Großübung zur Biometrie II
im WS 2000/2001

“Logistische Regression”

Stefan Schiffer
dr.stf@web.de
Claus Richterich
Claus@Richterich.net
Thomas Deselaers
Thomas@Deselaers.de

Betreuerin
Nicole Heussen
nicole.heussen@mbio.rwth-aachen.de

Inhaltsverzeichnis

1 Einleitung

Dieser Vortrag soll eine anwendungsorientierte Einführung in das Modellieren von Daten mit Hilfe der **logistischen Regression** geben, die zu einem der wichtigen Auswertungsverfahren der angewandten Statistik gehört. Viele Antwortvariablen in der Medizin sind dichotom (pro Beobachtungseinheit wird, eventuell in einem übertragenen Sinne, ein 'Erfolg' oder ein 'Mißerfolg' festgestellt). Man möchte zum Beispiel häufig herausfinden, ob ein bestimmtes Ereignis oder eine Krankheit auftritt oder nicht.

Die **binäre logistische Regression** ist eine **Regressionsanalyse**, bei der die abhängige Variable eine **dichotome** Ausprägung (kodiert mit 0 oder 1) hat, im Gegensatz zur **linearen Regression**, wo die abhängige Variable aus einem stetigen Zahlenbereich kommt.

Anhand eines Beispieldatensatzes aus einer Studie im **umweltepidemiologischen** Bereich, werden wir das Modell und die Interpretation der Parameter der **logistischen Regression** vorstellen.

2 Das Modell der (linearen) logistischen Regression

2.1 Motivation

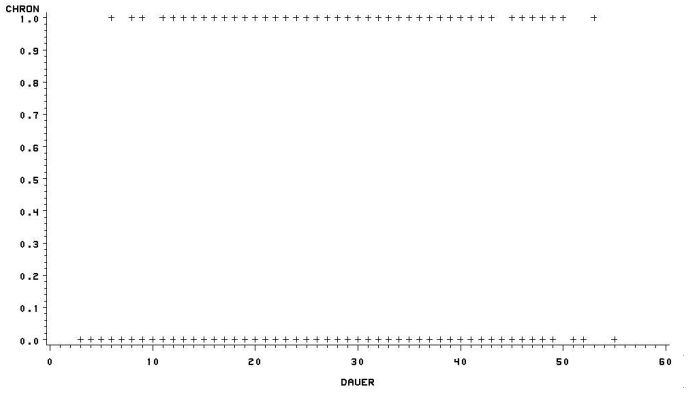
Wie schon behandelt, dient die Regressionsanalyse der Untersuchung des Einflusses von mehreren verschiedenen Faktoren auf ein Zielgröße, so wie es in unserem Datensatz die Faktoren Wohndauer, Umweltbelastung am Wohnort, Staubbelastung am Arbeitsplatz und Rauchverhalten in Bezug auf eine chronische Erkrankung der Bronchien sind. Der Zusammenhang zwischen Einflußfaktoren und Zielgröße wird hier durch den funktionalen Ansatz beschrieben:

$$Y = f(X_1, X_2, \ldots, X_m)$$

Die Ziele der Regressionanalyse sind:

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

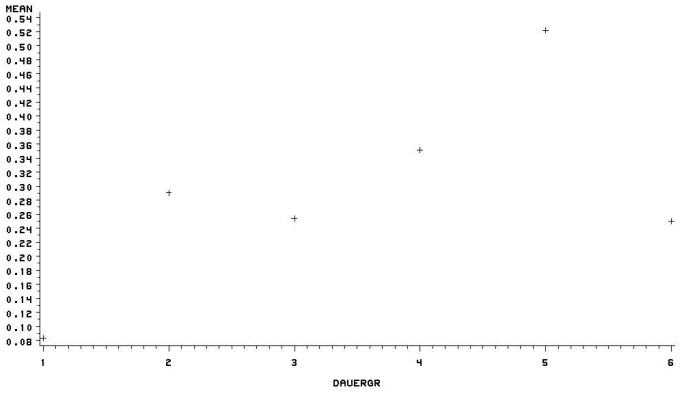
Lineare Regressionsmodelle können nur sinnvoll angwendet werden, wenn die abhängige Variable ein quantitatives Merkmal ist. Ist die Zielvariable qualitativ, so kommt die logistische Regression zum Einsatz. Dabei wird nicht die Zielvariable direkt sondern eine Funktion der Wahrscheinlichkeit, dass die Krankheit unter gegebenen Risikofaktoren auftritt, modelliert. Wären die Daten stetig anstelle von dichotom, so würden wir wahrscheinlich damit beginnen einen Scatterplot zu erstellen, um uns einen groben Überblick über die Daten zu verschaffen. Doch dies hilft uns in diesem Fall nicht sehr viel weiter, wie man am Scatterplot in Bild 2.1 erkennen kann.



Während dieses Bild den dichotomen Charakter der Ausprägung der Variable CHRON sehr deutlich zeigt, ergibt sich kein klares Bild über die Art des Zusammenhangs zwischen den beiden Variablen CHRON und DAUER,

Ein Problem in Bild 2.1 ist die große Streuung der Variable CHRON über den gesamten Wertebereich der Variable DAUER. Dies erschwert es den funktionellen Zusammenhang zwischen den beiden Variablen herauszufinden.

