

Parameterschätzen in logistischen Wachstumsmodellen

JOACHIM ENGEL, LUDWIGSBURG

Zusammenfassung: *Das Anpassen von Kurven an Streudiagrammdaten verbindet Inhalte aus Analysis und Stochastik in anwendungsbezogenen Kontexten. Im Aufsatz werden Methoden vorgestellt, wie mit Unterstützung dynamisch-interaktiver Software die Parameter in einem logistischen Wachstumsmodell geschätzt werden können.*

1 Einleitung

Fragestellungen aus der Biomathematik erweisen sich für einen anwendungsbezogenen Mathematikunterricht als attraktiv (siehe z. B. Ableitinger 2008), weil der Sachkontext oft auch ohne Spezialkenntnisse verständlich ist und die gesellschaftspolitische Relevanz von Themen aus den Bereichen Ökologie, Artenschutz oder demographischer Wandel Interesse wecken kann. Zugleich kann exemplarisch aufgezeigt werden, wie Mathematik zur Lösung von authentischen Problemen beiträgt. Die mathematische Analyse mag mitunter zu überraschenden Resultaten und zur Aufdeckung komplexer Strukturen (z. B. Räuber-Beute-Modelle) führen. Wachstumsprozesse bieten

neben interessanten zu entdeckenden innermathematischen Strukturen besondere Gelegenheiten, den gesamten Modellierungsprozess einschließlich Modellbildung und kritischer Evaluation ins Zentrum der Schülerüberlegungen zu stellen. Der Modellierungsschritt kann hierbei mit einem Streudiagramm der Daten beginnen, auf dessen Grundlage geeignete funktionale Modelle postuliert werden, die freilich auch vom Sachkontext her einer kritischen Prüfung standhalten müssen. Oft stehen bei der Modellierung von Wachstumsprozessen am Anfang eher strukturorientierte Analysen, z. B. indem auf der Grundlage theoretischer Überlegungen Annahmen über das lokale Änderungsverhalten eines Systems in Form von Differenzen- oder Differenzialgleichungen formuliert werden. Egal ob ein eher datenorientierter oder eher strukturorientierter Zugang gewählt wurde: Hat man sich auf eine bestimmte Funktionenklasse festgelegt, so müssen zur weiteren Spezifizierung des Modells Parameter geschätzt werden. Die Modellierung funktionaler Abhängigkeiten steht somit im Schnittfeld zwischen anwendungsbezogener Analysis und Stochastik.

2 Logistisches Wachstum

Wachstumskurven sind Kurven, die die Änderung eines Phänomens oder Sachverhalts über die Zeit beschreiben. Oft werden Wachstumsvorgänge durch eine Spezifizierung ihres lokalen Änderungsverhaltens in Form von Differenzen- oder Differenzialgleichungen formuliert. Unter didaktischen Gesichtspunkten nimmt dabei logistisches Wachstum eine besondere Rolle ein, weil es einerseits noch relativ leicht zu charakterisieren und zu motivieren ist, andererseits aber auch hinreichend komplex ist, um interessante Strukturen (bis hin zum Chaos im Feigenbaum-Diagramm) aufzuzeigen. Man unterscheidet zwischen einer diskreten Version logistischen Wachstums, bei der der Bevölkerungsstand nur zu festen (diskreten) Zeittakten erfasst wird und einem stetigen Wachstumsmodell, bei dem der Umfang der Bevölkerung als stetige Funktion $y(t)$ der reellen Variable t (der Zeit) angesehen wird.

Diskretes logistisches Wachstum mit Anfangswert y_0

$$y_{n+1} = y_n + ky_n(S - y_n) \quad (1)$$

Kontinuierliches logistisches Wachstum mit Anfangswert $y(t_0) = y_0$

$$y'(t) = ky(t)(S - y(t)) \quad (2)$$

Hierbei bezeichnet S die Obergrenze oder den Sättigungswert der Population, k ist ein Wachstumsfaktor und y_0 ist der Wert der Population zum Zeitpunkt t_0 . In der Regel ist $y_0 < S$.

