Nicklisch, Susanne

Potenziale der Datentransformation für die Strukturanpassung bei Odds-Ratio-Analysen

eingereicht als

DIPLOMARBEIT

an der

HOCHSCHULE MITTWEIDA (FH)

UNIVERSITY OF APPLIED SCIENCES

Fachbereich Mathematik

Chemnitz, 2009

Erstprüfer: Prof. Dr. rer. nat. Egbert Lindner
Zweitprüfer: Dr. rer. nat. Norman Bitterlich

Bibliographische Beschreibung:

Nicklisch, Susanne: Potenziale der Datentransformation für die Strukturanpassung bei Odds-Ratio-Analysen - 2009 - 63 S.
Hochschule Mittweida (FH), Fachbereich Mathematik, Diplomarbeit, 2009

Kurzreferat:

Ziel der Diplomarbeit ist es, Potenziale der Datentransformation für die Strukturanpassung bei Odds-Ratio-Analysen zu untersuchen. Es wird ein Verfahren erarbeitet, mit dessen Hilfe sich entscheiden lässt, inwieweit ein Datensatz für eine Modellierung durch die multiple logistische Regression geeignet ist. Dazu wird ein Maß eingeführt, das eine Bewertung über die Güte der Modellierung zulässt. Es ist zu untersuchen, ob durch eine nichtlineare Transformation eine Verbesserung dieser Maßzahl erreicht werden kann. Anschließend wird erläutert, wie transformierte Datensätze auf die Ausgangsdatensätze zurückgerechnet werden, wobei die ermittelten Risikoerhöhungen erhalten bleiben. Zum Schluss werden die Erkenntnisse auf ein Patientenkollektiv der arbeitsmedizinischen Untersuchung angewendet.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

DS	Datensatz
EXP	Exponentialfunktion
KI	Konfidenzintervall
LOG	natürlicher Logarithmus, Logarithmus zur Basis e (Eulersche Zahl)
LOGIT	Logit, Logarithmus eines Odds
MLR	multiple logistische Regression
oG	obere Grenze
OR	Odds Ratio
PJ	Packungsjahre
QMW	Quadratischer Mittelwert
RR	Risk Ratio
uG	untere Grenze
VFT	Vierfeldertafel

1 Einleitung

1.1 Motivation

Es ist in der Medizin üblich mithilfe von Odds-Ratio-Analysen und der Anwendung der multiplen logistischen Regression Patientenkollektive auf spezielle Risikofaktoren zu untersuchen. In dieser Arbeit wird die Odds-Ratio-Analyse für die Lungenkrebserkrankung in einem Patientenkollektiv der arbeitsmedizinischen Untersuchung[1] durchgeführt. Dabei werden die Einflussfaktoren *Alter* und *Rauchen* berücksichtigt. Wird der Parameter *Alter* in Subgruppen aufgeteilt, lässt sich unter Verwendung der Vierfeldertafeln zeigen, dass in jungen Jahren das Risiko an Lungenkrebs zu erkranken steigt und zunehmendes *Alter* risikohemmend erscheint. Im Falle einer globalen Untersuchung würde, aufgrund der Modellierung mithilfe der multiplen logistischen Regression, dieser Effekt ausgeglichen werden. So würde der Eindruck entstehen, dass das Risiko an Lungenkrebs zu erkranken insgesamt mit zunehmendem *Alter* sinkt. Die Ursachen dieses Problems liegen nicht bei den Daten, sondern bei der Modellanpassung der multiplen logistischen Regression. Deshalb sind Subgruppenanalysen erforderlich, die sich durch die Einflussfaktoren erstellen lassen[7].

Weiterhin kann auf der Basis der multiplen logistischen Regression keine sichere Risikobewertung durch die Einflussfaktoren vorgenommen werden, da die Änderungen im Datenmaterial durch Gruppierung und Skalierung Auswirkungen auf die Ergebnisse der Odds-Ratio-Analyse haben. Dazu kommt, dass die Aussagen abhängig von der Wahl der Subgruppen sind. All das führt zu einer Verschlechterung der statistischen Aussagekraft[6].

[1] medizinische Untersuchungen bezogen auf den Arbeitsplatz (z.B. Vorsorgeuntersuchungen, Eignungsuntersuchungen, Einstellungsuntersuchungen)

1.2 Aufgabe und Zielstellung

Im Rahmen dieser Arbeit soll ein Verfahren zur Bewertung der Modelladäquatheit eines Datensatzes entwickelt werden. Dazu sollen Datensätze mit einem bzw. zwei Parametern untersucht werden. Auf diese Weise kann das Verfahren später vom trivialen Fall auf den allgemeinen Fall übertragen werden. Eine Maßzahl soll eine Bewertung über die Güte der Modellierung zulassen, sodass eine Entscheidung über eine mögliche Verbesserung dieser Maßzahl durch eine Transformation der Daten getroffen werden kann. Es soll untersucht werden, ob eine lineare Transformation Auswirkungen auf die Modelladäquatheit des Datensatzes hat und ob eine nichtlineare Transformation zu deren Verbesserung beitragen kann. Abschließend sollen die transformierten Daten auf die Ausgangsdaten zurücktransformiert werden, ohne die erzielten Risikoerhöhungen zu verfälschen.

Zusammenfassend lässt sich die Aufgabe, unter Berücksichtigung aller genannten Kriterien, in folgende Teilaufgaben gliedern:

- Simulation von ein- bzw. zweiparametrischen Datensätzen
- Durchführung der multiplen logistischen Regression und der Berechnung der Vierfeldertafeln für die Datensätze
- Bewertung der Modelladäquatheit der Datensätze
- Transformation der Parameter der Datensätze
- Rücktransformation der Parameter der Datensätze

1.3 Gliederung der Arbeit

Die vorliegende Arbeit gliedert sich in sechs Kapitel. Angefangen mit Kapitel 2, werden zunächst die jeweiligen Themen der einzelnen Kapitel benannt.

Kapitel 2 Als wichtige Grundlage für diese Arbeit wird die Odds-Ratio-Analyse beschrieben. Dazu gehört der Unterschied zwischen Odds Ratio und Risk Ratio sowie die Darstellung beider Kennzahlen in Vierfeldertafeln. Anschließend wird der Zusammenhang zur multiplen logistischen Regression erläutert.

Kapitel 3 Dieses Kapitel dient der Simulation von unterschiedlichen ein- und zweiparametrischen Datensätzen für die anschließende Untersuchung.

Kapitel 4 Für die simulierten Datensätze wird die Anwendung der Odds-Ratio-Analyse sowie der multiplen logistischen Regression untersucht. So können Lösungsverfahren für die in Kapitel 1 benannten Probleme entwickelt werden.

Kapitel 5 Die gewonnenen Lösungsverfahren werden auf einen realen Datensatz, ein Patientenkollektiv der arbeitsmedizinischen Untersuchung, angewendet.

Kapitel 6 Das letzte Kapitel dient der Zusammenfassung der erreichten Ergebnisse und den weiterführenden Überlegungen zu dieser Arbeit.

2 Mathematische Grundlagen

2.1 Odds Ratio (OR) und Risk Ratio (RR)

In der Medizin findet die Analyse mittels OR Anwendung in retrospektiven[1] Fall-Kontroll-Studien[5]. Mithilfe einer solchen Studie wird eine Stichprobe bestehend aus erkrankten Personen (Fall) und gesunden Personen (Kontrolle) auf potenzielle Risikofaktoren untersucht.

Unter dem Begriff Odds bzw. Chance wird die Zahl der Ereignisse (Fälle) im Verhältnis zur Zahl der Nicht-Ereignisse (Kontrolle) innerhalb eines Kollektivs verstanden. In der folgenden Arbeit bedeutet Ereignis eine nachgewiesene BCa-Erkrankung[2] und Nicht-Ereignis eine Nicht-Erkrankung an Lungenkrebs[4].

Die Bezeichnung Chance könnte in der medizinischen Anwendung als ungeeignet für das Ereignis Erkrankung empfunden werden, deshalb wird in dieser Arbeit die englische Bezeichnung Odds verwendet.

Das Darstellen der Ereignisse und Nicht-Ereignisse erfolgt in einer Vierfeldertafel (VFT). In der abgebildeten VFT (Tabelle 2.1) bezeichnet die Gruppe A beispielsweise die Raucher und die Gruppe B die Nichtraucher in einem Kollektiv. Die Gruppe A ist folglich mit dem Risikofaktor *Rauchen* behaftet, der die Wahrscheinlichkeit für das Auftreten eines Bronchialkarzinoms beeinflusst. Bei einer Anzahl von $a + b$ Rauchern tritt bei a Rauchern das Ereignis einer nachgewiesenen BCa-Erkrankung ein.

[1] lat.: zurückblicken, Es wird ausgehend von der Gegenwart die Vergangenheit untersucht.
[2] auch: Bronchialkarzinom, Lungenkrebs

Es ergibt sich eine relative Häufigkeit

$$h(a)_{a+b} = \frac{a}{a+b},$$

die bei hinreichend großer Fallzahl der Auftretenswahrscheinlichkeit p entspricht. Folglich gibt die relative Häufigkeit

$$h(b)_{a+b} = \frac{b}{a+b}$$

das Eintreten des Nicht-Ereignisses bei einer Anzahl von b Rauchern an, was der Auftretenswahrscheinlichkeit $1-p$ entspricht. Nach der Definition des Odds bezeichnet $\frac{a}{b}$ das Verhältnis der Zahl der Ereignisse zur Zahl der Nicht-Ereignisse, das heißt, $\frac{p}{1-p}$.

Gruppe	**Krankheit**	
	ja	nein
A	a	b
B	c	d

Tabelle 2.1: Vierfeldertafel (VFT)

Das Odds Ratio (relative Chance) bewertet das Verhältnis zweier Gruppen und gibt eine Aussage über die Stärke des Zusammenhangs zwischen ihnen. Ein OR $= 1$ heißt, es gibt keinen Unterschied zwischen den beiden Gruppen, da das gleiche Quotenverhältnis vorliegt. Ein OR > 1 heißt, die Wahrscheinlichkeit an Lungenkrebs zu erkranken ist für Gruppe A größer als für Gruppe B . Entsprechend bedeutet ein OR < 1 für Gruppe A eine geringere Wahrscheinlichkeit an Lungenkrebs zu erkranken im Vergleich zu Gruppe B.

Der Begriff Risk bzw. Risiko beschreibt die relative Häufigkeit der nachgewiesenen BCa-Erkrankungen innerhalb eines Kollektivs ($\frac{a}{a+b}$). Wird das Verhältnis der Erkrankungswahrscheinlichkeit der risikobehafteten Gruppe zur Erkrankungswahrscheinlichkeit der nicht risikobehafteten Gruppe aufgestellt, so wird das Risk Ratio (relatives Risiko) genannt.

Die Tabelle 2.2 stellt die allgemeine Berechnungsvorschrift für Odds und Risk dar[1].

Gruppe	**erkrankt**	**nicht erkrankt**	**Anzahl**	**Risk**	**Odds**
A	p_A	n_A	$N_A = p_A + n_A$	$r_A = \frac{p_A}{N_A}$	$o_A = \frac{p_A}{n_A}$
B	p_B	n_B	$N_B = p_B + n_B$	$r_B = \frac{p_B}{N_B}$	$o_B = \frac{p_B}{n_B}$
	$p = p_A + p_B$	$n = n_A + n_B$	$N = N_A + N_B$ $= p + n$	RR $= \frac{r_A}{r_B}$	OR $= \frac{o_A}{o_B}$

Tabelle 2.2: Berechnungsvorschrift für Odds und Risk

Für die VFT in Tabelle 2.1 sieht die Berechnungsvorschrift folgendermaßen aus:

Berechnungsvorschrift[3]:
OR= $\frac{a}{b} : \frac{c}{d} \quad = \frac{ad}{bc}$ (OR für Gruppe B im Vergleich zu Gruppe A)
RR= $\frac{a}{a+b} : \frac{c}{c+d} = \frac{a \cdot (c+d)}{c \cdot (a+b)}$ (RR für Gruppe B im Vergleich zu Gruppe A)

Bei niedrigen Auftretenswahrscheinlichkeiten p liegen Odds und Risk in der gleichen Größenordnung, wobei das Risk stets etwas kleiner als das Odds ist.

Beispiel:
Bei Gruppe A sind von 100 Personen 20 erkrankt und 80 nicht erkrankt, dann beträgt das Risk 20% und das Odds 25% (Tabelle 2.3).

Gruppe	**erkrankt**	**nicht erkrankt**	**Anzahl**	**Risk**	**Odds**
A	$p_A = 20$	$n_A = 80$	$N_A = 100$	$\boldsymbol{r_A = 0,2}$	$\boldsymbol{o_A = 0,25}$
B	$p_B = 30$	$n_B = 70$	$N_B = 100$	$r_B = 0,3$	$o_B = 0,43$
	$p = 50$	$n = 150$	$N = 200$	RR $= 0,67$	OR $= 0,58$

Tabelle 2.3: Vergleich OR und RR

Ein Risk von 20% bedeutet, dass $\frac{1}{5}$ aller Patienten erkrankt ist, das heißt, es gibt viermal so viele Nicht-Erkrankte wie Erkrankte. Das Odds liegt bei 25%, also 1 : 4.

Das Risk kann nur Werte zwischen 0 und 1 (0% und 100%) annehmen, wobei das Odds wesentlich größer werden kann. Das ist besonders bei hohen Auftretenswahrscheinlichkeiten der Fall[8], wie in Abbildung 2.1 zu erkennen.

Abbildung 2.1: Odds und Risk in Abhängigkeit von der Auftretenswahrscheinlichkeit p

2.2 Das Modell der multiplen logistischen Regression (MLR)

Die Regressionsanalyse ist ein statistisches Verfahren zur Bestimmung eines Zusammenhangs zwischen einer stetigen Zielvariablen Y (abhängige Variable) und mehreren erklärenden Variablen $X_1, ..., X_m$ (unabhängige Variablen). Die einfache lineare Regression findet Anwendung, wenn nur eine erklärende Variable X vorliegt. Dann wird die folgende Geradengleichung verwendet.

$$Y = \alpha + \beta X$$

Im Fall mehrerer erklärender Variablen $X_1, ..., X_m$ wird von einer multiplen linearen Regression gesprochen, die durch folgende Gleichung beschrieben wird.

$$Y = \alpha + \beta_1 X_1 + ... + \beta_m X_m$$

Die multiple logistische Regression (MLR) kommt zum Einsatz, wenn die Zielvariable Y diskret (z.B. binär) ist. So könnte Y beschreiben, ob eine Person eine Krankheit hat oder nicht. Dann wird sie mit den beiden Werten 1 oder 0 codiert. Die Wahrscheinlichkeit p des Auftretens des Ereignisses Erkrankung wird durch $p = P(Y = 1)$ beschrieben[11].

Das Odds $\frac{p}{1-p}$ kann jeden beliebigen positiven Wert annehmen und für den Logarithmus des Odds, $c = \frac{p}{1-p}$, $\log(c) = logit(p)$ genannt, liegen die Werte in der Menge der reellen Zahlen[2]. Angenommen, es gibt eine lineare Beziehung zwischen den X_i mit $i = 1, ..., n$ und dem $logit(p)$, dann gilt:

$$\log(c) = \log(\tfrac{p}{1-p}) = \alpha + \sum_{i=1}^{n} \beta_i X_i$$

Die Parameter α und $\beta_1, ..., \beta_n$ der MLR lassen sich mit der Methode der Maximum-Likelihood-Schätzung ermitteln. Für die Berechnung der MLR wird in dieser Arbeit das Statistik-Programm PASW Statistics[3] verwendet.

