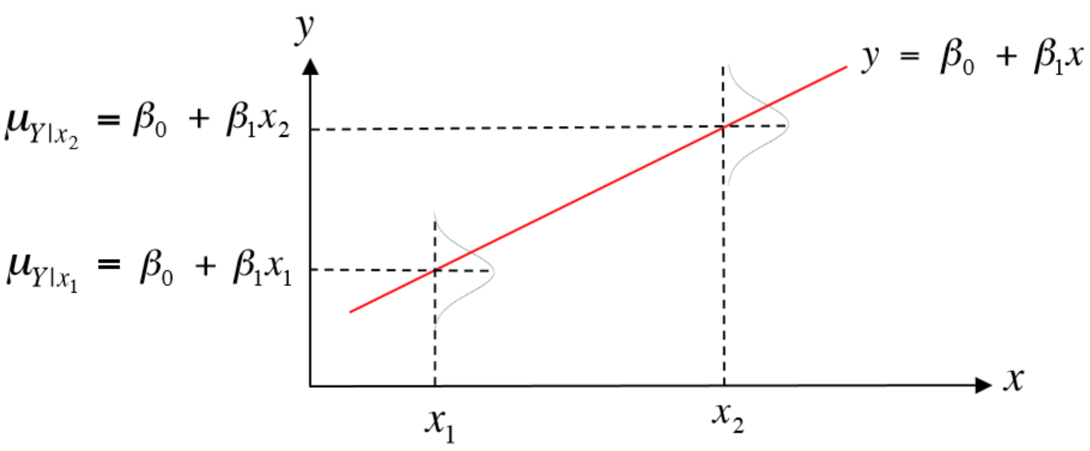
Linear regression (reminder)

- Linear regression is an approach for modelling dependent variable(y) and one or more explanatory variables (x).

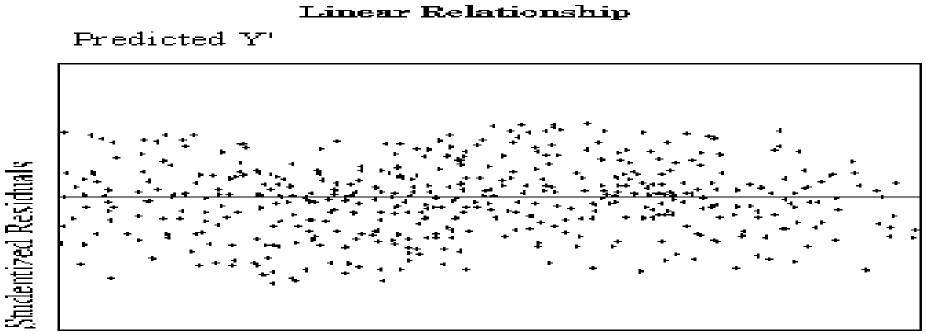
$$y = \beta_0 + \beta_1 x + \varepsilon$$

Assumptions:

$\varepsilon \sim N(0, \sigma^2)$ – iid (independently identically distributed)