Während im diskreten Fall die Differenzengleichung (1) gar nicht in eine explizite Formel für y_n überführt werden kann (und Anlass für interessante Chaos-Studien gibt), hat die logistische Differenzialgleichung (2) die explizite Lösung

$$y(t) = \frac{y_0 S}{y_0 + (S - y_0) \exp(-kSt)} \quad (3)$$

Die Funktion $y(t)$ hat den typischen sigmoiden Verlauf. Anfangs wächst die Kurve sehr schnell („fast“

exponentiell), dann nimmt die Wachstumsrate immer mehr ab, bis sich die Funktion dem Grenzwert S annähert.

Die unbekannt Parameter y_0 , k und S müssen zur näheren Spezifizierung des Modells in konkreten Anwendungen geschätzt werden. Zwar liegt zum Zeitpunkt t_0 ein Messwert, nämlich y_0 , vor. Jedoch außer im unrealistischen Fall eines perfekt passenden Modells (für das ja im Allgemeinen schon drei Messwerte zur Bestimmung der drei Parameter ausreichen würden) sind alle Messungen, y_0 eingeschlossen, mit Fehlern bzw. Abweichungen vom Modell behaftet. Modell und Daten sind nun mal nicht identisch. Daher muss auch der vom Modell vorhergesagte Anfangswert geschätzt werden. Auf eine notationelle Unterscheidung etwa in Form von y_0 und \hat{y}_0 soll aber, bei Anerkennung der damit verbundenen Gefahr der Einführung eines möglichen Missverständnisses, zugunsten der einfacheren Darstellung verzichtet werden.

Neuere Schulbücher und Unterrichtsmaterialien zur Analysis greifen den Themenbereich logistisches Wachstum auf, weil sich hier interessante Beispiele für einen datenorientierten Unterricht ergeben, der Inhalte aus Analysis und Stochastik verbindet, z. B. bei der Modellierung von Hefewachstum (Kohorst & Portscher 1999), Wachstum einer Sonnenblume (Hull & Langkamp 2001), der Entwicklung der PKW-Dichte in Deutschland, des Schienennetzes der Bahn oder der Nutzung und Verbreitung moderner Technologien wie z. B. Mobiltelefone oder Internetrechner. Abbildung 1 zeigt eine Aufgabenstellung zur Entwicklung der Bevölkerungszahlen in den USA zwischen 1790 und 1990 aus dem Schulbuch zum Leistungskurs in Analysis von Lambacher-Schweizer (2000).

Die inhaltliche Begründung für ein logistisches Wachstumsmodell mag sich von Kontext zu Kontext unterscheiden. In manchen Fällen wie z. B. der Verbreitung von Mobiltelefonen bezieht man sich auf die Annahme, dass die Ausbreitungsrate $y'(t)$ sowohl

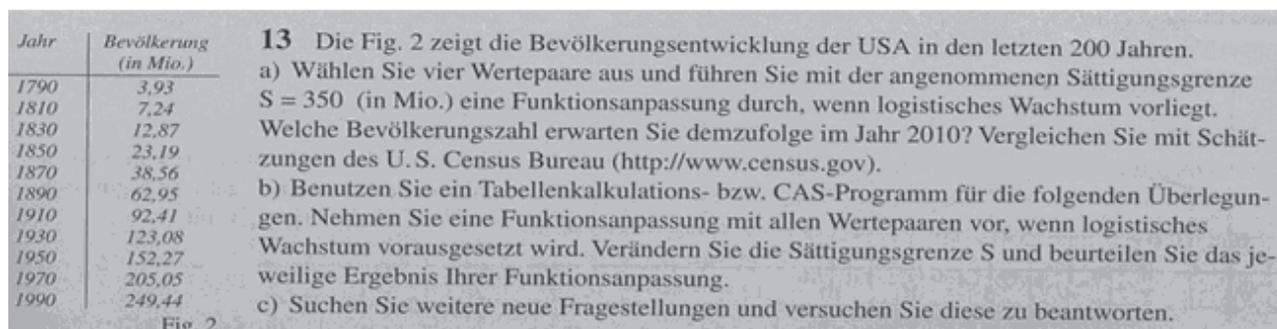


Abb. 1: Schulbuchaufgabe aus Lambacher-Schweizer, LK Analysis

proportional zum jetzigen Bestand $y(t)$ wie auch zur noch frei verfügbaren Kapazität (d. h. der Anzahl der Leute, die noch kein Telefon besitzen) $S - y$ sei. Ähnliche Überlegungen dienen auch als erste Annäherung bei der Modellierung des Ausbreitens von Krankheiten oder der Verbreitung eines Gerüchtes. In anderen Situationen wie z. B. bei der Entwicklung der Bevölkerungszahlen eines Landes oder der Entwicklung einer speziellen Tierpopulation über einen gewissen Zeitraum ist eine theoriegeleitete Begründung für ein logistisches Modell, wenn auch nicht unmöglich, so doch um einiges schwieriger zu erbringen. Eine Visualisierung der Daten im Streudiagramm, die man wohl in jedem Fall vornehmen sollte, mag dann tatsächlich die ersten Hinweise auf ein passendes funktionales Modell geben.

