Parameteridentifikation am Beispiel einer Giardia-Epidemie in Bergen (Norwegen) Hausarbeit zur Vorlesung Mathematische Modellierung mit Differentialgleichungen (Dr. Etienne Emmrich) Wintersemester 2004/05

Stud.Techn Terje Tofteberg
Norwegian University of Science and Technology

Stud.Math Are Losnegård
University of Bergen

4. April 2005

Zusammenfassung

In dieser Hausarbeit haben wir versucht parametrisierte Modelle an die Giardia-Epidemie in Bergen, Norwegen, Herbst 2004 anzupassen. Das SIR-Modell von Kermack und McKendrick wird diskutiert, sowie logistisches Wachstum. Die beiden Modelle beschreiben die Ausbreitung der Epidemie ohne große Abweichungen, aber das logistische Wachstums-Modell wird vorgezogen weil es nur zwei Parameter zu bestimmen gibt, im Gegensatz zu vier im SIR-Modell.

Wir haben untersucht, wie kleine Änderungen der Input-Daten zu Änderungen der Parameter führen. In dieser Aufgabe hatten wir Daten für die ganze Epidemiezeit und es stellte sich heraus, dass die Stabilität sehr hoch ist. Wenn wir nur Daten für den Anfangszeitraum hätten, haben wir gezeigt, dass die Stabilität wesentlich geringer wird. D.h. dass die Methode nicht geeignet ist, um Prognosen zu machen.

Jeg drakk den Skaal, som mig Ulriken skiænkte;
drikker den samme, I, som have Viin.
Hver som opriktig mod Fødeby tænkte,
lod denne Munterheds Skaal være sin.
Held for vort Bergen, vort Fødelands Held!
Giv alting maa blomstre fra Fiere - til Field.

Inhaltsverzeichnis

1 Giardia-Epidemie in Bergen, Norwegen

Von Oktober bis Dezember 2004 gab es in Bergen eine Epidemie der Amöbe *Giardia lamblia*. Diese Amöbe ist weit verbreitet in Asien, Afrika und in Süd-Amerika, aber normalerweise gibt es nicht mehr als 5 Fälle pro Jahr in einer Stadt wie Bergen, die zweitgrößte Stadt Norwegens mit 250 000 Einwohnern. Es wird angenommen, dass derzeit 200 millionen Menschen der Welt mit Giardia infiziert sind. Wenn eine Person mit Giardia infiziert wird, gelangen die Amöben in den Dünndarm. Nach einer Woche kommen Symptome wie Durchfall und Luftshcmerzen. Die Amöben können sich weiterverbreiten über Faeces. Diese Form der Amöben sind nicht so richtig problematisch, weil sie nicht lange überleben können ohne einen Träger. Wenn die Darmtätigkeit nach ein paar Tagen nachgelassen hat, weil das Magen-Darm-System des Opfers fast leer ist, gehen die Amöben in einen Schlafmodus. Sie bauen um sich eine dicke Wand und können in dieser Form monatelang im Wasser überleben, siehe *Skorping 2004* [2].

Der Ansteckungsherd in Bergen war wahrscheinlich ein Klärbecken eines Cafes am Gipfel Ulrikens, dem höchsten Berg Bergens. Die Amöben sind mit dem Wasser nach Svartediket, Bergens größtes Wasserbecken, gekommen.

Die Anzahl der Infizierten ist in Abbildung 1 gegeben. Die Zahlen in Abbildung 1 represäntieren die Patienten, die im Herbst 2004 zum Arzt gingen, siehe *Offentliche Information, Bergen Kommune* [6]. Die Zahl der tatsächlich Infizierten wird noch größer sein. Unser Ziel mit dieser Aufgabe ist, Parameter zu bestimmen, so dass wir eine Kurve bekommen, die mit diesen Daten bestmöglb h zusammenpasst.

2 SIR-Modell von Kermack-McKendrick

Für die Berechnung der Ausbreitung vieler Krankheiten kann man, wie es in *Emmrich* [4] gemacht wird, als Vereinfachung ein SIR-Modell benutzen. In diesem Modell unterteilt man die Bevölkerung in drei Klassen:

- S(t) - Susceptables : Die für die Krankheit anfälligen Personen zum Zeitpunkt t.
- I(t) - Infectives : Die Infektiösen Personen zum Zeitpunkt t.
- R(t) - Removed : Die erkrankten,verstorbenen oder sonst aus dem Ansteckungsverlauf ausgeschidenen Personen zum Zeitpunkt t.

In diesem SIR-Modell von Kermack und McKendrick bezeichnen wir die Infektionsrate mit α und die Erholungsrate mit β. Wir erhalten somit das

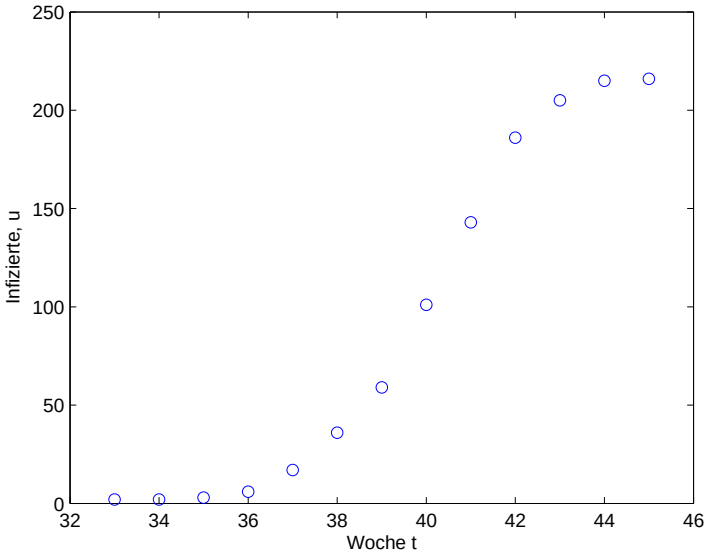

Abbildung 1: Giardia-Infizierte in Bergen, Herbst 2004

Differentialgleichungssystem:

$$S'(t) = -\alpha S(t)I(t) \tag{1a}$$

$$I'(t) = \quad \alpha S(t)I(t) - \beta I(t) \tag{1b}$$

$$R'(t) = \qquad\qquad \beta I(t) \tag{1c}$$

Die Gesamtpopulation ist $N(t) = S(t) + I(t) + R(t)$. Aus diesem Modell ergibt sich $S' + I' + R' = 0$. Also ist $N'(t) = 0$ und daraus folgt $N(t) = N(0)$. Wir setzen $\rho = \frac{\beta}{\alpha}$. Es folgt:

$$\frac{dI(t)}{dS(t)} = -1 + \frac{\rho}{S(t)}$$

$$\frac{dS(t)}{dR(t)} = -\frac{S(t)}{\rho} \quad .$$

Es handelt sich um separable Differentialgleichungen. Wenn man diese nun integriert, erhält man:

$$I(t) = -S(t) + \rho \ln S(t) + c_1$$

$$S(t) = c_2 e^{-R(t)/\rho}$$

wobei

$$c_1 = I(0) + S(0) - \rho \ln S(0) = N - \rho \ln S(0)$$

$$c_2 = S(0)$$

wenn wir $R(0) = 0$ annehmen. Somit ergibt sich

$$I(t) = N - S(t) - \rho \ln \frac{S(0)}{S}$$

$$S(t) = S(0)e^{-R(t)/\rho}$$

Normalerweise ist $I(0) << S(0)$, damit wir $S(0) \approx N$ setzen können. Wenn man eine Krankheit studiert und ein Modell entwerfen möchte, hat man normalerweise nur Daten für R(t). Also ist der Ausdruck für $R'(t)$ für uns das Interessanteste:

$$R'(t) = \beta I(t) = \beta(N - R(t) - S(t)) = \beta \left(N - R(t) - S(0)e^{-R(t)/\rho}\right) \quad (2)$$

