Seminar 27. - 30. September 1999 Basel
Region Oesterreich-Schweiz (ROeS) of the
International Biometric Society

Logistische Regressionsmodelle: Anwendung auf reale Daten und Simulationsergebnisse

Iris Zöllner, Eva Herbert

LGA Baden-Württemberg, Stuttgart

Bei der Analyse von Daten aus epidemiologischen Studien kommen häufig logistische Regressionsmodelle zur Anwendung. Die Ergebnisse werden meistens ohne Angaben zur Modellanpassung veröffentlicht, was dazu führt, daß die Bewertung der gefundenen Effekte ohne Einblick in die Originaldaten in vielen Fällen erschwert wird. Ein weiteres Problem vieler Studien besteht in der relativ großen Anzahl von betrachteten Einflußfaktoren, so daß auch "Zufallseffekte" zu erwarten sind.
Aus diesen Gründen wurden für eine umweltmedizinische Untersuchung zusätzlich zur Auswertung der Daten aus Elternfragebögen entsprechende zufällige Ergebnisse generiert und anschließend analog zu den Daten ausgewertet.
Simuliert wurden folgende Modellsituationen:

A: dichotome Zufallsvariablen
alle Variablen unabhängig (jeweils 1 Zielvariable, 10 Einflußvariable)

B: dichotome Zufallsvariablen
Korrelationen zwischen einigen
Einflußgrößen, Unabhängigkeit
zwischen Ziel- und Einflußgrößen (jeweils 1 Zielvariable, 10 Einflußvariable)

Die erzeugten Zufallsergebnisse waren binomialverteilt, wobei die Parameter sich an denen der real beobachteten Daten orientierten.

Beim Vergleich der Resultate der logistischen Regressionen wurden die Anzahl der gefundenen Effekte, die Höhe der OR-Schätzer und mehrere Parameter für die Modellanpassung betrachtet. Dabei entsprach die Anzahl der gefundenen Effekte pro Regressionsmodell bei den Zufallsdaten der Erwartung. In den "realen" Daten lag die Zahl höher, was ebenfalls nicht überraschte. Die OR-Schätzer unterschieden sich weniger in der Höhe als in der Form ihrer Verteilung. Weiter zeigte sich, daß die Parameter für die Modellanpassung nicht einheitlich reagieren. Anhand von R^2 , c und -2 LogL ließen sich die "Zufallsmodelle" am ehesten von den realen Datenmodellen trennen.

Logistische Regressionen in umweltepidemiologischen Untersuchungen
Anwendung auf reale Daten und Simulationsergebnisse

In der Analyse von Daten aus epidemiologischen Studien stellt die logistische Regression ein gängiges Verfahren dar. Untersucht werden Zusammenhänge zwischen einer oder mehreren Einflußvariablen und einer Zielvariablen. Modelliert wird die logit-transformierte Wahrscheinlichkeit für das Auftreten der Zielvariablen in Abhängigkeit von den Einflußgrößen. Im epidemiologischen Kontext soll das Modell eine Beziehung zwischen Risikofaktoren (und Confoundern) als möglichen Einflußgrößen und Krankheiten bzw. Symptomkomplexen als 'Zielgrößen' beschreiben.

Ein Problem vieler Studien besteht in der relativ großen Anzahl von betrachteten Einflußfaktoren, so daß bei Auswertungen durch die Vielzahl der immanenten Tests auch Zufallseffekte zu erwarten sind. Außerdem werden Studienergebnisse häufig ohne Angaben zur Modellanpassung veröffentlicht, was die Bewertung der gefundenen Effekte ohne Einblick in die Originaldaten in vielen Fällen erschwert.
Aus diesen Gründen wurden für eine umweltmedizinische Untersuchung, -zusätzlich zur Auswertung der Daten aus Elternfragebögen -, Zufallszahlen generiert und anschließend analog zu den Daten ausgewertet (SAS Version 6.12). Besondere Aufmerksamkeit galt dabei der Betrachtung von Odds Ratios und den Schätzern für die Güte der Anpassung des Gesamtmodells.