[3] bis 2009 unter dem Namen SPSS bekannt

Für konkrete Daten mit den Einflussgrößen $X_1, ..., X_n$ kann anschließend die Auftretenswahrscheinlichkeit p und das Odds c berechnet werden:

$$p = \frac{c}{1+c} = \frac{\exp(\alpha + \sum_{i=1}^{n} \beta_i X_i)}{1+\exp(\alpha + \sum_{i=1}^{n} \beta_i X_i)}$$

Es wird der einfache Fall einer VFT mit der Dimension $n = 1$ betrachtet:

erklärende Variable X	**Krankheit** ja	nein	**Odds**	**OR**
1	p_1	$1 - p_1$	$o_A = \frac{p_1}{1-p_1}$	
0	p_0	$1 - p_0$	$o_B = \frac{p_0}{1-p_0}$	$\frac{o_A}{o_B} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$

Tabelle 2.4: Einfacher Fall der VFT

Für die einfache logistische Regression lassen sich aus der VFT folgende zwei Gleichungen aufstellen:

$$\begin{aligned} (1) \quad X = 0: \quad & \log(\tfrac{p_0}{1-p_0}) = \alpha + \beta \cdot 0 \\ (2) \quad X = 1: \quad & \log(\tfrac{p_1}{1-p_1}) = \alpha + \beta \cdot 1 \end{aligned}$$

Wird in Gleichung (2) das α durch $\alpha = \log(\frac{p_0}{1-p_0})$ aus Gleichung (1) ersetzt, ergibt sich nach dem Umstellen für β:

$$\beta = \log(\tfrac{p_1/(1-p_1)}{p_0/(1-p_0)})$$

Folglich kann aus dem Regressionskoeffizienten einer einfachen logistischen Regression das OR direkt durch:

$$\text{OR} = \exp(\beta)$$

berechnet werden.

Nun wird auf den allgemeinen Fall eingegangen. Bei der Veränderung der Einflussgröße um eine Einheit ergibt sich für $X_1^* = X_1 + 1$[1]:

$$\begin{aligned} c^* &= \exp(\alpha + \beta_1(X_1 + 1) + \sum_{i=1}^{n} \beta_i X_i) \\ &= \exp(\beta_1 + \alpha + \sum_{i=1}^{n} \beta_i X_i) \\ &= \exp(\beta_1) \cdot \exp(\alpha + \sum_{i=1}^{n} \beta_i X_i) \\ &= \exp(\beta_1) \cdot c \\ \rightarrow \frac{c^*}{c} &= \exp(\beta_1) \end{aligned}$$

Hier beschreibt $\exp(\beta_1)$ den Faktor, um den sich das Odds bei der Erhöhung des Parameters X_1 um eine Einheit erhöht. Daraus folgt, dass die Exponentialwerte der Koeffizienten, die sich bei der Berechnung der MLR ergeben, das OR für die Erhöhung der Werte $X_1, ..., X_n$ um jeweils eine gegebene Größe liefern.

Allgemein gilt dieser Zusammenhang für die Odds c^* bzw. c von zwei Tupeln $(X_1^*, ..., X_n^*)$ bzw. $(X_1, ..., X_n)$:

$$\begin{aligned} \frac{c^*}{c} &= \frac{\exp(\alpha + \sum_{i=1}^{n} \beta_i \cdot X_i^*)}{\exp(\alpha + \sum_{i=1}^{n} \beta_i \cdot X_i)} \\ &= \exp(\sum_{i=1}^{n} \beta_i (X_i^* - X_i)) \\ &= \prod_{i=1}^{n+1} (\exp(\beta_i))^{(X_i^* - X_i)} \end{aligned}$$

Die ORs lassen sich also aus den Differenzen und Koeffizienten in den Eingangsgrößen berechnen.

3 Simulation von Datensätzen

Die Simulation der Datensätze wird mit Microsoft Excel™ vorgenommen. Sie dient der Untersuchung von vier Fällen. Es wird vom trivialen Fall mit einem Parameter ausgegangen. Zuerst wird für diesen gezeigt, wie die Modellanalyse bei einem für die Modellierung mithilfe der MLR geeigneten Datensatz durchgeführt wird. Anschließend wird dieses Verfahren auf einen für die Modellierung mittels MLR weniger modelladäquaten Datensatz übertragen. Ist die Vorgehensweise für den einparametrischen Fall abgeschlossen, wird diese auf den zweiparametrischen Fall reproduziert. Zuerst wird erneut die Modellanalyse für einen modelladäquaten Datensatz mit zwei Parametern durchgeführt. In Analogie wird anschließend das Verfahren auf einen zweiparametrischen weniger modelladäquaten Datensatz übertragen.

Bei der Simulation von Datensätzen können verschiedene Faktoren dazu beitragen, dass sich nicht für jedes α und β ein für die Modellierung mithilfe der MLR geeigneter Datensatz simulieren lässt. Eine große Rolle spielt dabei der Umfang der Stichprobe. Im Rahmen dieser Arbeit wurden die einparametrischen Datensätze auch für eine Stichprobenanzahl von 5000 getestet, wobei sich bei den Resultaten wesentlich geringere Schwankungen herausstellten. Da eine kleinere Stichprobenanzahl zwar die Simulation erschwert, sich damit aber trotzdem wesentliche Ergebnisse darstellen lassen, wurden die nachfolgenden Simulationen für 500 Fälle durchgeführt. Zudem ist eine Stichprobenanzahl in dieser Größenordnung realitätsnäher.

3.1 Einparametrischer modelladäquater Datensatz - DS(1)

Die Simulation eines einparametrischen modelladäquaten Datensatzes soll über die Eingabe von α und β auf der Basis der MLR durchgeführt werden. Als erster Parameter wird eine stetige Größe, die zwischen 20 und 70 liegt, gewählt. Im Folgenden wird sie zur besseren Anschauung als *Alter* interpretiert. Anschließend wird das *Alter* auf ein Intervall zwischen 0 und 1 linear transformiert und aus diesem Grund als $Alter_t$ bezeichnet.

Über die Eingabe von α und β wird die Auftretenswahrscheinlichkeit p berechnet. Dazu muss zuerst der Logarithmus des Odds bestimmt werden.

$$\log(c) = \alpha + \beta \cdot Alter$$

Anschließend wird der Ausdruck nach dem Odds c umgeformt:

$$c = \exp(\alpha + \beta \cdot Alter)$$

Jetzt kann die Auftretenswahrscheinlichkeit p berechnet werden. Laut MLR gilt:

$$p = \frac{c}{1+c}$$

Abschließend wird eine eindeutige Gruppenzuordnung vorgenommen. Es werden zwei Gruppen gebildet, Gruppe 1 und Gruppe 2. Gruppe 1 wird wie in Abschnitt 2.1 (Odds Ratio (OR) und Risk Ratio (RR)) als nicht erkrankt und Gruppe 2 als erkrankt interpretiert. Die Einteilung wird durch einen Vergleich der Auftretenswahrscheinlichkeit p mit einer gleichverteilten Zufallszahl bestimmt. Es wird von Zufall gesprochen, wenn das Ereignis nicht vorhersehbar ist. Die Zahl wird also rein zufällig aus einer Menge von Zahlen im Intervall $[0; 1]$ herausgegriffen. Ist die Auftretenswahrscheinlichkeit p kleiner als diese Zufallszahl, wird die Gruppe 1 zugeordnet, andernfalls die Gruppe 2.

Es wird ein Ausschnitt für die Berechnung eines derart modellierten Datensatzes für $\alpha = -1,4$ und $\beta = 0,02$ dargestellt (Tabelle 3.1):

Alter	***Alter$_t$***	***log(c)***	***c***	***p***	**Zufallszahl**	**Gruppe**
$20,9$	$0,018$	$-0,982$	$0,375$	$0,272$	$0,045$	2
$21,0$	$0,020$	$-0,980$	$0,375$	$0,273$	$0,719$	1
$21,1$	$0,022$	$-0,978$	$0,376$	$0,273$	$0,218$	2

Tabelle 3.1: Ausschnitt aus Berechnung für DS(1)

Es folgt die Berechnung der VFTs. Zuerst werden die VFTs für den Parameter *Alter* und die Gruppe erstellt. Der Wertebereich des Parameters *Alter* wird in Dekaden eingeteilt, sodass sich fünf Intervalle ergeben. Nun werden immer zwei nebeneinanderliegende Intervalle miteinander verglichen. Es ergibt sich, wie in Tabelle 2.2 (Berechnungsvorschrift für Odds und Risk) beschrieben, das OR für das Intervall mit den älteren Personen im Vergleich zu dem Intervall mit den jüngeren Personen.

Mithilfe des Statistik-Programms PASW Statistics lässt sich über die MLR für den erzeugten Datensatz ebenfalls ein OR ermitteln. Da die MLR eine gleichmäßige Struktur aufweist, beschreibt das OR die Risikoerhöhung je Einheit. Aus diesem Grund müssen die über die VFTs berechneten ORs alle in der gleichen Größenordnung liegen, das heißt, diese müssen sich möglichst wenig unterscheiden. Ist das der Fall, ist der vorliegende Datensatz für die Modellierung mithilfe der MLR geeignet.

Zum Schluss wird der Parameter *Alter$_t$* untersucht. Es werden die VFTs für *Alter$_t$* und Gruppe berechnet. Wegen der Linearität der Transformation ändert sich nichts an der Verteilung der Daten innerhalb des Intervalls $[0;1]$. Der Wertebereich des Parameters *Alter$_t$* wird ebenfalls in fünf Intervalle eingeteilt, wodurch sich eine Intervalllänge von $0,2$ ergibt. So stimmt die Anzahl der Personen in den Intervallen von *Alter* und *Alter$_t$* überein und bei der Berechnung der VFTs für das *Alter$_t$* und die Gruppe ergeben sich dieselben ORs wie bei den VFTs für das *Alter* und die Gruppe.

3.2 Einparametrischer weniger modelladäquater Datensatz - DS(2)

Es wird ein weniger modelladäquater Datensatz simuliert. Mithilfe dieses Datensatzes soll gezeigt werden, wie durch eine nichtlineare Transformation des Parameters ein modelladäquater Datensatz entsteht.

Es wird wieder der Parameter *Alter* gewählt, der im Intervall zwischen 20 und 70 liegt. Anschließend wird das *Alter* nichtlinear auf das Intervall $[0;1]$ transformiert, $Alter_t$ genannt. Die nichtlineare Transformation hat zur Folge, dass sich die Anzahl der Personen in den Intervallen bei gleichbleibender Intervalllänge ändert.

Das *Alter* könnte beispielsweise folgendermaßen transformiert werden:

$$Alter : X_1 \rightarrow (\tfrac{X_1-20}{50})^2 \in [0;1]$$

Es werden in diesem Fall für den transformierten Datensatz analog zum einparametrischen modelladäquaten Datensatz die Risikoerhöhungen aus den VFTs für das $Alter_t$ und die Gruppe berechnet. Wenn sich die vier ORs nur geringfügig unterscheiden, kann der transformierte Datensatz durch die MLR modelliert werden, das heißt, der transformierte Datensatz ist modelladäquat. Wird anschließend der Parameter $Alter_t$ auf den Parameter *Alter* zurücktransformiert, kann davon ausgegangen werden, dass der Ausgangsdatensatz nicht modelladäquat ist.

Ausschnitt aus der Berechnung eines Datensatzes für $\alpha = -2,1$ und $\beta = 4,0$ (Tabelle 3.2):

$\boldsymbol{Alter}$	$\boldsymbol{Alter_t}$	$\boldsymbol{log(c)}$	$\boldsymbol{c}$	$\boldsymbol{p}$	**Zufallszahl**	**Gruppe**
$57,9$	$0,575$	$0,198$	$1,219$	$0,549$	$0,812$	1
$58,0$	$0,578$	$0,210$	$1,234$	$0,552$	$0,627$	1
$58,1$	$0,581$	$0,223$	$1,249$	$0,555$	$0,047$	2

Tabelle 3.2: Ausschnitt aus Berechnung für DS(2)

In Abschnitt 4.2.2 (Modell(2) zu DS(2)) wird beschrieben, wie die Rückrechnung eines transformierten Datensatzes auf den Ausgangsdatensatz erfolgt.

3.3 Zweiparametrischer modelladäquater Datensatz - DS(3)

Bei einem zweiparametrischen Datensatz ist die Gruppenzuordnung von zwei Parametern abhängig. Es soll hier über die Eingabe von α, β_1 und β_2 auf Basis der MLR ein modelladäquater Datensatz simuliert werden. Der erste Parameter bleibt das *Alter*. Dieser wird anschließend linear auf das Intervall zwischen 0 und 1 transformiert und mit $Alter_t$ bezeichnet. Es kommt eine zweite stetige Größe zwischen 0 und 60 hinzu. Sie wird als Packungsjahre (PJ) interpretiert.

Unter dem Begriff Packungsjahre wird eine Einheit verstanden, die die inhalierte Rauch-Dosis eines Zigaretten-Rauchers beschreibt. Die Anzahl der Packungsjahre ergibt sich aus der Zahl der täglich konsumierten Zigarettenpackungen multipliziert mit der Zahl der Raucherjahre[10].

Wie das *Alter*, werden die PJ linear auf das Intervall $[0; 1]$ transformiert. Aus diesem Grund wird es analog zum $Alter_t$ mit PJ_t bezeichnet.

Über die Eingabe von α, β_1 und β_2 wird die Auftretenswahrscheinlichkeit p berechnet. Dazu wird als Erstes der Logarithmus des Odds bestimmt:

$$log(c) = \alpha + \beta_1 \cdot Alter + \beta_2 \cdot PJ$$

Nun wird die Gleichung nach dem Odds c umgeformt:

$$c = \exp(\alpha + \beta_1 \cdot Alter + \beta_2 \cdot PJ)$$

Anschließend kann die Auftretenswahrscheinlichkeit p berechnet werden:

$$p = \frac{c}{1+c}$$

Nun kann durch einen Vergleich zwischen der Auftretenswahrscheinlichkeit p und einer Zufallszahl die Gruppenzuordnung vorgenommen werden. Ist die Auftretenswahrscheinlichkeit kleiner als die gleichverteilte Zufallszahl, wird Gruppe 1 zugeordnet, andernfalls Gruppe 2.
Die Berechnung erfolgt wieder für 500 Fälle, sodass sich ein Kollektiv mit 500 Personen ergibt.

In der nachfolgenden Tabelle 3.3 wird ein Ausschnitt aus der Berechnung eines derart modellierten Datensatzes für $\alpha = -2,0$, $\beta_1 = 0,03$ und $\beta_2 = 0,02$ dargestellt.

$Alter$	$Alter_t$	PJ	PJ_t	$log(c)$	c	p	Zufallszahl	Gruppe
30,8	0,216	17	0,283	−0,736	0,479	0,324	0,112	2
30,9	0,218	48	0,800	−0,113	0,893	0,472	0,129	2
31,0	0,220	59	0,983	0,110	1,116	0,527	0,996	1

Tabelle 3.3: Ausschnitt aus Berechnung für DS(3)

Es folgt die Berechnung der VFTs für das *Alter*, die PJ und die Gruppe. Anders als bei einem einparametrischen Datensatz werden bei den zweiparametrischen Datensätzen über die VFTs zunächst keine ORs berechnet, sondern die Odds für je ein Intervall des Wertebereiches des ersten Parameters und ein Intervall des Wertebereiches des zweiten Parameters dargestellt. Der Wertebereich des Parameters *Alter* wird in vier gleichlange Intervalle der Länge 12,5 und der Wertebereich der PJ in drei gleichlange Intervalle der Länge 20 eingeteilt. Nun wird, angefangen bei den beiden kleinsten Intervallen für *Alter* und PJ und die entsprechende Gruppe die Anzahl der sich darin befindlichen Personen bestimmt und anschließend die Odds berechnet. Die zwölf Odds können in einem dreidimensionalen Säulendiagramm dargestellt werden (Abbildung 3.1).