3 Parameterschätzen im logistischen Wachstumsmodell

Wir gehen im Folgenden von Daten $(t_1, y_1), \dots, (t_n, y_n)$ aus, an die wir ein logistisches Modell anpassen wollen. Dazu müssen drei Parameter, nämlich S , y_0 und k , geschätzt werden. Mit geeigneter Software könnte man vielleicht erwägen, drei Schieberegler einzuführen und solange die Werte zu verändern, bis eine nach Augenmaß zufrieden stellende Anpassung gefunden ist. Als erste heuristische Annäherung mag dieses Vorgehen auch genügen, die gleichzeitige Variation von drei Stellgrößen kann aber auch schnell in die Irre führen. Sieht man vom graduellen Anpassen von Schieberegler ab, so sind beim Anpassen von nichtlinearen Modellen an Daten prinzipiell zwei Vorgehensweisen denkbar:

1. Linearisierung: Eine Methode zur Parameterschätzung, die zumindest in einigen (aber bei weitem nicht in allen) nichtlinearen Modellen funktioniert, besteht darin, die Daten geeignet zu transformieren, so dass die transformierten Daten eine lineare Struktur besitzen. Dann kann in die transformierten Daten eine Gerade, z. B. per kleinsten-Quadrate-Methode, eingepasst werden. Per Rücktransformation erhält man dann ein Modell für die ursprünglichen Daten.
2. Nichtlineare Regression: Alternativ kann man versuchen, direkt ein kleinsten-Quadrate-Kriterium zu minimieren.

3.1 Linearisierung

Im Prinzip ist man beim logistischen Modell in der glücklichen Lage, dass eine Linearisierung möglich ist. Aus Gleichung (3) folgt nämlich

$$\ln\left(\frac{1}{y(t)} - \frac{1}{S}\right) = \ln\left(\frac{S - y_0}{S y_0}\right) - kSt, \quad (4)$$

woraus geschlossen werden kann, dass – falls das logistische Modell gilt und S angemessen spezifiziert wurde – das Streudiagramm der transformierten Daten (t_i, y_i^*) mit

$$y_i^* = \ln\left(\frac{1}{y_i} - \frac{1}{S}\right) \quad (5)$$

eine lineare Struktur hat. Mit Hilfe einer Geradenanpassung im Streudiagramm $(t_1, y_1^*), \dots, (t_n, y_n^*)$

$$y^* = mt + b$$

lassen sich die fehlenden Parameter y_0 und k aus dem Achsenabschnitt und der Steigung der eingepassten (kleinsten-Quadrate-) Gerade schätzen. Per Rücktransformation folgt für die Parameter k und y_0 im ursprünglichen nichtlinearen Modell

$$k = -\frac{m}{S}, \quad y_0 = \frac{S}{1 + S \exp(b)}.$$

Die dargestellte Vorgehensweise hat aber leider noch einen entscheidenden Mangel: Wir haben so getan, als würden wir die Obergrenze S kennen. Abgesehen von Beispielen wie etwa der einen See zunehmend bedeckenden Algenpest ist dies nur in ganz wenigen Anwendungen der Fall. Oft ist gerade die Bestimmung der Obergrenze S die inhaltlich am meisten interessierende Frage („Wie viele Personen werden von einer Epidemie betroffen?“). Im Schulbuch Lambacher-Schweizer (2000) wird S einfach vorgegeben (siehe Abbildung 1). Damit ist die am meisten interessierende Frage ohne jegliche Begründung beantwortet. Mit Hilfe dynamisch-interaktiver Software wie FATHOM lässt sich auch mit schultauglichen Mitteln die Obergrenze S aus den Daten schätzen. Wir illustrieren die Vorgehensweise anhand des Beispiels der Entwicklung der US-Bevölkerung zwischen 1790 und 1990, die mit Hilfe eines logistischen Modells ausgezeichnet beschrieben werden kann. Die Daten sind in Abbildung 2 in einem Streudiagramm wiedergegeben. Eine genauere Betrachtung zeigt einen starken Anstieg zwischen 1820 und 1900 und nachlassende Wachstumsintensität im 20. Jahrhundert. Dies weist auf eine logistische Funktion als geeignetes Modell hin. Allerdings fällt die weitgehende Stagnation zur Zeit des 2. Weltkrieges auf. Diese Überlegungen zeigen auch, dass die Entwicklung von menschlichen Bevölkerungen keinen festen Naturgesetzen folgt, sondern von Menschen verursachte Vorgänge sind.

