Logistische Regression

Großübung Biometrie II

bei Nicole Heussen.

am Lehrstuhl für Biometrie der RWTH Aachen

von

Claus Richterich, Stefan Schiffer und Thomas Deselaers

Motivation

- funktionalen Zusammenhang erkennen
- Darstellung der Ursache Wirkung Beziehung
- Schätzen der Parameter einer bekannten funktionalen Beziehung
- Interpolation fehlender Werte
- Prognose zukünftiger Werte

“Unsere” Daten

umweltepidemiologische Studie

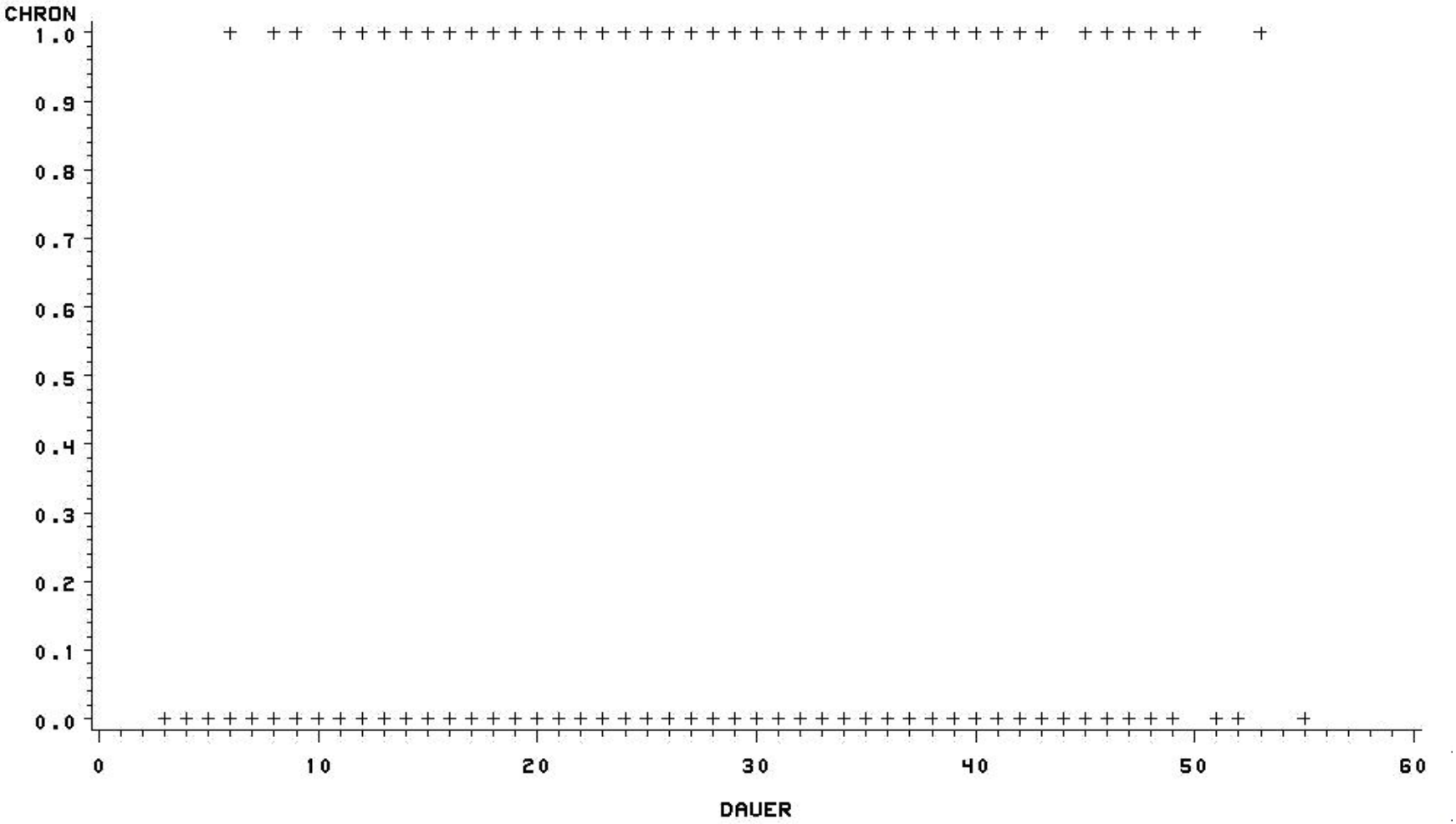
- chronische Erkrankung der Bronchien (dichotom [0/1])
- mehrere potentielle Einflussfaktoren (dichotom und stetig)