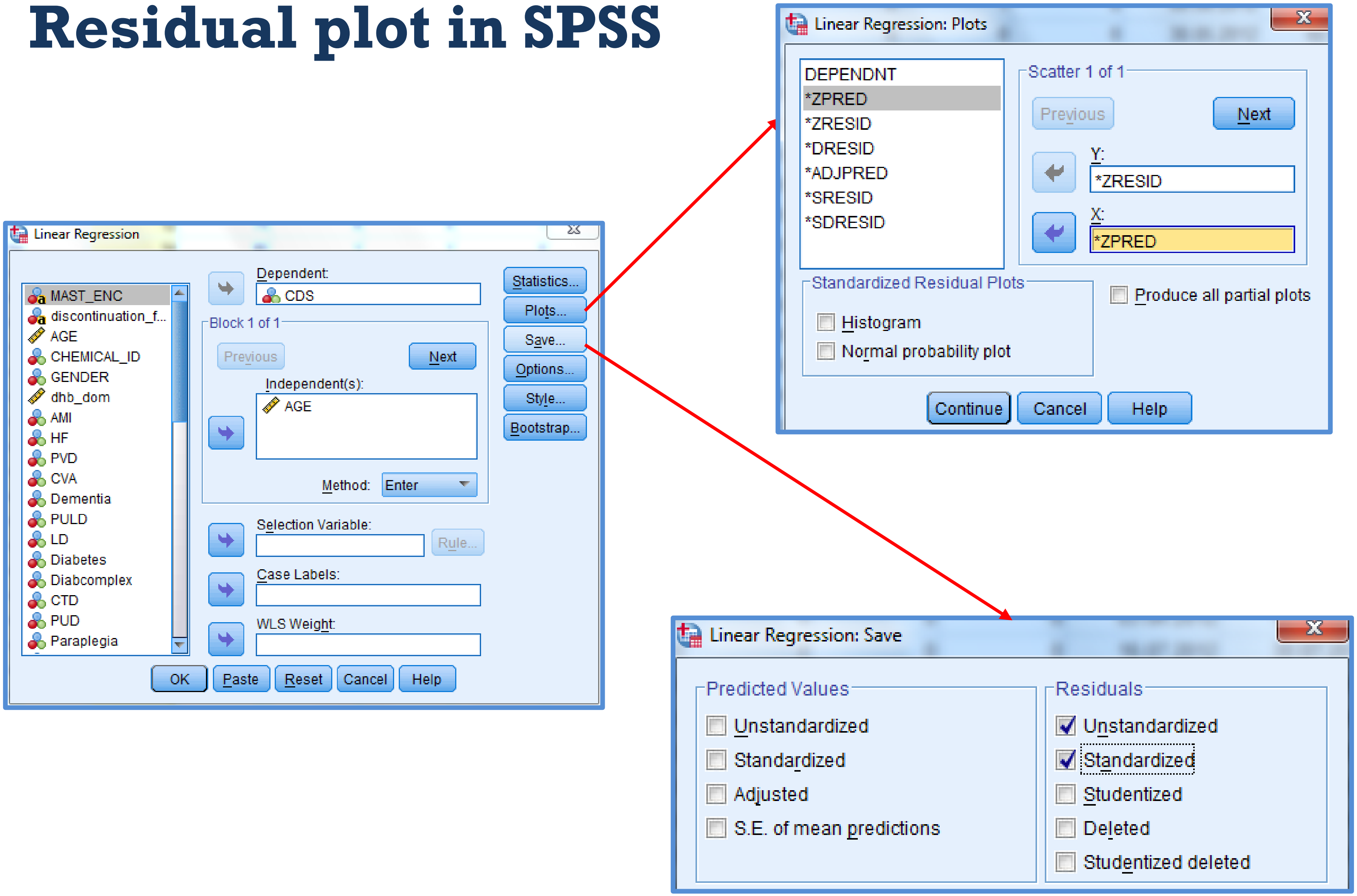
Checking linear Assumptions



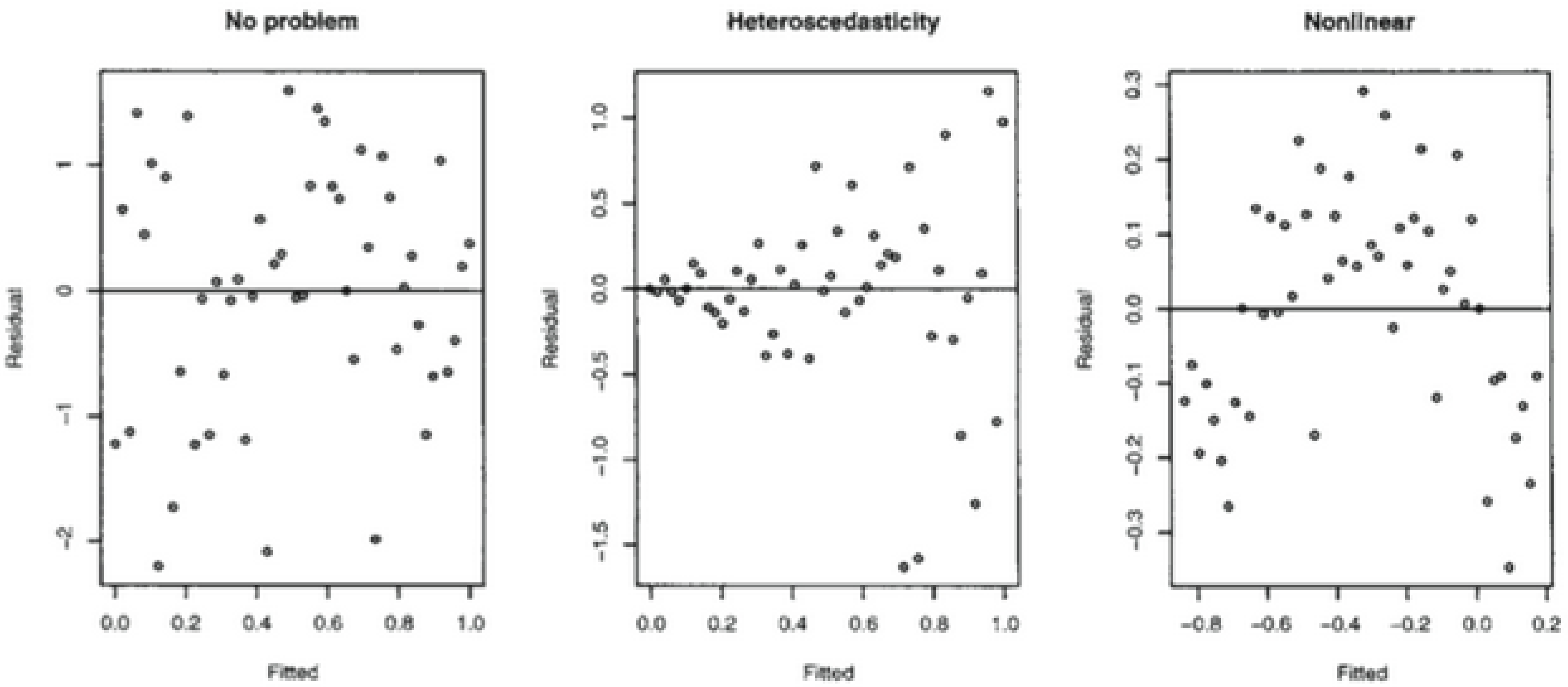
iid- residual plot ($\varepsilon \; vs \; \hat{y}$) can be inspect to check that assumptions are met.