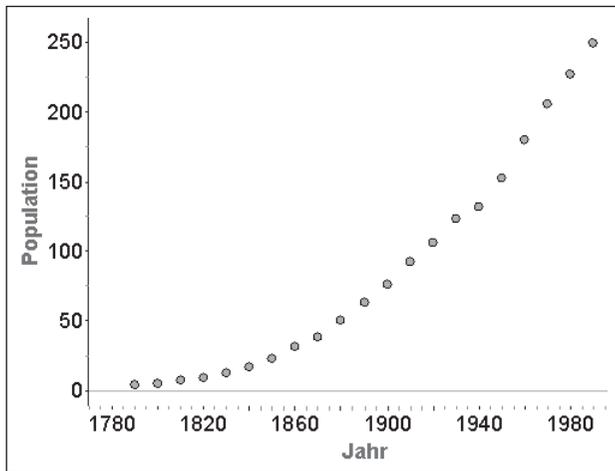


Abb. 2: Bevölkerungsentwicklung der USA (in Millionen) zwischen 1790 und 1990

Bei der folgenden Modellierung kommen wir ohne Vorgabe der Obergrenze S aus. Wir definieren einen Schieberegler mit Namen S und führen – basierend auf diesem S – die Transformation (3) durch. Die transformierten Daten $(t_1, y_1^*), \dots, (t_n, y_n^*)$ stellen wir in einem Streudiagramm dar, zeichnen die kleinst-Quadrate-Gerade ein und berechnen die Summe der Abweichungsquadrate (siehe Abbildung 3). Diese Aktionen sind in FATHOM per einfachem Klick im Kontextmenu schnell verfügbar (siehe Biehler et al. 2006). Jetzt verändern wir den Wert von S im Schieberegler graduell, bis wir eine kleinst-mögliche Summe der Abweichungsquadrate erzielt haben. Da hier nur eine einzige Stellgröße verändert wird, führt dieser Schritt zu einem zuverlässigen und robusten Ergebnis.

Mit diesem Wert von S werden die Daten so gut es nur irgend geht (im Sinne des k-Q-Kriteriums) in einer lineare Struktur überführt. Den so erhaltenen Wert für S nehmen wir als Schätzwert der Populationsobergrenze. Der Rest (Schätzung von k und y_0) erfolgt dann wie oben beschrieben.

Auf die vorliegenden Daten zur Entwicklung der US-Bevölkerung zwischen 1790 und 1990 angewandt ergibt sich das in Abbildung 3 (linke Darstellung) dargestellte Bild. Bei einem Wert von $S = 317$ [Millionen] ist die Summe der Abweichungsquadrate (annähernd) minimal (Summe der Abweichungsquadrate 0,2094). Aus der in das Streudiagramm der transformierten Daten eingepassten Geraden mit der Gleichung

$$y^* = -0,0272 t + 47,2$$

und $S = 317$ als Sättigungsgrenze lassen sich jetzt mittels Vergleich mit der Gleichung (4) Schätzwerte für k und y_0 herleiten.

$$k = -\frac{-0,02719}{S} = 0,00006845,$$

$$y_0 = \frac{S}{1 + S \exp(-0,0272 \cdot 1790 + 47,2)} = 4,367$$

Jetzt kann die eingepasste logistische Funktion in das ursprüngliche Streudiagramm eingezeichnet werden. Zur Modellvalidierung erstellen wir ein Residuendiagramm (Abbildung 3, rechte Darstellung). Die Summe der Abweichungsquadrate beträgt hier 695,5.