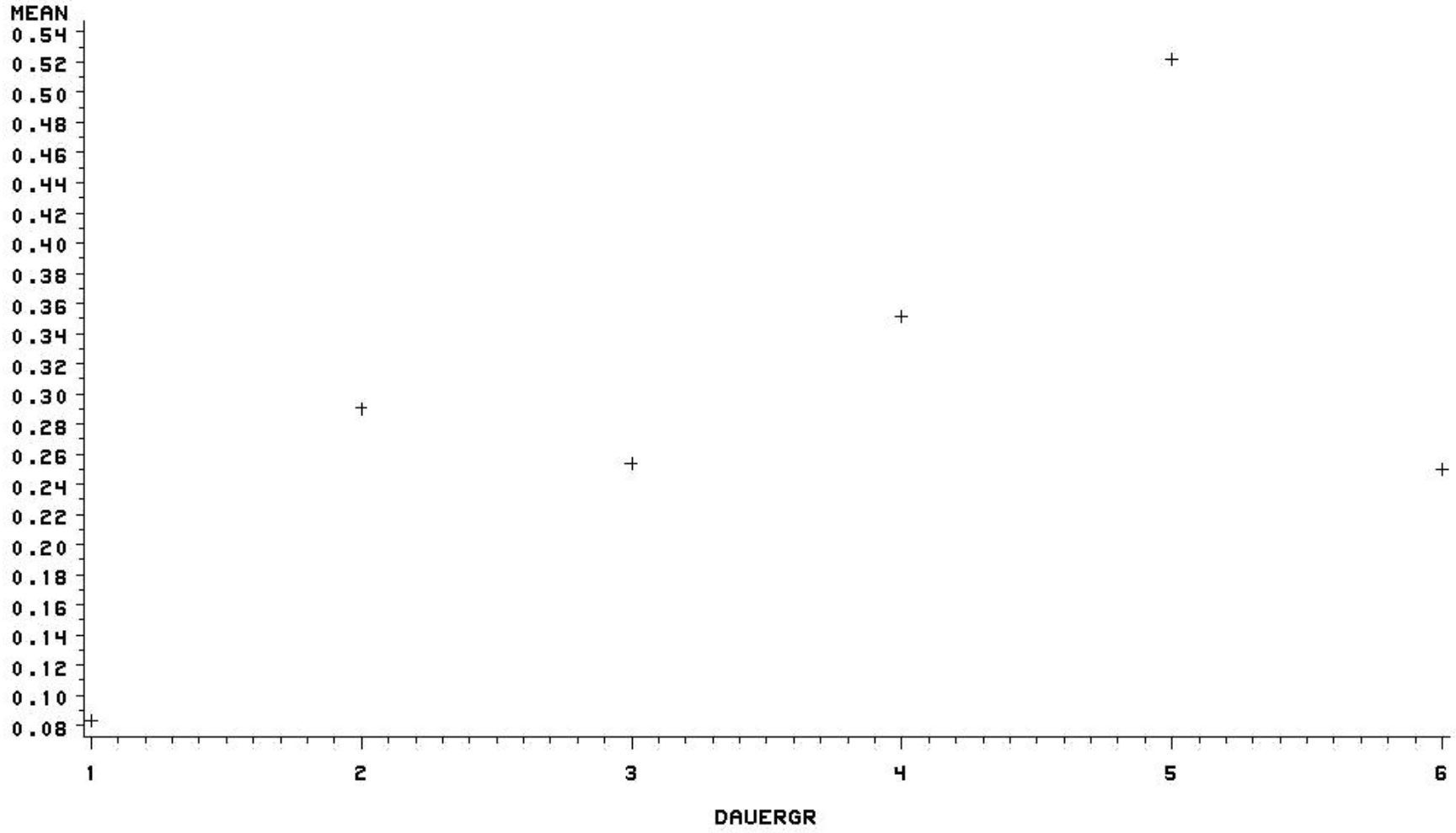
Warum logistische Regression?

Dichotome Merkmale

- nicht mit linearer Regression erfassbar

- keine Modellierung der abhängigen Variablen sondern einer Fkt. der Wahrscheinlichkeit, dass die Krankheit unter geg. Risikofaktoren auftritt.