- Constant variance- Scattering is a constant magnitude
- Normal data- few outliers, systematic spared above and below the axis
- Liner relationship- No curve in the residual plot

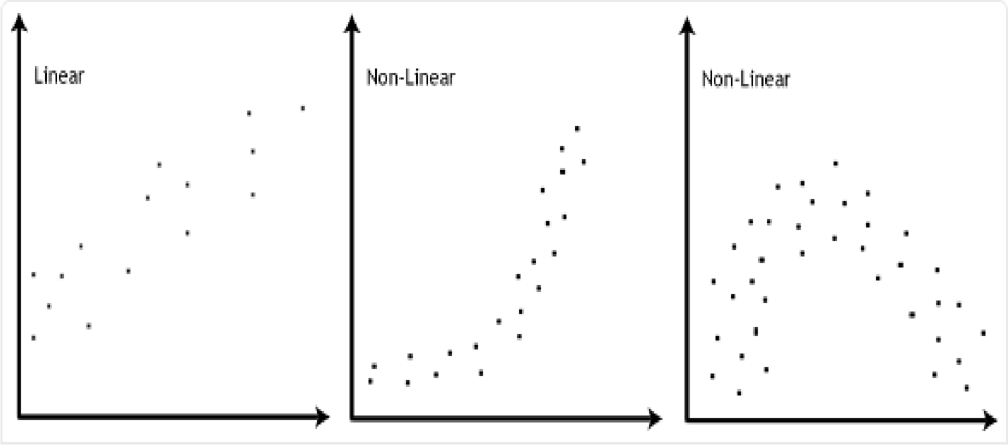
Residual plot in SPSS



Residual plots in SPSS



Linear vs non linear



Linear

- Linear scatter plot
- No curves in residual plot
- Correlation between variable is significant

Non-linear

- Curves in scatter plot
- Curves in residual plot
- No significant correlation between variables

Non linear regression

- Non linear regression arises when predictors and response follows particular function form.

$$y = f(\beta, x) + \varepsilon$$

Examples

$y = \beta^2 x + \varepsilon$ - non linear

$y = \beta x^2 + \varepsilon$ - linear

$y = \frac{1}{\beta} x + \varepsilon$ - non linear

$y = \beta \frac{1}{x} + \varepsilon$ - linear

$y = e^{\beta x} + \varepsilon$ - non linear

$y = \beta \ln x + \varepsilon$ - linear

$y = \frac{1}{1+\beta x} + \varepsilon$ - non linear

Transformation

- Some nonlinear regression problems can be moved to a linear domain by a suitable transformation of the model formulation.

- Four common transformations to induce linearity are: logarithmic transformation, square root transformation, inverse transformation and the square transformation

Examples

- $y = \ e^{\beta x}$ $\Longrightarrow$ $\ln y = \beta x \quad \text{if } y \geq 0$

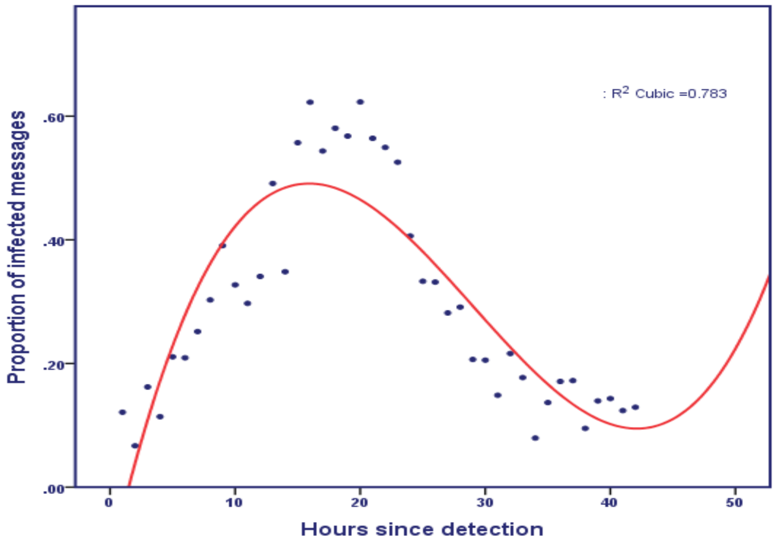
- $y = \frac{1}{1+\beta x}$ $\Longrightarrow$ $\frac{1}{y} - 1 = \beta x \quad \text{if } y \ \neq 0$