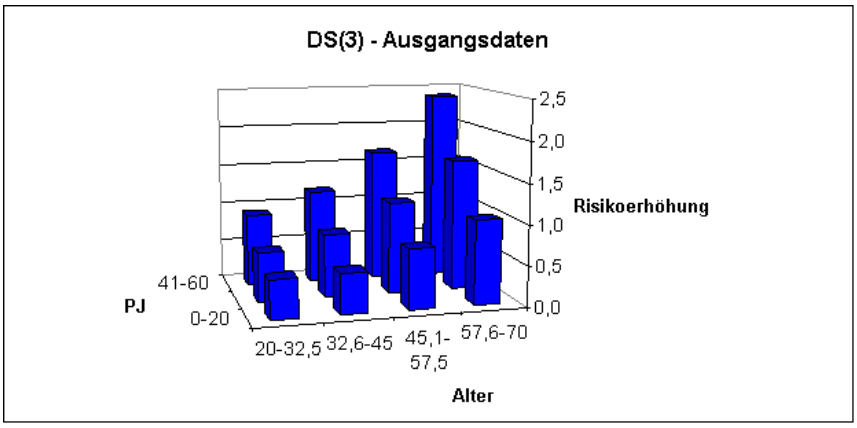

Abbildung 3.1: Säulendiagramm zur Darstellung der Odds

Ein Datensatz, der durch die MLR modelliert wurde, hat in einem Säulendiagramm eine charakteristische Form. Die Säulen steigen gleichmäßig an. Wird jede Säulenreihe für sich betrachtet, ergibt sich eine Form annähernd dem Kurvenverlauf der Exponentialfunktion.

Anschließend wird jedes Odds durch das Odds der beiden kleinsten Intervalle für *Alter* und PJ dividiert. So lassen sich die Risikoerhöhungen bezogen auf das kleinste Odds bestimmen. In Abschnitt 4.4.1 (Modell(3) zu DS(3)) wird beschrieben, wie sich mithilfe des Statistik-Programms PASW Statistics die Risikoerhöhungen bezogen auf die beiden kleinsten Intervalle für *Alter* und PJ berechnen lassen. Diese werden anschließend mit den Risikoerhöhungen aus den berechneten Odds verglichen. Stimmen die Risikoerhöhungen überein, liegt ein modelladäquater Datensatz vor.

Als Letztes werden über die VFTs für $Alter_t$, PJ_t und Gruppe die Odds für die transformierten Parameter berechnet. Wie bei einem einparametrischen modelladäquaten Datensatz bleibt durch die Linearität der Transformation der beiden Parameter die Verteilung der Daten innerhalb der Intervalle erhalten. Der Wertebereich des Parameters $Alter_t$ wird wie der Wertebereich des Parameters *Alter* in vier gleichlange Intervalle eingeteilt, die hier die Länge $0,25$ besitzen. Der Wertebereich des Parameters PJ_t wird wie der Wertebereich der PJ in drei gleichlange Intervalle mit der Intervalllänge $0,33$ aufgeteilt. So stimmt die Anzahl der Personen in den Intervallen der Parameter und der transformierten Parameter überein. Bei der Berechnung der Odds über die VFTs für $Alter_t$, PJ_t und Gruppe ergeben sich dieselben Odds wie bei den VFTs für *Alter*, PJ und Gruppe.

3.4 Zweiparametrischer weniger modelladäquater Datensatz - DS(4)

Bei einem realen zweiparametrischen Datensatz kann sich herausstellen, dass dieser weniger modelladäquat für die Modellierung mithilfe der MLR ist. Es soll gezeigt werden, dass durch die Transformation der beiden Parameter ein modelladäquater Datensatz erzeugt wird. Dabei müssen gegebenenfalls beide Parameter nichtlinear transformiert werden.

Für die Simulation eines solchen Datensatzes erfolgt die Vorgehensweise analog zu einem einparametrischen weniger modelladäquaten Datensatz. Die Parameter des Datensatzes werden beliebig nichtlinear transformiert. Anschließend werden α, β_1 und β_2 für den transformierten Datensatz so gewählt, dass dieser modelladäquat ist. Der transformierte Datensatz wird auf den Ausgangsdatensatz zurücktransformiert. Es entsteht ein Datensatz, der weniger modelladäquat ist, da die Gruppenzuordnung auf den transformierten Datensatz zurückzuführen ist.

Wie bei einem zweiparametrischen modelladäquaten Datensatz wird als erster Parameter das *Alter* gewählt, welches zwischen 20 und 70 liegt. Das *Alter* wird nichtlinear auf das Intervall $[0; 1]$ transformiert, $Alter_t$ genannt. Als zweiter Parameter kommen die PJ im Intervall von 0 bis 60 hinzu. Dieser Parameter wird nichtlinear auf das Intervall zwischen 0 und 1 transformiert und als PJ_t bezeichnet. Beispielsweise könnten *Alter* und PJ wie folgt transformiert werden:

$$\begin{aligned} Alter : X_1 &\rightarrow (\tfrac{X_1-20}{50})^2 \in [0;1] \\ PJ : \quad X_2 &\rightarrow \sqrt{\tfrac{X_2}{60}} \qquad \in [0;1] \end{aligned}$$

So ändert sich wie bei einem einparametrischen Datensatz in beiden Fällen die Verteilung der Daten und somit die Anzahl der Personen in den Intervallen bei gleichbleibender Intervalllänge von $0,25$ ($Alter_t$) und $0,33$ (PJ_t).

Nach der Festlegung der Transformationen werden mithilfe der VFTs analog zum zweiparametrischen modelladäquaten Datensatz die Odds für $Alter_t$, PJ_t und Gruppe berechnet.

Weisen die zwölf Odds die in Abschnitt 3.3 (Zweiparametrischer modelladäquater Datensatz - DS(3)) beschriebene charakteristische Form im dreidimensionalen Säulendiagramm auf, können die Risikoerhöhungen bezogen auf die beiden kleinsten Intervalle für $Alter_t$ und PJ_t berechnet werden.

Wird bei dem transformierten Datensatz mit dem Statistik-Programm PASW Statistics die Modellierung mittels MLR durchgeführt, werden erneut β_1 und β_2 ermittelt. So lassen sich die Risikoerhöhungen bezogen auf die beiden kleinsten Intervalle für $Alter_t$ und PJ_t berechnen und mit denen aus den VFTs für $Alter_t$, PJ_t und die Gruppe vergleichen. Mithilfe einer Maßzahl lässt sich anschließend bestimmen, ob der transformierte Datensatz modelladäquat ist. In diesem Fall können die Parameter $Alter_t$ und PJ_t zurücktransformiert werden und es kann davon ausgegangen werden, dass die Modellierung des Ausgangsdatensatzes weniger modelladäquat ist.

Ausschnitt aus der Berechnung für einen Datensatz für $\alpha = -2,7$, $\beta_1 = 0,04$ und $\beta_2 = 0,02$ (Tabelle 3.4):

$\boldsymbol{Alter}$	$\boldsymbol{Alter_t}$	$\boldsymbol{PJ}$	$\boldsymbol{PJ_t}$	$\boldsymbol{log(c)}$	$\boldsymbol{c}$	$\boldsymbol{p}$	**Zufallszahl**	**Gruppe**
$44,3$	$0,236$	21	$0,592$	$-0,508$	$0,602$	$0,376$	$0,237$	2
$44,4$	$0,238$	2	$0,183$	$-0,884$	$0,413$	$0,292$	$0,118$	2
$44,5$	$0,240$	32	$0,730$	$-0,280$	$0,756$	$0,430$	$0,655$	1

Tabelle 3.4: Ausschnitt aus Berechnung für DS(4)

In Abschnitt 4.4.2 (Modell(4) zu DS(4)) wird beschrieben, wie die Rückrechnung eines transformierten Datensatzes auf den Ausgangsdatensatz erfolgt.

4 Analyse von simulierten Datensätzen

4.1 Simulierte Datensätze mit einem Risikofaktor

Es werden zwei simulierte Datensätze auf die Anpassung mittels MLR untersucht. Das Kollektiv umfasst jeweils 500 Personen, für die angegeben ist, ob ein Ereignis eintritt oder nicht. Bei DS(1) tritt bei 195 Personen das Ereignis ein und bei 305 Personen tritt das Ereignis nicht ein. Bei DS(2) tritt bei 190 Personen das Ereignis ein und bei 310 Personen nicht. Eine stetige Größe für die Patienten liegt zwischen 20 und 70 und wird als *Alter* interpretiert.

Unter Anwendung des Statistik-Programms PASW Statistics lässt sich über die MLR das OR für jeden Datensatz berechnen. Für DS(1) beträgt das OR $1,013$ und für DS(2) liegt das OR bei $1,098$. Beide ORs gelten für den Abstand einer Einheit. Da die stetige Größe hier als *Alter* interpretiert wird, kann von dem Abstand eines Jahres gesprochen werden. Es werden beide Datensätze in Subgruppen eingeteilt, wobei immer zehn Jahre zusammengefasst werden. Anschließend werden jeweils zwei nebeneinander liegende Intervalle miteinander verglichen und das OR für Gruppe B, die Gruppe der älteren Personen, im Vergleich zu Gruppe A, die Gruppe der jüngeren Personen, ermittelt. Bei der Betrachtung der Intervalle von zehn Jahren, werden die durch die VFTs ermittelten ORs (siehe Tabelle 4.1) mit der zehnten Potenz der ORs aus der MLR verglichen, um Aussagen über die Güte der Modellierung der beiden Datensätze treffen zu können.

Altersgruppe	**DS(1)** Anzahl (A+B)	OR	**DS(2)** Anzahl (A+B)	OR
A: 20-30 B: 31-40	200	1,142	200	1,502
A: 31-40 B: 41-50	200	1,137	200	1,806
A: 41-50 B: 51-60	200	1,133	200	2,929
A: 51-60 B: 61-70	200	1,130	200	4,205
OR (für 10 Jahre)		1,138		2,547

Tabelle 4.1: Vergleich der ORs zweier Datensätze

Der Vergleich der ORs der VFTs mit den durch die MLR berechneten ORs zeigt, dass der DS(1) durch die MLR besser modelliert wird als der DS(2). Zur Veranschaulichung dient die Abbildung 4.1.

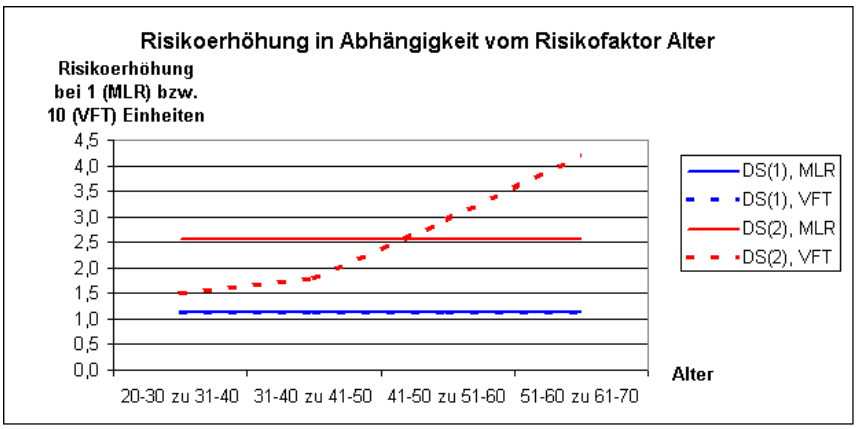

Abbildung 4.1: Vergleich der ORs zweier Datensätze

Der DS(1) wird durch die beiden blauen Kurven und der DS(2) durch die beiden roten Kurven dargestellt. Dabei bezeichnen die durchgehenden Kurven jeweils das mit der MLR berechnete OR. Die dick, gestrichelt eingezeichneten Kurven verdeutlichen die mittels VFTs berechneten ORs. Hierbei muss beachtet werden, dass die ORs der MLR jeweils für den Abstand eines Jahres berechnet wurden. Die ORs der VFTs wurden für Dekaden berechnet, weshalb nur Punkte abgebildet werden, die aber zur besseren Anschauung durch gestrichelte Linien verbunden wurden.

Die Kurven stellen die Risikoerhöhung in Abhängigkeit vom Risikofaktor dar.

Bei der Darstellung beider Berechnungsmöglichkeiten des OR des DS(1) ist wie in Tabelle 4.1 zu erkennen, dass sich das über die MLR berechnete OR durchgängig durch die Berechnung der in Dekaden eingeteilten ORs mittels VFTs erklären lässt. Beide blauen Kurven stimmen in etwa miteinander überein. Das bedeutet, dass das OR unabhängig von der Höhe des Risikofaktors ist. Interpretiert heißt das, ein 20-Jähriger besitzt die gleiche jährliche Risikoerhöhung wie ein 70-Jähriger.

Bei der Betrachtung der beiden roten Kurven fällt auf, dass es eine deutliche Abweichung der Kurven voneinander gibt. Die VFTs liefern keine Erklärung für das mittels MLR ermittelte OR. Das OR ist hier abhängig von der Lage des Risikofaktors. Eine Interpretation der dick, gestrichelt eingezeichneten roten Kurve kann hier sein, dass ein 20-Jähriger eine auffallend geringere Risikoerhöhung pro Lebensjahr aufweist als ein 70-Jähriger.

4.2 Modellberechnung

Im Folgenden wird zum DS(1) ein Modell(1) beschrieben. Das Modell(1) beinhaltet zum einen den rechnerischen Nachweis für die Eignung der Modellierung des Datensatzes durch die MLR und zum anderen wird gezeigt, dass eine lineare Transformation des Datensatzes keine Auswirkung auf die Modellierung des Datensatzes mithilfe der MLR hat. In Analogie zum Modell(1) wird ein Modell(2) auf den DS(2) angewendet. Dabei wird verdeutlicht, dass die Modellierung des DS(2) durch die MLR weniger modelladäquat ist. Der DS(2) wird durch eine vorgegebene nichtlineare Transformation in einen Datensatz überführt, für den sich die erneute Modellierung durch die MLR als geeignet erweist. Anschließend gilt es, die Ergebnisse des transformierten Datensatzes auf den Ursprungsdatensatz zurückzuführen.

4.2.1 Modell(1) zu DS(1)

Das Verfahren wird zur Übersichtlichkeit in fünf Schritte unterteilt.

Schritt 1
Für den DS(1) wird die MLR durchgeführt. Dabei wird ein Faktor der Risikoerhöhung für den Abstand Δ eines Jahres berechnet:

$$\beta = 1,013 \qquad \Delta_\beta = 1 \text{ Jahr}$$

Schritt 2
Der DS(1) wird in Dekaden eingeteilt, die sich mithilfe der VFTs vergleichen lassen. Auf diese Weise lässt sich ebenfalls ein Faktor für die Risikoerhöhung berechnen, welcher hier mit β^* bezeichnet wird. Durch die Einteilung in Dekaden gilt β^* für den Abstand von zehn Jahren. Hier wird der Mittelwert für die in Tabelle 4.1 aufgeführten ORs des DS(1) verwendet.

$$\beta^* = 1,135 \qquad \Delta_{\beta^*} = 10 \text{ Jahre}$$

Um β und β^* zueinander in Beziehung setzen zu können, werden beide für den Abstand von zehn Jahren betrachtet.

Dazu wird die zehnte Potenz von β gebildet.

$$\beta^* \approx \beta^{10}$$
$$1,135 \approx 1,138$$

Der theoretisch erwartete Zusammenhang zwischen β^* der VFTs und β der MLR wird bestätigt. Die MLR lässt sich durch die VFTs erklären, weshalb sie als geeignete Modellierung des DS(1) bezeichnet werden kann.