Wir stellen dabei fest, dass es während der Zeit des 2. Weltkrieges noch zu beachtlichen Abweichun-

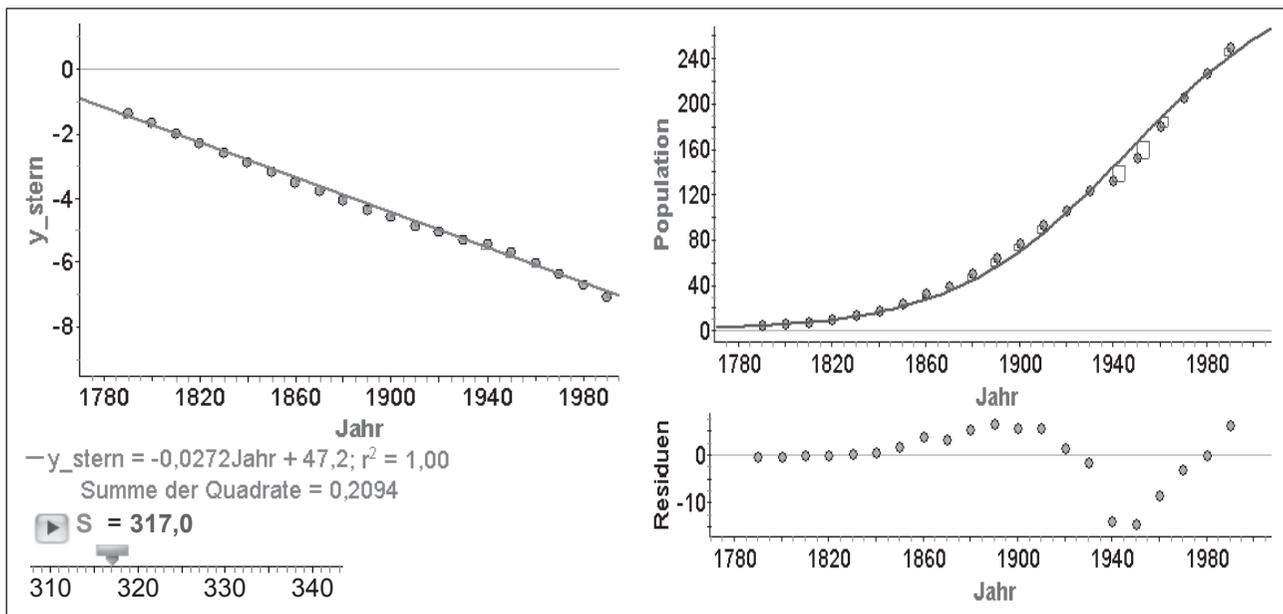


Abb. 3: Linearisierte Daten zur Bevölkerung der USA zwischen 1790 und 1990, wobei die Obergrenze S die Summe der quadrierten Abstände minimiert.

gen vom logistischen Modell kommt, die aber wohl durch die Besonderheit dieser Jahre auch erklärbar sind. Wenn man sich vor Augen führt, durch welche menschen-gemachten Verursachungen es zu Bevölkerungszuwächsen in einem traditionellen Einwandererland wie den USA kommt (Hungersnöte in Europa, politische und religiöse Verfolgung, Krieg), so ist es erstaunlich, dass sich die Bevölkerungsentwicklung überhaupt mit Hilfe eines mathematischen Modells wie der logistischen Wachstumsfunktion beschreiben lässt.

Die durchgeführten Aktionen lassen sich übrigens auch ohne ein Programm wie FATHOM ausführen. Die Sättigungsgrenze S wurde bestimmt als per Augenmaß bestimmtes Minimum der Funktion

$$\varphi(S) = \sum_{i=1}^n \left[\ln\left(\frac{1}{y_i} - \frac{1}{S}\right) - (mt_i + b) \right]^2,$$

wobei m und b selbst wieder auf komplizierte Art von S abhängen. Wir verzichten auf die rechtechnisch sehr aufwändigen Details, weisen aber darauf hin, dass man auch mit Hilfe mathematischer Algorithmen das Minimum der Funktion φ suchen kann. Hier zeigen sich überzeugend die Annehmlichkeiten einer dynamisch interaktiven und benutzerfreundlichen Software wie FATHOM.