Curve Estimation

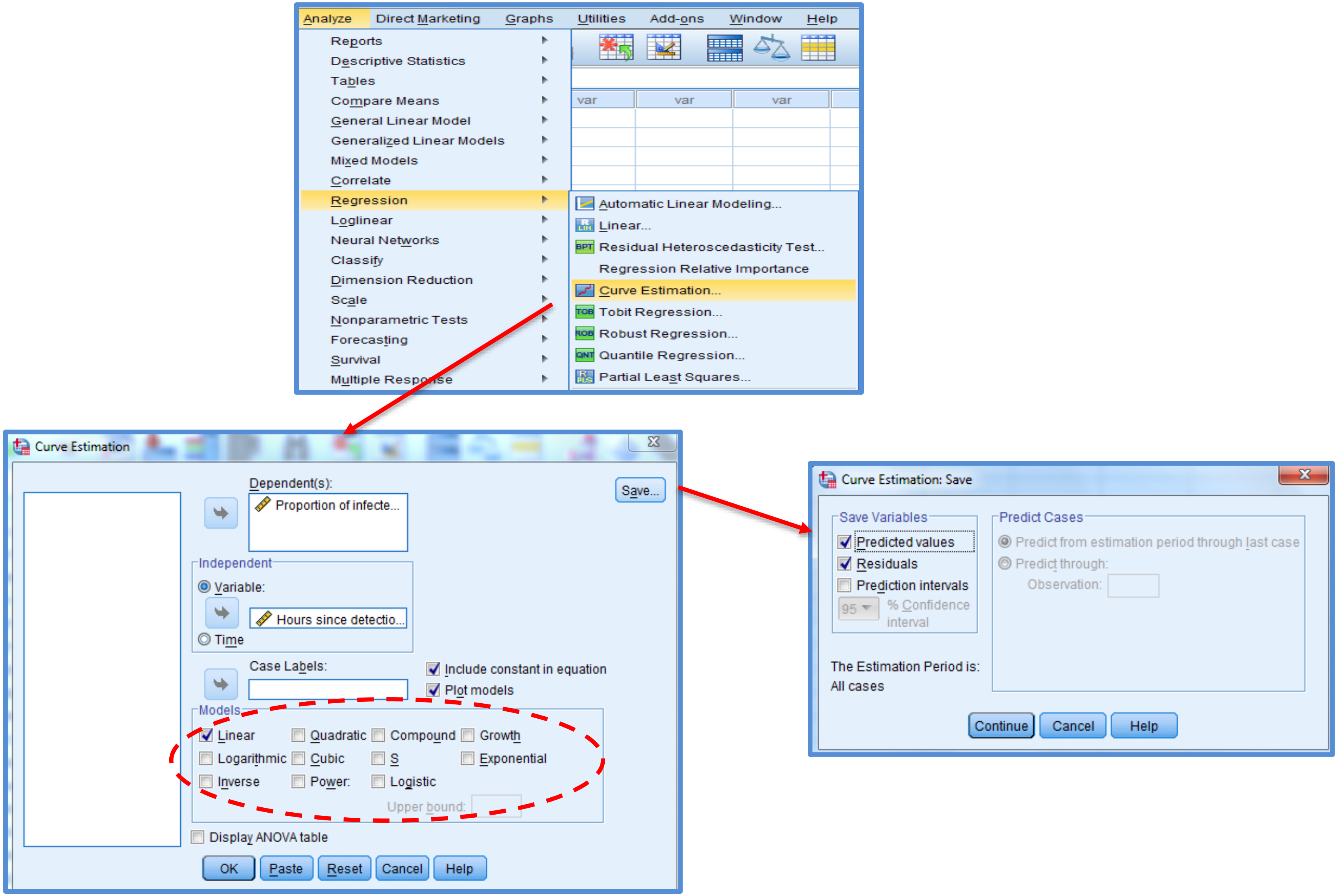
Curve fitting is the process of constructing a curve, or mathematical function, that has the best fit to a series of data points.

Example – Viral growth model

- An internet service provider (ISP) is determining the effects of a virus on its networks. As part of this effort, they have tracked the (approximate) percentage of e-mail traffic on its networks over time, from the moment of discovery until the threat was contained.



Curve Estimation- Cont.



Output

Model Summary and Parameter Estimates

Dependent Variable: Proportion of infected messages

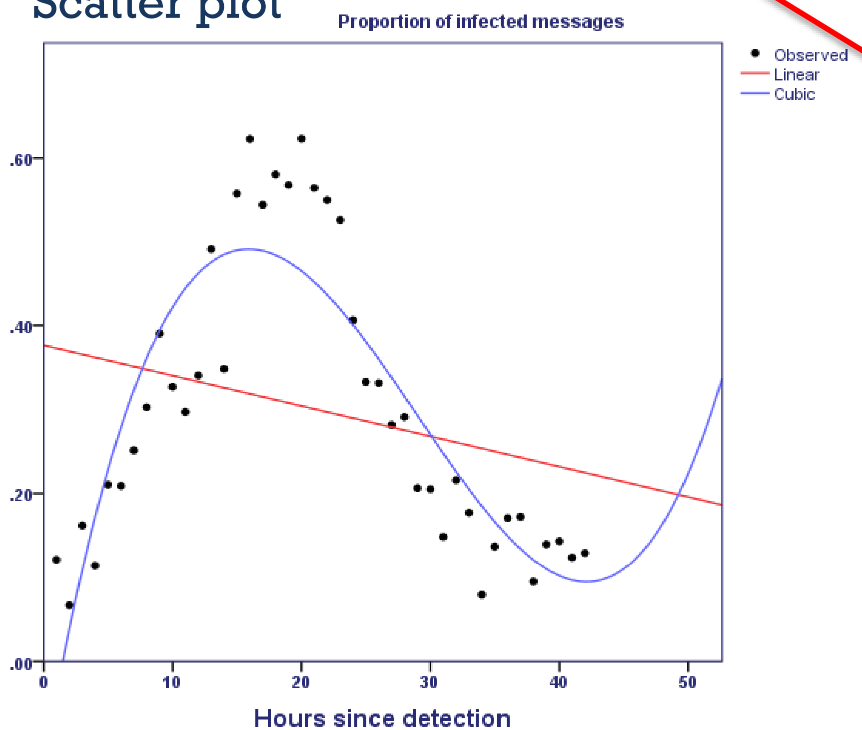
Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.066	2.844	1	40	.100	.377	-.004		
Cubic	.783	45.736	3	38	.000	-.123	.088	-.004	4.399E-5

The independent variable is Hours since detection.