Diesen Ausdruck können wir anders schreiben. Man macht eine Taylorentwicklung von $e^{-\frac{R}{\rho}}$. Wir nehmen an dass $R(t) << \rho$ ist, und machen die Entwicklung um 0:

$$e^{-R(t)/\rho} = 1 - \frac{R(t)}{\rho} + \frac{1}{2}\left(\frac{R(t)}{\rho}\right)^2 + \dots$$

Damit erhalten wir für $R'(t)$:

$$R'(t) \approx \beta \left(N - R(t) - N\left(1 - \frac{R(t)}{\rho} + \frac{1}{2}\left(\frac{R(t)}{\rho}\right)^2\right)\right)$$

$$\approx \beta \left(R(t)\left(\frac{N}{\rho} - 1\right) - \frac{1}{2}\frac{NR(t)^2}{\rho^2}\right)$$

$$\approx \beta \left(\frac{N}{\rho} - 1\right) R(t) \left(1 - \frac{R(t)}{\frac{2\rho^2}{N}(\frac{N}{\rho} - 1)}\right)$$

Diesen Ausdruck können wir später benutzen, wenn wir die Fehlerquadrat-Methode für die Identifikation der Parameter anwenden werden. Wir haben drei Parameter, aber in unserem Ausdruck für R' können wir $\beta(\frac{N}{\rho} - 1)$ und $\frac{2\rho^2}{N}(\frac{N}{\rho} - 1)$ als bzw. b_1 und b_2 schreiben. Daraus ergibt sich:

$$R'(t) \approx b_1 R(t)\left(1 - \frac{R(t)}{b_2}\right) \quad . \quad (3)$$

3 Logistisches Wachstum

Ein einfaches Modell für die Populationsdynamik ist das logistische Wachstum. Es wird, nach *Bohl* [3], durch die Verhulstgleichung

$$u'(t) = ru(t)\left(1 - \frac{u(t)}{K}\right) \tag{4}$$

beschrieben. Man erhält genau dieselbe Gleichung wie (3), also wenn man eine Approximation für $R'(t)$ im SIR-Modell durchführt. K ist hier der Wert, nach dem $u(t)$ für $t \to \infty$ strebt. Dies wird ersichtlich weil $u'(t) = 0$ wenn $u(t) = K$. Weiter ist r ein Proportionalitätsfaktor. Wie in *Emmrich* [4] möchten wir jetzt $u(t)$ finden:

$$\frac{du}{dt} = ru(t)\left(1 - \frac{u(t)}{K}\right) = ru(t)\left(\frac{K - u(t)}{K}\right)$$

Nach Trennung der Veränderlichen erhält man:

$$\frac{Kdu}{u(t)(K - u(t))} = rdt \quad . \tag{5}$$

Partialbruchzerlegung liefert:

$$\frac{K}{u(t)(K - u(t))} = \frac{A}{u(t)} + \frac{B}{K - u(t)} \quad .$$

Hier ist

$$A(K - u(t)) + Bu(t) = K$$

also $A = 1$ und $B = 1$. Integration von (5) fuhrt auf von:

$$\int \left(\frac{1}{u(t)} + \frac{1}{K - u(t)}\right) du = rdt,$$

so dass

$$\ln \frac{u(t)}{K - u(t)} = rt + c \quad .$$

Daraus ergibt sich

$$\frac{u(t)}{K - u(t)} = e^{rt+c} = e^{rt}e^{c} \quad .$$

Für $t = 0$ haben wir

$$\frac{u_0}{K - u_0} = e^{c} \quad .$$

Somit erhalten wir

$$\frac{K - u(t)}{u(t)} = e^{-rt} \frac{K - u_0}{u_0}$$

Weiter geht es mit

$$\frac{K}{u(t)} = e^{-rt} \frac{K - u_0}{u_0} + 1 = \frac{e^{-rt}(K - u_0) + u_0}{u_0}$$

Schliesslich folgt

$$u(t) = \frac{K}{1 + \left(\frac{K}{u_0} - 1\right) e^{-rt}} \quad .$$

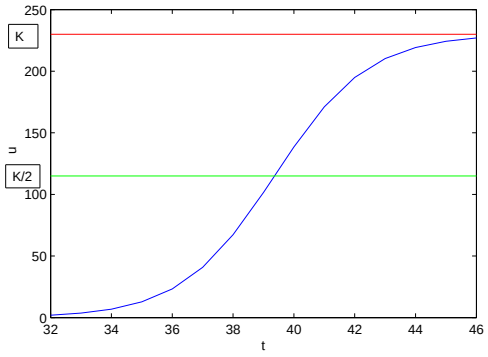

Abbildung 2: Logistisches Wachstum

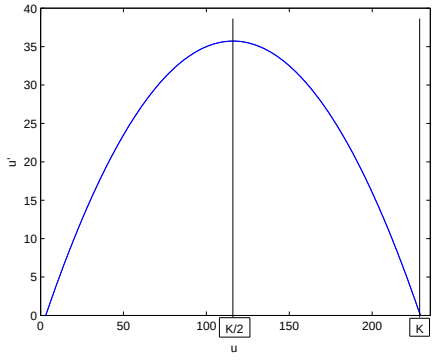

Abbildung 3: Logistisches Wachstum. Hier ist u'(t) eine Funktion von u(t).