Eine übliche Methode die Gewichtung der Streuung zu reduzieren ohne die Struktur des Zusammenhangs zu zerstören, ist es, die stetige Variable (hier DAUER) in Intervalle aufzuteilen und dann den Mittelwert der abhängigen Variable getrennt für jedes Intervall zu berechnen. Das Ergebnis hiervon sieht man in Bild 2.1, welches den Erwartungswert der Wahrscheinlichkeit eine chronische Erkrankung der Bronchien zu bekommen gegen die Wohndauer, für diejenigen Personen darstellt, die an ihrem Wohnort einer starken Umweltbelastung ausgesetzt sind.



Die Betrachtung des Bildes verdeutlicht den Zusammenhang zwischen der Wohndauer und dem Erwartungswert der Wahrscheinlichkeit einer chronische Erkrankung der Bronchien.

Das Bild, was hier entsteht erscheint ähnlich zu denen, die man erhalten könnte, wenn man eine lineare Regression durchführt, jedoch bestehen hier Unterschiede.

Meistens findet ein **lineare Regression** ihre Anwendung, wenn es um die Darstellung einer Einflußbeziehung geht, jedoch können wir nicht direkt die abhängige Variable modellieren, sondern wollen die Wahrscheinlichkeit modellieren, dass die Variable einen bestimmten Wert annimmt.

Also wollen wir hier das relative Auftreten chronischer Bronchienerkrankungen untersuchen. Dabei erhalten wir das prozentuale Auftreten der Erkrankung in einer Population.

2.2 Einführung des Modells

Anstelle der Erkrankungswahrscheinlichkeit P modelliert man das Odds-Ratio $OR = \frac{P}{1-P}$. Dies gibt die maximale Wettquote an.

Nun wird aber nicht das Odds-Ratio direkt modelliert, sondern die logarithmierten Werte.

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

.

Diese Transformation heißt **logit-Transformation**.

Mittels der oben eingeführten **logit-Transformation** wird das Intervall $[0, 1]$ auf den gesamten Zahlenstrahl abgebildet, wie man in Tabelle 2.2 sehen kann. Damit wird jeder Erkrankungswahrscheinlichkeit eine reele Zahle zugeordnet, die zwischen $-\infty$ und ∞ liegt.

p	0	0.01	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.99	1
logit	$-\infty$	-4.60	-2.94	-2.20	-1.1	0.00	1.10	2.20	2.94	4.60	∞

Wenn man von den L-Werten auf die Wahrscheinlichkeit zurückrechnen will, so erfolgt dies mit Hilfe der **expit-Transformation**:

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

.

Da wir hier das Auftreten einer chronischen Erkrankung der Bronchien als eine Wahrscheinlichkeit p betrachten wollen, dürfen dann die Vorhersagewerte das Intervall $[0, 1]$ nicht verlassen. Das logistische Regressionsmodell kann dadurch definiert werden, dass man $logit(P)$ als Zielvariable eines (multiplen) linearen Regressionsmodells auffasst.

$$logit(p) = \alpha_0 + \alpha_1 \cdot X_1 \ldots \alpha_m \cdot X_m$$

Diese Form ist identisch zu der Form

$$OR = exp(\alpha_0) \cdot exp(X_1)^{\alpha_1} \cdot \ldots \cdot exp(X_m)^{\alpha_m}$$

denn

$$\begin{aligned} OR &= exp(log(OR)) \\ &= exp(log(\tfrac{P}{1-P})) \\ &= exp(logit(P)) \\ &= exp(\alpha_0 + \alpha_1 \cdot X_1 + \ldots + \alpha_m \cdot X_m) \\ &= exp(\alpha_0) \cdot exp(X_1)^{\alpha_1} \cdot \ldots \cdot exp(X_m)^{\alpha_m} \end{aligned}$$

Auflösen nach P ergibt:

$$P = \frac{exp\alpha_0 + \alpha_1 \cdot X_1 + \ldots + \alpha_m \cdot X_m}{1 + exp\alpha_0 + \alpha_1 \cdot X_1 + \ldots + \alpha_m \cdot X_m}$$

,

so dass man die Wahrscheinlichkeit für das Eintreten des Zielereignisses berechnen kann:

X_1 bis X_m sind hierbei m Risikofaktoren als Einflußvariablen und P ist, wie oben entwickelt, die Wahrscheinlichkeit, dass die interessierende Krankheit unter Bedingung der Risikofaktoren auftritt, also das Risiko:

$$P = P(K = 1 \mid X_1 = x_1 \ldots X_m = x_m)$$

,

wobei $X_i = x_i$ den Umstand bedeutet, dass der Risikofaktor X_i, etwa das Gewicht, genau den Wert x_i (z.B. 13 Kilo) annimmt.

Die Risikofaktoren können sowohl stetig als auch dichotom sein.