P value< 0.05 means model is significant

Higher the R^2 better the model fit

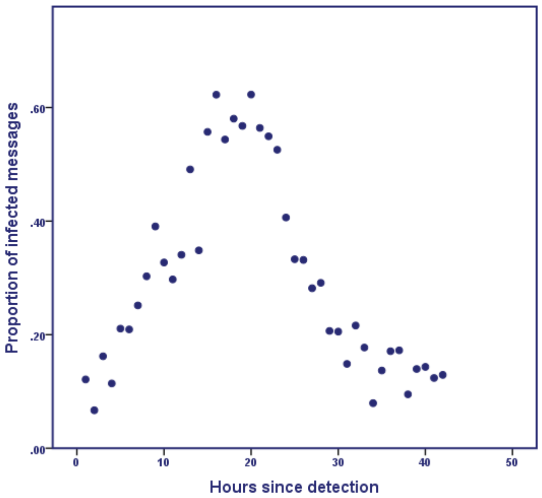
Scatter plot



Segmentation

We can split the graph in to segments and fit a segmented model.

Example – Viral growth model



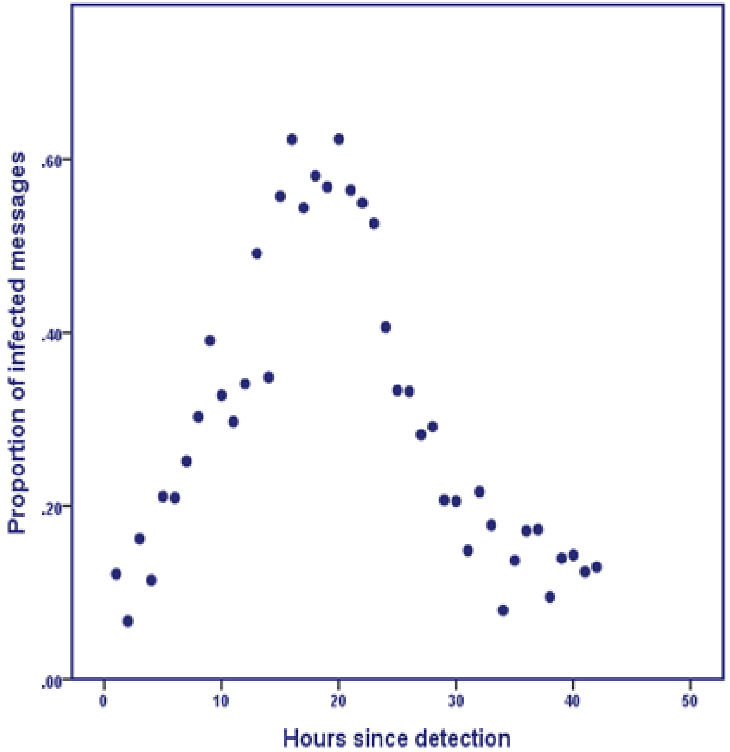
We can fit a logistic equation for the first 19 hours and an asymptotic regression for the remaining hours should provide a good fit and interpretability over the entire time period.

Logistic model and choosing starting values

$$y = \frac{\beta_1}{1 + \beta_2 e^{-\beta_3 x}}$$

Starting values

- β_1- upper value of growth (0.65)
- β_2- ratio upper value and lowest value (0.65/0.13=5)
- β_3- estimated slop between points in plot.
 (0.6-0.12/19-3)=0.03