4 Regression

4.1 Warum den quadratischen Fehler minimieren?

Wir haben ein Datenset (x_i, y_i) und ein Modell f mit m variablen Parametern $a_1 \dots a_m$, welches beschreibt wie y von x abhängt:

$$y = f(x; a_1 \dots a_m)$$

Um eine Kurve an dieses Datenset anzupassen, minimiert man den quadratischen Fehler:

$$\sum_{i=1}^{n} (y_i - f(x_i))^2$$

Warum minimiert man immer diese Größe und nicht ein anderes Maß der Abweichung, zum Beispiel den Betrag:

$$\sum_{i=1}^{n} |y_i - f(x_i)|$$

Wir nehmen hier an dass für jede Messung von y, gibt es eine Abweichung des Modells und dass diese Abweichung mit Erwartungswert 0 und Standardabweichung σ normalverteilt ist. D.h. wenn wir annehmen dass unser Modell richtig ist, ist die Wahrscheinlichkeit dass messung i von y liegt im Intervall $[y_i, y_i + \Delta y]$ gleich

$$\begin{aligned} P_i :&= P(y_i \in [y_i, y_i + \Delta y] | x, a_1 \dots a_m) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ \frac{-\left(y_i - f(x_i; a_1 \dots a_m)\right)^2}{2\sigma^2} \right\} \Delta y \end{aligned}$$

Δy kann hier beliebig klein gemacht werden. Die Totale Wahrscheinlichkeit dass alle die Messungen $y_i, i = 1 \dots n$ von y liegt in den Intervallen $[y_i, y_i + \Delta y]$ ist dann

$$P_{tot}(y_1 \dots y_n | x_1 \dots x_n, a_1 \dots a_m) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ \frac{-(y_i - f(x_i; a_1 \dots a_m))^2}{2\sigma^2} \right\} \Delta y, \tag{6}$$

Dann nutzen wir hier ein „Maximum Likelyhood“-Methode und sagen, dass unsere beste Schätzung für die Parameter $a_1 \dots a_n$ sind die, die die Wahrsceinlichkeit in Gleichung 6 maximieren. Weil der Logarithmus eines

Arguments eine monotone wachsende Funktion ist, können wir auch den Logarithmus maximieren:

$$\ln(P_{tot}) = n \ln\left(\frac{\Delta y}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(x_i; a_1 \dots a_m))^2$$

Hier ist n, σ und Δy nur konstante Größen. Das heißt, $\ln(P_{tot})$ wird durch die Minimierung des quadratischen Fehlers maximiert.

4.2 Lineare Regression

Sei y eine Variable, die linear von p Variablen $\mathbf{x} = (x_1 x_2 \dots x_p)^T$ abhängt,

$$y = \mathbf{x^T b} \tag{7}$$

wobei $\mathbf{b}$ eine Parametervektor ist. Von dieser Variablen haben wir n Messungen ($\mathbf{x_i}$, y_i). Wir möchten aus diesen Daten eine Schätzung des Parametervektors $\mathbf{b}$ machen. Normalerweise sind die Messungen mit einem Messfehler oder anderen Ungenauigkeiten behaftet, so dass jede Messung beschrieben werden kann durch:

$$y_i = \mathbf{x_i^T b} + e_i$$

wobei e_i ein Rauschterm ist. Wir nehmen an, dass das Rauschen einer Normalverteilung mit Erwartungswert 0 und Standardabweichung σ, unabhängig von $\mathbf{x_i}$ folgt. Wir nehmen auch an dass wir die unabhängigen Variablen $\mathbf{x_i}$ exakt bestimmen können. Wie in 4.1 gezeigt, möchten wir jetzt die Parameter $\mathbf{b}$ so bestimmen dass der quadratische Fehler so gering wie möglich wird. Für jeden Datenpunkt ist der Fehler v_i, eine Funktion von unserer Schätzung $\hat{\mathbf{b}}$ für $\mathbf{b}$:

$$v_i(\hat{\mathbf{b}}) = y_i - \mathbf{x_i^T}\hat{\mathbf{b}}$$

Der Totale quadratische Fehler V, ist dann gegeben durch

$$\begin{aligned} V &= \mathbf{v^T v} \\ &= (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) \\ &= \mathbf{y^T y} - \mathbf{y^T X}\hat{\mathbf{b}} - \hat{\mathbf{b}}^\mathbf{T}\mathbf{X^T y} + \hat{\mathbf{b}}^\mathbf{T}\mathbf{X^T X}\hat{\mathbf{b}} \end{aligned}$$

wobei

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \text{ und } \mathbf{X} = \begin{pmatrix} \mathbf{x_1^T} \\ \mathbf{x_2^T} \\ \vdots \\ \mathbf{x_n^T} \end{pmatrix}.$$

Um die Parametern $\hat{\mathbf{b}}$ zu bestimmen, machen wir die Ableitung von V, bezüglich $\hat{\mathbf{b}}$ und setzen gleich 0.

$$\frac{\partial V}{\partial \hat{\mathbf{b}}} = -2\mathbf{X^T y} + 2\mathbf{X^T X}\hat{\mathbf{b}} = 0$$

Mit dem Ergebnis

$$\hat{\mathbf{b}} = (\mathbf{X^T X})^{-1}\mathbf{X^T y} \tag{8}$$

Dieses $\hat{\mathbf{b}}$ ist unsere beste Schätzung für die Parameter $\mathbf{b}$. Die Matrix $\mathbf{X^T X}$ ist positiv definit und invertierbar wenn $\mathbf{X}$ eine reelle n-mal-p-Matrix mit $rang(\mathbf{X}) \geq p$ ist. Das heißt, die Anzahl der paarweise verschidene Messungen muss größer oder gleich sein als p.

5 Linearisierung eines nichtlineares Problems

Viele Probleme sind ursprunglich nichtlinear, trotzdem ist es manchmal möglich die ober beschreibenen Verfahren zu benutzen. Im Folgenden sind drei verschiedene Fälle skizziert.

1. Die Funktion y ist ein Polynom.

   $$y = b_0 + b_1 x + b_2 x^2 + \ldots + b_p x^p$$

   Hier führen wir eine neue Matrix ein, die sogenannte Vondermonde Matrix.

   $$\hat{\mathbf{X}} := \begin{pmatrix} 1 & x_1 & \ldots & x_1^p \\ 1 & x_2 & \ldots & x_2^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^p \end{pmatrix}$$

   Wir behandeln hier jede Potenz von x als eine neue Variable und wie können die Parameter wie in 4.2 berechnen aus

   $$\hat{\mathbf{b}} = (\hat{\mathbf{X}}^\mathbf{T}\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^\mathbf{T}\mathbf{y} \tag{9}$$

2. Wir können y nicht linear in $\mathbf{x}$ darstellen, aber die Parameter $\mathbf{b}$ gehen noch immer linear in das Problem ein. Zum Beispiel haben wir

$$y = b_1 f_1(x_1) + b_2 f_2(x_2) + \ldots + b_p f_p(x_p)$$

Hier können wir neue Variablen $\tilde{\mathbf{x}}$ einführen um die Gleichung in die selbe Form wie (7) zu bringen. Wir nehmen dann

$$\tilde{x_i} = f_i(x_i) \qquad\qquad y = \mathbf{b^T}\tilde{\mathbf{x}}$$

und bekommen wie in 4.2 mit einer minimierung des quadratisches Fehlers

$$\hat{\mathbf{b}} = (\tilde{\mathbf{X^T}}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X^T}}\mathbf{y}$$

3. Wir haben y gegeben als eine injektive Funktion, g von $\mathbf{x^T b}$

$$y = g(\mathbf{x^T b})$$

Dann ist $g^{-1}(y)$ linear in den Parametern,

$$g^{-1}(y) = \mathbf{x^T b}$$

und wir bestimmen die Parameter durch

$$\hat{\mathbf{b}} = (\mathbf{X^T X})^{-1}\mathbf{X^T} g^{-1}(\mathbf{y})$$

Wir notieren hier dass wir nicht mehr den qudratratischen Fehler $\sum_{i=1}^{n}(g(\mathbf{x^T b}) - y_i)^2$ minimieren, sondern $\sum_{i=1}^{n}(\mathbf{x^T b} - g^{-1}(y_i))^2$.