Charakterisierung der Daten und Modellsituationen

Es wurden ausschließlich Variablen mit dichotomer Ausprägung (0, 1) betrachtet.

Datenmodell A	simulierte Zufallsvariable	alle Variablen unabhängig
Datenmodell B	simulierte Zufallsvariable	einzelne Ziel- und Einfluß-variablen jeweils *untereinander* korreliert, aber Unabhängigkeit zwischen Ziel- und Einflußgrößen

***Datenmodell* PPBEO** Daten aus dem Pilotprojekt Beobachtungsgesundheitsämter

Tabelle 1: **Datenmodell A**

Erzeugen von unabhängigen Zufallsvariablen binomialverteilt - N = 1000

Zielvariable	**p = P(zi=1)**	**Einflußvariable**	**p = P(xi=1)**
z1	0,05	x1	0,10
z2	0,10	x2	0,15
z3	0,15	x3	0,20
z4	0,20	x4	0,25
z5	0,25	x5	0,25
z6	0,30	x6	0,30
z7	0,35	x7	0,45
z8	0,40	x8	0,50
		x9	0,60
p = Auftretenswahrscheinlichkeit einer Variablen		x10	0,70

Tabelle 2: **Datenmodell B**

Erzeugen von abhängigen Zufallsvariablen binomialverteilt - N = 1000

Zielvariable	**p = P(zi=1)**	**Einflußvariable**	**p = P(xi=1)**
z1	0,05	x1	0,10
z2	0,10	x2	0,15
z3	0,15	x3	0,20
z4	0,20	x4	0,25
z5	0,25	x5	0,25
z6	0,30	x6	0,30
z7	0,35	x7	0,45
z8	0,40	x8	0,50
		x9	0,60
p = Auftretenswahrscheinlichkeit einer Variablen		x10	0,70

Abhängig erzeugt: (z1 - z4 - z7) (z2 - z5 - z8) (z3 - z6)
(x1 - x5 - x9 - x10) (x2 - x6) (x3 - x7) (x4 - x8)

Für jede der in Tabelle 1 und 2 aufgeführten Modellsituationen wurden 30 Realisierungen mit N = 1000 Beobachtungen generiert.

Bei den empirischen Daten handelt es sich um Datensätze aus einem umweltmedizinischen Pilotprojekt des Landesgesundheitsamts Baden-Württemberg. Die Daten stammen aus drei Untersuchungsabschnitten in den Winterhalbjahren 1992/93, 1993/94 und 1994/95. Sie wurden zum einen getrennt ausgewertet (PPBEO92, PPBEO93, PPBEO94; N ≅ 700) und zum anderen zur Auswertung in einem Datensatz zusammengefaßt (PPBE925, N ≅ 2000).

Tabelle 3: **PPBEO-Daten**

Variablen und beobachtete Auftretenswahrscheinlichkeiten

Zielvariable	**p = P(zi=1)**	**Einflußvariable**	**p = P(xi=1)**
AALL	0,21	gend	0,50
ASTHMA	0,05	passivr	0,48
AASTH	0,05	status	0,45
APSEUDO	0,16	tier	0,56
AKEUCH	0,29	schimmel	0,13
ALUNG	0,12	bedroom	0,40
AHEU	0,08	allfam	0,33
ASTOBRO	0,12	heizung	0,21
		ortma	0,27
p = Auftretenswahrscheinlichkeit einer Variablen		ortke	0,41