2.3 Zusammenhang zwischen den Koeffizienten und dem Odds-Ratio

Der Zusammenhang zwischen den Koeffizienten und den Odds Ratio macht die logistische Regression so geeignet für die Auswertung epidemiologische Studien. Um dies darzustellen, beschränken wir uns zunächst auf ein **Modell mit lediglich einem Risikofaktor** X_1 mit zwei Ausprägungen (0 und 1), was der Analyse einer **Vierfeldertafel** entspricht.

$$logit(P) = \alpha_0 + \alpha_1 X_1$$

Wir betrachten nun den Logarithmus des Odds Ratios innerhalb der Vierfeldertafel. Wir erhalten:

$$\begin{aligned} ln(OR) &= ln[\tfrac{Odds(P(K=1|X=1))}{Odds(P(K=1|X=0))}] \\ &= ln[Odds(P(K=1 \mid X=1))] - ln[Odds(P(K=1 \mid X=0))] \\ &= logit[P(K=1 \mid X=1)] - logit[P(K=1 \mid X=0)] \\ &= (\alpha_0 + \alpha_1 1) - (\alpha_0 + \alpha_1 0) \\ &= \alpha_1 \end{aligned}$$

Wie man sieht, stellt sich das Odds Ratio durch die Logarithmierung als eine Differenz von zwei logit-Funktionen da. Diese Differenz wiederum kann durch den Regressionskoeffizienten α_1 ausdrücken. Für das Odds Ratio ergibt sich damit:

$$OR = exp(\alpha_1)$$

Und auch der Parameter α_0 des logistischen Modells ergibt sich zu:

$$\alpha_0 = logit(P(K=1 \mid X=0))$$

Das Risiko der Nichtexponierten $P_{10} = P(K=1 \mid X=0)$ kann mittels Transformation des Parameters α_0 dargestellt werden.

$$P_{10} = P(K=1 \mid X=0) = \frac{exp(\alpha_0)}{1+\alpha_0}$$

Durch diese Tranformationen ist es möglich, die Parameter α_0 und α_1 des logistischen Modells wieder epidemiologisch zu interpretieren.

2.4 Das logistische Modell und stetige Risikofaktoren

Bis hierher haben wir das logistische Modell/Odds-Ratio nur für binäre Einflußvariablen betrachtet. Dabei ist das Odds-Ratio das Verhältniss zwischen den Krankheitschancen bei Exponierten ($X_1 = 1$) zu Nicht-Exponierten ($X_1 = 0$).

Mit Hilfe des logistischen Modells können Odds-Ratios für beliebige Arten von Expositionsleveln bestimmt werden. Hierbei werden die Exposition x_1 einer bestimmten Person mit der Exposition x_2 einer anderen Person verglichen. Dies sieht wie folgt aus:

$$\begin{aligned} ln(OR(x_1, x_2)) &= logit[P(K = 1 \mid X = x_1)] - logit[P(K = 1 \mid X = x_2)] \\ &= (\alpha_0 + \alpha_1 x_1) - (\alpha_0 + \alpha_1 x_2) \\ &= \alpha_1(x_1 - x_2)\,, \end{aligned}$$

so dass $OR(x_1, x_2) = exp[\alpha_1(x_1 - x_0)]$.

2.5 Dosis - Wirkung Beziehung

Wenn der Einfluß eines Risikofaktors stetig gemessen wird, so kann eine Beziehung zwischen Exposition und Krankheit auch genau quantifiziert werden. Die Modellierung dieser Dosis - Wirkung Beziehung bedarf einiger Vorüberlegungen.

So besteht die eigentliche Wirkung einer Exposition in der Erhöhung des Risikos, an der Krankheit zu leiden. Der Verlauf der zugehörigen Erkrankungswahrscheinlichkeit würde wohl ein s-förmiger sein, bei zunächst geringer Exposition ist das Krankheitsrisiko kaum erhöht, dann steigt es an, bei starker Exposition flacht sie wieder ab, d.h. dass bei schon extrem hoher Exposition ein weitere Erhöhung kaum weitere Erkrankungen verursacht. Diese Verteilung entspricht der logistischen Verteilungsfunktion, die dem logistischen Modell seinen Namen gab.

Betrachtet man nun den von x_1 abhängigen Odds Ratio, so erhält man bei stetigen Risikofaktoren den schon bekannten exponentiellen Zusammenhang

$$OR(x_1) = exp(\alpha x_1).$$

In der Praxis kommt es oft vor, dass die Dosis - Wirkung Beziehung nicht den gesamten Verlauf einer s-Kurve beschreibt, sondern vielmehr einen Ausschnitt daraus. Meist liegt das daran, dass extrem starke Expositionen gar nicht auftauchen. Das läßt die Frage auftauchen, ob vielleicht auch andere Dosis - Wirkung Beziehungen betrachtet werden sollten und können, z.B. auch nicht lineare. Dies entspricht der Frage, ob $logit(P)$ auch durch andere Funktionen als eine Gerade angenähert werden kann. Dies ist der Fall, so dass auch Modellierungen wie $ln[OR](x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 {x_2}^2$ oder $ln[OR](x) = \alpha_0 + \alpha_1 ln(x_1 + 1)$ mit Hilfe des logistischen Ansatzes gebildet werden können.

2.6 Interpretation der Koeffizienten

Nun sollen die Koeffizienten der Modellgleichung diskutiert werden. Dabei bedeutet:

α_0 log odds für ein Individuum mit $X = 0$

α_i Veränderung des log-odds bei Veränderung der entsprechenden Variable um eine Einheit.