CHRON
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
DAUER


MEAN
0.54
0.52
0.50
0.48
0.46
0.44
0.42
0.40
0.38
0.36
0.34
0.32
0.30
0.28
0.26
0.24
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.08
1
2
3
4
5
6
DAUERGR

Odds-Ratio

Statt der Wahrscheinlichkeit P wird das **Odds-Ratio** benutzt

$$OR(P) = \frac{P}{1-P}$$

Das Odds-Ratio bezeichnet den Quotienten der Wahrscheinlichkeit und ihrer Gegenwahrscheinlichkeit.

Logit-Transformation

Wahrscheinlichkeiten sind aus dem Intervall $[0, 1]$, Odds-Ratios aus dem Intervall $[0, \infty]$.

Definition

$$logit(P) = log(OR(P)) = log(\frac{P}{1-P}) = log(P) - log(1-P)$$

heißt logit-Transformation und bildet dies auf den gesammten Reelen Zahlenstrahl $[-\infty, \infty]$ ab.

Logit-Transformation 2

Abbildung auf den gesammten Zahlenstrahl

P	0	0.01	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.99	1
$OR(P)$	0	0.01	0.052	0.11	0.33	1	3	9	19	99	∞
$logit(P)$	$-\infty$	-4.60	-2.94	-2.20	-1.1	0.00	1.10	2.20	2.94	4.60	∞

Expit-Transformation

Ist die Umkehrfunktion zur Logit-Transformation

$$P = \frac{exp(L)}{1 + exp(L)}$$

wobei

$$L = logit(P)$$

so dass man vom $logit(P)$ direkt wieder auf die Wahrscheinlichkeit zurück-rechnen kann.

Logistische Regression

Das Modell der logistischen Regression benutzt $logit(P)$ als Zielvariable einer linearen Regression.

Wdh.: Lineare Regression

$$Y = \alpha_0 + \alpha_1 \cdot x$$

Logistische Funktion

Also:

$$logit(P) = \alpha_0 + \alpha_1 \cdot x$$

Auflösen nach P ergibt:

$$P(Y \mid X_1) = \frac{exp(\alpha_0 + \alpha_1 X_1)}{1 + exp(\alpha_0 + \alpha_1 X_1)}$$

Modellbildung

Zur Modellbildung müssen alle denkbaren Vorraussetzungen, Annahmen und Möglichkeiten im Vorraus geklärt werden.

Um einen Zusammenhang zwischen Ziel- und Einflussfaktoren zu bestimmen bildet man ein geeignetes Modell.

Man könnte alle gemessenen Einflüsse in das Modell mit aufnehmen, dies ist jedoch oft nicht sinnvoll, da Variablen irrelevant sein können.

Modellbildungsverfahren

- inhaltlich begründet
- vorwärtige Variablenselektion
- rückwärtige Variablenselektion
- schrittweise Variablenselektion

Qualität eines Modells

Die Qualität eines Modells testet man mit der **Likelihood-Quotienten-Statistik**.

Dabei ist die **Nullhypothese**:

Der Faktor X_i hat keinen Einfluss auf die Zielvariable.

$$P(K \mid X_i = x_i) = P(K \mid X_i \neq x_i)$$

d.h. das Odds-Ratio von X_i ist $OR(X_i) = 1$

dann hat der Regressionskoeffizient α_i den Wert 0.

Maximum-Likelihood-Quotienten-Test

$$D = 2[l(\hat{P}) - l(\tilde{P})] = 2\sum P_{ij} log \frac{P_{ij}}{n_i \tilde{P}_{ij}}$$

D ist dann approximativ χ^2 -verteilt mit $(r-1)\cdot(s-1)$ Freiheitsgraden.

Confounding

Confounder = Faktor, der auf die Zielvariable wirkt aber nicht Ziel der Untersuchung ist.

Confounding = Berücksichtigung von Confoundern.

Störende Einflüsse analysieren und kontrollieren.

Man nimmt die gefundenen Confounder mit in das logistische Modell auf, um ihre Störwirkung angemessen berücksichtigen zu können.

Statistik oder die Alchemie der Neuzeit

Fragestellungen

- Deskription der Daten
- Einfluss der Umweltbelastung
- Einfluss des Rauchens
- Umweltbelastung als Risikofaktor
- Welche Variablen fehlen
- Bestimmung des besten Modells

Die Variablen

- CHRON
- STAUB
- RAUCH
- DAUER
- UMWELT

Einen Ausschnitt aus den Daten

```
0       0.2     1       5       1
0       0.25    1       8       1
0       0.25    1       4       1
0       0.25    1       8       1
1       0.25    1       8       1
...
```

Erster Überblick

Von den 623 Personen sind : 448 Raucher,
170 am Wohnort einer starken Umweltbelastung ausgesetzt,
136 haben eine chronische Erkrankung der Bronchien