6 Approximation für die Ableitung einer Funktion

Wenn man logistisches Wachstum annimmt, besagt, nach (4), die Verhulst-Gleichung:

$$u'(t) = ru(t)\left(1 - \frac{u(t)}{K}\right)$$

Um die Fehlerquadrat-Methode für diese Gleichung anzuwenden, brauchen wir zunächst Daten für $u'(t)$. Im Folgenden werden hier zwei Verfahren diskutiert.

6.1 Lineare Regression

Die Idee ist, mit Hilfe von linearer Regression, $u'(t_i)$ von drei Punkten zu bestimmen, siehe Abbildung (4). Die Annahme ist, nach *Quarteroni u.a* [5], die gleiche wie in (7). Man muss eine lineare Regression für jeden Punkt durchführen. Die Parameter werden bestimmt wie in (?):

$$\hat{\mathbf{b_i}} = (\mathbf{X_i^T X_i})^{-1} \mathbf{X_i^T u_i}$$

wobei

$$\hat{\mathbf{b_i}} = \begin{pmatrix} \hat{b_{i,1}} \\ \hat{b_{i,2}} \end{pmatrix}, \ \mathbf{u_i} = \begin{pmatrix} u(t_{i-1}) \\ u(t_i) \\ u(t_{i+1}) \end{pmatrix} \text{ und } \mathbf{X_i} = \begin{pmatrix} 1 & t_{i-1} \\ 1 & t_i \\ 1 & t_{i+1} \end{pmatrix}.$$

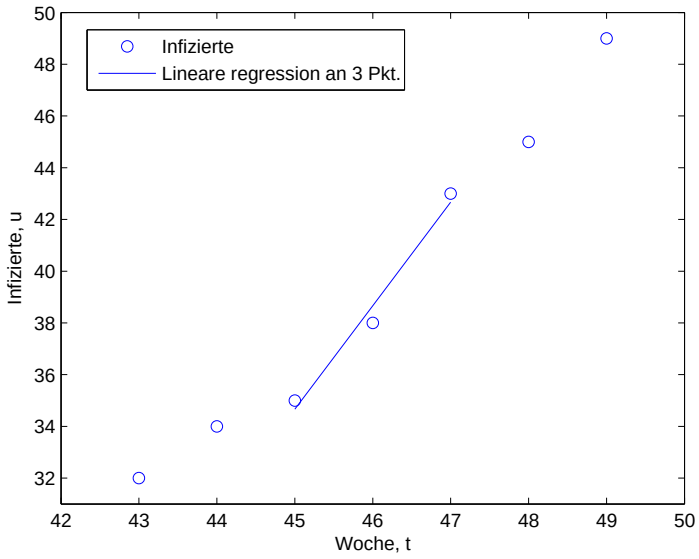

Abbildung 4: Lineare Regression angewandt an 3 Punkten

Zu jedem u_i haben wir jetzt eine Approximation für die Ableitung:

$$u'(t_i) = \hat{b_{i,2}}$$

Für $\mathbf{u_1}$ und $\mathbf{u_n}$ wird $\hat{\mathbf{b_1}}$ bzw. $\hat{\mathbf{b_n}}$ mit Hilfe von zwei Punkten begerechnet.

6.2 Taylor-Entwicklung

Eine andere Methode u_i' zu bestimmen, ist, nach *Emmrich* [4], die mit Hilfe der Taylor-Entwicklung von den zwei folgenden Punkten u_{i+1} und u_{i+2}. Damit erhalten wir eine $\mathbf{O(\Delta t^2)}$-Approximation.

$$u_i = u(t_i)$$

$$u_{i+1} = u(t_i + \Delta t) = u_i + u_i'\Delta t + u_i''\frac{(\Delta t)^2}{2} + O(\Delta t^3)$$

$$u_{i+2} = u(t_i + 2\Delta t) = u_i + u_i'(2\Delta t) + u_i''\frac{(2\Delta t)^2}{2} + O(\Delta t^3)$$

wobei $\Delta t = t_{i+1} - t_i$. Wir brauchen jetzt

$$u_{i+1} - u_i = u_i'\Delta t + u_i''\frac{(\Delta t)^2}{2} + O(\Delta t^3)$$

$$u_{i+2} - u_i = 2u_i'\Delta t + 2u_i''(\Delta t)^2 + O(\Delta t^3) \quad .$$

Daraus folgt

$$u_{i+2} - u_i - 4(u_{i+1} - u_i) = -2u_i'\Delta_t + O(\Delta t^3) \quad .$$

Mit Vereinfachung

$$u_i' = \frac{1}{\Delta t}\left(-\frac{3}{2}u_i + 2u_{i+1} - \frac{1}{2}u_{i+2}\right) + O(\Delta t^2) \quad .$$

Es ist angenommen das unsere Daten genauer wird als t wächst. Darum benutzen wir diese vorwärz genommene Differenz.

7 Bestimmung der Parameter in der Verhulst-Gleichung

Wir haben jetzt eine Approximation für $u'(t_i)$ und Daten für $u(t_i)$. Jetzt wenden wir die Fehlerquadrat-Methode mit diesen Daten für die Verhulst-Gleichung (4) an:

$$u'(t) = ru(t) - \frac{r}{K}u(t)^2$$

Es gibt zwei Parameter zu bestimmen, R und K. Weil es jetzt kein lineares Verhältnis gibt, werden diese wie in () bestimmt. Jetzt gilt:

$$\hat{\mathbf{b}} = \begin{pmatrix} R \\ -\frac{R}{K} \end{pmatrix}, \mathbf{y} = \begin{pmatrix} u'(t_1) \\ u'(t_2) \\ \vdots \\ u'(t_n) \end{pmatrix} \text{ und } \hat{\mathbf{X}} = \begin{pmatrix} u(t_1) & u^2(t_1) \\ u(t_2) & u^2(t_2) \\ \vdots & \vdots \\ u(t_n) & u^2(t_n) \end{pmatrix}.$$

8 Behandlung der Giardia-Epidemie

In dieser Aufgabe werden wir zwei verschiedene Modellen benutzen, um den Verlauf der Giardia-Epedemie in Bergen zu beschreiben. Die zwei Modelle sind ein logistisches Wachstums-Modell und ein SIR-Modell. Es kann gezeigt werden, dass das SIR-Modell mit Taylor-Approximation erster Ordnung, in einem logistischen Modell umgeschrieben werden kann.