Wenn das Odds-Ratio größer als 1 ist, ist die Wahrscheinlichkeit zu erkranken für die exponierten Personen größer als für die nicht Exponierten. z.B. bei einem $OR = 1.6$ ist es 60% wahrscheinlicher zu erkranken.

α_0
- $\alpha_0 = 0$: Wenn der Y-Achsenabschnitt 0 ist, so wissen wir, dass für $X_1 = 0$ der logit-Wert 0 beträgt, d.h. dass das Risiko 1 ist und so die unabhängige Variable X_1 den Wert $p = 0.5$ annehmen muss.
- $\alpha_0 < 0$: Hier ist für $X_1 = 0$ das Risiko < 1 und damit $p < 0.5$.
- $\alpha_0 > 0$: Hier ist für $X_1 = 0$ das Risiko > 1 und damit $p > 0.5$.

α_i
- $\alpha_i = 0$: X_i hat keinen (logistisch beschreibbaren) Einfluß auf die abhängige Variable
- $\alpha_i < 0$: Es **sinkt** die Auftretenswahrscheinlichkeit p mit wachsendem X_i.
- $\alpha_i > 0$: Es **steigt** die Auftretenswahrscheinlichkeit p mit wachsendem X_i.

2.7 Qualität eines Modells

Natürlich ist interessant, wie gut oder wie schlecht ein postuliertes Modell ist. Eine Antwort auf diese Frage kann man unter Verwendung von statistischen Tests ermitteln. Der wohl angesehenste ist hierbei der **Likelihood-Quotienten-Test**.

Als Nullhypothese wird hierbei angenommen, dass der Faktor keinen Einfluß auf das Modell hat, d.h. das wahre Odds-Ratio des Faktors hat den Wert 1, woraus folgt, dass der Regressionskoeffizient α_i den Wert 0 hat.

3 Modellbildung, Confounding

Bis jetzt sind wir beim Einsatz des logistischen Modells davon ausgegangen, dass die konkrete Fragestellung zu einer konkreten Formulierung einer Ursache - Wirkung Beziehung führte, die den logit der Erkrankungswahrscheinlichkeit einer Zielvariable in Verbindung mit einem Vektor von n Einflußvariablen setzt. Jedoch sollten inhaltliche und statistische Fragestellungen nie unabhängig voneinander behandelt werden.

3.1 Modellbildung

Die Festlegung des richtigen Modells im Einzelfall bedarf einer gemeinsamen Abstimmung aller Beteiligten Institutionen **vor** Studienbeginn. Dabei müssen alle denkbaren Voraussetzungen, Annahmen und Möglichkeiten des Zusammenwirkens der Risikofaktoren und der Zielvariable geklärt werden.

Um einen Zusammenhang zwischen Zielvariable und Einflußfaktoren zu bestimmen bildet man ein geeignetes Modell. Theoretisch können dabei alle gemessenen Einflüsse in das Modell aufgenommen werden, jedoch ist dies oft nicht sinnvoll, da Variablen irrelevant sein können.

Dieses Modell soll dann den unbekannten Zusammenhang erklären, wobei hier nicht auf Einzelschicksale sondern vielmehr auf den durchschnittlichen Effekt Wert gelegt wird. Ein gutes Modell arbeitet die wesentlichen Strukturen der Zusammenwirkung heraus. Auf dem Weg zum "richtigen" Modell kann man verschiedene Strategien verfolgen:

inhaltlich begründet Aufgrund empirischer Beweggründe oder fundierten Halbwissens kann man von Anfang an Variablen als unbedingt notwendig oder irrelevant klassifizieren. Dieses Verfahren zur Modellbildung ist den anderen Verfahren vorzuziehen.

vorwärtige Variablenselektion Man geht von einer leeren Menge von Einflußvariablen aus und fügt sukzessive weitere Einflußvariablen hinzu.

rückwärtige Variablenselektion Man geht anfangs davon aus, dass alle Variablen Einfluß haben und nimmt dann inkrementell Variablen weg.

schrittweise Variablenselektion Während bei der vorwärtigen Variablenselektion Variablen, die einmal in das Modell aufgenommen wurden, stets in diesem Modell bleiben, entscheidet das schrittweise Verfahren nach jeder neu aufgenommenen Variable für alle im Modell enthaltenen Variablen erneut, ob diese tatsächlich das Einschlusskriterium erfüllen.

Mit dem Likelihood-Quotienten-Test bestimmt man die Qualität eines Modells, indem man für einzelne Parameter überprüft, ob sie Einflußvariablen für das Modell sind.

3.2 Confounding

Confounding ist eine Methode um störende Einflüsse zu analysieren und kontrollieren. Die Berücksichtigung von **Counfoundern** (störenden Einflüssen) ist eine Motivation das logistische Modell anzupassen. Man muss allerdings sehr sorgfältig überlegen, welche Variablen als Confounder in das Modell mit aufgenommen werden sollen oder nicht. Diese Entscheidung

wird davon getragen, ob a) die Daten ein Confounding erkennen lassen und b) das ganze inhaltlich verträglich ist.

Confounding ist von sehr großer Bedeutung, da es leicht zu Fehlinterpretationen kommen kann, wenn man dies vernachlässigt. Andererseits muss man auch beachten, dass eine sehr große Zahl von ins Modell aufgenommenen Confoundern die Varianz zu einer stark erhöhten Varianz des Schätzers führt.