Schritt 3
Der Risikofaktor X_1, das *Alter*, wird durch eine lineare Transformation auf das Intervall $[0;1]$ abgebildet. Die Transformation ergibt sich durch die beiden Grenzen, wobei die untere bei 20 und die obere bei 70 liegt. Die Differenz beider Grenzen hat einen Wert von 50.

$$X_1 \rightarrow (\tfrac{X_1-20}{50}) \in [0;1]$$

Für den transformierten DS(1) wird erneut die MLR durchgeführt. Dabei berechnet sich ein Faktor der Risikoerhöhung für den Abstand einer Einheit, welcher aufgrund der Transformation im Folgenden mit β_t bezeichnet wird.

$$\beta_t = 1,883 \qquad \Delta_{\beta_t} = 1$$

Schritt 4
Mithilfe der VFTs wird für den transformierten DS(1) β_t^* berechnet. Dafür werden jeweils Intervalle der Länge $0,2$ gebildet, die sich durch die erneute Einteilung in fünf gleichlange Intervalle ergeben.

$$\beta_t^* = 1,135 \qquad \Delta_{\beta_t^*} = 0,2$$

Hier zeigt sich, dass β_t^* des transformierten DS(1) mit β^* des ursprünglichen DS(1) übereinstimmt.

$$\beta_t^* = \beta^*$$

Bei dem Vergleich von β_t^* mit β_t ergibt sich der gleiche Zusammenhang wie beim ursprünglichen DS(1).

$$\beta_t^* \approx \beta_t^{0,2}$$
$$1,135 \approx 1,135$$

Beide Werte wurden auf drei Nachkommastellen gerundet, weshalb die minimalen Abweichungen nicht ersichtlich sind. Die Werte sind sehr genau, obwohl bei einer Potenz von zehn Rundungsfehler auftreten.
Es wird gezeigt, wie vielen Jahren im ursprünglichen DS(1) eine Einheit im transformierten DS(1) entspricht.

$$\Delta_{\beta_t} \mathrel{\widehat{=}} 50 \cdot \Delta_\beta$$

Daraus ergibt sich für β_t:

$$\beta_t \approx \beta^{50}$$

Schritt 5
Nun lässt sich β_t des transformierten DS(1) auf β^* des ursprünglichen DS(1) zurückführen. Es ergibt sich die fünfte Wurzel aus der Intervalllänge des ursprünglichen DS(1) von zehn Jahren.

$$\sqrt[5]{\beta_t} \approx \beta^*$$

4.2.2 Modell(2) zu DS(2)

Schritt 1
Mithilfe des Statistik-Programms PASW Statistics wird für den DS(2) die MLR durchgeführt. Dabei wird ein Faktor der Risikoerhöhung für den Abstand Δ eines Jahres berechnet:

$$\beta = 1,098 \qquad \Delta_\beta = 1 \text{ Jahr}$$

Schritt 2

Der DS(2) wird wieder in Dekaden eingeteilt, die sich mithilfe der VFTs vergleichen lassen. Auf diese Weise lässt sich ebenfalls ein Faktor für die Risikoerhöhung berechnen, welcher hier mit β^* bezeichnet wird. Durch die Einteilung in Dekaden gilt β^* für den Abstand von zehn Jahren. Allerdings kann in diesem Fall nicht der Mittelwert aus den in Tabelle 4.1 aufgeführten ORs für den DS(2) gebildet werden, da die Werte der ORs monoton ansteigen. Zur Veranschaulichung dient die Abbildung 4.2.

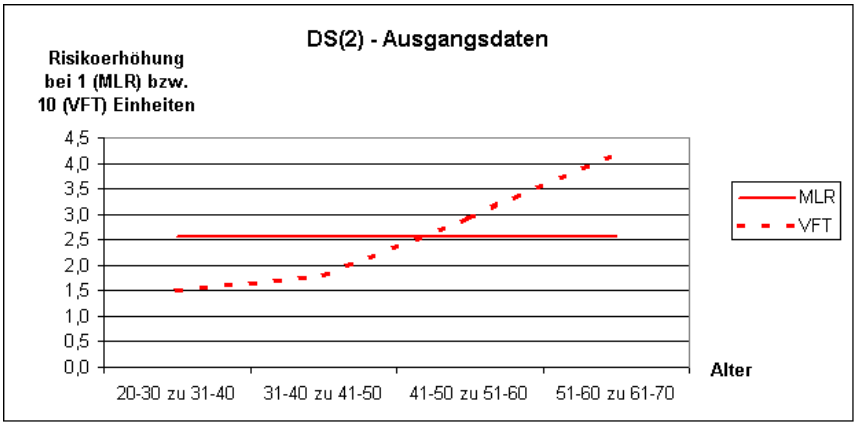

Abbildung 4.2: ORs für DS(2)

$$\Delta_{\beta^*} = 10 \text{ Jahre}$$

In diesem Fall ist es wenig sinnvoll, β^* und β zueinander in Beziehung zu setzen, da es sich bei der Abbildung von β^* um einen nichtlinearen Kurvenverlauf handelt im Gegensatz zur Darstellung von β. Es ist deutlich erkennbar, dass kein unmittelbarer Zusammenhang zwischen dem nichtlinearen Kurvenverlauf der ORs der VFTs und dem β der MLR besteht. Die MLR lässt sich nicht durch die VFTs erklären, weshalb sie als weniger modelladäquate Modellierung des DS(2) bezeichnet werden kann.

Schritt 3

Das *Alter* wird durch die vorgegebene nichtlineare Transformation auf das Intervall $[0; 1]$ abgebildet.

$$X_1 \rightarrow (\tfrac{X_1 - 20}{50})^2 \in [0; 1]$$

Für den transformierten DS(2) wird erneut die MLR durchgeführt. Dabei berechnet sich ein Faktor der Risikoerhöhung für den Abstand einer Einheit, welcher aufgrund der Transformation im Folgenden mit β_t bezeichnet wird.

$$\beta_t = 83,610 \qquad \Delta_{\beta_t} = 1$$

Schritt 4
Mithilfe der VFTs wird für den transformierten DS(2) β_t^* berechnet. Dafür werden jeweils Intervalle der Länge 0,2 gebildet, die sich durch die erneute Einteilung in fünf gleichlange Intervalle ergeben. Jetzt lässt sich aus den ermittelten ORs ein Mittelwert bilden. Die Abbildung 4.3 zeigt, dass durch die Transformation die deutlichen Abweichungen vom OR der MLR reduziert wurden. Die MLR ist geeignet für die Modellierung der durch die Transformation entstandenen Daten und lässt sich nun durch die VFTs erklären.

$$\beta_t^* = 2,428 \qquad \Delta_{\beta_t^*} = 0,2$$

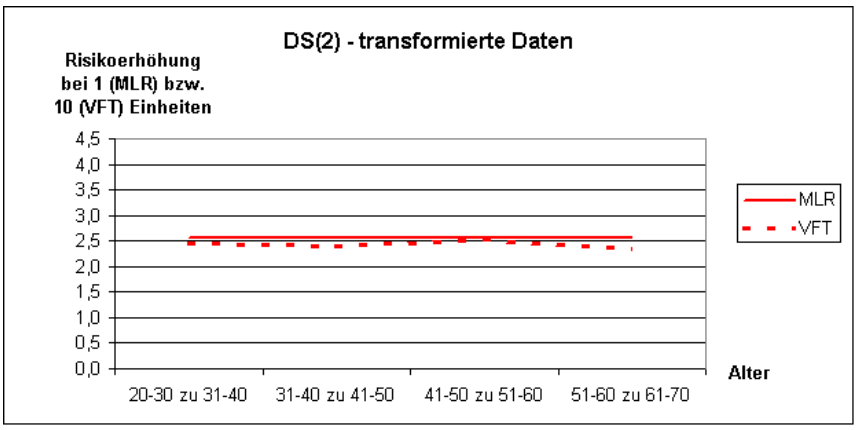

Abbildung 4.3: ORs für transformierten DS(2)

Schritt 5
Im Folgenden wird am Beispiel des OR der ersten beiden Intervalle [20-30] und [31-40] die Rückrechnung auf die Ausgangsdaten erläutert. Mithilfe der VFTs wurde für die beiden Intervalle ein OR berechnet:

$$\beta^*_{[20-30]-[31-40]} \approx 1,502$$

Nun gilt es, über die MLR den gleichen Wert zu erhalten. Dafür wird jeweils der quadratische Mittelwert (QMW) für beide Intervalle genutzt (siehe Tabelle 4.2).

$$QMW = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}$$

Intervall	$[20-30]$	$[31-40]$
QMW	$25,661$	$35,616$

Tabelle 4.2: Quadratischer Mittelwert für DS(2)

Für die MLR gilt folgende Ausgangsgleichung:

$$Y = \log(c) = \alpha + \beta X$$

Wird diese Gleichung nach dem Odds c umgeformt, folgt:

$$c = \exp(\alpha + \beta X)$$

Mithilfe der MLR ergeben sich demnach für beide Werte folgende Darstellungen:

$$c_{25,661} = \exp(\alpha + \beta_t(\tfrac{25,661-20}{50})^2)$$

$$c_{35,616} = \exp(\alpha + \beta_t(\tfrac{35,616-20}{50})^2)$$

Um das OR zu ermitteln wird anschließend die Differenz beider Exponentialfunktionen gebildet.

$$c_{35,616-25,661} = \exp(\beta_t((\tfrac{15,616}{50})^2 - (\tfrac{5,661}{50})^2))$$

Nun lässt sich β_t des transformierten DS(2) auf β^* des ursprünglichen DS(2) zurückführen. Es ergibt sich für β^*:

$$c_{35,616-25,661} = \exp(4,426(\tfrac{211,813}{2500})) \approx 1,455 = \beta^*_{[20-30]-[31-40]}$$

In der nachfolgenden Tabelle 4.3 sind die Berechnungen aller ORs, die mithilfe der VFTs ermittelt wurden, aufgeführt:

Altersgruppe	**VFT**	$\boldsymbol{\beta^*}$
A: 20-30 B: 31-40	$1,502$	$1,455$
A: 31-40 B: 41-50	$1,806$	$2,070$
A: 41-50 B: 51-60	$2,929$	$2,948$
A: 51-60 B: 61-70	$4,205$	$4,199$

Tabelle 4.3: Vergleich der ORs aus den VFT mit den ORs der MLR

Um ein Maß für die Abweichungen zwischen den β^* der VFTs und den β, die sich über die MLR berechnen lassen, festzulegen, wird die mittlere quadratische Abweichung s^2 über die folgende Formel eingeführt:

$$s^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left|X_i - \widehat{X_i}\right|^2}{\max(\widehat{X_i})}$$

X_i sind die Werte der β^* der VFTs und $\widehat{X}$ sind die β aus der Berechnung der MLR. So ergibt sich für $n = 4$ im obigen Beispiel für die mittlere quadratische Abweichung Folgendes:

$$s^2 = 0,003$$

Die Abweichung ist minimal, das heißt die MLR lässt sich durch die VFTs erklären.

4.3 Simulierte Datensätze mit zwei Risikofaktoren

Es werden die beiden simulierten Datensätze mit zwei Risikofaktoren auf die Anpassung mittels MLR untersucht. Der DS(3) umfasst ein Kollektiv von 500 Personen, für die erneut festgelegt ist, ob ein Ereignis eintritt oder nicht. Die beiden Risikofaktoren sind stetige Größen und unabhängig voneinander. Ein Risikofaktor liegt zwischen 20 und 70 und wird zur besseren Anschauung als *Alter* interpretiert. Die Werte des anderen Risikofaktors liegen im Intervall von 0 bis 60 und werden als PJ interpretiert.

Bei DS(3) tritt bei 248 Personen das Ereignis ein und bei 252 Personen tritt das Ereignis nicht ein. In DS(4) tritt bei 214 Personen das Ereignis ein und bei 286 Personen nicht.

4.4 Modellberechnung

In diesem Abschnitt wird zum DS(3) ein Modell(3) beschrieben. Dieses Modell soll zeigen, dass die Vorgehensweise bei einem einparametrischen weniger modelladäquaten DS(1) auch für einen Datensatz mit zwei Parametern zutrifft. Die Bestätigung hierfür liefert der rechnerische Nachweis. Zudem wird der DS(3) genau wie der einparametrische DS(1) durch eine lineare Transformation verändert, um zu zeigen, dass dies bei einem Datensatz mit zwei Risikofaktoren keine Auswirkungen auf das Ergebnis hat. Das Modell(4) beschreibt anschließend einen zweiparametrischen weniger modelladäquaten Datensatz, der sich mithilfe einer nichtlinearen Transformation in einen modelladäquaten Datensatz umformen lässt.

4.4.1 Modell(3) zu DS(3)

Die Modellanalyse für den zweiparametrischen modelladäquaten DS(3) wird wie bei den vorhergehenden Modellen zur Übersichtlichkeit in fünf Schritte unterteilt.

Schritt 1
Unter Anwendung des Statistik-Programms PASW Statistics lassen sich über die MLR für den DS(3) Faktoren der Risikoerhöhung für den Abstand Δ eines Jahres und die Konstanten B_A und B_P jeweils für das *Alter* und die PJ berechnen:

$$\begin{array}{llll} \textit{Alter}: \beta_A = 1,026 & B_A = 0,025 & \Delta_{\beta_A} = 1 \text{ Jahr} \\ PJ: \quad \beta_P = 1,020 & B_P = 0,019 & \Delta_{\beta_P} = 1 \text{ Jahr} \end{array}$$

Schritt 2
Die Wertebereiche der beiden Parameter des DS(3) werden in Intervalle eingeteilt. Die Länge der Intervalle des Parameters *Alter* beträgt $12,5$ und die des Parameters PJ 20. Mithilfe der VFTs werden nun die Odds für jedes Altersintervall und PJ-Intervall gebildet.

In der folgenden Tabelle 4.4 sind die Odds für den DS(3) aufgeführt.

$\boldsymbol{Alter/PJ}$	$\mathbf{0-20}$	$\mathbf{21-40}$	$\mathbf{41-60}$
$\mathbf{20-32,5}$	$0,483$	$0,609$	$0,917$
$\mathbf{32,6-45}$	$0,500$	$0,783$	$1,182$
$\mathbf{45,1-57,5}$	$0,750$	$1,150$	$1,667$
$\mathbf{57,6-70}$	$1,048$	$1,643$	$2,385$

Tabelle 4.4: VFT: Odds für DS(3)

Anschließend werden die Odds grafisch dargestellt. In der Abbildung 4.4 ist erkennbar, dass sich der DS(3) für eine Modellierung mittels MLR eignen könnte, da sich bei der Betrachtung jeder einzelnen Säulenreihe, ein Anstieg vergleichbar mit einer Exponentialfunktion ausmachen lässt.

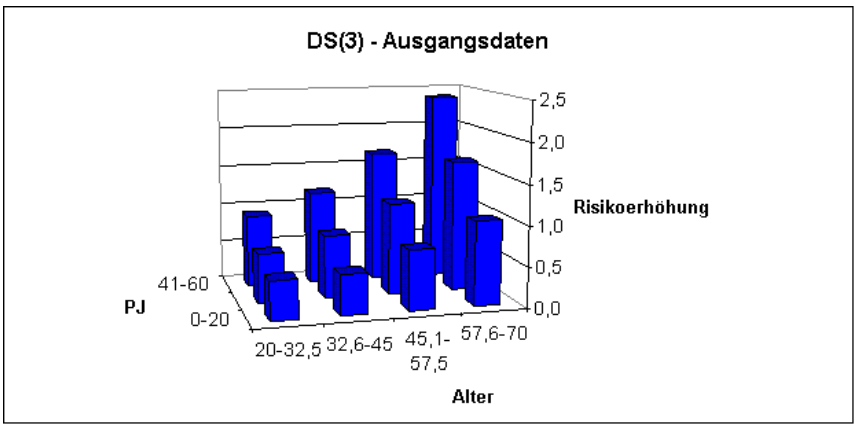

Abbildung 4.4: VFT: Odds für DS(3)

Im Gegensatz zu einem einparametrischen Datensatz können hier keine ORs von zwei nebeneinanderliegenden Gruppen berechnet werden. Jedes Odds hat mindestens drei benachbarte Odds, wie in Abbildung 4.4 erkennbar, sodass die Darstellung der Risikoerhöhungen in einer zweidimensionalen Abbildung nicht möglich ist. Aus diesem Grund werden bei einem zweiparametrischen Datensatz die Werte aus den VFTs mit denen der MLR unter der Betrachtung der Risikoerhöhung bezüglich eines Referenzpunktes verglichen. Das kann beispielsweise das kleinste Odds sein.

So berechnet sich folgende Tabelle 4.5.