8.1 Logistisches Wachstum

Wie man die Ableitung einer Funktion u annähern kann, ist beschrieben in Abschnitt 6. Wir bezeichnen mit u(t) die kummulative Anzahl Giardia-Infizierter, die im Herbst 2004 zum Arzt gingen. In Abbildung 5 ist u'(t) gegeben als eine Funktion von u(t). Wir sehen in dieser Abbildung, dass der Unterschied zwischen den zwei Verfahrenen, die wir benutzen um die Ableitung anzunähern, in einzelnen Punkten manchmal sehr groß ist, aber die Polynome trotzdem ähnlich sind. Dass heißt, wir bekommen ähnliche Werte für die Parameter im logistischen Wachstum mit den zwei verschiedenen Verfahren. Das wäre im Allgemeine nicht der Fall. Die zwei Verfahren könnten zu völlig verschiedenen Parametern leiten. Die Ableitung, die wir mit einer 3-Punkt Annäherung zu einer Geraden gefunden haben, folgt unsere Polynom besser als eine Vorwärts-Taylor-Entwicklung; deshalb benutzen wir diese Daten weiter.

8.2 Übereinstimmung mit experimentellen Daten

Wir sehen in Abbildung 6, dass die experimentellen Daten gut zu dem Modell des logistischen Wachstums passen. Die Zeitskala ist wenig verschoben, aber die Form der Kurve passt sehr gut. Die Unstimmigkeit

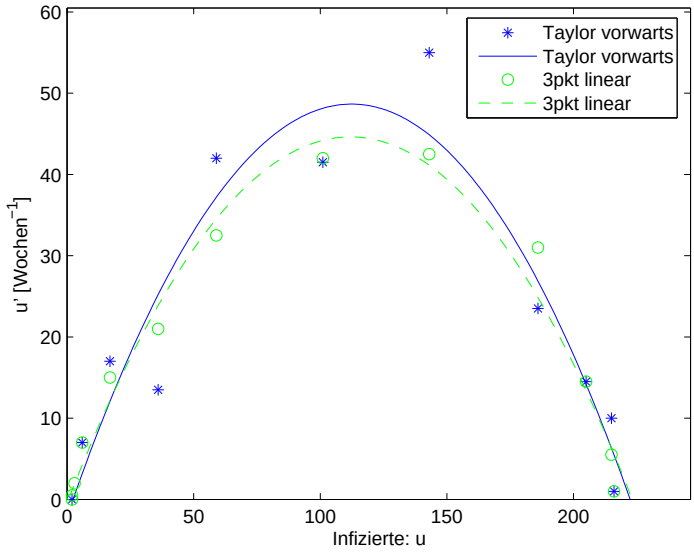

Abbildung 5: u'(t) als eine Funktion von u(t). Die Ableitung ist mit zwei verschiedenen Methoden berechnet.

in der Zeit können wir mit einer Unsicherheit in der bestimmung des Epedemieausbruchs erklären.

8.3 Direkte Anwendung des SIR-Modells

Mit hilfe der MATLAB-Funktion *lsqcurvefit* [1] haben wir die in Gleichung 2 auftretenen Parameter β, $S(0)$ und ρ direkt gefunden, die Ergebnis ist in Abbildung 7 gezeigt. Dann haben wir die Funktion *ode45* benutzt um die Kurve in Abbildung 8 zu Konstruieren, *ode45* benutzt ein Runge-Kutta-Verfahren um eine ordinäre Differentialgleichung mit Anfangswerten zu lösen.

8.4 Stabilität des Modells

Soll ein Modell stabil sein, dürfen kleine Änderungen in den Input-Daten nur zu kleinen Änderungen in den Parameter führen. Um diese Eigenschaft unseres Modells zu untersuchen, haben wir für jede Woche ein Rauschen eingeführt. Das heißt, in Woche i waren es u_i registrierte Infizierte in Bergen. Wir führen jetzt ein normalverteiltes Rauschen e_i mit dem Erwartungswert 0 und der Standardabweichung 1 ein.

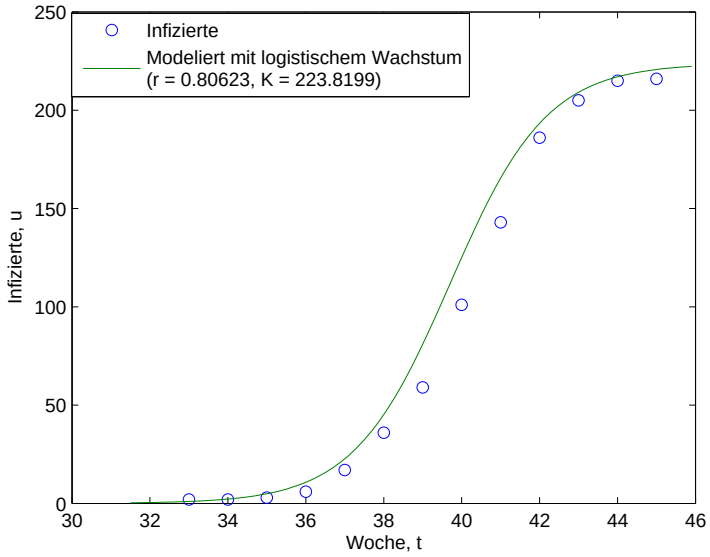

Abbildung 6: Logistisches Wachstum zusammen mit experimentellen Daten. Die Parameter r und K sind durch eine multilineare Anpassung der Verhulst-Gleichung gefunden

$$\tilde{u_i} = u_i + e_i$$

und wir berechnen die neuen Parameter $\tilde{r}$ und $\tilde{K}$, begründet auf den neuen Daten $\tilde{\mathbf{u}}$. Die relative Schwankung in den berechneten Parametern ist in Abbildung 9 für 50 verschiedene Sets von Rauschtermen gezeigt. Die relative Abweichung f ist durch die folgende Gleichung berechnet:

$$f_r = \frac{\tilde{r} - r}{r} \text{ und } f_K = \frac{\tilde{K} - K}{K}$$

Wir sehen hier, dass die Schwankungen in den Parametern unter 1% für K und 3 % für r ist, und wir können sagen, dass unser Modell stabil ist, was im allgemeinen nicht der Fall für logistisches Wachstum ist. Hier haben wir Daten für den ganzen sigmoiden Teil der Kurve. Hätten wir nur Daten für einen Teil der Kurve gehabt, wäre die Situation anders. Es würde dann oft viele Paare von Parametern geben, die die Daten gut beschreiben könnten. So ist es normalerweise, wenn man eine sich entwickelnde Epidemie beschreiben möchte. Dann hat man