Man nimmt Confounder als Einflußvariablen in das Modell auf, um sie im Modell zu berücksichtigen. Wird der interessierende Haupteinflußfaktor als Exposition des Parameters x_i betrachtet, dann können alle anderen Parameter Confounder oder zusätzliche Einflußfaktoren sein. Insgesammt erreicht man durch die Berücksichtigung von Confoundern eine adjustierte Aussage bezüglich der Einflussfaktoren.

4 Lösung der Aufgaben

4.1 Deskription der Daten

Beginnen Sie mit einer sinnvollen Deskription der Daten, beschreiben Sie den Datensatz auch stratifiziert nach den Ausprägungen der qualitativen Merkmale.(`PROC FREQ, PROC MEANS, PROC UNIVARIATE`)

Bei den Daten handelt es sich, wie bereits oben erwähnt, um 623 Datensätze aus einer umweltepidemiologischen Studie über Atemwegserkrankungen in industriell geprägten Regionen.

Die Daten sind in folgender Form gespeichert, wobei die einzelnen Spalten folgende Bedeutung haben:

`CHRON` Diese Variable ist eine dichotome Variable, die angibt, ob die untersuchte Person eine chronische Erkrankung der Bronchien hat (1) oder nicht (0).

`STAUB` Diese Variable gibt die Staubbelastung am Arbeitsplatz in mg/m^3 an.

`RAUCH` Diese Variable ist eine dichotome Variable, die angibt, ob die untersuchte Person Raucher ist (1) oder nicht (0).

`DAUER` Diese Variable gibt in Jahren an, wie lange die Person bereits am untersuchten Wohnort lebt.

`UMWELT` Diese Variable ist eine dichotome Variable, die angibt, ob die untersuchte Person an ihrem Wohnort einer starken (2) oder einer schwachen (1) Umweltbelastung ausgesetzt ist.

Ein Ausschnitt aus der Datei mit den Messwerten:

```
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
...
```

Um sich einen ersten Einblick in die Daten zu verschaffen, haben wir ersteinmal einen Überblick erstellt, wie die Werte verteilt sind. Hierzu haben wir die beiden "stetigen" Variablen DAUER und STAUB in kleine, jeweils gleichgroße Intervalle eingeteilt.

Diese Daten lassen wir von SAS auswerten:

```
PROC FREQ DATA=logreg;
  TABLES chron staubgr rauch umwelt dauergr;
RUN;
```

Als Ergebnis erhalten wir, folgende Tabellen:

Für die Variable CHRON, wobei eine 1 bedeutet, dass die Person eine chronische Erkrankung der Bronchien hat, und 0, dass die Person keine chronische Erkrankung der Bronchien hat:

CHRON	Anzahl	Prozent
0	487	78.2
1	136	21.8

Man sieht also, dass sehr viel mehr Personen nicht erkrankt sind, als Personen an einer chronischen Erkrankung leiden.

Für die Variable RAUCH, wobei eine 1 bedeutet, dass die Person Raucher ist, und eine 0, dass die Person Nichtraucher ist:

RAUCH	Anzahl	Prozent
0	175	28.1
1	448	71.9

Also sind ca. 72% der Personen, die an der Studie teilgenommen haben Raucher.

Für die Variable Umwelt, wobei eine 1 bedeutet, dass die Person an ihrem Wohnort keiner starken Umweltbelastung ausgesetzt ist, und eine 2, dass die Person an ihrem Wohnort einer starken Umweltbelastung ausgesetzt ist:

UMWELT	Anzahl	Prozent
1	453	72.7
2	170	27.3

Die Variable DAUER haben wir in 6 gleichgroße Intervalle eingeteilt. Die ersten Gruppe enthält alle Personen, die 0 bis 10 Jahre an dem Ort wohnen, die Zweite enthält alle Personen, die 11-20 Jahre an dem Ort wohnen,... die 6. Gruppe enthält alle Personen, die mehr als 50 Jahre an ihrem Wohnort wohnen.

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

Die Variable STAUB haben wir in 5 Gruppen eingeteilt. Die erste Gruppe enthält alle Personen, die 0 bis 2 mg/m^3 Staubbelastung am Arbeitsplatz haben. In der zweiten Gruppe sind die Personen, die 2 bis 4 mg/m^3 Staubbelastung haben. In der dritten Gruppe sind die Personen, die $4-6$ mg/m^3 Staubbelastung haben, in der vierten Gruppe sind die Personen, die $6-8$ mg/m^3 Staubbelastung haben und in der fünften Gruppe sind die Leute, die eine noch höhere Belastung durch Staub an ihrem Arbeitsplatz haben.

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

Außerdem haben wir für die beiden stetigen Variablen einige Kenngrößen berechnet:

Name	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

Um sich einen ersten Eindruck zu verschaffen, wie groß der Einfluß des Rauchens und wie groß der Einfluß der Umweltbelastung am Wohnort auf das Auftreten einer chronischen Erkrankung der Bronchien ist, sind hier exemplarisch drei Tabellen und Diagramme aufgeführt.