$\boldsymbol{Alter/PJ}$	$\mathbf{0-20}$	$\mathbf{21-40}$	$\mathbf{41-60}$
$\mathbf{20-32,5}$	$1,000$	$1,261$	$1,899$
$\mathbf{32,6-45}$	$1,036$	$1,621$	$2,448$
$\mathbf{45,1-57,5}$	$1,554$	$2,382$	$3,452$
$\mathbf{57,6-70}$	$2,170$	$3,403$	$4,940$

Tabelle 4.5: VFT: Risikoerhöhung bezogen auf Wertepaar $(20-32,5;0-20)$ für DS(3)

Nun werden die berechneten Daten der Tabelle 4.5 erneut grafisch dargestellt. Erwartungsgemäß bleibt die ansteigende Form der Abbildung erhalten.

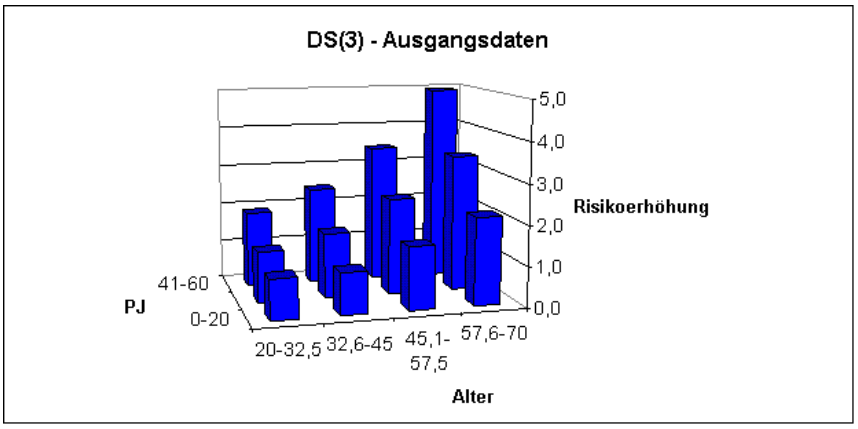

Abbildung 4.5: VFT: Risikoerhöhung bezogen auf Wertepaar $(20-32,5;0-20)$ für DS(3)

Um die Risikoerhöhungen aus den VFTs mit denen der MLR vergleichen zu können, werden nun die Daten aus der MLR in die gleiche Form wie die aus den VFTs gebracht. Es wird für jedes Intervall der Wertebereiche der Parameter *Alter* und PJ das arithmetische Mittel gebildet.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Damit die Intervalle des Wertebereiches der beiden Parameter äquidistant sind und ganze Zahlen enthalten, werden die Intervallgrenzen mit Kommastelle gerundet.

So ergeben sich folgende Werte:

Alter	$\bar{X}$	***PJ***	$\bar{X}$
$[20-32]$	26	$[0-20]$	10
$[33-45]$	39	$[21-40]$	31
$[46-58]$	52	$[41-60]$	51
$[59-70]$	64		

Tabelle 4.6: Arithmetisches Mittel für DS(3)

Über die MLR lässt sich nun ein Faktor der Risikoerhöhung für jedes mittlere *Alter* und mittlere PJ berechnen. Für das B_A und B_P werden die aus dem Statistik-Programm PASW Statistics berechneten Konstanten eingesetzt.

$$F(A,P) = \exp(B_A \cdot (\bar{X}_{1(A)} - \bar{X}_{0(A)}) + B_P \cdot (\bar{X}_{1(P)} - \bar{X}_{0(P)}))$$

Wird diese Berechnung für jedes arithmetische Mittel (immer bezogen auf das kleinste arithmetische Mittel $\bar{X}_0$) durchgeführt, ergibt sich für die MLR folgende Tabelle:

Alter/PJ	**10**	**31**	**51**
26	$1,000$	$1,490$	$2,179$
39	$1,384$	$2,063$	$3,016$
52	$1,916$	$2,855$	$4,175$
64	$2,586$	$3,854$	$5,635$

Tabelle 4.7: MLR: Risikoerhöhung bezogen auf Wertepaar $(26; 10)$ für DS(3)

Um an dieser Stelle ein Maß für die Abweichungen zwischen den Daten, die sich durch die VFTs ermitteln lassen, und den Daten, die sich über die MLR berechnen lassen, festzulegen, wird wie bei einem einparametrischen Datensatz die mittlere quadratische Abweichung s^2 für n Odds über die folgende Formel bestimmt:

$$s^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left|X_i - \widehat{X_i}\right|^2}{\max(\widehat{X_i})}$$

In diesem Fall bezeichnet X_i die Werte der VFTs und $\widehat{X}$ die Werte aus der Berechnung der MLR.

So ergibt sich für $n = 12$ im obigen Beispiel für die mittlere quadratische Abweichung Folgendes:

$$s^2 = 0,037$$

Nun können für den Parameter *Alter* ab 26 und den Parameter PJ ab 10 flächendeckend Risikoerhöhungen bezogen auf den Referenzpunkt (26;10) berechnet werden. Es entsteht dabei ein Oberflächendiagramm (Abbildung 4.6), in dem die zwölf Balken aus der Berechnung der VFTs (Abbildung 4.5) eingefügt sind. Sie sind durch geringe Abweichungen in dem Oberflächendiagramm erkennbar. Das heißt, die MLR lässt sich wie bei einem einparametrischen Datensatz durch die VFTs erklären, weshalb sie in diesem Fall als geeignete Modellierung des DS(3) bezeichnet werden kann.

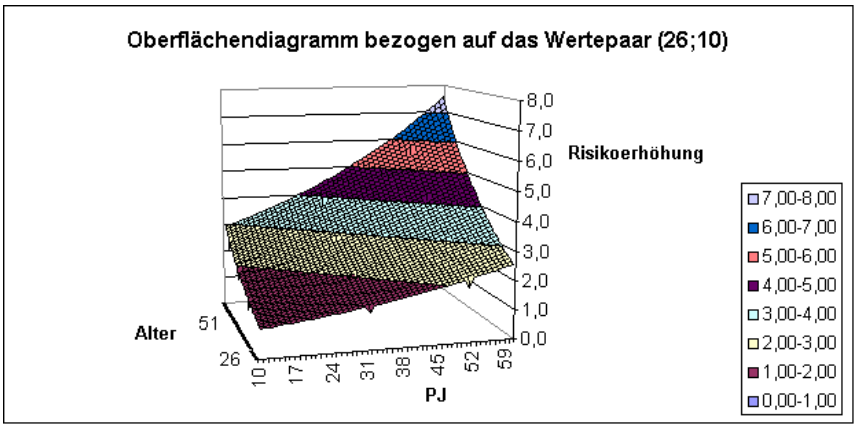

Abbildung 4.6: Oberflächendiagramm: Risikoerhöhung bezogen auf Wertepaar $(26; 10)$ für DS(3)

Es ist gezeigt, dass die flächendeckende Berechnung der Risikoerhöhungen auf der MLR beruht. Wird der Bezugspunkt auf (20;0) gesetzt, beziehen sich alle Risikoerhöhungen auf einen 20-jährigen Nichtraucher. Da der neue Bezugspunkt niedriger ist als der Bezugspunkt (26;10), werden die Risikoerhöhungen höher sein als in der vorigen Abbildung 4.6. In der Abbildung 4.7 lassen sich nun Risikoerhöhungen für jedes *Alter* zwischen 20 und 70 und jedes PJ zwischen 0 und 60 eindeutig ablesen. Eine detailliertere Abbildung befindet sich im Anhang A.

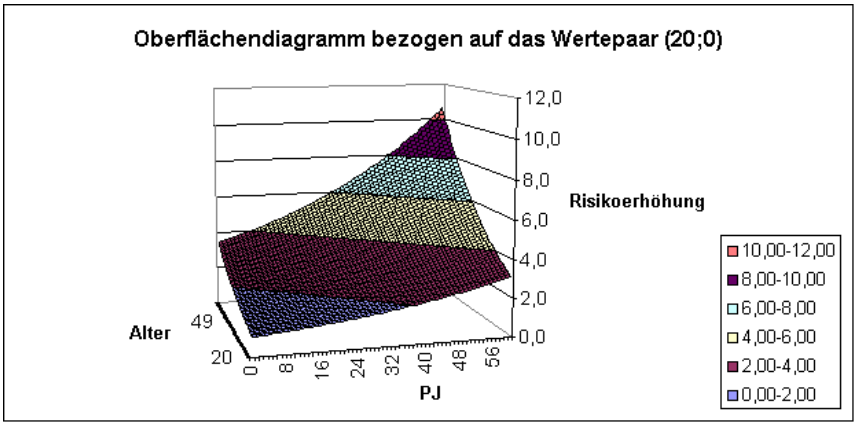

Abbildung 4.7: Oberflächendiagramm: Risikoerhöhung bezogen auf Wertepaar $(20; 0)$ für DS(3)

Schritt 3

Beide Risikofaktoren *Alter* (X_1) und PJ (X_2) werden durch eine lineare Transformation jeweils auf das Intervall $[0; 1]$ abgebildet. Wie im einparametrischen Fall ergeben sich die Grenzen aufgrund der Intervalle. Das *Alter* liegt zwischen 20 und 70 und die PJ zwischen 0 und 60.

$$\begin{aligned} Alter : X_1 &\rightarrow (\tfrac{X_1-20}{50}) \in [0;1] \\ PJ : \quad X_2 &\rightarrow (\tfrac{X_2}{60}) \quad\;\; \in [0;1] \end{aligned}$$

Für den transformierten DS(3) wird noch einmal die MLR berechnet. Dabei werden neue Faktoren der Risikoerhöhung für den Abstand Δ eines Jahres und neue Konstanten $B_{t(A)}$ und $B_{t(P)}$ jeweils für die Parameter $Alter_t$ und PJ_t berechnet. Aufgrund der Transformation werden die Bezeichnungen zusätzlich mit einem t indiziert:

$$\begin{aligned} Alter_t : \beta_{t(A)} &= 3,543 \qquad B_{t(A)} = 1,265 \qquad \Delta_{\beta_{t(A)}} = 1 \text{ Jahr} \\ PJ_t : \quad \beta_{t(P)} &= 3,208 \qquad B_{t(P)} = 1,166 \qquad \Delta_{\beta_{t(P)}} = 1 \text{ Jahr} \end{aligned}$$

Schritt 4

Es werden für den transformierten DS(3) die Odds über die VFTs berechnet.

Dafür werden für die Wertebereiche des Parameters $Alter_t$ Intervalle der Länge $0,25$ und des Parameters PJ_t Intervalle der Länge $0,33$ gebildet. Die Intervalllängen ergeben sich durch die erneute Einteilung in vier bzw. drei gleichlange Intervalle.

Es gilt:

$$\begin{aligned} Alter &: \Delta_{\beta_{t(A)}} \mathrel{\widehat{=}} 50 \cdot \Delta_{\beta(A)} \\ PJ &: \quad \Delta_{\beta_{t(P)}} \mathrel{\widehat{=}} 60 \cdot \Delta_{\beta(P)} \end{aligned}$$

Es ergibt sich für β_t:

$$\begin{aligned} Alter_t &: \beta_{t(A)} \approx \beta_A^{50} \\ &\qquad 3,543 \approx 3,609 \\ PJ_t &: \quad \beta_{t(P)} \approx \beta_P^{60} \\ &\qquad 3,208 \approx 3,281 \end{aligned}$$

Die Abweichungen zwischen den Faktoren der Risikoerhöhung für den Abstand Δ eines Jahres sind gering, obwohl Rundungsfehler aufgrund der hohen Potenzen auftreten.

4.4.2 Modell(4) zu DS(4)

Schritt 1
Mithilfe des Statistik-Programms PASW Statistics lassen sich über die MLR für den DS(4) Faktoren der Risikoerhöhung für den Abstand Δ eines Jahres und die Konstanten B_A und B_P für das $Alter$ und die PJ ermitteln:

$$\begin{aligned} Alter &: \beta_A = 1,036 \qquad B_A = 0,036 \qquad \Delta_{\beta_A} = 1 \text{ Jahr} \\ PJ &: \quad \beta_P = 1,025 \qquad B_P = 0,025 \qquad \Delta_{\beta_P} = 1 \text{ Jahr} \end{aligned}$$

Schritt 2
Es werden für beide Parameter $Alter$ und PJ die Wertebereiche in Intervalle eingeteilt. Für das $Alter$ beträgt die Intervalllänge $12,5$ bei vier gleichlangen Intervallen und für die PJ ist die Intervalllänge 20 bei drei gleichlangen Intervallen.

Über die VFTs werden nun die Odds für jedes Altersintervall und PJ-Intervall berechnet.

$\boldsymbol{Alter/PJ}$	$\mathbf{0-20}$	$\mathbf{21-40}$	$\mathbf{41-60}$
$\mathbf{20-32,5}$	$0,200$	$0,500$	$0,548$
$\mathbf{32,6-45}$	$0,407$	$0,720$	$0,833$
$\mathbf{45,1-57,5}$	$0,500$	$0,789$	$1,150$
$\mathbf{57,6-70}$	$0,828$	$1,545$	$3,778$

Tabelle 4.8: VFT: Odds für DS(4)

In der Abbildung 4.8 wird die Tabelle 4.8 grafisch dargestellt. Hier lässt sich erkennen, dass der DS(4) eine gleichmäßig ansteigende Struktur aufweist. Bei der Betrachtung jeder einzelnen Säulenreihe ist eine Form vergleichbar mit dem Kurvenverlauf einer Exponentialfunktion erkennbar. Allerdings lässt der steile Anstieg des Odds für das Alter im Intervall $[57,6; 70]$ und die PJ im Intervall $[41; 60]$ erahnen, dass der DS(4) für die Modellierung mittels MLR weniger geeignet ist.

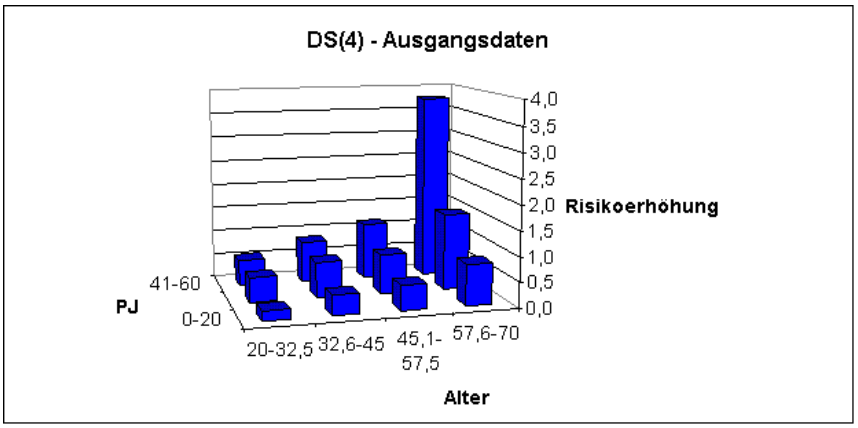

Abbildung 4.8: VFT: Odds für DS(4)

Um dies zu bestätigen werden die Odds der VFTs in Bezug auf einen Referenzpunkt, beispielsweise das kleinste Odds, betrachtet.

So lassen sich die Risikoerhöhungen bezogen auf das kleinste Odds angeben, um sie dann mit denen der MLR vergleichen zu können (siehe Tabelle 4.9).

$\boldsymbol{Alter/PJ}$	$\mathbf{0-20}$	$\mathbf{21-40}$	$\mathbf{41-60}$
$\mathbf{20-32,5}$	$1,000$	$2,500$	$2,742$
$\mathbf{32,6-45}$	$2,037$	$3,600$	$4,167$
$\mathbf{45,1-57,5}$	$2,500$	$3,947$	$5,750$
$\mathbf{57,6-70}$	$4,138$	$7,727$	$18,889$

Tabelle 4.9: VFT: Risikoerhöhung bezogen auf Wertepaar $(20-32,5;0-20)$ für DS(4)

Es werden die in der Tabelle berechneten Werte grafisch dargestellt.