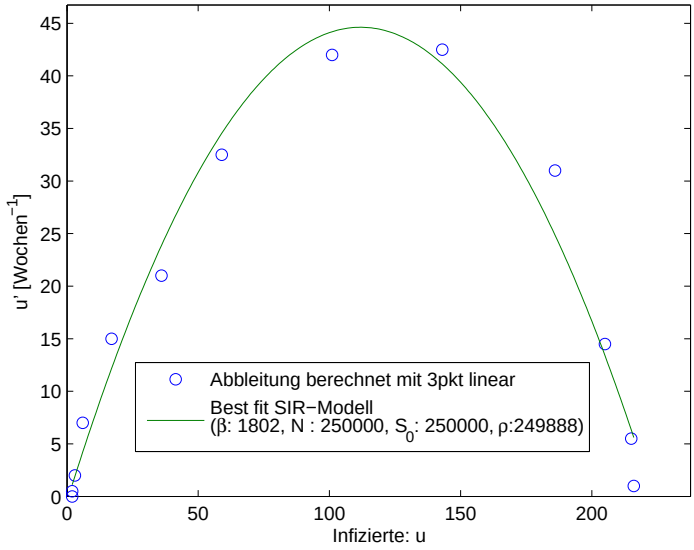

Abbildung 7: Die Abbleitung wird approximiert mit dem SIR-Modell

nur Daten zu Beginn der Epidemie und damit ist die Bestimmung der Parameter viel unsicherer. Um diese Problem zu beschriben haben wir das Verfahren wiederholt mit nur die Daten für die 7 erste Wochen. Dann bekommen wir die Ergebnis beschriben in

8.5 Direkte Bestimmung der Parameter von der Kurve

Wir nehmen an, dass unsere Epidemie einen Logistisches Wachstum folgt. Die Differentialgleichung können wir sofort lösen und erhälten

$$u(t) = \frac{K}{1 + \left(\frac{K}{u_0} - 1\right) e^{-rt}}, u(0) = u_0.$$

Wir können sagen, dass unsere beste Schätzung für die Parameter r und K die Kombination ist, die die quadratische Abweichung zwischen unseren Messdaten und dem Modell so gering wie möglich macht.

$$(r, K) = \arg \left\{ \min_{r,K} \sum_i \left(\frac{K}{1 + \left(\frac{K}{u_0} - 1\right) e^{-rt_i}} - u_i \right)^2 \right\}$$

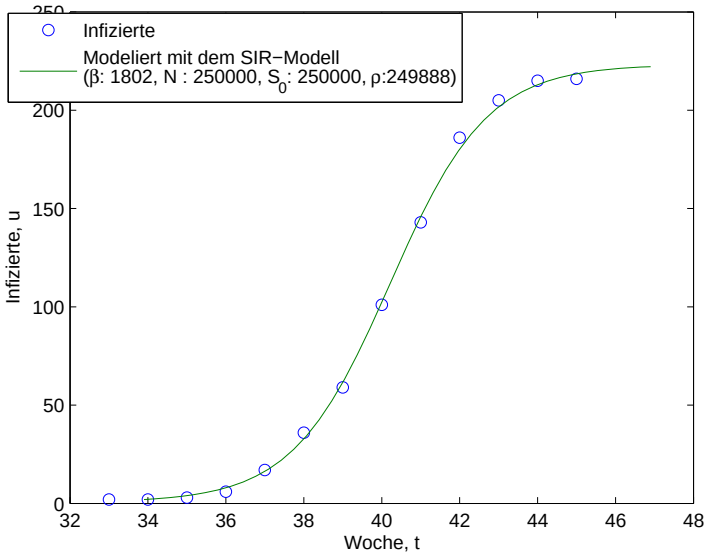

Abbildung 8: Anzahl der Infizierte modelliert mit dem SIR-Modell

Es gibt verschiedene Verfahren, die man benutzen kann, um dieses Minimierungsproblem zu lösen. Zum Beispiel hat Matlab eine Funktion fminunc(), die ein Lokales Minimum einer Funktion in mehreren Variabeln findet. Diese Funktion ist ein Teil des „Optimization Toolbox“ und sie benutzt ein „Steepest -decline“ Verfahren. Wir haben diese Funktion benutzen um die Parametern r und K zu finden. Die Lösung ist in Abbildung 11 gezeigt.

Hier sehen wir, dass die Abweichung zwischen Modell und Messdaten klein ist, trotzdem beschreibt diese Kurve unsere Daten nicht ausreichend. Zum Beispiel haben unsere Daten ein leicht höheres Maximum als 220, aber die modelierten Daten steigen bis auf 236 an, liegen also um etwas 7 % höher.

8.6 Anfangszeit der Epidemie

Bis jetzt haben wir nur die zwei Parameter r und K geendert. Ein Problem ist dann, dass wir nicht genau wissen, wann die Epidemie angefangen ist und zu welchem Zeitpunkt welche Anfangsbedingungen zu stellen ist. Am Anfang der Epidemie gibt es nicht viele infizierte Personen, und die Anzahl der Infizerten kann nicht als kontinuierlich angenommen werden. Deshalb ist es unwahrscheinlich, dass die Epidemie ein SIR-Modell folgt. Erst wenn es

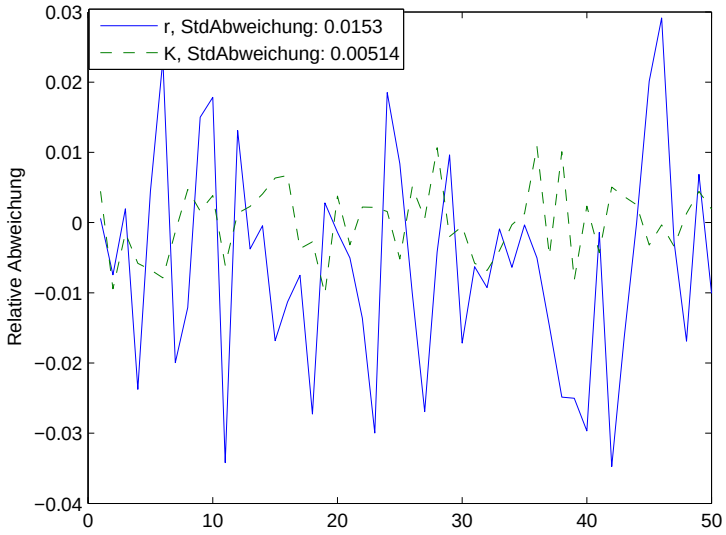

Abbildung 9: Relative Schwankung in den Parametern r und K wenn man einen Rauschterm mit den input-Daten u anwendet

eine größere Gruppe von Infizierte gibt können wir annehmen, dass unsere Problem beschrieben werden kann durch ein SIR-Modell. Deshalb ist es sinnvoll, auch die Anfangszeit zu variieren um die Parameter bestmöglich zu bestimmen. Das heißt, wir wollen das folgende Minimalisierungsproblem zu lösen:

$$(r, K, \Delta t) = \arg \left\{ \min_{r,K,\Delta t} \sum_i \left(\frac{K}{1 + \left(\frac{K}{u_0} - 1\right) e^{-r(t_i + \Delta t)}} - u_i \right)^2 \right\}$$

Dieses Problem können wir direkt in MATLAB lösen durch die Verfahren beschrieben im Kapitel 8.5. Die Ergebnis ist in Abbildung 12 gezeigt.