3.2 Nichtlineare Regression

Als Alternative zum Transformationsansatz kann man auch darauf abzielen, direkt ein kleinstes-Quadrat-Kriterium zu minimieren. Gesucht sind dabei Werte für S , k und y_0 , so dass

$$f(y_0, S, k) = \sum_{i=1}^n \left[y_i - \frac{y_0 S}{y_0 + (S - y_0) \exp(-kSt)} \right]^2$$

minimal wird. Die Lösung dieses multivariaten nicht-linearen Minimierungsproblems ist recht anspruchsvoll und verlangt aufwändige numerische Verfahren wie z. B. den Gauß-Newton-Algorithmus, einer mehrdimensionalen Verallgemeinerung des Newton-Verfahrens zur iterativen Berechnung von Nullstellen univariater reeller Funktionen. Wie auch das einfachere univariate Newton-Verfahren ist beim multivariaten Gauß-Newton-Verfahren die Konvergenz der Iterationsfolge keineswegs gesichert und hängt stark von der Wahl geeigneter Anfangsschätzer für die Parameter ab.

Derartige Minimierungsaufgaben verlangen nach leistungsfähiger Software. Das Programmsystem R zur statistischen Datenanalyse und zur graphischen Darstellung von Daten ist nicht nur als Open Source Software unter <http://www.r-project.org> frei verfü-

bar, sondern auch sehr flexibel und leistungsstark. Die Routine `nls` erlaubt die flexible Definition von Funktionenklassen, die an vorgegebene Datensätze angepasst werden, indem per Gauß-Newton Verfahren die Parameter geschätzt werden. Ein Befehl wie

```
nls(Pop~y0*S/(y0+(S-y0)*exp(-k*S*Jahr)),
start=list(y0=5,S=330,k=0.000006))
```

führt im vorliegenden Fall logistischen Wachstums den iterativen Gauß-Newton-Algorithmus aus. Die Daten stecken in den Variablen `Pop` bzw. `Jahr`.

Hier erhalten wir mit den Startwerten 5, 330, 0,00006 die Parameterschätzungen

$$y_0 = 6,957, \quad S = 385,9 \quad \text{und} \quad k = 0,0000587.$$

Ein Programmaufruf mit anderen, auch stark abweichenden Anfangswerten, führt zum gleichen Ergebnis. Die Summe der Abstandsquadrate beträgt hier nur 341,1. Abbildung 4 zeigt die eingepasste logistische Funktion mitsamt Residuendiagramm. Man erkennt die leicht verbesserte Modellanpassung, obgleich systematische Abweichungen zwischen Daten und Modell in der ersten Hälfte des 20. Jahrhunderts auch hier unverkennbar sind.

Für das Jahr 2010 sagt dieses Modell eine Bevölkerungsgröße von 281 Millionen voraus. Die tatsächliche Größe der Bevölkerung im November 2009 liegt jedoch bei 307 Millionen Einwohnern (siehe www.census.gov/population/www). Gerade seit 1990 hat die Bevölkerung stärker zugenommen als vom logistischen Modell vorhergesagt.

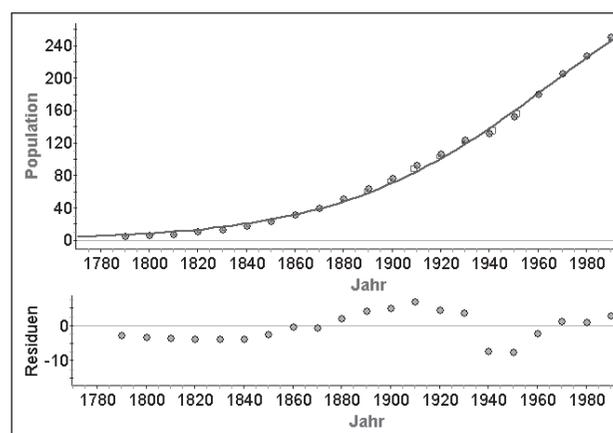


Abb. 4: US-Bevölkerungsdaten und per nicht-linearer Regression eingepasste logistische Funktion

4 Zusammenfassung

Es ist nicht überraschend, dass die beiden hier betrachteten Methoden – der Ansatz über Linearisierung der Daten wie auch die direkte Minimierung

eines kleinste-Quadrate-Kriteriums – zu sehr ähnlichen Resultaten führen. Gleichzeitig gelingt es im vorliegenden Beispiel der nichtlinearen Regression, die Summe der quadrierten Abweichungen um mehr als 50 % gegenüber dem Transformationsansatz zu reduzieren. Die resultierenden Kurven sind also nicht identisch.