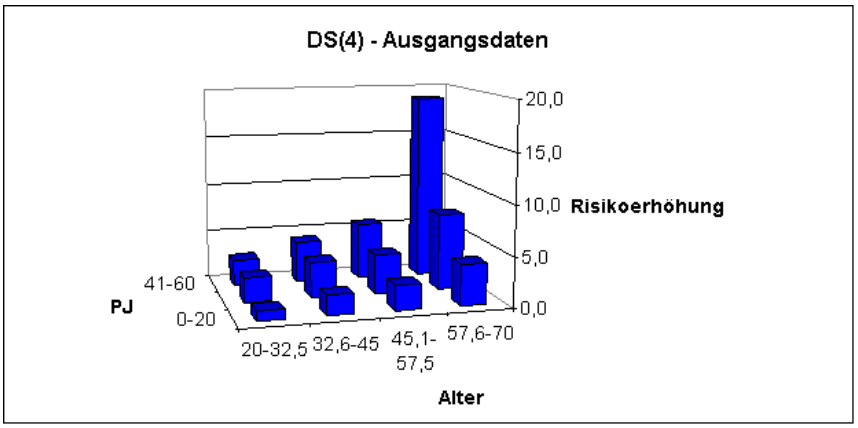

Abbildung 4.9: VFT: Risikoerhöhung bezogen auf Wertepaar $(20-32,5;0-20)$ für DS(4)

Damit die Risikoerhöhungen aus den VFTs mit denen der MLR verglichen werden können, werden die Daten aus der MLR in die gleiche Form gebracht. Wie bei einem zweiparametrischen modelladäquaten Datensatz werden deshalb für die Intervalle der Parameter *Alter* und PJ der arithmetische Mittelwert gebildet. Damit die Intervalle der Wertebereiche der beiden Parameter äquidistant sind und ganze Zahlen enthalten, werden die Intervallgrenzen mit Kommastelle gerundet.

Die Tabelle 4.10 zeigt die arithmetischen Mittelwerte der Intervalle.

Alter	$\bar{X}$	***PJ***	$\bar{X}$
$[20-32]$	26	$[0-20]$	10
$[33-45]$	39	$[21-40]$	31
$[46-58]$	52	$[41-60]$	51
$[59-70]$	64		

Tabelle 4.10: Arithmetisches Mittel für DS(4)

Anschließend lässt sich über die MLR ein Faktor der Risikoerhöhung für jedes mittlere *Alter* und mittlere PJ wie bei einem zweiparametrischen modelladäquaten Datensatz ermitteln (siehe Abschnitt 4.4.1 (Modell(3) zu DS(3))).

So ergeben sich mithilfe der MLR folgende Werte:

Alter/PJ	**10**	**31**	**51**
26	$1,000$	$1,690$	$2,787$
39	$1,597$	$2,699$	$4,450$
52	$2,550$	$4,310$	$7,106$
64	$3,927$	$6,639$	$10,946$

Tabelle 4.11: MLR: Risikoerhöhung bezogen auf Wertepaar $(26; 10)$ für DS(4)

Um das Resultat mit dem zweiparametrischen modelladäquaten DS(3) (Abschnitt 4.4.1 (Modell(3) zu DS(3))) vergleichen zu können, wird nun die mittlere quadratische Abweichung zwischen den Risikoerhöhungen der VFTs und den Werten, die über die MLR ermittelt wurden, berechnet.
In diesem Fall ergibt sich für s^2:

$$s^2 = 0,518$$

Die mittlere quadratische Abweichung für den DS(3) lag bei $s^2 = 0,037$. Damit ist die Abweichung für den DS(4) 14 Mal höher.

Um die Erkenntnis der Nichteignung des DS(4) durch die Modellierung mittels MLR noch zu verstärken, werden nun für den Parameter *Alter* ab 26 und den Parameter PJ ab 10 flächendeckend Risikoerhöhungen bezogen auf den Referenzpunkt (26;10) bestimmt. Dabei entsteht ein Oberflächendiagramm (Abbildung 4.10). Zum Vergleich sind hier die zwölf Balken der Risikoerhöhungen der VFTs (Abbildung 4.9) eingetragen. Es lässt sich erkennen, dass die Risikoerhöhungen der VFTs (Abbildung 4.9) erwartungsgemäß nicht mit denen der MLR übereinstimmen. Die MLR lässt sich in diesem Fall nicht durch die VFTs erklären.

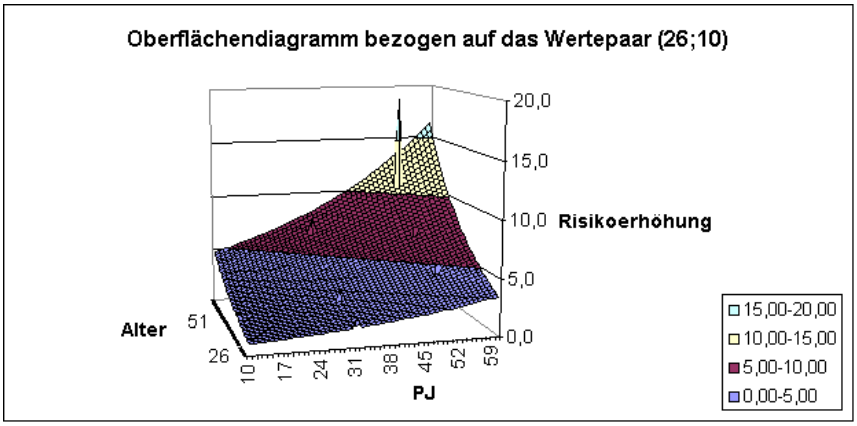

Abbildung 4.10: Oberflächendiagramm: Risikoerhöhung bezogen auf Wertepaar $(26; 10)$ für DS(4)

Es ist nun bewiesen, dass der DS(4) weniger geeignet für die Modellierung mittels MLR ist. Das heißt, es ist zu untersuchen, ob sich durch eine nichtlineare Transformation die Parameter so verändern lassen, dass sich der transformierte DS(4) durch die MLR modellieren lässt.

Schritt 3

Die Parameter *Alter* (X_1) und PJ (X_2) werden durch eine nichtlineare Transformation auf das Intervall zwischen 0 und 1 abgebildet. Die Grenzen ergeben sich aufgrund der Intervalle. Das *Alter* liegt zwischen 20 und 70. Die PJ liegen zwischen 0 und 60.

$$\begin{aligned} Alter : X_1 &\rightarrow (\tfrac{X_1-20}{50})^2 \in [0;1] \\ PJ : \quad X_2 &\rightarrow \sqrt{\tfrac{X_2}{60}} \qquad \in [0;1] \end{aligned}$$

Unter Anwendung des Statistik-Programms PASW Statistics wird für den transformierten DS(4) erneut die MLR durchgeführt. Dabei wird wieder ein Faktor der Risikoerhöhung für den Abstand Δ eines Jahres sowie die Konstanten $B_{t(A)}$ und $B_{t(P)}$ jeweils für den Parameter $Alter_t$ und den Parameter PJ_t berechnet. Aufgrund der Transformation werden die Bezeichnungen zusätzlich mit einem t indiziert.

$$\begin{array}{lll} Alter_t : \beta_{t(A)} = 5,444 & B_{t(A)} = 1,695 & \Delta_{\beta_{t(A)}} = 1 \text{ Jahr} \\ PJ_t : \quad \beta_{t(P)} = 6,468 & B_{t(P)} = 1,867 & \Delta_{\beta_{t(P)}} = 1 \text{ Jahr} \end{array}$$

Schritt 4

Es folgt die Berechnung der Odds über die VFTs für den transformierten DS(4). Für die Parameter $Alter_t$ und PJ_t werden Intervalle gebildet. Der Wertebereich des Parameters $Alter_t$ wird in vier gleichlange Intervalle der Länge 0,25 und der des Parameters PJ_t in drei gleichlange Intervalle der Länge 0,33 unterteilt. Nun werden über die VFTs die Odds ermittelt. In der nachfolgenden Tabelle 4.12 sind die Odds für den transformierten DS(4) aufgeführt.

$\boldsymbol{Alter_t / PJ_t}$	$\mathbf{0 - 0,33}$	$\mathbf{0,331 - 0,67}$	$\mathbf{0,671 - 1}$
$\mathbf{0 - 0,25}$	$0,217$	$0,370$	$0,674$
$\mathbf{0,251 - 0,5}$	$0,300$	$0,500$	$1,160$
$\mathbf{0,51 - 0,75}$	$0,571$	$0,750$	$1,733$
$\mathbf{0,751 - 1}$	$1,000$	$1,333$	$2,889$

Tabelle 4.12: VFT: Odds für transformierten DS(4)

In der Abbildung 4.11 ist erkennbar, dass aus dem DS(4) durch die nichtlinearen Transformationen der beiden Parameter ein für die Modellierung durch die MLR geeigneter Datensatz geworden ist.

Die Abbildung zeigt die typisch gleichmäßige Struktur, die auch bei dem modelladäquaten DS(3) zu erkennen war.

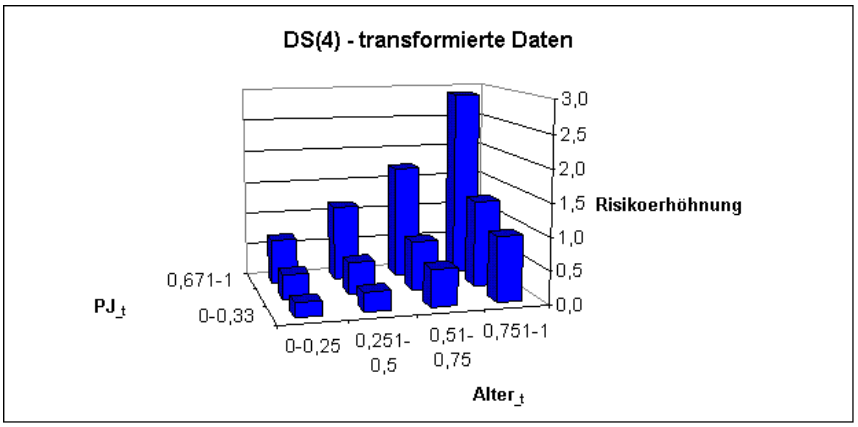

Abbildung 4.11: VFT: Odds für transformierten DS(4)

Nun werden die Risikoerhöhungen in Bezug auf das kleinste Odds ermittelt. Dazu werden alle zwölf Odds durch das kleinste Odds dividiert und es ergeben sich folgende Werte:

$\boldsymbol{Alter_t/PJ_t}$	$\mathbf{0-0,33}$	$\mathbf{0,331-0,67}$	$\mathbf{0,671-1}$
$\mathbf{0-0,25}$	$1,000$	$1,704$	$3,101$
$\mathbf{0,251-0,5}$	$1,380$	$2,300$	$5,336$
$\mathbf{0,51-0,75}$	$2,629$	$3,450$	$7,973$
$\mathbf{0,751-1}$	$4,600$	$6,133$	$13,289$

Tabelle 4.13: VFT: Risikoerhöhung bezogen auf Wertepaar $(0-0,25; 0-0,33)$ für transformierten DS(4)

In der folgenden Abbildung 4.12 werden die berechneten Werte der Tabelle grafisch dargestellt.

Wie bei einem zweiparametrischen modelladäquaten Datensatz bleibt erwartungsgemäß die gleichmäßig ansteigende Struktur erhalten.

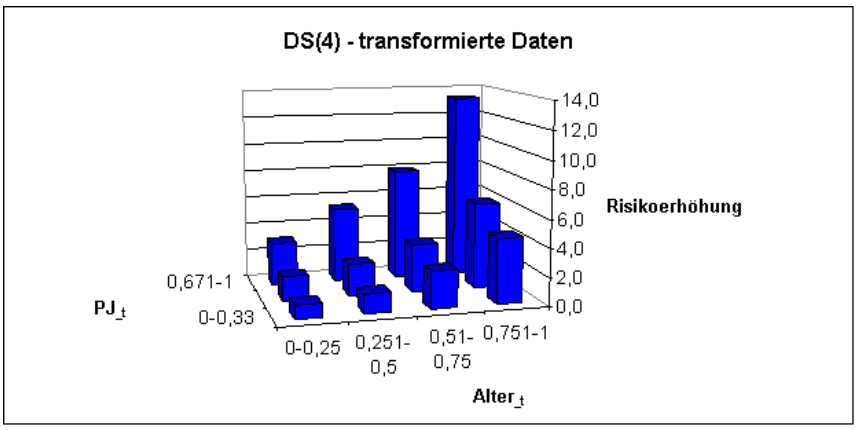

Abbildung 4.12: VFT: Risikoerhöhung bezogen auf Wertepaar $(0-0,25; 0-0,33)$ für transformierten DS(4)

Nun werden die Risikoerhöhungen aus den VFTs mit denen der MLR verglichen. Dazu müssen die Daten aus der MLR erneut in die gleiche Form gebracht werden. Hierfür wird für jedes Intervall der Parameter $Alter_t$ und PJ_t das arithmetische Mittel gebildet (auf zwei Nachkommastellen gerundet) und anschließend über die MLR der Faktor der Risikoerhöhung für jedes mittlere $Alter_t$ und mittlere PJ_t berechnet. Das erfolgt analog zur Berechnung der Werte mittels MLR des modelladäquaten DS(3) wie in Abschnitt 4.4.1 (Modell(3) zu DS(3)) beschrieben.

Das arithmetische Mittel für alle Intervalle wird in der folgenden Tabelle 4.14 dargestellt.

$\boldsymbol{Alter_t}$	$\boldsymbol{\bar{X}}$	$\boldsymbol{PJ_t}$	$\boldsymbol{\bar{X}}$
$[0-0,24]$	$0,12$	$[0-0,33]$	$0,17$
$[0,241-0,5]$	$0,38$	$[0,331-0,67]$	$0,51$
$[0,51-0,76]$	$0,69$	$[0,671-1]$	$0,84$
$[0,761-1]$	$0,88$		

Tabelle 4.14: Arithmetisches Mittel für transformierten DS(4)

Dadurch ergeben sich folgende Risikoerhöhungen:

$\boldsymbol{Alter_t/PJ_t}$	$\mathbf{0,17}$	$\mathbf{0,51}$	$\mathbf{0,84}$
$\mathbf{0,12}$	$1,000$	$1,887$	$3,493$
$\mathbf{0,38}$	$1,554$	$2,931$	$5,428$
$\mathbf{0,69}$	$2,628$	$4,958$	$9,180$
$\mathbf{0,88}$	$3,626$	$6,841$	$12,668$

Tabelle 4.15: MLR: Risikoerhöhung bezogen auf Wertepaar $(0,12; 0,17)$ für transformierten DS(4)

An dieser Stelle wird erneut die mittlere quadratische Abweichung berechnet. So lässt sich beurteilen, ob die Transformation besser für die Modellierung mittels MLR geeignet ist oder nicht. Es ergibt sich für s^2:

$$s^2 = 0,041$$

Im Gegensatz zur mittleren quadratischen Abweichung der Ausgangsdaten liegt dieser s^2-Wert in der Größenordnung der Abweichung von DS(3) mit $s^2 = 0,037$. Die vorliegende Transformation ist besser für die Modellierung geeignet. So kann für den transformierten DS(4) die MLR durchgeführt werden.

Nun kann mithilfe der MLR ein flächendeckendes Diagramm erstellt werden (siehe Abbildung 4.13). Das Verfahren wird in Abschnitt 4.4.1 (Modell(3) zu DS(3)) beschrieben. Es ist erkennbar, dass die Risikoerhöhungen aus den VFTs (Abbildung 4.12) mit denen der MLR übereinstimmen.

Die MLR lässt sich durch die VFTs erklären.