Odds ratio

In der Epidemiologie hat die*[1] „Odds Ratio“ (OR) als Schätzer für relative Risiken eine große Bedeutung. Als Odds einer Wahrscheinlichkeit ist der Ausdruck

$$\text{Odds}(P) = \frac{P}{1 - P}$$

definiert. Es gibt die Chance an, mit der ein Ereignis eintritt. Die Chance (Odds), unter Exposition zu erkranken, dividiert durch die Chance (Odds), bei fehlender Exposition zu erkranken, ergibt die Odds Ratio (Chancenverhältnis). Bei niedriger Prävalenz einer Krankheit entspricht die OR etwa dem relativen Risiko. Im Rahmen der logistischen Regression kann die OR direkt aus den Parametern hergeleitet werden. Sei $x = (x_1 , x_2 , \ldots x_m)$ und

$$\text{logit}\,(P\,(Y = 1 \mid X = x)) = ß_0 + ß_1x_1 + ß_2x_2 + \ldots + ß_mx_m$$

dann gilt: $\exp(ß_i) = OR\,(x_i)$ für i=1,...,m

Die ML-Schätzer der Parameter $ß_0 - ß_m$ ergeben sich durch Maximierung der Likelihoodfunktion

$L\,(ß_0 \ldots ß_m) = \Pi\, P\,(k_i = j \mid X_i^{(1)}, \ldots , X_i^{(m)})$ i = 1, ..., n Individuen. Dabei ist

j = 1 , wenn eine Person erkrankt ist, und j = 0, wenn keine Erkrankung vorliegt.

Um Aussagen über die Genauigkeit der Schätzung der ORs und damit Informationen über die Gültigkeit einer Risikoaussage in der Zielpopulation zu erhalten, wurden die Konfidenzintervalle (KI) der ORs betrachtet. Die Konfidenzintervalle überdecken mit einer Wahrscheinlichkeit von 95% die unbekannten ORs. Schließt ein Konfidenzintervall die 1 aus, geht man in der epidemiologischen Praxis häufig von "gefundenen Effekten" aus.

Maßzahlen für die Güte der Anpassung (bzw. der ‘Vorhersage’) des Gesamtmodells

Im Rahmen der Untersuchungen wurden auch Parameter zur Güte der Anpassung (bzw. der Vorhersage) des Gesamtmodells in Bezug auf die Daten betrachtet. Ausgewählt wurden folgende Parameter:

- Deviance
- Pearson Statistik
- -2 log L
- R^2 (verallgemeinertes Bestimmtheitsmaß)
- Hosmer & Lemeshow Goodness of Fit Statistik
- c (Rangkorrelationsindex)

[1] * auch: das

Ergebnisse der Anpassung logistischer Regressionsmodelle in den untersuchten Modellsituationen

Beim Vergleich der Resultate logistischer Regressionen wurden die Anzahl der "gefundenen Effekte" (KI der OR überdeckt 1,0 nicht), die Höhe der OR-Schätzer und mehrere Parameter für die Modellanpassung betrachtet.

Das Verhältnis von “gefundenen Effekten" (KI des OR überdeckt 1,0 nicht) zu getesteten Modellen wurde für jede Modellsituation bestimmt. In den zufälligen Datensätzen (A und B) wurden 0,475 bzw. 0,479 “signifikante“ Odds Ratios pro Modellanpassung gefunden. Mit anderen Worten, unter uneingeschränkten (A) sowie unter eingeschränkten (B) Zufallsbedingungen wurde für jedes 2. Modell durchschnittlich ein Effekt gefunden. Diese Größenordnung war bei einer Irrtumswahrscheinlichkeit von α=0,05 zu erwarten.

Regressionsmodell mit 10 Einflußvariablen	$\Rightarrow$	10 „immanente“ Tests
240 getestete Modelle	$\Rightarrow$	2400 Tests
Irrtumswahrscheinlichkeit α=0,05	$\Rightarrow$	etwa 120 „falsch positive“ Tester-gebnisse erwartet

In der Auswertung der Daten aus dem PPBEO wurden pro Modell 2,2 Effekte gefunden.