Viele nichtlineare Modelle können durch geeignete Transformationen linearisiert werden. Nach Geradenanpassung und nachfolgender Rücktransformation lässt sich dann ein Modell für die Ursprungsdaten finden. Allerdings ist man oft auch nicht in der glücklichen Lage, eine passende Transformation verfügbar zu haben. Entweder ist eine passende Transformation selbst noch von unbekanntem Parametern abhängig. Im vorliegenden Fall logistischen Wachstums haben wir einen Weg aufgezeigt, wie man sich hier mit Unterstützung von Technologie helfen kann. In komplexeren Situationen – etwa wenn geeignete Transformationen nicht nur von einem, sondern von mehreren Parametern abhängen – wird dieser Weg nicht mehr zu beschreiten sein. In manchen Situationen ist eine Linearisierung auch aus strukturellen Gründen gar nicht möglich, weil die Daten nicht monoton auf eine lineare Skala abgebildet werden können. Beispiele hierfür sind Wellenbewegungen, Sonnenaufgangszeiten und Daten, die stark von der Tageszeit oder Jahreszeit abhängen. Periodische Strukturen können nicht streng monoton in eine lineare Struktur abgebildet werden. Anstatt des Transformations-Rücktransformations-Ansatzes kann man dann versuchen, ein Außenkriterium wie die Summe der quadratischen Abstände zwischen Modell und Daten direkt zu minimieren. Verfahren der nichtlinearen Regression verlangen in der Regel den Einsatz anspruchsvoller softwaregestützter numerischer Verfahren. Bei Problemstellungen mit wenigen Parametern mag man geneigt sein, optimale Parameter auch per Schieberegler durch Ausprobieren anzunähern, wobei Software zur Illustration eingesetzt werden kann. Jedoch sollte man beachten: Die Durchführbarkeit eines Vorgehens ist ein Aspekt, die Angemessenheit eine andere Sache. Wenn wir stochastisch-funktionale Zusammenhänge modellieren, modellieren wir immer gleichzeitig Signal und Rauschen, Struktur und zufällige Abweichung. Durch Linearisierung der Daten verändern wir nicht nur die Struktur (d. h. die

funktionale Abhängigkeit wird linear), sondern auch die Abweichungen zwischen Modell und Struktur. Daher ist es nicht verwunderlich, dass wir in Situationen, in denen beide Methoden anwendbar sind, unterschiedliche Resultate erhalten, je nachdem ob eine Funktionsanpassung per Transformationsansatz oder per nichtlinearer Regression erfolgt.

Weitere interessante Beispiele von realen Datensätzen mit logistischer Struktur, an denen die hier besprochenen Methoden eingesetzt und erprobt werden können, finden sich bei Engel (2009) und auf der Homepage des Quantitative Environmental Learning Projects von Hull und Langkamp (2001).

Literatur

- ABLEITINGER, C. (2008): Diskrete biomathematische Modelle im Schulunterricht – Chancen aus der Sicht der Mathematikdidaktik. Dissertation, Universität Wien.
- BIEHLER, R.; HOFMANN, T.; MAXARA, C.; PRÖMEL, A. (2006): Fathom 2. Eine Einführung. Heidelberg: Springer
- ENGEL, J. (2009): Anwendungsorientierte Mathematik: Von Daten zur Funktion. Springer: Heidelberg.
- HULL, J. & LANGKAMP, G. (2001): Quantitative Environmental Learning Project. <http://www.seattlecentral.org/qelp> (Zugriff: 5.10.2009).
- KOHORST, HELMUT & PORTSCHELLER, PHILIPP (1999): Vom exponentiellen zum logistischen Wachstum. Wozu Hefe nicht alles gut ist. In: *mathematiklehren* 97.
- LAMBACHER-SCHWEIZER (2000): LK Analysis, Ausgabe Baden-Württemberg. Stuttgart: Klett.

Anschrift des Verfassers

Joachim Engel
Institut für Mathematik und Informatik
Pädagogische Hochschule Ludwigsburg
Reuteallee 46
71634 Ludwigsburg
engel@ph-ludwigsburg.de