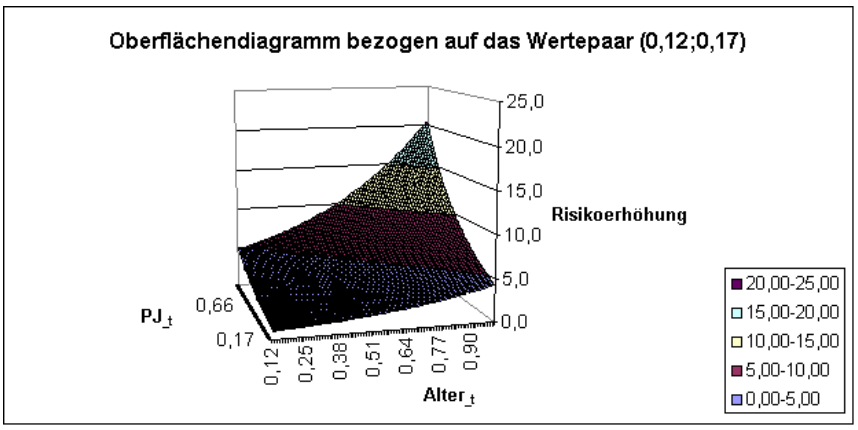

Abbildung 4.13: Oberflächendiagramm: Risikoerhöhung bezogen auf Wertepaar $(0,12;0,17)$ für transformierten DS(4)

Anschließend werden die Risikoerhöhungen bezogen auf den Referenzpunkt $(0;0)$ berechnet. Es ergibt sich folgendes Oberflächendiagramm.

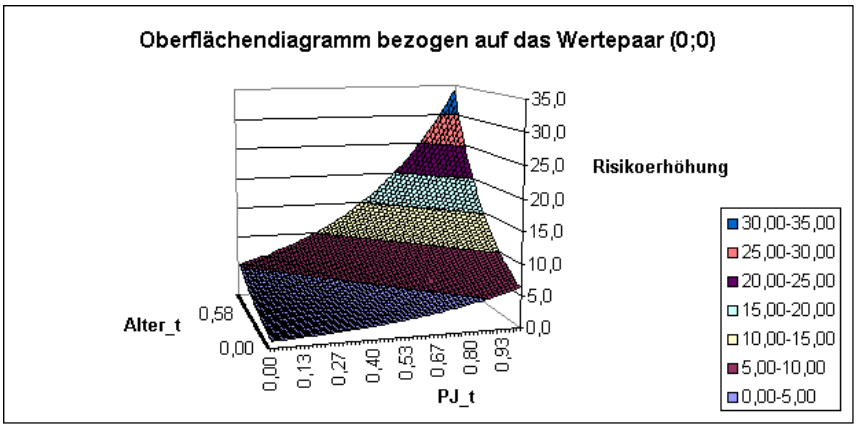

Abbildung 4.14: Oberflächendiagramm: Risikoerhöhung bezogen auf Wertepaar $(0;0)$ für transformierten DS(4)

Schritt 5

Es wird die Rückrechnung von dem transformierten DS(4) auf den DS(4) erläutert. Dazu wird das Oberflächendiagramm (Abbildung 4.14) benötigt. In diesem Diagramm sind alle Risikoerhöhungen des transformierten DS(4) in Bezug auf (0;0) aufgeführt.

Nun wird jede Risikoerhöhung in ein neues Oberflächendiagramm eingetragen.

Dabei werden die Parameter $Alter_t$ und PJ_t entsprechend der nichtlinearen Transformation auf $Alter$ und PJ zurückgerechnet.

$$\begin{aligned} Alter &= \sqrt{Alter_t} \cdot 50 + 20 \\ PJ \quad &= PJ_t^2 \cdot 60 \end{aligned}$$

Werden alle Risikoerhöhungen auf die Parameter $Alter$ zwischen 20 und 70 und PJ zwischen 0 und 60 übertragen, ergibt sich folgendes Oberflächendiagramm (Abbildung 4.15). Eine detailliertere Abbildung befindet sich im Anhang B.

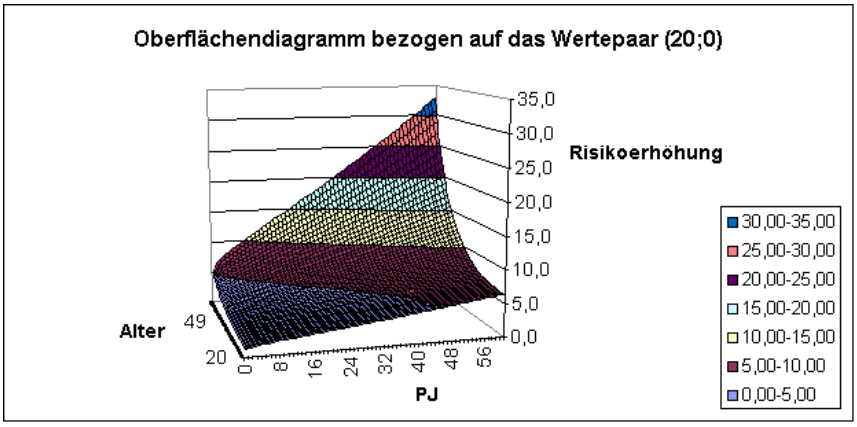

Abbildung 4.15: Oberflächendiagramm: Risikoerhöhung bezogen auf Wertepaar $(20; 0)$ für DS(4)

Die Abbildung 4.15 zeigt die Modellierung der Ausgangsdaten (Abbildung 4.9) mittels MLR. Es ist der steile Anstieg des Odds für das Alter im Intervall $[57,6; 70]$ und die PJ im Intervall $[41; 60]$ erkennbar. Die Datentransformation ermöglicht eine Verbesserung der Modellierung.

Berechnungsbeispiel:
Für $Alter_t = 0,02$ und $PJ_t = 0,45$ lässt sich im Oberflächendiagramm (Abbildung 4.14) eine Risikoerhöhung von $2,40$ ablesen. Werden $Alter_t$ und PJ_t nun nach obigen Formeln zurücktransformiert, kann die Risikoerhöhung $2,40$ für $Alter = 27,07$ und $PJ = 12,15$ eingetragen werden. In dem Oberflächendiagramm (Abbildung 4.15) ist für $Alter = 27$ und $PJ = 12$ eine Risikoerhöhung von $2,38$ abzulesen. Es können geringe Abweichungen aus Rundungsfehlern resultieren.

4.5 Fazit

Es wurde anhand des DS(1) gezeigt, dass sich bei einem modelladäquaten Datensatz mit einem Parameter die MLR durch die VFTs plausibel erklären lässt. Außerdem konnte nachgewiesen werden, dass die Linearität einer Transformation keine Auswirkungen auf die Resultate hat. Zudem ließ sich die Risikoerhöhung des transformierten Datensatzes auf die Risikoerhöhung des Ausgangsdatensatzes zurückrechnen.

Der zweite einparametrische Datensatz, der DS(2), ließ Folgendes erkennen. Wenn sich die MLR nicht durch die VFTs erklären lässt, kann untersucht werden, ob eine nichtlineare Transformation eine Verbesserung der Modelladäquatheit mit sich bringt. Anschließend konnten die Risikoerhöhungen des transformierten Datensatzes auf die Risikoerhöhungen des Ausgangsdatensatzes zurücktransformiert werden.

Nun wurden analog dazu die zweiparametrischen Datensätze untersucht. Der DS(3) ergab das Gleiche wie der DS(1). Ist die MLR plausibel durch die VFTs erklärbar, so lassen sich beide Parameter linear transformieren, ohne dass sich diese Vorgehensweise auf die Ergebnisse auswirkt. Anschließend konnten die beiden transformierten Parameter zurücktransformiert werden, wobei die Risikoerhöhungen erhalten blieben.

Zum Schluss wurde mit DS(4) analog zu DS(2) festgestellt, wenn die MLR nicht durch die VFTs erklärbar ist, kann eine nichtlineare Transformation verwendet werden. Dabei muss untersucht werden, ob diese Transformation eine Verbesserung der mittleren quadratischen Abweichung erzielt. Ist das der Fall, kann der transformierte Datensatz auf die Ausgangsdaten zurückgerechnet werden ohne die Risikoerhöhungen zu beeinflussen.

Gerade der DS(4) hat gezeigt, dass es sich anhand der Ausgangsdaten nicht einschätzen lässt, ob ein Datensatz modelladäquat ist oder nicht. Es sollte stets die mittlere quadratische Abweichung als Maß dafür überprüft werden. Jeder einzelne Parameter beeinflusst die Risikoerhöhung. Damit diese Einflüsse nicht verloren gehen, muss das Modell angepasst werden. Zur Überprüfung dient auch der Vergleich des abschließenden Oberflächendiagramms bezogen auf das Wertepaar $(20; 0)$ mit den Risikoerhöhungen der Ausgangsdaten, die über die VFTs berechnet wurden. Diese Risikoerhöhungen müssen sich in dem Oberflächendiagramm widerspiegeln.

4.6 Bemerkung zur Signifikanz

Bei der Modellierung mittels MLR unter Anwendung des Statistik Programms PASW Statistics werden für alle vier Datensätze folgende Werte berechnet. Dabei werden die p-Werte[1], die ORs (Exp(B)) und die zugehörigen 95%-Konfidenzintervalle (KI) angegeben, wobei sich das KI aus einer unteren Grenze (uG) und einer oberen Grenze (oG) zusammensetzt.

Datensatz	**Parameter**	**p**	**Exp(B)**	**KI(uG)**	**KI(oG)**
DS(1)	$Alter$	$0,048$	$1,013$	$1,000$	$1,026$
Transformierter DS(1)	$Alter_t$	$0,048$	$1,883$	$1,005$	$3,525$
DS(2)	$Alter$	$0,000$	$1,098$	$1,078$	$1,117$
Transformierter DS(2)	$Alter_t$	$0,000$	$83,610$	$37,148$	$188,185$
DS(3)	$Alter$	$0,000$	$1,026$	$1,013$	$1,039$
	PJ	$0,000$	$1,020$	$1,009$	$1,030$
Transformierter DS(3)	$Alter_t$	$0,000$	$3,543$	$1,879$	$6,679$
	PJ_t	$0,000$	$3,208$	$1,701$	$6,052$
DS(4)	$Alter$	$0,000$	$1,036$	$1,023$	$1,050$
	PJ	$0,000$	$1,025$	$1,014$	$1,036$
Transformierter DS(4)	$Alter_t$	$0,000$	$5,444$	$2,869$	$10,329$
	PJ_t	$0,000$	$6,468$	$2,883$	$14,511$

Tabelle 4.16: Signifikanz der Datensätze

Es ist zu beachten, dass alle p-Werte kleiner als $0,05$ sind. Das heißt, dass der Einfluss aller Parameter statistisch abgesichert werden kann. Allerdings wird damit keine Aussage darüber getroffen, ob der entsprechende Datensatz für die Modellierung mithilfe der MLR geeignet ist.

[1] auch: Irrtumswahrscheinlichkeit, Signifikanzniveau

5 Patientenkollektiv der arbeitsmedizinischen Untersuchung

Es wird für ein Patientenkollektiv der arbeitsmedizinischen Untersuchung die MLR durchgeführt. Das Kollektiv besteht aus 610 Personen, von denen 169 Personen an Lungenkrebs erkrankt sind und 441 Personen nicht[9].

Bevor mit Schritt 1 begonnen wird, sollte der Verlauf der Risikoerhöhung in Abhängigkeit von je einem der beiden Parameter betrachtet werden. So lässt sich erkennen, wie sich *Alter* und PJ getrennt voneinander auf die Risikoerhöhung auswirken. Der Grund ist, dass die Effekte, die hier auftreten, erhalten bleiben sollen, wenn die Risikoerhöhung in Abhängigkeit von beiden Parametern analysiert wird. Das vereinfacht die Frage nach der Einteilung der Wertebereiche von *Alter* und PJ in Intervalle.

Angefangen mit dem Parameter *Alter* ergibt sich folgende Abbildung.

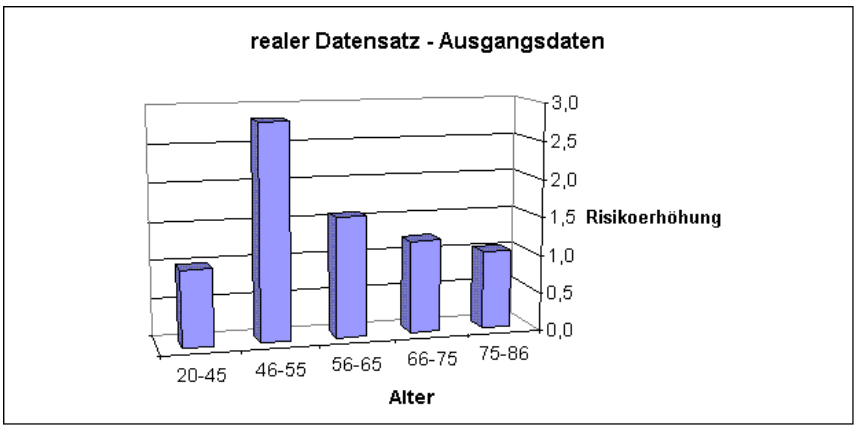

Abbildung 5.1: Risikoerhöhung in Abhängigkeit vom Parameter *Alter* für realen DS

Laut Abbildung 5.1 steigt bis zum *Alter* 55 das Risiko an Lungenkrebs zu erkranken, um anschließend mit zunehmendem *Alter* wieder zu sinken.

Eine mögliche Erklärung dafür ist, dass ab *Alter* 55 schon einige Personen an der Erkrankung gestorben sind und deshalb in der Statistik nicht mehr auftauchen.

Nun wird die Abhängigkeit der Risikoerhöhung vom Parameter PJ gezeigt.

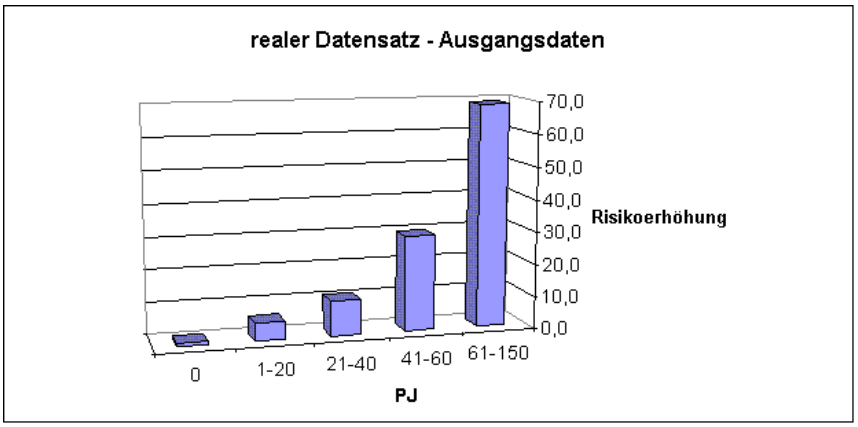

Abbildung 5.2: Risikoerhöhung in Abhängigkeit vom Parameter PJ für realen DS

Der Verlauf der Risikoerhöhung in Abbildung 5.2 ist wie erwartet. Mit zunehmenden PJ steigt das Risiko an Lungenkrebs zu erkranken.

Nach der Betrachtung der Auswirkungen der beiden einzelnen Parameter auf die Risikoerhöhung werden die Intervalle der Wertebereiche der Parameter für die folgende Untersuchung so gewählt, dass die Effekte, die sich gezeigt haben, möglichst erhalten bleiben.

Nun kann mit den eigentlichen Schritten der Modellanalyse begonnen werden.

Schritt 1
Die Durchführung der MLR mittels Statistik-Programms PASW Statistics:

$$\begin{array}{lll} Alter: \beta_A = 0,968 & B_A = -0,032 & \Delta_{\beta_A} = 1 \text{ Jahr} \\ PJ: \quad \beta_P = 1,045 & B_P = \quad 0,044 & \Delta_{\beta_P} = 1 \text{ Jahr} \end{array}$$

Schritt 2
Es werden die Odds für die Ausgangsdaten des realen DS berechnet (siehe Tabelle 5.1)

$\boldsymbol{Alter/PJ}$	$\mathbf{0-10}$	$\mathbf{11-24}$	$\mathbf{25-45}$	$\mathbf{46-150}$
$\mathbf{20-46}$	$0,125$	$0,500$	$1,000$	$DIV/0$
$\mathbf{47-65}$	$0,140$	$0,545$	$0,652$	$2,625$
$\mathbf{66-86}$	$0,061$	$0,344$	$0,441$	$1,261$

Tabelle 5.1: VFT: Odds für realen DS

Bei der weiteren Berechnung werden die Odds, für die die Anzahl der Personen in den entsprechenden Intervallen zu gering ist, nicht betrachtet.

