Universitäts-Rechenzentrum Trier

AWS.SPSS.16

Trier, den 2.6.1998

Nichtlineare Regression mit SPSS

Herausgeber: Universitäts-Rechenzentrum Trier
Universitätsring 15
D-54286 Trier
Tel.: (0651) 201-3417, Fax.: (0651) 3921
Leiter: Prof. Dr.-Ing. Manfred Paul
Autor: Bernhard Baltes-Götz
Mail: baltes@uni-trier.de

Vorwort

In diesem Manuskript werden elementare Begriffe und Verfahren der nichtlinearen Regressionsanalyse in Theorie und Praxis behandelt. Auf der Basis einer statistischen und mathematischen Grundausbildung (zu Begriffen wie *Parameter*, *Signifikanztest*, *Ableitung* etc.) sollten die Erläuterungen zur Begründung von Analyseschritten und zur Interpretation der Ergebnisse nachvollziehbar sein.

Als Software kommt SPSS 6.1 für Windows zum Einsatz, jedoch können praktisch alle vorgestellten Verfahren auch mit jüngeren SPSS-Versionen unter Windows, MacOS oder Linux realisiert werden.

Das Manuskript ist als PDF-Dokument zusammen mit den im Kurs benutzen Dateien auf dem Webserver der Universität Trier von der Startseite (http://www.uni-trier.de/) ausgehend folgendermaßen zu finden:

Rechenzentrum > Studierende > EDV-Dokumentationen >
Statistik > Nichtlineare Regression mit SPSS

Hinweise auf Unzulänglichkeiten im Manuskript werden mit Dank entgegen genommen

1 Die lineare Regression

Der Anwendungsbereich *linearer* Modelle ist durchaus größer, als es auf den ersten Blick erscheint. Sie müssen nämlich nur **linear in den Parametern** sein, d.h. die abhängige Variable Y muß durch eine Parameter-gewichtete Summe der beliebig individuell transformierten Regressoren X_1, X_1,..., X_m und einen additiven Fehlerterm ε erklärbar sein:

$$Y = \beta_0 + \beta_1 f_1(X_1) + \beta_2 f_2(X_2) + ... + \beta_m f_m(X_m) + \varepsilon, \quad \text{mit } E(\varepsilon) = 0 \tag{1}$$

In diesem Sinne ist z.B. das folgende Regressionsmodell mit parabelförmiger Gleichung linear in den beiden Parametern β_0 und β_1 ($f(X) = X^2$):

$$Y = \beta_0 + \beta_1 X^2 + \varepsilon$$

Viele Modelle sind auf den ersten Blick nicht von der Form (1), können aber durch geeignete Transformationen auf diese Form gebracht werden, z.B.:

a) $Y = e^{\beta_0 + \beta_1 X + \varepsilon} \underset{Y > 0}{\Leftrightarrow} \ln(Y) = \beta_0 + \beta_1 X + \varepsilon$

b) $Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \varepsilon \underset{Y, \beta, X_1, X_2, \varepsilon > 0}{\Leftrightarrow} \ln(Y) = \ln(\beta_0) + \beta_1 \ln(X_1) + \beta_2 \ln(X_2) + \ln(\varepsilon)$

In der linearisierten Variante des letzten Modells tritt der gesuchte Parameter β_0 zwar in logarithmierter Form auf, doch ist mit $\ln(\beta_0)$ sofort auch β_0 bekannt: Wir können auf den Schätzer für $\ln(\beta_0)$ die Exponentialfunktion anwenden, um einen Schätzer für β_0 zu gewinnen.
Ist eine Linearisierung möglich, sollte diese stets vorgenommen werden, damit die einfachen, nicht-iterativen, Schätzmethoden des linearen Modells verwendet werden können.
Die häufig anzutreffende Praxis, lineare Regressionsmodelle als Approximation zu verwenden, wenn theoretisch fundierte Modelle fehlen, ist zur Not akzeptabel, aber nicht unbedingt empfehlenswert. In diesem Kurs werden wir das Beispiel einer scheinbar quadratischen Regression kennenlernen, der eigentlich ein logistisches Modell zugrunde liegt. In dieser Situation bieten die Parameterschätzungen zum korrekten Model erheblich mehr Information über den zugrunde liegenden Sachverhalt, d.h. über den Apparat, der die Daten erzeugt hat. Sie erlauben z.B. eher eine Extrapolation (siehe unten) oder eine Klärung der Frage, was bei einer Intervention zur Veränderung gewisser Parameter im empirischen System passieren wird.