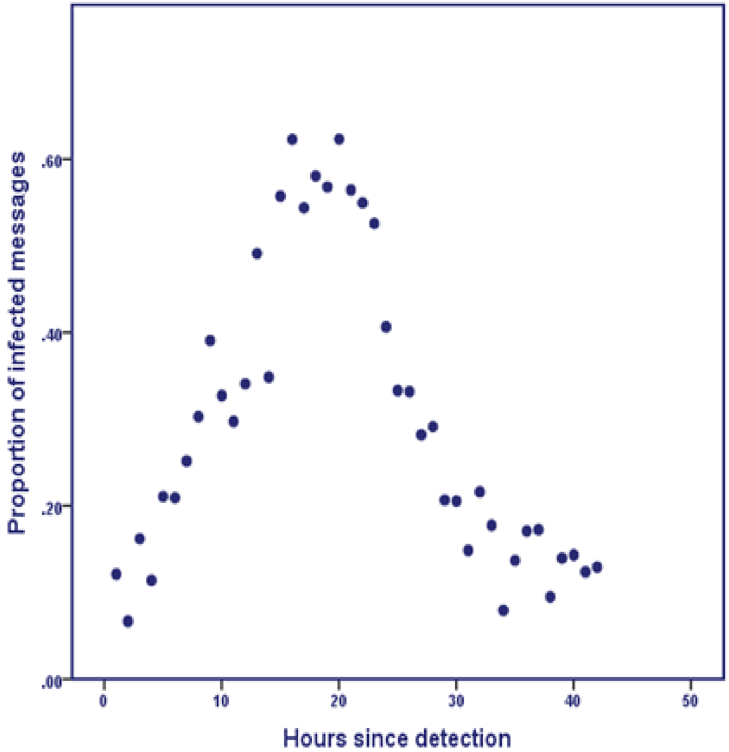


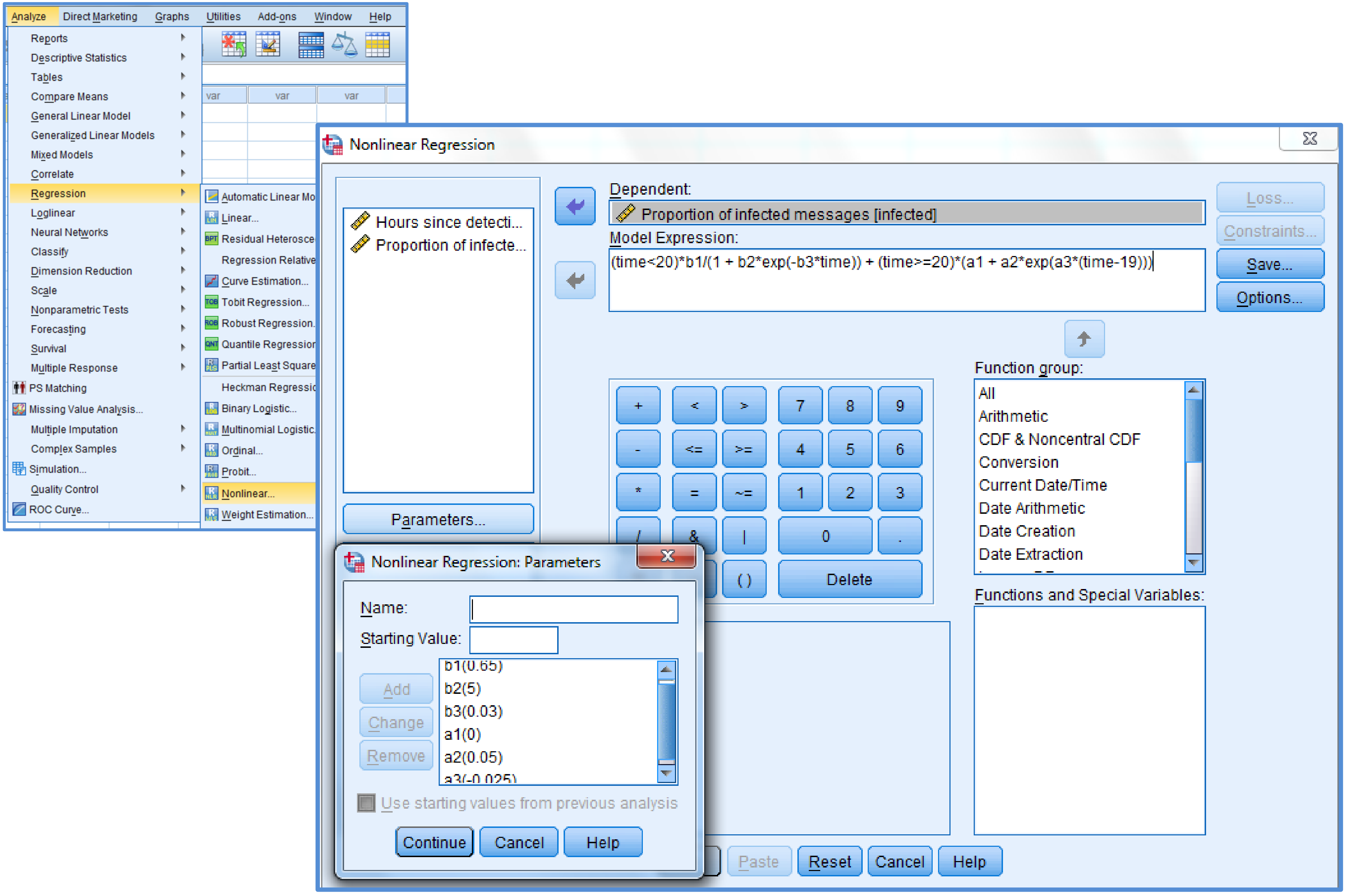
Asymptotic regression model

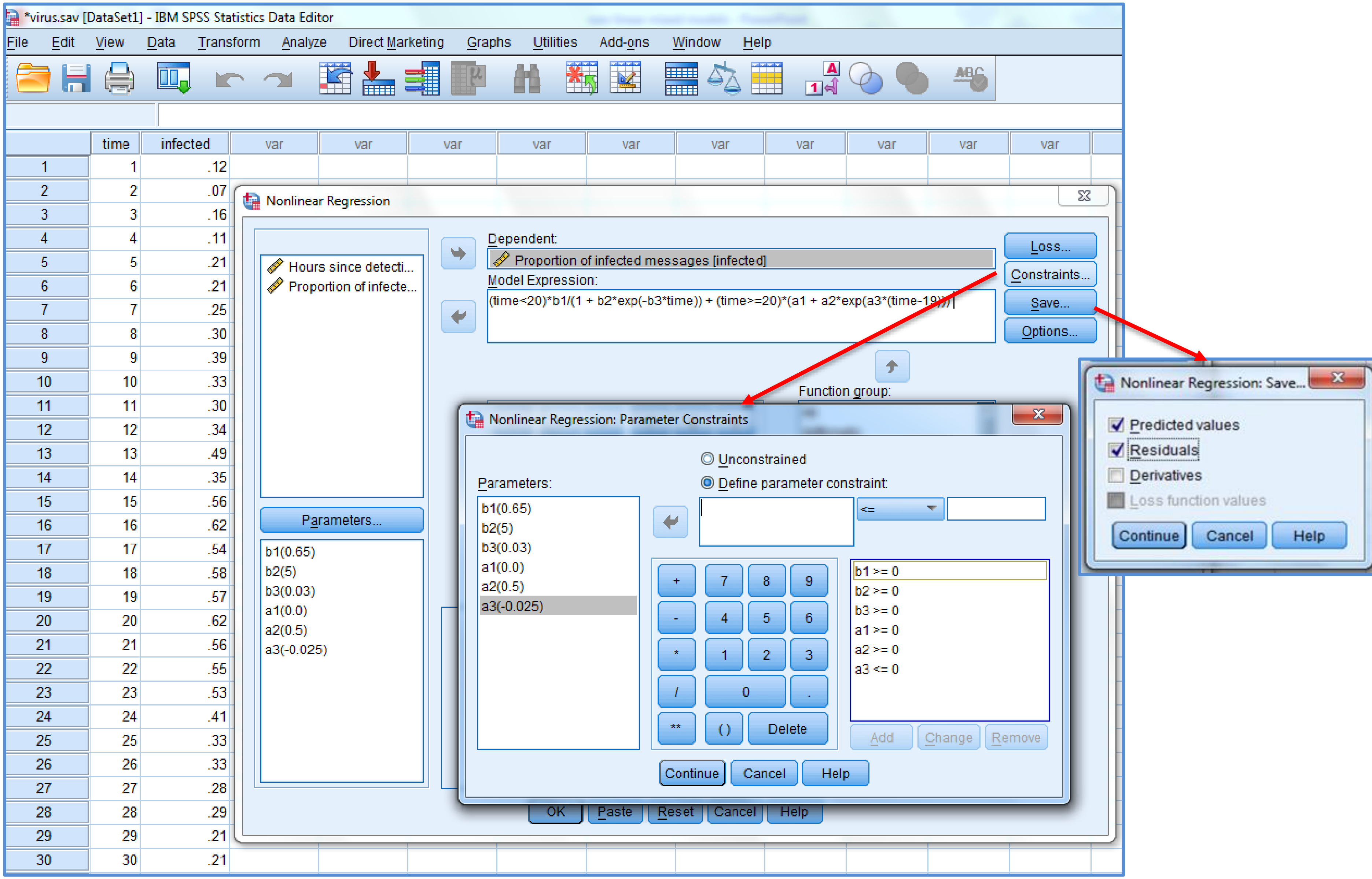
$$y = \theta_1 + \theta_2 e^{\theta_3 x}$$

Starting values

- θ_1- lowest value (0)
- θ_2- difference upper value and lowest value (0.6)
- β_3- estimated slop between points in plot.
  (0.6-0.1/20-40)=-0.025




Nonlinear Regression
Dependent:
Proportion of infected messages [infected]
Model Expression:
(time<20)*b1/(1 + b2*exp(-b3*time)) + (time>=20)*(a1 + a2*exp(a3*(time-19)))
Hours since detecti...
Proportion of infecte...
Parameters...
Loss...
Constraints...
Save...
Options...
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Date Creation
Date Extraction
Functions and Special Variables:
Delete
Paste
Reset
Cancel
Help
Nonlinear Regression: Parameters
Name:
Starting Value:
b1(0.65)
b2(5)
b3(0.03)
a1(0)
a2(0.05)
a3(-0.025)
Add
Change
Remove
Use starting values from previous analysis
Continue
Cancel
Help
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
PS Matching
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
ROC Curve...
Linear...
Curve Estimation...
Tobit Regression...
Binary Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...