In Abbildung 1a und b ist die Verteilung der OR-Werte, für die das Konfidenzintervall die 1 nicht einschloß, in den betrachteten Modellsituationen A und B sowie für die realen Daten (PPBEO) dargestellt. Abbildung 1c zeigt den Vergleich für die Modellsituation A und die PPBEO-Daten.

‘Signifikante’ OR-Schätzer bis zu Werten um 3,0 treten auch unter Zufallsbedingungen (Situation A und B) auf. Die Abhängigkeit der Einflußvariablen untereinander beeinflußt die Höhe der beobachteten ("signifikanten") OR-Schätzer. (s. Abb. 1a)

Im Gesamtvergleich unterschieden sich die OR-Schätzer allerdings weniger in der Höhe als in der Form ihrer Verteilung. Für die realen Daten fanden sich seltener signifikante OR-Schätzungen kleiner als 1, was durch die gezielte Auswahl der Variablen (hier: Risikofaktoren) erklärbar ist. Andererseits liegen OR-Werte um 2 durchaus im Bereich der in den Modellsituationen A und B gefundenen Verteilung.

Abb. 1a: **Verteilung der signifikanten ORs (Modellsituationen A und B)**

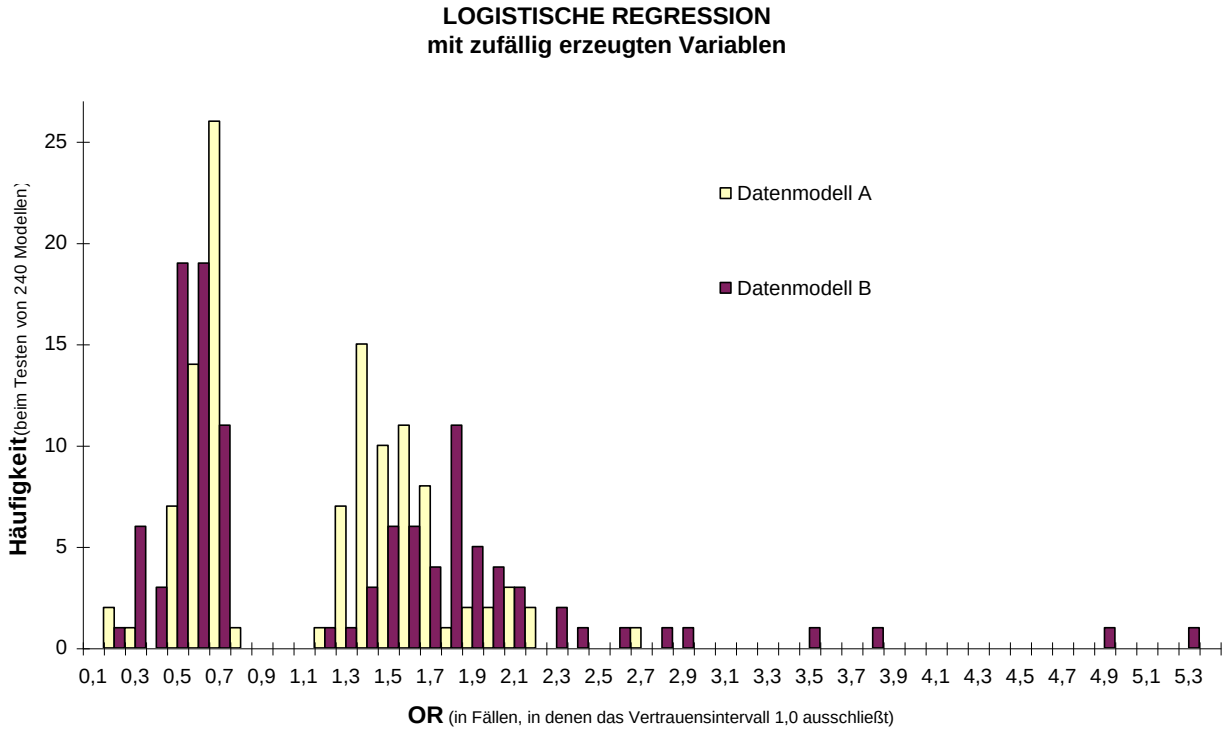


Abb. 1b: **Verteilung der signifikanten ORs (PPBEO-Daten)**

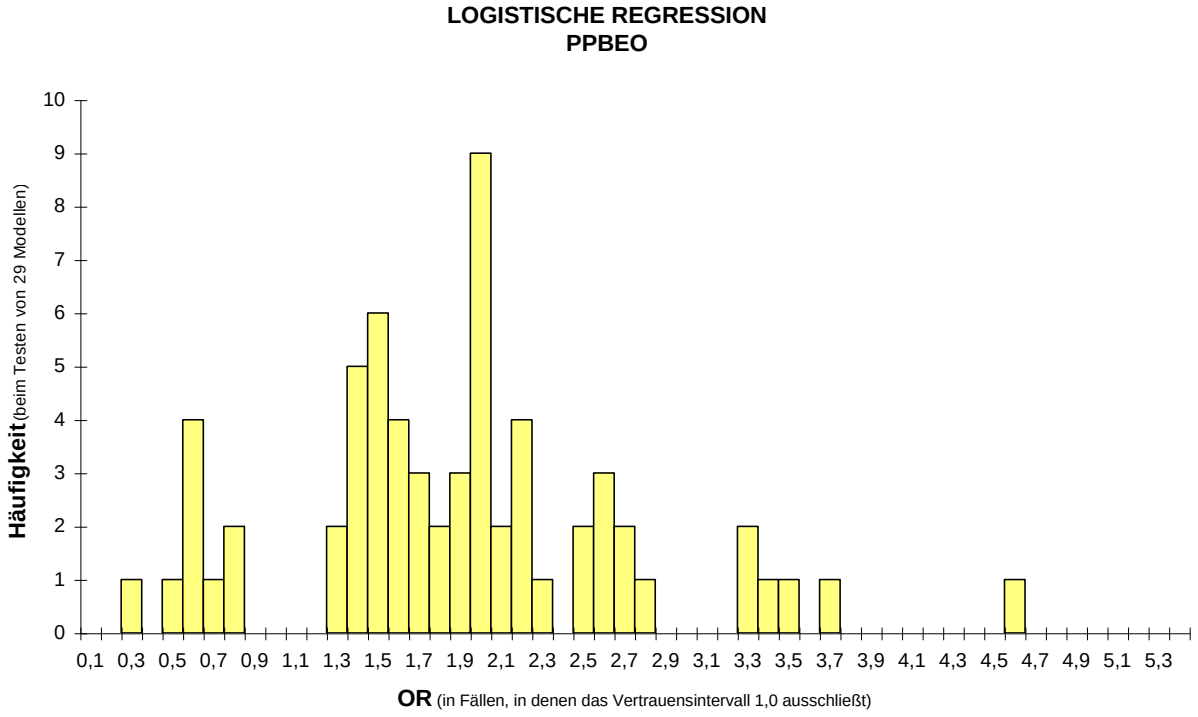
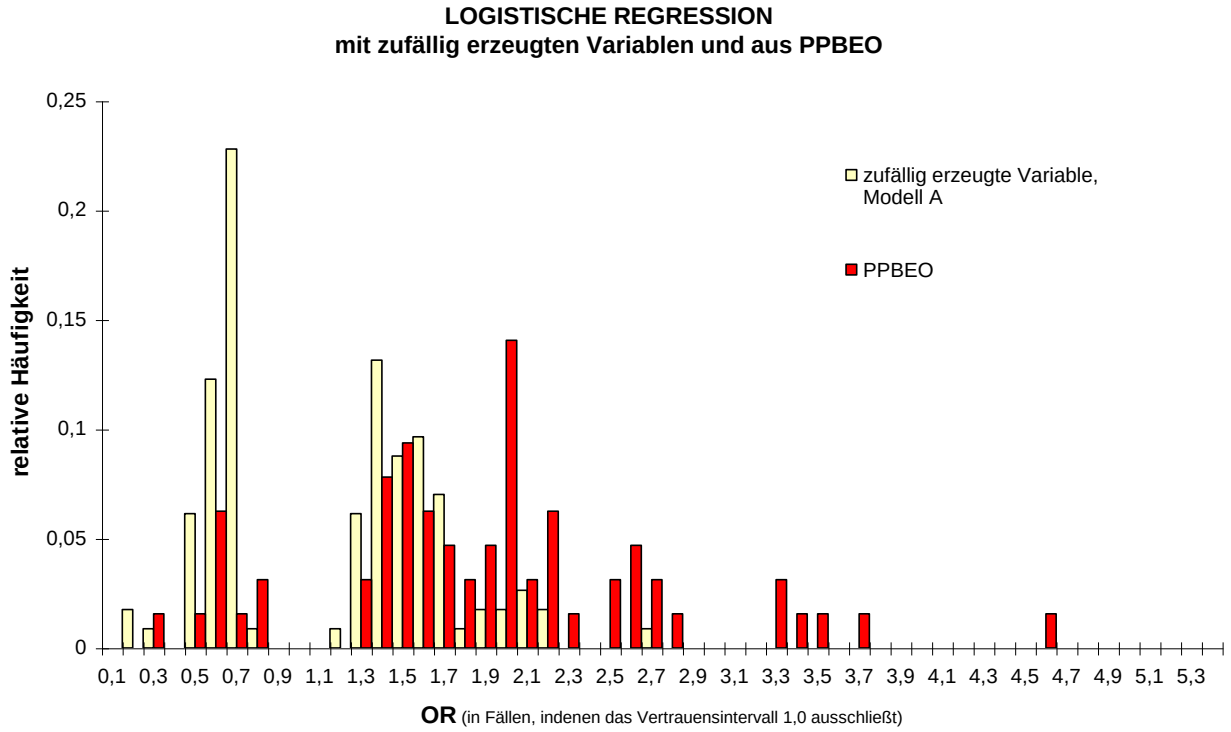