Oder wir können das in Abbildung 6 gezeigte Ergebnis so verschieben, dass wir eine optimale Lösung bekommen. So haben wir es gemacht und das Ergebnis ist in Abbildung 13 gezeigt.

9 Schlussvolgerung

Wir haben gesehen, dass man mit unseren Daten von Bergen, die Epidemie bestimmend Parameter ganz gut bestimmen kann. Wir haben Daten über

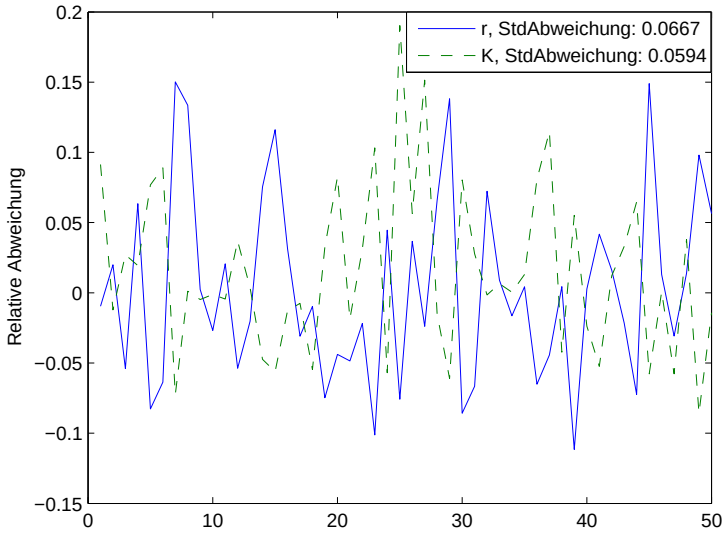

Abbildung 10: Relative Schwankung in den Parametern r und K wenn man einen Rauschterm mit den input-Daten u anwendet. Nur die 7 ersten Wochen der Epedemie sind

die gesamte Zeitperiode des Krankheitsverlaufs. Weiterhin haben wir gezeigt, dass man auch Schätzung für die Anfangszeit der Beobachtung bestimmen kann.

Problematisch ist Folgendes:

- Man hat nur Daten für einen beschränkten Zeitraum T. Je kleiner T, desto größer wird die Unsicherheit des Parameters.
- Die Anfangszeit der Beobachtung ist unbekannt. In Kombination mit beschränkten Daten wäre das ein Problem. Mit guten Daten haben wir gesehen, dass man eine gute Schätzung diesen Zeitraum bestimmen kann.
- Die Ausbreitung der Krankheit ist komplex. Wenn es mehrere Wege der Ansteckung gibt, und sogar unbekannte, muss man vielleicht ein anderes Modell benutzen.

Wir haben auch gezeigt, dass man mit einer Annahme der Unsicherheit der Daten die Änderung der Parameter untersuchen kann. Dies zeigt wie stabil unser Modell ist. Das heisst, wir wissen wie gut unser Modell für unsere Daten ist. Normalerweise möchte man so früh wie möglich wissen, wie sich der

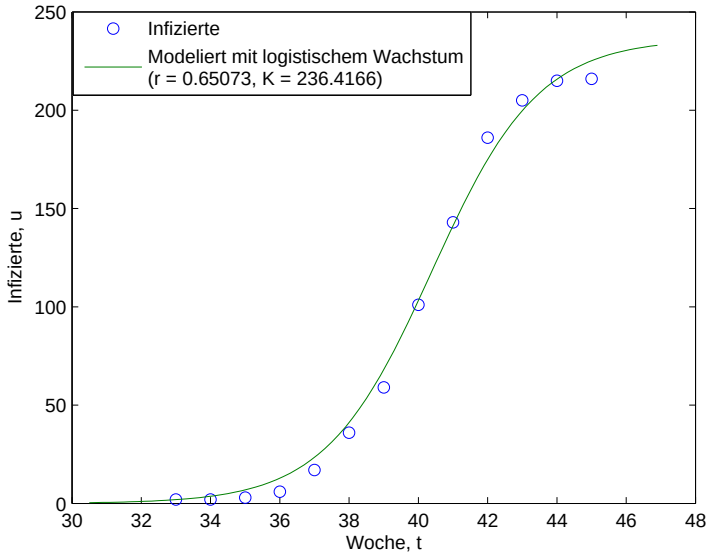

Abbildung 11: Die Parameter des logistischen Wachstums, gefunden mit einer Minimierungs-Routine in MATLAB

Krankheitsverlauf verhält. Dann hat man oft das Probleme wie oben bereits beschrieben, und muss dann früh feststellen, welches Modell brauchbar ist.

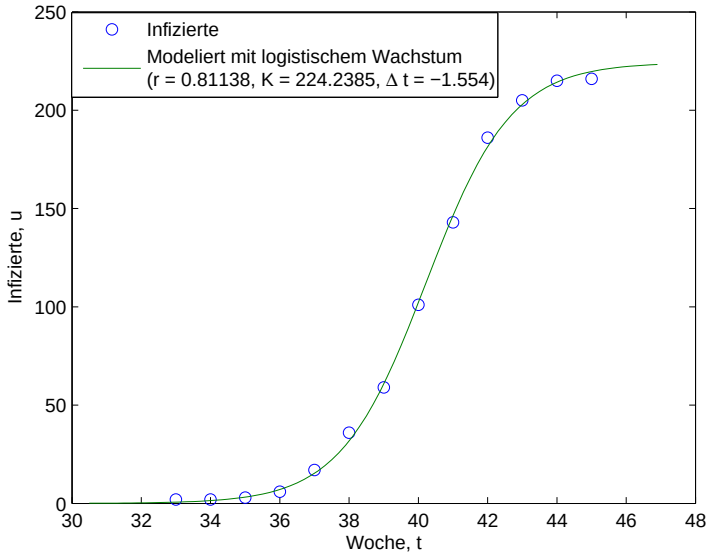

Abbildung 12: Die Parameter des logistischen Wachstums, gefunden mit optimalisierung in MATLAB

Literatur

[1] *MATLAB 7.0.0, release 14.* The Mathworks Inc, 2004.

[2] Skorping Arne. *Kronikk Bergens Tidende: Giardia, hva er det?* 16. November 2004, http://www.bt.no/meninger/kronikk/article309343.

[3] Bohl E. *Mathematik in der Biologie.* Springer-Verlag, Berlin - Heidelberg, 2001.

[4] Emmrich E. *Vorlesungen in Mathematische Modellierung mit Differentialgleichungen.* Technische Universität, Berlin, Wintersemester 04/05.

[5] Quertoni A. Sacco R. Saleri F. *Numerische Mathematik 2.* Springer-Verlag, Berlin - Heidelberg, 2002.

[6] Bergen Kommune Offentlig Informasjon. *Giardia-smitten, grafisk fremstilling.* http://www.bergen.kommune.no/info/.