2 Das Modell der nichtlinearen Regression

Häufig werden von der Theorie oder von der Empirie Regressionsmodelle diktiert, die sich nicht auf die Form (1) bringen lassen. Mit solchen Modellen wollen wir uns in diesem Kurs näher beschäftigen.
Treten in einem Modell die Regressoren X_1, ..., X_m sowie die Parameter θ_1, ..., θ_p auf, dann lautet das allgemeine Modell der nichtlinearen Regression:

$$Y = f(X_1,...,X_m;\theta_1,...,\theta_p) + \varepsilon \quad \text{mit } E(\varepsilon) = 0 \tag{2}$$

Im *Stichprobenmodell* mit den Beobachtungsvariablen Y_1,..., Y_n und den zugehörigen Fehlervariablen ε_1,...,ε_n wird wie bei der linearen Regression angenommen:

- $\text{Var}(\varepsilon_\nu) = \sigma^2, \nu = 1,...,n$ (Varianzhomogenität der Fehler)
- $\text{Korr}(\varepsilon_\nu, \varepsilon_\kappa) = 0, \quad \nu \neq \kappa$ (Unkorreliertheit der Fehler)

Meist setzt man auch noch voraus:

- $\varepsilon_\nu \sim N(0,\sigma^2), \nu = 1,...,n$ (Normalverteilung der Fehler)

3 Beispiel für ein echt-nichtlineares Modell: Logistisches Populationswachstum

Wir wollen das Wachstum einer Gattung von Parasiten in einem Wirtstier betrachten, wobei wir über die Fortpflanzungsbiologie einige vereinfachende Annahmen machen wollen:

- Die aktuelle Populationsgröße als Funktion der Zeit *t* soll mit *S*(*t*) bezeichnet werden.
- Das Populations*wachstum* in der Zeit, also die erste Ableitung der Funktion *S*(*t*) ist proportional zur aktuellen Populationsgröße *S*(*t*), d.h. je mehr Parasiten vorhanden sind, desto stärker ist der Zuwachs.
- Die maximale Populationsgröße ist aufgrund knapper Ressourcen auf die endliche Zahl α beschränkt, die asymptotisch erreicht wird.
- Das Wachstum der Population wird durch die allmähliche Verknappung der Ressourcen gehemmt. Neben dem eben besprochenen Einfluß der aktuellen Populationsgröße soll auch die verbleibende relative Wachstumsreserve:

$$\frac{(\alpha - S(t))}{\alpha}$$

  als Faktor auf das Populationswachstum einwirken.

Für das Wachstum, d.h. für die Ableitung der Populationsgröße *S*(*t*) nach der Zeit *t*, wollen wir also insgesamt die folgende Differentialgleichung mit dem Proportionalitätsfaktor *w* postulieren:

$$\frac{dS(t)}{dt} = wS(t)\frac{(\alpha - S(t))}{\alpha}$$

Aus dieser Differentialgleichung gewinnen wir durch Integrieren eine Lösungsfunktion *S*(*t*):

$$S(t) = \frac{\alpha}{1 + e^{\beta - wt}} \tag{3}$$

Wir wollen uns durch Ableiten nach der Quotientenregel $\left(\left(\frac{1}{f}(x)\right)' = \frac{-f'(x)}{f^2(x)}\right)$ kurz vergewissern, daß unsere Funktion tatsächlich die obige Differentialgleichung erfüllt:

$$\begin{aligned}
\frac{dS(t)}{dt} &= \left[\frac{\alpha}{1+e^{\beta-wt}}\right]' = \frac{\alpha(e^{\beta-wt}\, w)}{\left(1+e^{\beta-wt}\right)^2} = \frac{w}{\alpha}\frac{\alpha^2\, e^{\beta-wt}}{\left(1+e^{\beta-wt}\right)^2} \\
&= \frac{w}{\alpha}\frac{\alpha}{\left(1+e^{\beta-wt}\right)}\frac{\alpha(1+e^{\beta-wt}-1)}{\left(1+e^{\beta-wt}\right)} = \frac{w}{\alpha}S(t)\frac{\alpha(1+e^{\beta-wt})-\alpha}{\left(1+e^{\beta-wt}\right)} \\
&= \frac{w}{\alpha}S(t)\left(\alpha - \frac{\alpha}{\left(1+e^{\beta-wt}\right)}\right) = wS(t)\frac{(\alpha - S(t))}{\alpha}
\end{aligned}$$