*virus.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
time infected var var var var var var var var var var
1 1 .12
2 2 .07
3 3 .16
4 4 .11
5 5 .21
6 6 .21
7 7 .25
8 8 .30
9 9 .39
10 10 .33
11 11 .30
12 12 .34
13 13 .49
14 14 .35
15 15 .56
16 16 .62
17 17 .54
18 18 .58
19 19 .57
20 20 .62
21 21 .56
22 22 .55
23 23 .53
24 24 .41
25 25 .33
26 26 .33
27 27 .28
28 28 .29
29 29 .21
30 30 .21
Nonlinear Regression
Hours since detecti...
Proportion of infecte...
Dependent:
Proportion of infected messages [infected]
Model Expression:
(time<20)*b1/(1 + b2*exp(-b3*time)) + (time>=20)*(a1 + a2*exp(a3*(time-19)))
Loss...
Constraints...
Save...
Options...
Function group:
Parameters...
b1(0.65)
b2(5)
b3(0.03)
a1(0.0)
a2(0.5)
a3(-0.025)
OK Paste Reset Cancel Help
Nonlinear Regression: Parameter Constraints
Unconstrained
Define parameter constraint:
Parameters:
b1(0.65)
b2(5)
b3(0.03)
a1(0.0)
a2(0.5)
a3(-0.025)
<=
+ 7 8 9
- 4 5 6
* 1 2 3
/ 0 .
** ( ) Delete
b1 >= 0
b2 >= 0
b3 >= 0
a1 >= 0
a2 >= 0
a3 <= 0
Add Change Remove
Continue Cancel Help
Nonlinear Regression: Save...
Predicted values
Residuals
Derivatives
Loss function values
Continue Cancel Help

Output

Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
b1	.734	.127	.477	.991
b2	7.428	1.375	4.638	10.217
b3	.184	.040	.103	.265
a1	.091	.030	.030	.153
a2	.661	.044	.572	.750
a3	-.150	.027	-.205	-.095

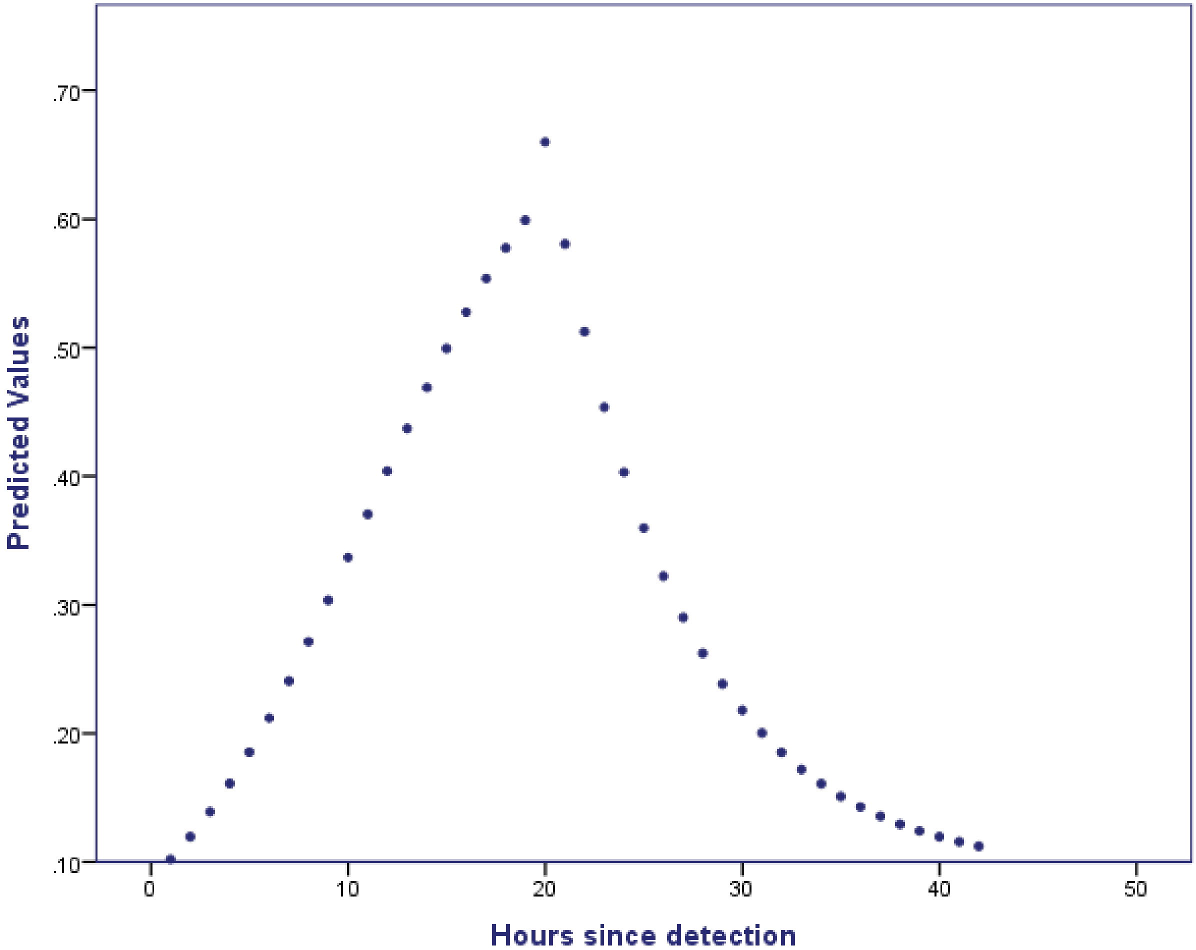
ANOVA[a]

Source	Sum of Squares	df	Mean Squares
Regression	4.884	6	.814
Residual	.082	36	.002
Uncorrected Total	4.966	42	
Corrected Total	1.212	41	

Dependent variable: Proportion of infected messages

a. R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .933.

Output



.70
.60
.50
.40
.30
.20
.10
Predicted Values
0
10
20
30
40
50
Hours since detection

Thanks for listening