TABLE OF CHRON BY RAUCH

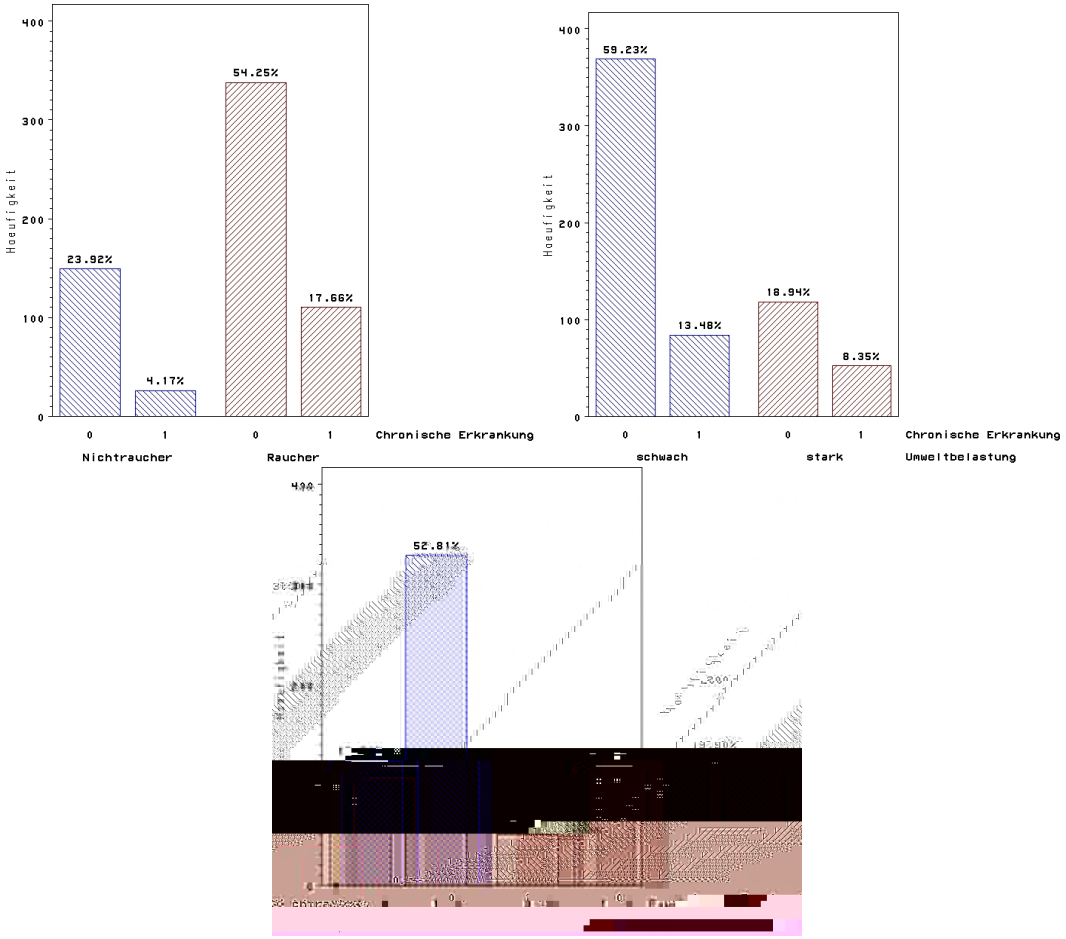
CHRON / Frequency / Percent	RAUCH 0	RAUCH 1	Total
0	149	338	487
	23.92	54.25	78.17
1	26	110	136
	4.17	17.66	21.83
Total	175	448	623
	28.09	71.91	100.00

TABLE OF CHRON BY UMWELT

CHRON / Frequency / Percent	UMWELT 1	UMWELT 2	Total
0	369	118	487
	59.23	18.94	78.17
1	84	52	136
	13.48	8.35	21.83
Total	453	170	623
	72.71	27.29	100.00

TABLE OF RAUCH BY UMWELT

RAUCH / Frequency / Percent	UMWELT 1	UMWELT 2	Total
0	124	51	175
	19.90	8.19	28.09
1	329	119	448
	52.81	19.10	71.91
Total	453	170	623
	72.71	27.29	100.00



Die obigen Darstellungen zeigen, wieviele der Raucher bzw. wieviele der stark umweltbelasteten Personen eine chronische Erkrankung der Bronchien haben.

4.2 Einfluß der Umweltbelastung auf CHRON

Welchen Einfluß besitzt die Umweltbelastung am Wohnort auf das Auftreten einer chronischen Erkrankung der Bronchien? Wie ist der Einfluß zu bewerten, wenn Sie weitere potentielle Risikofaktoren in die Analyse mit einbeziehen? (`PROC FREQ, PROC LOGISTIC`)

Da es sich hier bei beiden zu untersuchenden Variablen um dichotome Merkmale handelt, bietet sich eine Vierfeldertafel an, um den Einfluß zu beschreiben:

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

Was zu folgender Ausgabe führt:

```
     TABLE OF CHRON BY UMWELT

CHRON     UMWELT

Frequency|
Percent  |
Row Pct  |
Col Pct  |       1|       2|  Total
---------+--------+--------+
       0 |    369 |    118 |    487
         |  59.23 |  18.94 |  78.17
         |  75.77 |  24.23 |
         |  81.46 |  69.41 |
---------+--------+--------+
       1 |     84 |     52 |    136
         |  13.48 |   8.35 |  21.83
         |  61.76 |  38.24 |
         |  18.54 |  30.59 |
---------+--------+--------+
Total         453      170      623
            72.71    27.29   100.00


        STATISTICS FOR TABLE OF CHRON BY UMWELT

Statistic                     DF     Value        Prob
------------------------------------------------------
Chi-Square                     1    10.510       0.001
Likelihood Ratio Chi-Square    1    10.020       0.002
Continuity Adj. Chi-Square     1     9.816       0.002
Mantel-Haenszel Chi-Square     1    10.493       0.001
Fisher's Exact Test (Left)                       0.999
                    (Right)                   1.08E-03
                    (2-Tail)                  1.57E-03
Phi Coefficient                     0.130
Contingency Coefficient             0.129
Cramer's V                          0.130

                  Sample Size = 623
```

Demzufolge ist also der Einfluß der Umweltbelastung auf die Tatsache, ob jemand eine chronische Erkrankung der Bronchien bekommt statistisch signifikant, da der p-Wert deutlich kleiner als 0.05 ist.

Weiterhin haben wir eine Logistische Regression mit den gegebenen Daten gemacht, um eine

Regressionsgleichung, die den Einfluß der Variable UMWELT auf die Variable CHRON modelliert zu erhalten.

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

Als Ergebnis erhält man folgende Regressionsgleichung:

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

Wobei für die Variable Umwelt das Odds-Ratio $OR = 1.936$ beträgt. Somit ist also die Wahrscheinlichkeit eine chronsiche Bronchialerkrankung zu erleiden für stark umweltbelastete Personen fast doppelt so groß. Das 95%-Konfidenzintervall zum Odds-Ratio ist $[1.293, 2.897]$.

Wir testen, ob die Variable STAUB einen statistisch signifikanten Einfluss auf die Variable CHRON hat. Dies machen wir mit dem Wilcoxon-Test. In SAS geht dies mit der Prozedur:

```
proc npar1way data=logreg wilcoxon;
  class chron;
  var staub;
run;
```

Hierbei ergibt sich eine p-Wert von 0.6205. Demnach hat die Variable STAUB keinen Einfluß.

Das gleiche Verfahren wird angewandt um zu testen, ob die Variable DAUER einen Einfluß hat, ergibt einen p-Wert von 0.001. Dies besagt also, dass die Variable DAUER einen statistisch signifikanten Einfluß auf CHRON hat.

Wie man später sehen wird, hat auch die Variable RAUCH einen statistisch signifikanten Einfluss auf CHRON.

Weitere Einflussvariablen berücksichtigen wir im Kapitel 4.6 bei der Bildung des besten Modells.

Die Variablen UMWELT, RAUCH und DAUER haben somit einen signifikanten Einfluss auf CHRON. STAUB jedoch nicht.

4.3 Einfluß von RAUCH auf CHRON

Wie hoch ist die Chance einer bronchialen Erkrankung bei einem Raucher gegenüber einem Nichtraucher, wenn man das Rauchen als alleinigen Risikofaktor betrachtet? Wie relativiert

sich dieses Chance, wenn zusätzliche Risikofaktoren betrachtet werde. (`PROC FREQ, PROC LOGISTIC`)

Diese Frage entspricht beinahe der Frage in der letzten Aufgabestellung, dementsprechend ist das Verfahren sehr ähnlich. Zusätzlich wird allerdings das Odds-Ratio für Raucher gegen Nichtraucher berechnet.

In SAS kann man die mit `PROC FREQ` mit der Option `RELRISK` berechnen:

```
proc freq data=logreg;
  tables chron*rauch /cmh relrisk nocol norow chisq;
run;
```

Hierbei ergibt sich für den Chi-Quadrat-Test ein p-Wert von 0.008. Das bedeutet also, dass Rauchen einen statistisch signifikanten Einfluss auf die Wahrscheinlichkeit eine chronische Erkrankung der Bronchien zu bekommen hat.

Als Odds-Ratio ergibt sie $OR = 1.865$. Also ist die Gefahr für einen Raucher eine chronische Erkrankung der Bronchien zu bekommen ca. 1,865 mal so groß wie die eines Nichtrauchers.

Als Regressionsgleichung ergibt sich:

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

Wobei für die Variable `RAUCH` das Odds-Ratio $OR = 1.865$ beträgt. Somit ist also die Wahrscheinlichkeit eine chronsiche Bronchialerkrankung zu erleiden für einen Raucher 85% größer als für einen Nichtraucher. Das 95%-Konfidenzintervall zum Odds-Ratio ist $[1.167, 2.981]$.

Weitere Einflussvariablen berücksichtigen wir im Kapitel 4.6 bei der Bildung des besten Modells. Der Einfluss der anderen Variablen ist hier genauso wie in der vorherigen Aufgabenstellung zu bewerten.

4.4 Nach wieviel Jahren wird die Umweltbelastung zum Risikofaktor

Offensichtlich ist die Zeit, die jemand an einem belasteten Ort verbracht hat ein wichtiger Einflußfaktor auf die Entwicklung einer bronchialen Erkrankung. Bestimmen Sie den Zeitpunkt, bis zu dem man umziehen sollte, damit die Dauer des Aufenthaltes nicht zu einem Risikofaktor wird.