Der technische Parameter β resultiert aus einer gewissen Unterbestimmtheit unseres Problems und kann durch Wahl eines festen Funktionswertes *S*(0) beseitigt werden:

$$S(0) = \frac{\alpha}{1+e^{\beta}}.$$

In unserem Beispiel ist dies die anfängliche Größe der Parasitenpopulation. In der erfundenen Studie, mit der wir uns im weiteren Verlauf beschäftigen wollen, soll dieser Startwert durch eine entsprechende experimentelle Manipulation auf 50 festgesetzt werden soll. Wir können uns z.B. vorstellen, daß zu Beginn

des Experimentes in jedem betrachteten Wirtstier 50 Parasitenexemplare „ausgesetzt“ werden. (Biologen mögen diese Zahl noch durch eine passende Zehnerpotenz in eine realistische Größenordnung bringen.)
Wir können also den Parameter β aufgrund unserer Anfangswertbedingung folgendermaßen ersetzen:

$$S(0) = 50 = \frac{\alpha}{1+e^{\beta}} \Leftrightarrow e^{\beta} = \frac{\alpha}{50} - 1 \Leftrightarrow \beta = \ln\left(\frac{\alpha}{50} - 1\right) \quad \alpha > 0$$

Diesen Ausdruck setzen wir nun in die Gleichung (3) ein:

$$S(t) = \frac{\alpha}{1+e^{\ln\left(\frac{\alpha}{50}-1\right)-wt}} = \frac{\alpha}{1+\left[\left(\frac{\alpha}{50}-1\right)e^{-wt}\right]} \qquad (4)$$

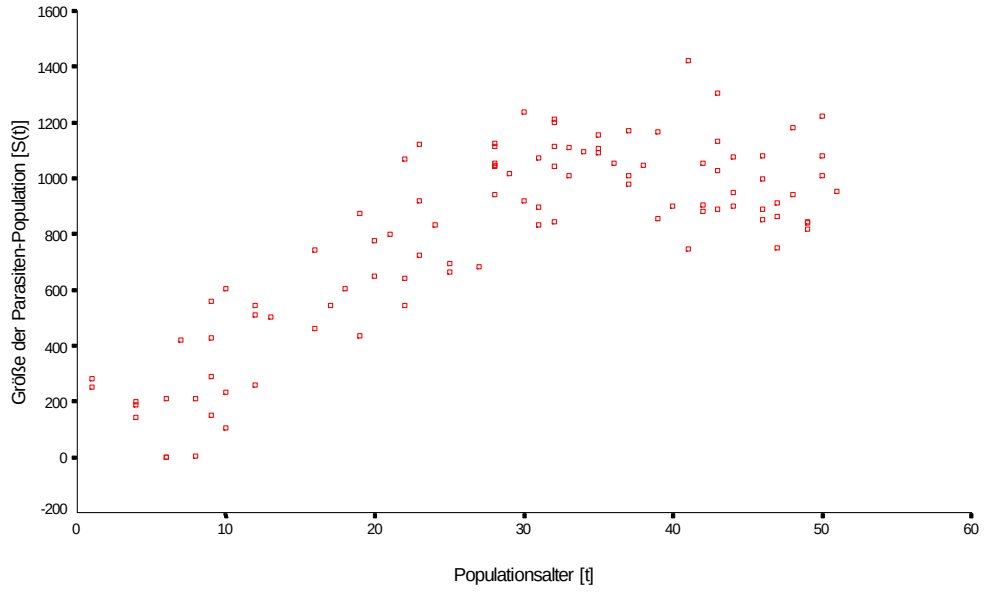
Einige Eigenschaften der Funktion $S(t)$ und der zu untersuchenden künstlichen Daten:

- Für $t \to \infty$ geht der Funktionswert $S(t)$ bei $w > 0$ gegen $\frac{\alpha}{1+0} = \alpha$.

  In unserer Anwendung ist dies die maximale Größe der Parasitenpopulation. Bei der Erzeugung künstlicher Parasiten-Populationszahlen wollen wir α auf 1000 festlegen. Wir können uns z.B. vorstellen, daß in jedem Wirtstier (= Fall) aus der untersuchten Population maximal 1000 Parasiten leben können.