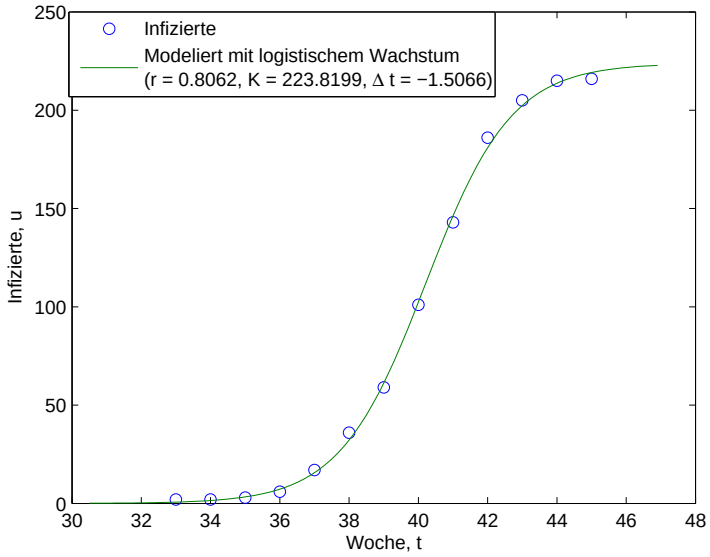

Abbildung 13: Die Parameter des logistischen Wachstums, die parameter r und K gefunden durch die Abbleitung und Δt gefunden durch optimalisierung in MATLAB

```
% modell.m
% Terje Tofteberg, 01.02.05
%
% Diese Program berechnet die Parameter r und K in der Verhulst-Gleichung,
% die Best mit einen Daten-Vektor kummulativ passen.
%

clear all;
close all;
load data; % ein vektor tilfeller mit neue registrierte Zufälle jede Woche
woche = 33:45; % wochen  2004 mit registrierten infizierten Pasienten

kummulativ = zeros(size(tilfeller));
kummulativ(1) = tilfeller(1);
for i = 2:length(tilfeller)
    kummulativ(i) = kummulativ(i-1) + tilfeller(i);
end
plot(woche,kummulativ,'o');
%title('Registrierte Giardia-infizierten Pasienten in Bergen Kommune,
% Norwegen')
```

```
xlabel('Woche t')
ylabel('Infizierte, u')

% Zwei verschiedene möglichkeiten um die Ableitung zu bestimmen
ableitungTaylor = ableitTaylor(kummulativ);
ableitung3Pkt   = ableit3Pkt(kummulativ);

% Abbau von der Vandermonde-Matrix
gradPol = 2
X = ones(length(kummulativ),gradPol+1);
for i = 1: gradPol
    for j = 1:gradPol+1-i
        X(:,j) = kummulativ'.*X(:,j);
    end
end

% Linear Regression, berechnet die Polynomkoeffizienten b2, b1 und b0 in
% dem Polynom u' = b2*u^2 + b1*u + b0
polynomTaylor = inv(X'*X)*X'*ableitungTaylor';
polynom3Pkt   = inv(X'*X)*X'*ableitung3Pkt';

u0 = 1;
[r K] = parameterRausch(kummulativ,0);
t = -1.5:0.2:13;
modellert = zeros(size(t));

for i = 1:length(t)
    modellert(i) = logistikk(r,K,u0,t(i));
end

figure
plot(woche,kummulativ,'o',t+33, modellert);

%title('Registrerte tilfeller av Giardia-smitte i Bergen Kommune')
xlabel('Woche, t')
ylabel('Infizierte, u')
legend('Infizierte',['Modeliert mit logistischem Wachstum (r = ' ..
    num2str(r) ', K = ' num2str(K) ')'])

figure
kum = 1:K;
```

```
plot(kummulativ, ableitungTaylor,'b*',kum,polyval(polynomTaylor,kum)..
    ,'b-', kummulativ, ableitung3Pkt,'go',kum,polyval(polynom3Pkt,kum),'g--')

xlabel('Infizierte: u')
ylabel('u'' [Wochen^{-1}]')
legend('Taylor vorwarts','Taylor vorwarts','3pkt linear','3pkt linear')
axis([0 K*1.1 0 max(ableitungTaylor)*1.1]);

m = 50;
for i = 1:m
    [rRausch(i) KRausch(i)] = parameterRausch(kummulativ,1);
end
figure
x = 1:m;

plot(x,(rRausch-r)/r,'-',x,(KRausch-K)/K,'--')
%title('Relative Abweichung in die Parametern r und K, bestimmten von
% rauschaddierten Dateien.')
ylabel('Relative Abweichung')
legend(['r, StdAbweichung: ' num2str(std((rRausch-r)/r),'%0.3g')],['K, ..
    StdAbweichung: ' num2str(std((KRausch-K)/K),'%0.3g')])
```

```
%ableitTaylor.m
%Terje Tofteberg, 22.01.05
%
%Diese Program berechnet die Ableitung du, einer diskreten Funktion u.
%Die Annährung zu du in jedem Punkt xi ist berechnet von einer
%quadratischer Taylor-Entwicklung der Funktion u.
%%

function du = ableitTaylor(u)
n = length(u);
du = zeros(size(u));
for i = 1:n-2
    du(i) = -1.5* u(i) + 2*u(i+1) - 0.5*u(i+2);
end
du(n-1) = -u(n-2) + u(n-1);
du(n) = -u(n-1) + u(n);
return
```

```
%ableit3PKt.m
%Terje Tofteberg, 22.01.05
%
%Diese Program berechnet die Ableitung du, einer diskreten Funktion u.
%Die Annährung zu du in jedem Punkt xi ist eine lineare Regression berechnet von
%(xi-1,yi-1), (xi,yi),(xi+1,yi+1).
%%
function du = ableit3Pkt(u);
du = zeros(size(u));
du(1) = u(2) -u(1);
for i = 2:length(u) -1
    X = [1 i-1;
         1 i ;
         1 i+1];
    y = [u(i-1) u(i) u(i+1)]';
    b = inv(X'*X)*X'*y;
    du(i) = b(2);
end
du(length(u)) = u(length(u)) - u(length(u)-1);
return
```

```
% Terje Tofteberg 22.01.05
% parameterRausch
%
% Berechnet die Parameter r und K von einem Logistischen Wachstum von
% einem funktion u
% u' = ru(1-u/K)
% aber mit einem introdusierten Rauschterm.

function [r,K] = parameterRausch(u,std)
%Wir addieren ein Rauschterm in jede Datenpunkt. Der Rausch ist
%Normalverteilt mit Erwartungswert 0 und Standardabweichung 1.
rausch = std*randn(size(u));
u = u + rausch;

ableitung3Pkt   = ableit3Pkt(u);
polynom3Pkt   = polyfit(u, ableitung3Pkt,2);

r = polynom3Pkt(2);
K = -r/polynom3Pkt(1);

return
```

```
function verdi = logistikk(r,K,u0,t)
verdi = K /(1+(K/u0-1)*exp(-r*t));

return
```