Abb. 1c: **Verteilung der signifikanten ORs (Modellsituation A und PPBEO-Daten)**



Umrechnung von Odds Ratio (OR) zu relativem Risiko (RR)

Die Odds Ratio (OR) ist nur dann eine gute Schätzung für das relative Risiko (RR), wenn die Inzidenz oder Prävalenz einer Erkrankung/eines Symptoms in der Studienpopulation niedrig ist (unter 10 %). Je größer die Auftretenswahrscheinlichkeit der Erkrankung, desto stärker überschätzt (falls OR >1) oder unterschätzt (falls OR <1) die OR das relative Risiko.

Da einige der in dem PPBEO-Projekt untersuchten Zielgrößen auch Prävalenzen über 10 % aufweisen, wurden aus den (signifikanten) OR-Schätzungen mit Hilfe der von ZHANG und KAI (JAMA 1998) angegebenen Umrechnungsformel die entsprechenden RR-Werte und KI bestimmt. In Abb. 2a sind für die PPBEO-Daten die "signifikanten" OR gegen die Inzidenz bei den bezüglich des jeweiligen Einflußfaktors Nicht-Exponierten aufgetragen. Für die Daten aus dem Zufallsmodell A wurden die "signifikanten" OR gegen die Inzidenz (Auftretenswahrscheinlichkeit) der Zielvariablen aufgetragen.Gleichzeitig enthält die Grafik die Linien gleichen relativen Risikos, die sich aus der Umrechnungsformel zwischen OR und RR von ZHANG & KAI (JAMA 1998) ergeben:

$$RR = \frac{OR}{(1 - P_0) + (P_0 \times OR)} \quad ,$$

wobei P_0 = Inzidenz unter den Nicht-Exponierten.