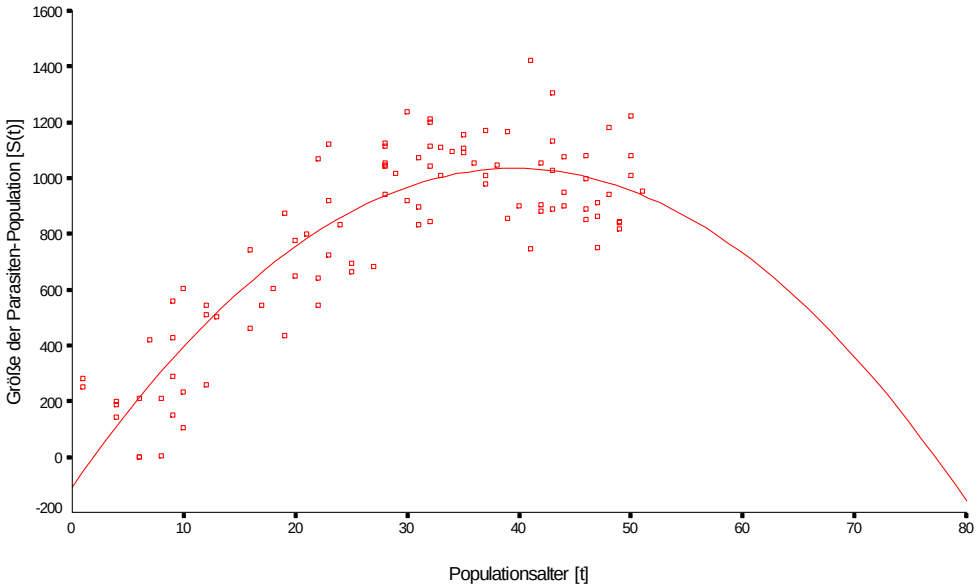
- w ist der Proportionalitätsfaktor für der Einfuß der Produktgröße $S(t)\frac{(\alpha - S(t))}{\alpha}$ auf das Populationswachstum. In unserem Beispiel soll w auf 0,2 festgelegt werden.

Mit der Gleichung (4) und den obigen Parameterfestlegungen wurden künstliche Daten zu 100 Wirtstieren erzeugt, die alle mit 50 Parasitenexemplaren gestartet waren und unterschiedlich lange Beobachtungsintervalle absolviert haben (zwischen 1 und 51 Monate).
Die 100 Realisationen der abhängigen Variablen wurden außerdem mit normalverteilten, unabhängigen und varianzhomogenen Fehlern versehen.
Damit erhalten wir für die Regression der Populationsgröße (Variable S) auf das Populationsalter (Variable T) das folgende Streuungsdiagramm:



Der Zusammenhang wirkt sehr „quadratisch“, und tatsächlich erhält man in der Regression von S auf T^2 ein beeindruckendes adjustiertes R^2 von 0,77. Die Extrapolation des quadratischen Zusammenhangs bringt je-

doch keine brauchbare Prognose für die zu erwartende Entwicklung der Populationsgrößen bei einer Verlängerung des Beobachtungsintervalls:



Es ist zu hoffen, daß sich daß nun mit Hilfe der nichtlinearen Regression zu schätzende Modell sinnvoller benimmt.
Sie finden den Datensatz in der SPSS-Datendatei **para.sav** an der im Vorwort vereinbarten Stelle.

4 Parameterschätzung bei der nichtlinearen Regression

Bei den Schätzalgorithmen zur nichtlinearen Regression werden die Parameterschätzungen ebenfalls so bestimmt, daß die Summe der quadrierten Modellresiduen minimal wird:

$$S(\Theta) := S(\theta_1,...,\theta_p) := \sum_{i=1}^{n}[Y_i - f(X_{1i},...,X_{mi};\theta_1,...,\theta_p)]^2 \xrightarrow{!} \text{Min}$$

Bei diesem Minimierungsproblem werden die beobachteten y- und x_j-Werte als konstant angenommen, so daß die Fehlerquadratsumme nur von den Parameterwerten $\theta_1,\ldots,\theta_p$, zusammengefaßt im Vektor Θ, abhängt. Die Lösung $(\hat{\theta}_1,\ldots,\hat{\theta}_p)$ kann nicht, wie bei der linearen Regression, in geschlossener Form angegeben werden, sondern muß in einem itererativen Verfahren gesucht werden.