Um zu bestimmen, zu welchem Zeitpunkt man spätestens umziehen sollte, damit die Dauer des Aufenthaltes nicht zu einem Risikofaktor wird, muss man sich zuerst die Frage stellen, ab welcher Wahrscheinlichkeit, eine chronische Erkrankung der Bronchien zu erleiden, man davon

sprechen kann, dass es sich hierbei um einen Risikofaktor handelt. Wir gehen hier mal davon aus, dass es keine Risikofaktor ist, solange die Wahrscheinlichkeit unter 50% bleibt. Also stellt sich die Frage, ab welcher Dauer des Aufenthaltes es wahrscheinlicher als 50% wird, dass man eine chronische Erkrankung der Bronchien bekommt. Um dies zu berechnen braucht man die Regressionsgleichung, die die Wahrscheinlichkeit eine chronische Erkrankung der Bronchien zu bekommen als Funktion der Wohndauer.

Die hierzu verwendete SAS-Funktion sieht folgendermaßen aus:

```
PROC LOGISTIC DESCENDING DATA=logreg;
  MODEL chron = dauer ;
  BY umwelt;
RUN;
```

Aus der Ausgabe hiervon lesen wir die beiden Werte a und b für die Gleichung

$$logit(\hat{P}) = a + dauer \cdot b$$

Diese sind $a = -1.5824$ und $b = 0.0275$ womit sich die Gleichung zu

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

.

Dies läßt sich auflösen zu

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

Also ist ab einer Wohndauer von 57.54 Jahren bei hoher Umweltbelastung am Wohnort die Umweltbelastung als Risikofaktor ernst zu nehmen.

4.5 Welche Variablen fehlen im Modell?

Welche nicht im Modell enthaltene Variable könnte den Effekt der Dauer des Aufenthaltes verursachen?

Im direkten Zusammenhang mit der DAUER steht das Alter der Person, da Personen, die erst 10 Jahre alt sind, nicht länger als 10 Jahre an einem Ort wohnen können. Weiterhin kann man sagen, dass Personen, die schon sehr lange an einem Ort wohnen schon recht alt sein müssen. Wenn nun das Alter der Person einen direkten Einfluß auf die Anfälligkeit gegenüber

chronischen Bronchienerkrankungen hätte, so ist dieser Effekt in der Studie als möglicher **Confounder** nicht berücksichtigt.

Weiterhin ist in den Daten nicht enthalten, wie hoch die Umweltbelastung an dem Ort war, wo die Person wohnte, bevor sie an den Ort zog, wo sie wohnte, als der Test durchgeführt wurde. Dies ist besonders bei Leuten, die erst seit einer kurzen Zeit an ihrem aktuellen Wohnort wohnen von großer Bedeutung.

Auch die Vorbelastung durch frühere Arbeitsplätze ist nicht in der Studie enthalten. Wenn jemand früher als Bergmann gearbeitet hat, diesen Beruf aber z.B. aufgrund einer Silikose aufgeben mußte, so ist er sicherlich in einer anderen Risikogruppe, als jemand, der immer im Büro gearbeitet hat. Genausowenig wird berücksichtigt, seit wie langer Zeit die Person ihren evtl. sehr stark staubbelasteten Arbeitsplatz bereits hat.

Auch die Zeit, seit der eine Person Raucher ist, ist nicht in der Studie festgehalten.

Da die Studie als Querschnittsstudie nur ein sehr kurzes Zeitintervall umfaßt, berücksichtigt sie in keiner Weise die "Vorbelastung"der Probanden. Vielleicht wäre hier eine Langzeitstudie angebrachter.

4.6 Bestimmung des “besten” Modells

Bestimmen Sie mit Hilfe von Modellsuchverfahren das ‚beste’ Modell. (`PROC LOGISTIC`)

Wie bereits oben gezeigt üben die Variablen `Umwelt`, `RAUCH` und `DAUER` einen statistisch signifikanten Einfluß auf `CHRON` aus. Daher haben wir einmal alle möglichen Regressionsgleichungen mit beliebigen Kombinationen dieser Variablen berechnet. Da es nich sinnvoll ist, diese alle als Gleichung anzugeben, haben wir sie tabellarisch aufgelistet.

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

Mit SAS kann man mit

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

nach dem besten Modell suchen. Als Ergebnis erhält man, dass die Variablen `UMWELT`, `RAUCH` und `DAUER` das beste Modell zu dieser Studie bilden. Also kann man die Regressionsgleichung aus obiger Tabelle ablesen:

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

.

5 Literaturangaben

- Kreienbrock, L., Schach, S.; Epidemiologische Methoden; Gustav Fischer Verlag, 1995.

- SAS Institute; Logistic regression; a self-learning text; Springer-Verlag 1994.
- David W. Hosmer & Stanley Lemeshow - Applied Logistic Regression
- H.G. Lipinski - Einführung in die medizinische Informatik
- Günther Gediga; (Münster: LIT-Verlag, 1998) http://www.psycho.uni-osnabrueck.de/skala/ch05.htm
- Armin Koch - http://www.urz.uni-heidelberg.de/statistik/sas-ah/B/LogistischeRegression.html