Abb. 2a: Signifikante ORs in Abhängigkeit von der Inzidenz bei den bezüglich des jeweiligen Einflußfaktors Nicht-Exponierten (PPBEO-Daten) und in Bezug zu Linien gleichen relativen Risikos. Jedes Quadrat entspricht einer OR-Schätzung, deren Konfidenzintervall 1 nicht überdeckt (einem gefundenen Effekt)

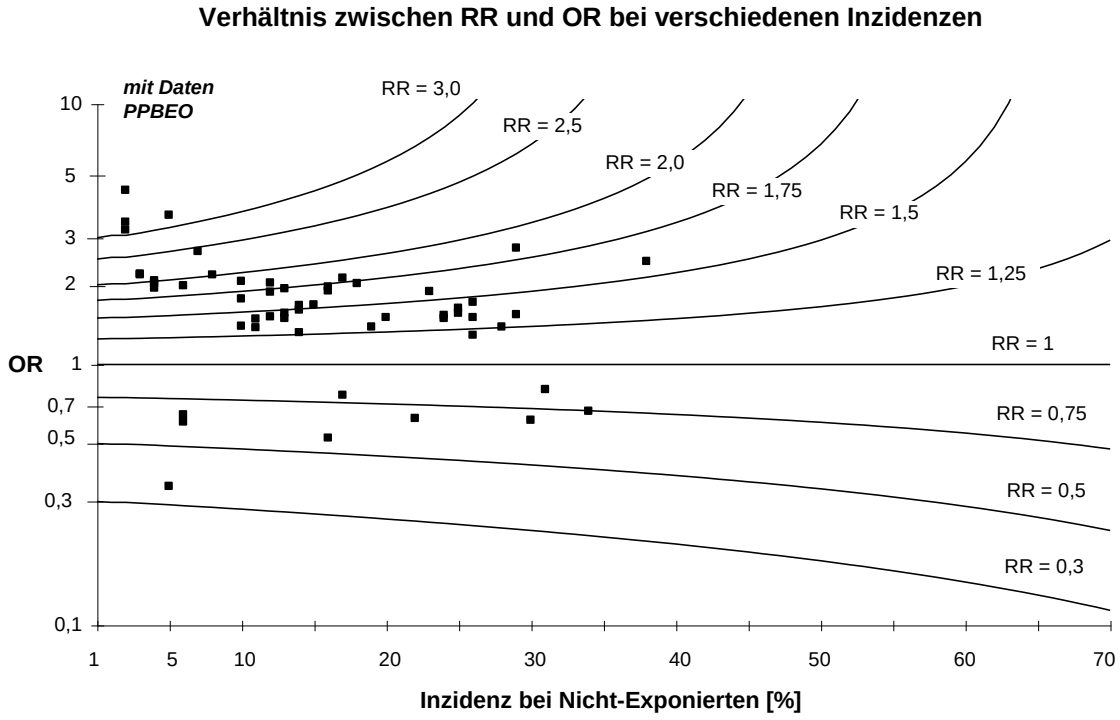
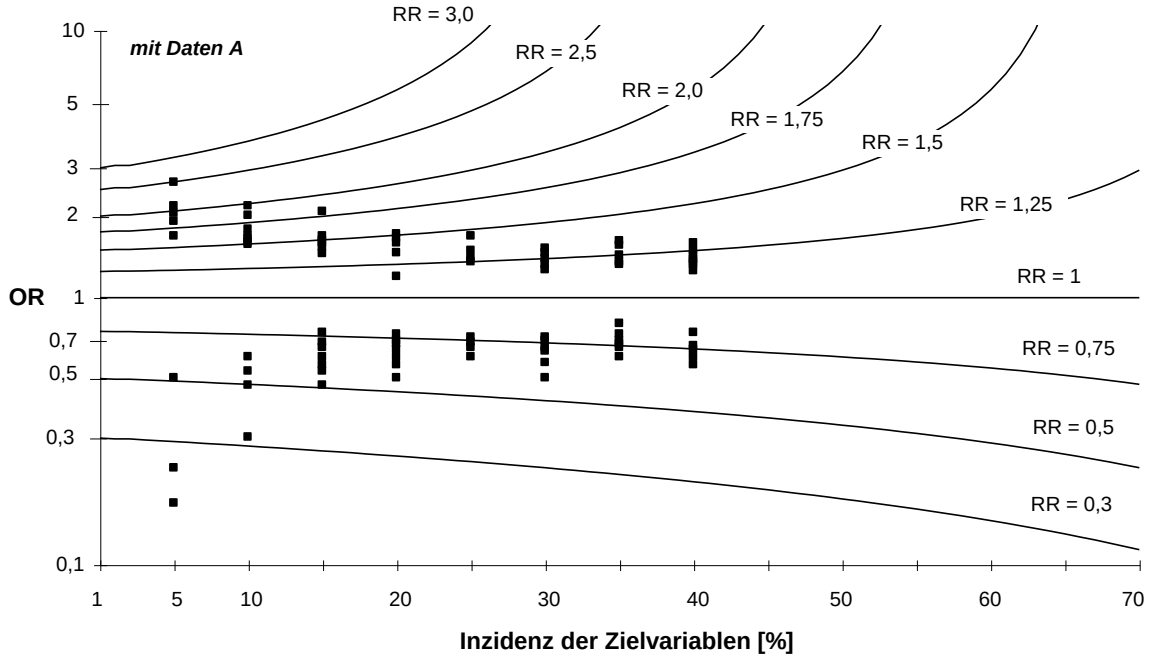