4.1 Startwerte

Die iterativen Schätzverfahren der nonlinearen Regressionsanalyse benötigen für jeden Parameter einen Starwert. Dieser sollte möglichst nahe beim optimalen Schätzwert liegen, damit das Verfahren diesen findet und dabei möglichst schnell konvergiert. Bei schlecht gewählten Startwerten können unerfreuliche Ergebnisse auftreten:

- Das Verfahren landet in einem lokalen Minimum.
- Das Verfahren produziert inhaltlich unmögliche Schätzungen.
- Das Verfahren konvergiert überhaupt nicht.

In der Rolle eines echten Forschers, der unser Wissen über die künstliche Population natürlich *nicht* besitzt, können Startwerte für die Parameter α und w aufgrund folgender Überlegungen gewonnen werden:

- α gibt die asymptotische Populationsgröße an, so daß wir aus obiger Abbildung leicht z.B. den Startwert 1400 ablesen könnten.

- Auch bei der Ermittlung des zweiten Startwertes benutzen wir die beobachteten Daten. Für $t = 10$ erhalten wir ungefähr die mittlerer Populationsgröße 250, also gilt mit unserer Gleichung (4):

$$250 = \frac{1400}{1+\left[\left(\frac{1400}{50}-1\right)e^{-10w}\right]} \Leftrightarrow 27\,e^{-10w} = \frac{1400}{250} - 1 \Leftrightarrow -10w = \text{-}1{,}769780562509 \Leftrightarrow w \approx 0{,}18$$

Zur Festlegung der Startwerte können auch beliebige andere Informationsquellen herangezogen werden, z.B. Erfahrungen aus früheren Studien.

4.2 Der Levenberg-Marquardt-Algorithmus

Bei dem von SPSS per Voreinstellung verwendeten **Levenberg-Marquardt**-Algorithmus wird in jedem Iterationsschritt eine neue Richtung für die Suche nach dem Minimum von $S(\Theta)$ über dem p-dimensionalen Parameterraum unter Verwendung zweier Informationsquellen festgelegt (siehe Draper & Smith, 1981, S. 471):

- Es wird der Vektor

$$\left(-\frac{d\,S(\Theta)}{d\,\theta_1} \quad -\frac{d\,S(\Theta)}{d\,\theta_2} \quad \cdots \quad -\frac{d\,S(\Theta)}{d\,\theta_p}\right)$$

  mit den negativen partiellen Ableitungen der Funktion $S(\Theta)$ nach allen Parametern bestimmt. Dieser Vektor gibt die Richtung des steilsten Abstiegs im "Gebirge" der Funktion $S(\Theta)$ über dem p-dimensionalen Parameterraum an.
- Es wird eine lokale Linearisierung des Problems in einer Umgebung um die aktuellen Parameterschätzungen Θ_v vorgenommen (Taylor-Entwicklung um Θ_v). Daraus werden mit den klassischen Methoden der linearen Regression neue Parameterschätzungen gewonnen, so daß sich eine weitere mögliche Richtung für das weitere Vorrücken im Parameterraum ergibt.

Die beiden resultierenden Bewegungsrichtungen werden vom Levenberg-Marquardt-Algorithmus interpoliert. Außerdem wird eine Schrittweite gewählt.
Das Verfahren endet (hoffentlich), wenn sich die Schätzungen für die Parameter bzw. für die Fehlerquadratsumme nach einem geeigneten Kriterium stabilisieren.

Für unser Beispiel erhalten wir folgenden Verlauf der Levenberg-Marquardt-Schätzverfahrens:

```
There are 100 cases.  There is enough memory for them all.

 Iteration  Residual SS       ALPHA           W

     1      9366346,898  1400,00000  ,180000000
     1.1    2497569,202  1008,60779  ,192084647
     2      2497569,202  1008,60779  ,192084647
     2.1    2432277,027  1017,63050  ,202979989
     3      2432277,027  1017,63050  ,202979989
     3.1    2432193,201  1018,49183  ,203185513
     4      2432193,201  1018,49183  ,203185513
     4.1    2432193,201  1018,48953  ,203185363

Run stopped after 8 model evaluations and 4 derivative evaluations.
Iterations have been stopped because the relative reduction between successive
residual sums of squares is at most SSCON = 1,000E-08
```