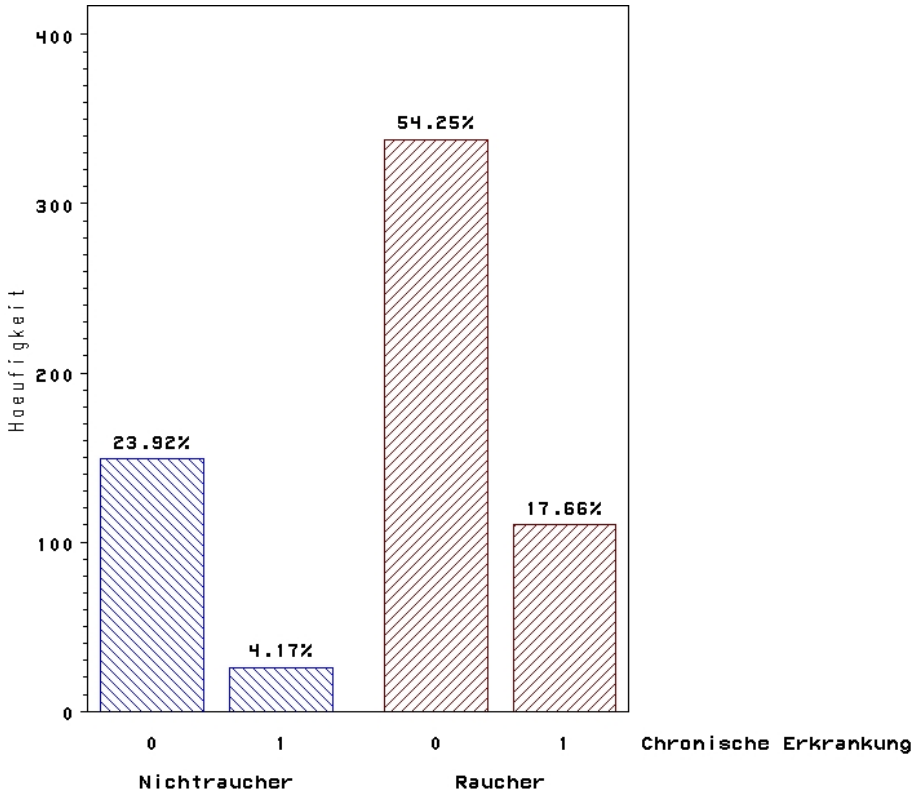
Überblick

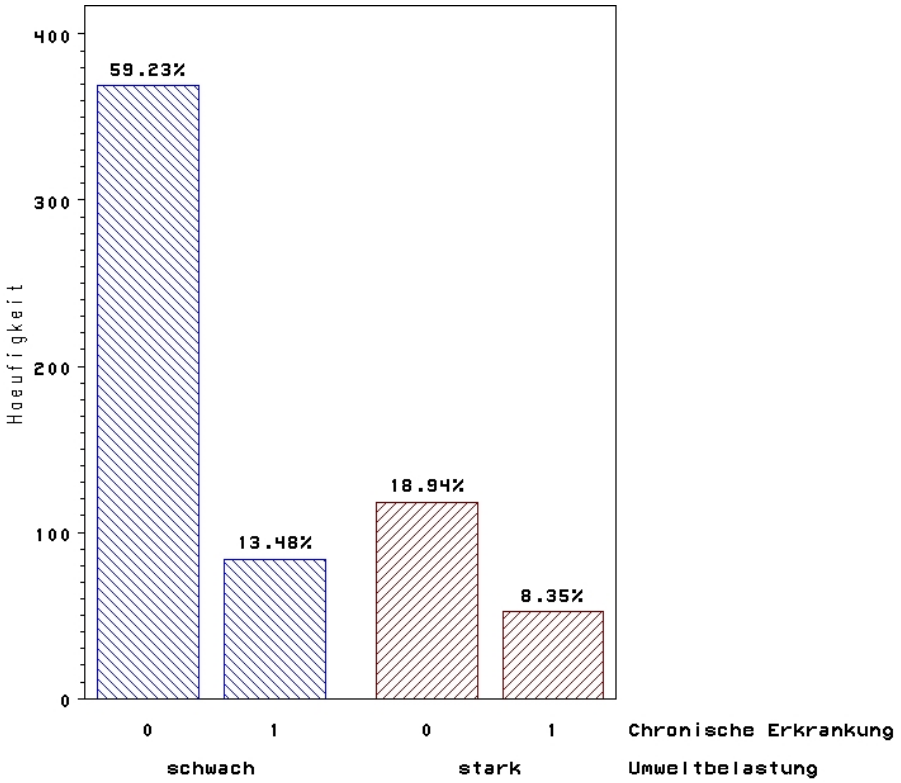
DAUERGR	Anzahl	Prozent
1	50	8.0
2	171	27.4
3	181	29.1
4	148	23.8
5	67	10.8
6	6	1.0