Abb. 2b: "Signifikante" ORs in Abhängigkeit von der Inzidenz der Zielvariablen (Zufällige Daten: Modell A) mit Bezug zu Linien gleichen relativen Risikos. Jedes Quadrat entspricht einer OR-Schätzung, deren Konfidenzintervall 1 nicht überdeckt.



Güte der Modellanpassung/Vorhersage

Im Rahmen der Untersuchungen wurden auch Parameter zur Güte der Anpassung des Gesamtmodells an die Daten betrachtet. Es zeigte sich, daß die ausgewählten Parameter für die Modellanpassung nicht einheitlich reagieren. Anhand von R^2 , c und -2 log L ließen sich die "Zufallsmodelle" am ehesten von den realen Daten trennen.

Schlußfolgerungen

Für die Praxis epidemiologischer Studien erscheint beachtenswert, daß ‘signifikante’ Odds Ratios bis zu Werten um 3,0 auch unter Zufallsbedingungen auftreten können. Auch der gefundene Effekt, daß die Abhängigkeit der Einflußvariablen untereinander die Höhe der beobachteten ("signifikanten") OR-Schätzungen beeinflußt, sollte bei der Anwendung logistischer Regressionsmodelle und der Interpretation ‘adjustierter’ ORs Berücksichtigung finden. Anstrebenswert wäre auch die Angabe eines (oder mehrerer) Maße zur Modellanpassung in Publikationen.

Literatur:

Breslow, N. E., Day, W. (1980): Statistical Methods in Cancer Research. Vol. 1-The Analysis of Case-Control Studies, Lyon: IARC Schientific Publication No. 32

Hosmer, D.W,Jr., Lemeshow, S. (1989): Applied Logistic Regression.
New York: John Wiley&Sons

Faus-Keßler, Th., Brüske-Hohlfeld, I., Scherb, H., Tritschler, J., Weigelt, E. (1992): Einführung in die arbeitsmedizinische Epidemiologie. Dortmund

Kreienbrock, L., Schach, S. (1995): Epidemiologische Methoden. Stuttgart Jena

Zhang, J., Kai, F. Y. (1998): What’s the Relative Risk?
JAMA, Vol. 280, 1690-1691