Mit der erfolgreichen Konvergenz ist in unserem Beispiel eine wesentliche technische Hürde genommen. Wenn die Konvergenz nicht gelingt, kommen u.a. die folgenden Maßnahmen in Frage:

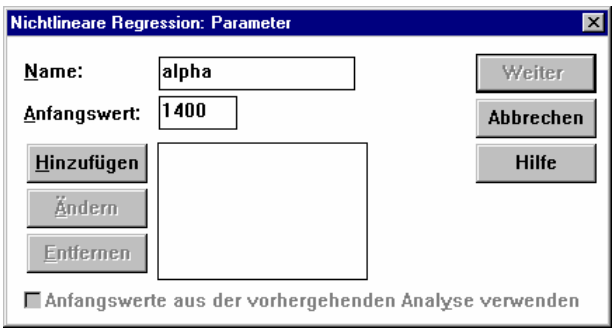
- Bessere Startwerte wählen
- Ein alternatives Schätzverfahren wählen
- Das Modell verbessern
- Andere Daten verwenden

5 Anforderung einer nichtlinearen Regression

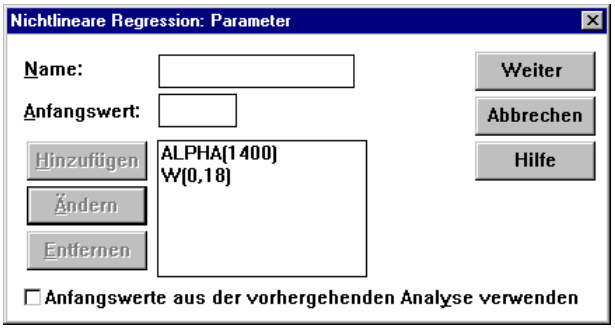
Nun sollten Sie lernen, wie eine nichtlineare Regression in SPSS angefordert wird. Starten Sie nötigenfalls SPSS und öffnen Sie die oben angegebene Datendatei. Die Dialogbox zur Spezifikation einer nichtlinearen Regression wird geöffnet mit:

Statistik > Regression > Nichtlinear...

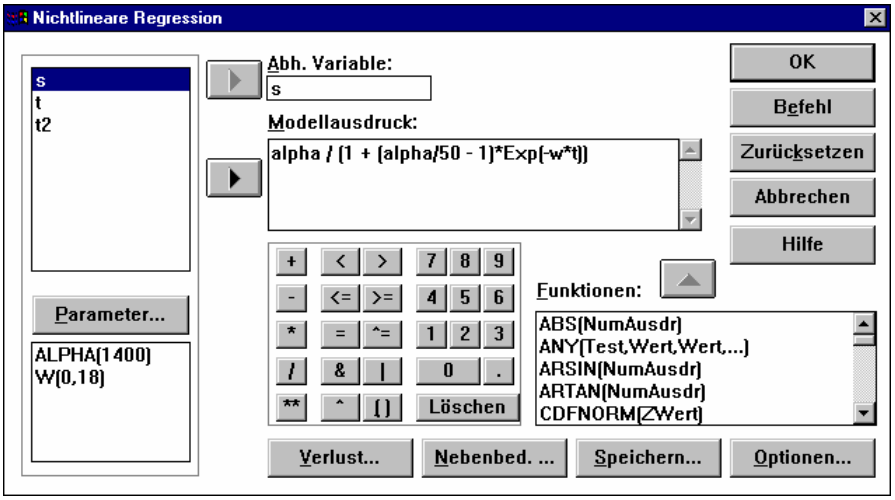
Legen Sie zunächst **S** als abhängige Variable fest, und öffnen Sie dann mit einem Mausklick auf **Parameter...** die Dialogbox zur Spezifikation der Startwerte. Tragen Sie den Namen und den Wert des ersten Parameters ein:



und klicken Sie auf den Schalter **Hinzufügen**. Wenn Sie alle Parameter angemeldet haben, sieht die Parameter-Subdialogbox folgendermaßen aus:



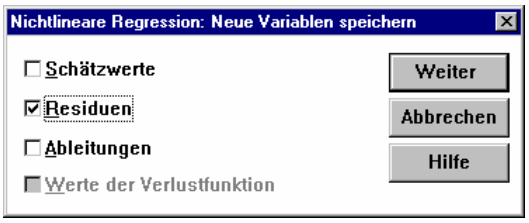
Machen Sie dann **Weiter** und ergänzen Sie in der Hauptdialogbox die Formel der Regressionsfunktion:



Im Modellausdruck dürfen auftreten:

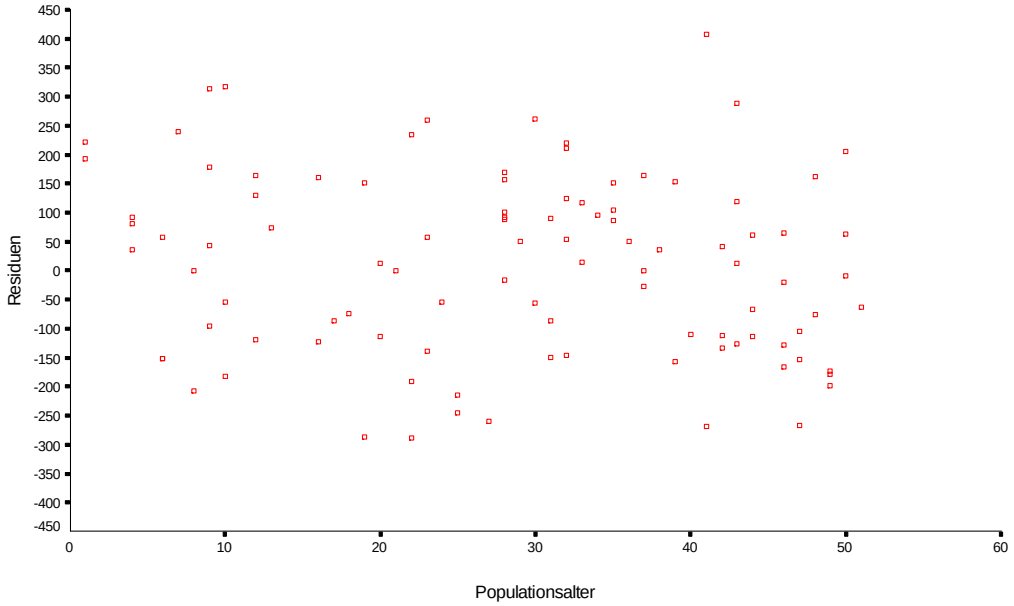
- Unabhängige Variablen
- Definierte Parameter
- Arithmetische Operatoren und Funktionen

Um später die Residuen unseres Modells untersuchen zu können, veranlassen wir noch in der **Optionen**-Subdialogbox, daß diese in die Arbeitsdatei gesichert werden:



5.1 Residuenanalyse

Ist das unterstellte Modell korrekt, dann sollten die Residuen für alle Prädiktorenkombinationen mit homogener Varianz um Null variieren. Diese Bedingung ist bei unseren Daten gut erfüllt:



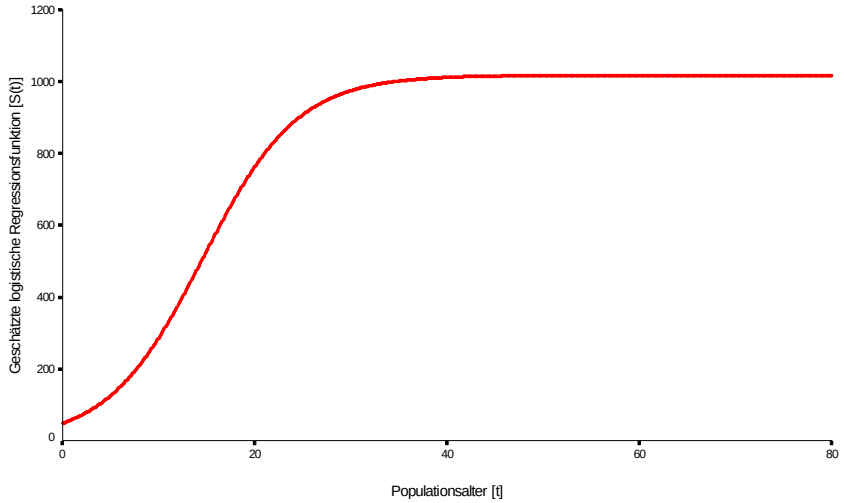
Daher schließen wir auf die Gültigkeit unseres Modells.