STAUB	Anzahl	Prozent
1	412	66.1
2	14	2.2
3	125	20.1
4	62	10.0
5	10	1.6

Überblick 2

Kenngröße	DAUER	STAUB
Mittelwert	24.91	2.35
Std.Abweichung	11.45	2.58
Varianz	131.10	6.63
Maximalwert	55	15.04
75%-Quantil	33	4.95
50%-Quantil	25	0.71
25%-Quantil	15	0.42
Minimalwert	3	0.01


Haeufigkeit
400
300
200
100
0
23.92%
4.17%
54.25%
17.66%
0
1
0
1
Chronische Erkrankung
Nichtraucher
Raucher


Haeufigkeit
400
300
200
100
0
59.23%
13.48%
18.94%
8.35%
0
1
0
1
Chronische Erkrankung
schwach
stark
Umweltbelastung

Einfluss der Umweltbelastung

```
proc freq data=logreg;
  tables chron*umwelt /chisq;
run;
```

STATISTICS FOR TABLE OF CHRON BY UMWELT

Statistic	DF	Value	Prob
Chi-Square	1	10.510	0.001

...

Logistische Regression

```
proc logistic descending data=logreg;
  model chron = umwelt;
run;
```

Ergibt folgende Regressionsgleichung:

$$logit(P) = -2.1405 + 0.6605 \cdot UMWELT$$

als Odds-Ratio ergibt sich $OR = 1.936$.

Einfluss der anderen Variablen

- STAUB hat keinen statistisch nachweisbaren Einfluss
- DAUER hat Einfluss
- RAUCH hat Einfluss

Einfluss des Rauchens

```
proc logistic descending data=logreg;
  model chron = rauch;
run;
```

Ergibt folgende Regressionsgleichung:

$$logit(P) = -1.74 + 0.6233 \cdot RAUCH$$

als Odds-Ratio ergibt sich $OR = 1.865$.

Umweltbelastung als Risikofaktor

Logistische Regression mit DAUER als Einflussfaktor ergibt

$$logit(\hat{P}) = -1.5824 + dauer \cdot 0.0275$$

als Regressionsgleichung.

Auflösen nach Dauer:

$$dauer = \frac{logit(\hat{P}) + 1.5824}{0.0275} = \frac{logit(0.5) + 1.5824}{0.0275} = 57.54$$

Ab einer Wohndauer von 57.54 Jahren Dauer ist die Wahrscheinlichkeit eine chronische Bronchienerkrankung zu bekommen höher als 0.5

Mögliche nichtberücksichtigte Confounder

- Alter
- Umweltbelastung am vorherigen Wohnort
- Frühere Arbeitsplätze
- Dauer des Rauchens
- genetische Vorbelastung

Name	y-Abschnitt	UMWELT	RAUCH	DAUER
Wert OR KI(OR)	−2.1405 – 	0.6605 1.936 [1.293, 2.897]		
Wert OR KI(OR)	−2.6605	0.6854 1.985 [1.321, 2.981]	0.6533 1.922 [1.197, 3.085]	
Wert OR KI(OR)	−1.7458		0.6233 1.865 [1.167, 2.981]	
Wert OR KI(OR)	−2.4079			0.0428 1.044 [1.026, 1.062]
Wert OR KI(OR)	−3.0953	0.5670 1.763 [1.167, 2.663]		0.0407 1.042 [1.024, 1.060]
Wert OR KI(OR)	−2.9321		0.6599 1.935 [1.198, 3.123]	0.0438 1.045 [1.027, 1.063]
Wert OR KI(OR)	−3.6602	0.5886 1.802 [1.188, 2.731]	0.6819 1.978 [1.220, 3.205]	0.0415 1.042 [1.024, 1.061]

Modellbildung mit SAS

```
proc logistic descending data=logreg;
  model chron = rauch umwelt staub dauer /selection=stepwise;
run;
```

Damit ergibt sich

$$logit(P) = -3.6602 + 0.5886 \cdot UMWELT + 0.6819 \cdot RAUCH + 0.0415 \cdot DAUER$$

.

als bestes Modell.