5.2 Ergebnisse zur globalen Modellbeurteilung (Determinationskoeffizient)

Wir erhalten bei der nichtlinearen Regressionsanalyse wie bei der linearen Variante zur globalen Modellbeurteilung Quadratsummen und einen Determinationskoeffizienten. Allerdings können diese Größen nicht inferenzstatistisch beurteilt werden. So ist z.B. die mittlere Fehlerquadratsumme auch bei gültigem Modell keine unverzerrte Schätzung der Fehlervarianz σ^2 im Modell (2), weshalb kein F-Test zur globalen Nullhypothese möglich ist. Für unser Beispiel erhalten wir:

```
Nonlinear Regression Summary Statistics     Dependent Variable S

  Source                 DF  Sum of Squares  Mean Square

  Regression              2  72729348,7989  36364674,3995
  Residual               98  2432193,20106    24818,29797
  Uncorrected Total     100  75161542,0000

  (Corrected Total)      99  11350202,7600

  R squared = 1 - Residual SS / Corrected SS =     ,78571
```

Trotz der gebotenen Vorsicht kann der (nicht adjustierte) Determinationskoeffizient von 0,79 als Beleg für eine gelungene Modellierung betrachtet werden. Er ist allerdings kaum höher als die nicht adjustierte Variante des Determinationskoeffizient für die lineare Regression von S auf T^2 (0,78).
Unser logistisches Modell hat aber einen entscheidenden Vorzug: Es ist korrekt. Dies zeigt sich u.a. darin, daß es sinnvolle Extrapolationen für die zu erwartende, weitere Populationsentwicklung produziert. Die folgende Abbildung zeigt die logistische Funktion mit den aus unserer Stichprobe geschätzten Parametern:



5.3 Parameterschätzungen und approximative Konfidenzintervalle

Bei einer nichtlinearen Regression können lediglich asymptotische Standardfehler bzw. Vertrauensintervalle bestimmt werden, die nur in großen Stichproben zuverlässig sind. Alle Ergebnisse beruhen auf der lokalen Linearisierung des Problems in einer Umgebung um die geschätzten Parameterwerte (siehe Draper & Smith, 1981, S. 462ff). Darauf werden die klassischen Methoden der linearen Regressionsanalyse angewendet.
Für unser Beispiel erhalten wir recht präzise Schätzungen:

```
                                          Asymptotic 95 %
                          Asymptotic     Confidence Interval
  Parameter   Estimate    Std. Error     Lower         Upper

  ALPHA     1018,4895299 22,642607431 973,55601237 1063,4230474
  W           ,203185363   ,010001779   ,183337158   ,223033568
```

Alle Konfidenzintervalle sind weit vom Wert Null entfernt, so daß ein Forscher aufgrund dieser Ergebnisse mit einiger Sicherheit auf signifikante Parameter schließen könnte.

5.4 Asymptotische Korrelationsmatrix der Parameterschätzer

Die ebenfalls nur asymptotisch korrekte Korrelationsmatrix der Parameterschätzungen kann eine **Überparametrisierung** des Modells bzw. ein Konditionierungsproblem aufdecken, erkennbar an sehr hohen Korrelationen zwischen verschiedenen Parameterschätzungen. Ursache können überflüssige Modellpa-

rameter sein, aber auch ungünstige Datenverhältnisse, die keine Identifikation aller Parameter ermöglichen (siehe Draper & Smith, 1981, S. 466ff).
In unserem Beispiel treten keine Konditionierungsprobleme auf:

Asymptotic Correlation Matrix of the Parameter Estimates

	ALPHA	W
ALPHA	1,0000	-,4768
W	-,4768	1,0000

6 Weitere Optionen zur nichtlinearen Regression in SPSS

SPSS bietet u.a. noch die folgenden Möglichkeiten bei der nichtlinearen Regression an:

- Ein alternatives Schätzverfahren: Quadratische Programmierung
- Schätzung von Standardfehlern und Vertrauensintervallen mit der Bootstrap-Methode
- Einschränkungen des Parameterraums (nur mit der Quadratische Programmierung)
- Segmentierungen des Parameterraums („Strukturbrüche“)

7 Literatur

Draper, N.R. & Smith, H. (1981). Applied regression analysis. New York: Wiley.
Norušis, M. J. (1994). *SPSS Advanced Statistics 6.1*. Chicago, IL: SPSS.