In der Abbildung 5.3 werden die Odds grafisch dargestellt.

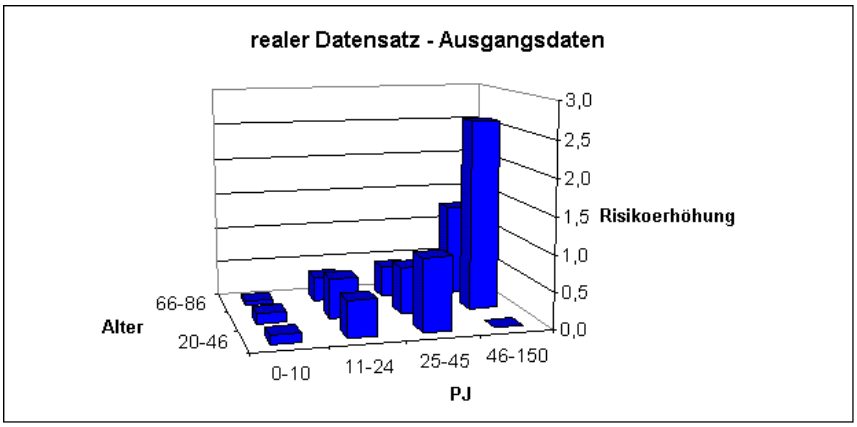

Abbildung 5.3: VFT: Odds für realen DS

Die Abbildung 5.3 hat die erwartete Form. Die PJ steigen mit zunehmender Anzahl an. Wird das *Alter* für die PJ $[0-10]$ und $[11-24]$ betrachtet, ist erkennbar, dass das Odds erst ansteigt und anschließend wieder fällt. Nur die beiden Odds für das *Alter* $[20-46]$ und die beiden PJ $[25-45]$ und $[46-150]$ spiegeln die erwarteten Risikoerhöhungen nicht wider. Das hat die Erklärung, dass die Anzahl der Personen an diesen beiden Stellen zu gering ist.

Als Nächstes werden die Risikoerhöhungen aus den Odds berechnet (siehe Tabelle 5.2).

$\boldsymbol{Alter/PJ}$	$\mathbf{0-10}$	$\mathbf{11-24}$	$\mathbf{25-45}$	$\mathbf{46-150}$
$\mathbf{20-46}$	$0,000$	$0,000$	$0,000$	$0,000$
$\mathbf{47-65}$	$2,272$	$8,883$	$10,621$	$42,750$
$\mathbf{66-86}$	$1,000$	$5,607$	$7,177$	$20,534$

Tabelle 5.2: VFT: Risikoerhöhung bezogen auf Wertepaar $(66-86; 0-10)$ für realen DS

Bei der Einteilung der Wertebereiche der Parameter *Alter* und *PJ* in Intervalle kann es vorkommen, dass bei der Berechnung der VFTs Intervalle auftreten, die statistisch gesehen zu gering besetzt und daher nicht aussagekräftig sind. Dann sind für die Berechnung der Risikoerhöhungen aus den Odds nur die Odds zu verwenden, die sich aus einer Anzahl ≥ 20 Personen ergeben. Die übrigen Risikoerhöhungen werden 0 gesetzt.

Die Abbildung 5.4 zeigt die Risikoerhöhungen bezogen auf das Wertepaar $(66-86; 0-10)$:

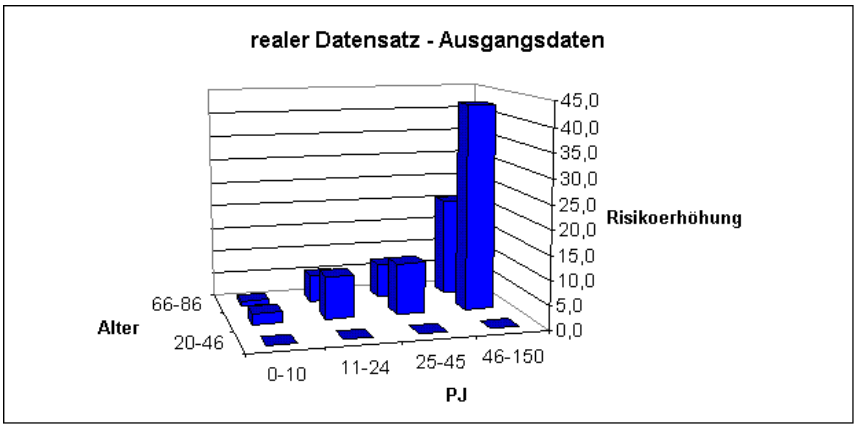

Abbildung 5.4: VFT: Risikoerhöhung bezogen auf Wertepaar $(66-86; 0-10)$ für realen DS

Nun werden die arithmetischen Mittelwerte für alle Intervalle berechnet.

Alter	$\bar{X}$	***PJ***	$\bar{X}$
$[20-46]$	33	$[0-10]$	5
$[47-65]$	56	$[11-24]$	18
$[66-86]$	76	$[25-45]$	35
		$[46-150]$	98

Tabelle 5.3: Arithmetisches Mittel für realen DS

Es folgt die Berechnung der Risikoerhöhungen mithilfe der MLR (siehe Tabelle 5.4).

Alter/PJ	**5**	**18**	**35**	**98**
33	$0,000$	$0,000$	$0,000$	$0,000$
56	$1,896$	$3,360$	$7,099$	$113,522$
76	$1,000$	$1,772$	$3,743$	$59,859$

Tabelle 5.4: MLR: Risikoerhöhung bezogen auf Wertepaar $(76;5)$ für realen DS

Anschließend wird die mittlere quadratische Abweichung s^2 ermittelt:

$$s^2 = 7,295$$

Nun wird die flächendeckende Berechnung der MLR grafisch dargestellt:

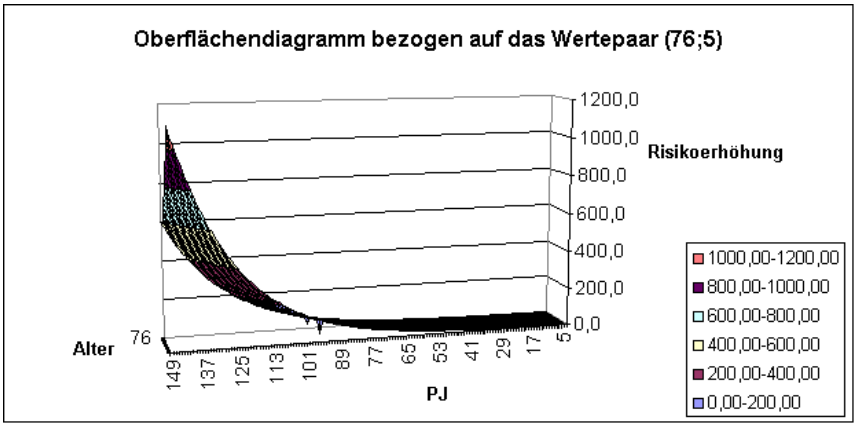

Abbildung 5.5: Oberflächendiagramm: Risikoerhöhung bezogen auf Wertepaar $(76;5)$ für realen DS

Die MLR für den Datensatz lässt sich nicht durch die VFTs erklären.

Es stellt sich die Frage, ob das Problem der Intervalleinteilung umgangen werden kann, wenn die Wertebereiche der Parameter in zwei Intervalle eingeteilt werden. Bei der Darstellung in einem Säulendiagramm würde eine Form entstehen, wie es für die Modellierung mittels MLR charakteristisch ist.

Der zweite Schritt wird für die Einteilung in je zwei Intervalle der Wertebereiche der Parameter erneut vorgenommen, um auch diesen Fall zu untersuchen.

Die Berechnung der Odds für die Ausgangsdaten:

$\boldsymbol{Alter/PJ}$	$\mathbf{0-20}$	$\mathbf{21-150}$
$\mathbf{20-60}$	$0,183$	$1,097$
$\mathbf{61-86}$	$0,175$	$0,694$

Tabelle 5.5: VFT: Odds für realen DS (zwei Intervalle)

Es werden die Odds grafisch dargestellt:

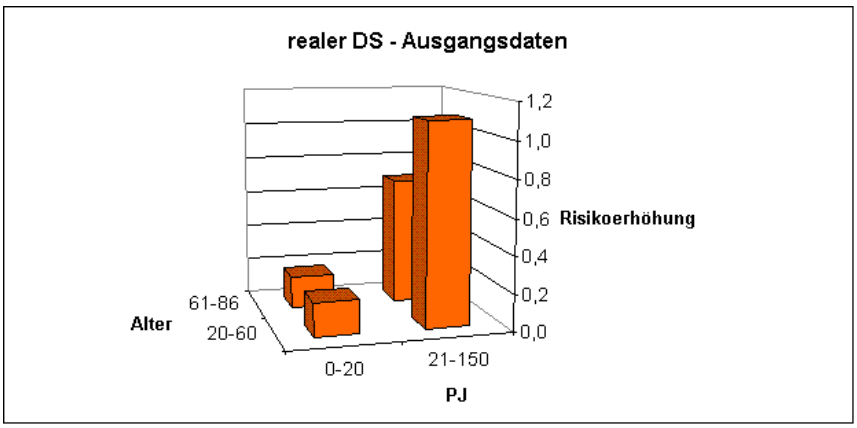

Abbildung 5.6: VFT: Odds für realen DS (zwei Intervalle)

Die Risikoerhöhungen aus den Odds werden in der Tabelle 5.6 dargestellt.

$\boldsymbol{Alter/PJ}$	$\mathbf{0-20}$	$\mathbf{21-150}$
$\mathbf{20-60}$	$1,050$	$6,279$
$\mathbf{61-86}$	$1,000$	$3,974$

Tabelle 5.6: VFT: Risikoerhöhung bezogen auf Wertepaar $(61-86; 0-20)$ für realen DS (zwei Intervalle)

Die Abbildung 5.7 zeigt die grafische Darstellung der Odds für je zwei Intervalle.

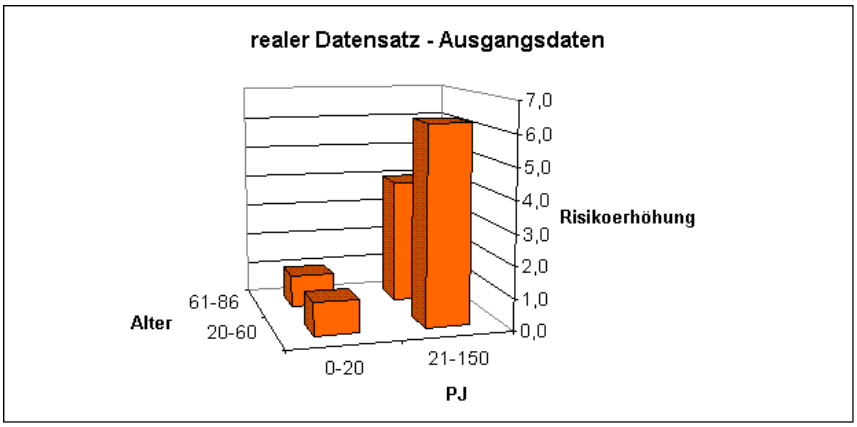

Abbildung 5.7: VFT: Risikoerhöhung bezogen auf Wertepaar $(61-86; 0-20)$ für realen DS (zwei Intervalle)

Es werden die arithmetischen Mittelwerte für alle Intervalle berechnet (siehe Tabelle 5.7)

Alter	$\bar{X}$	***PJ***	$\bar{X}$
$[20-60]$	40	$[0-20]$	10
$[61-86]$	74	$[21-150]$	86

Tabelle 5.7: Arithmetisches Mittel für realen DS (zwei Intervalle)

In der Tabelle 5.8 folgt die Berechnung der Risikoerhöhungen mithilfe der MLR.

Alter/PJ	**6**	**18**
33	2,968	84,099
56	1,000	28,332

Tabelle 5.8: MLR: Risikoerhöhung bezogen auf Wertepaar $(74; 10)$ für realen DS (zwei Intervalle)

Es wird die mittlere quadratische Abweichung s^2 berechnet:

$$s^2 = 19,777$$

Der s^2-Wert ist fast drei Mal so hoch wie der vorherige. Es ist also ein Trugschluss, dass ein $2 * 2$-Säulendiagramm einen besseren s^2-Wert liefert, weil das Diagramm den Eindruck erweckt, die gewünschte Form zu haben.

Aus diesem Grund wird nun mit Schritt 3 für die ursprüngliche Einteilung der Wertebereiche der Parameter fortgesetzt.

Schritt 3
Die Parameter *Alter* (X_1) und PJ (X_2) werden durch eine nichtlineare Transformation auf das Intervall zwischen 0 und 1 abgebildet. Die Transformation des Parameters *Alter* ergibt sich, da bei $Alter = 53$ das Maximum liegt und das $Alter < 53$ sowie $Alter > 53$ monoton steigt bzw. fällt. Die PJ werden linear transformiert, da die Ausgangsdaten schon die ansteigende Form besitzen.

$$\begin{aligned} Alter : X_1 &\rightarrow 1 - (\tfrac{|53 - X_1|}{33}) \in [0;1] \\ PJ : \quad X_2 &\rightarrow \tfrac{X_2}{150} \qquad\qquad \in [0;1] \end{aligned}$$

Die Durchführung der MLR mittels Statistik-Programm PASW Statistics:

$$\begin{aligned} Alter : \beta_{t(A)} = \quad 5,776 \qquad & B_{t(A)} = 1,754 \qquad & \Delta_{\beta_{t(A)}} = 1 \text{ Jahr} \\ PJ : \quad \beta_{t(P)} = 713,218 \qquad & B_{t(P)} = 6,570 \qquad & \Delta_{\beta_{t(P)}} = 1 \text{ Jahr} \end{aligned}$$

Schritt 4
Die Berechnung der Odds für die transformierten Daten:

$\boldsymbol{Alter_t / PJ_t}$	$\mathbf{0 - 0,067}$	$\mathbf{0,0671 - 0,161}$	$\mathbf{0,1611 - 0,301}$	$\mathbf{0,3011 - 1}$
$\mathbf{0 - 0,395}$	$0,038$	$0,156$	$0,188$	$1,222$
$\mathbf{0,3951 - 0,683}$	$0,089$	$0,550$	$0,500$	$1,733$
$\mathbf{0,6831 - 1}$	$0,152$	$0,542$	$0,750$	$2,143$

Tabelle 5.9: VFT: Odds für transformierten realen DS

Die Abbildung 5.8 zeigt die Odds für den transformierten realen DS.

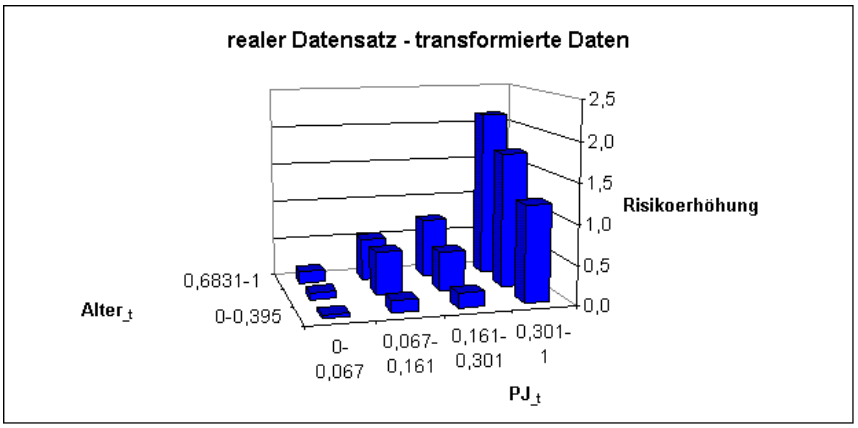

Abbildung 5.8: VFT: Odds für transformierten realen DS

Es werden die Risikoerhöhungen aus den Odds für den transformierten realen DS berechnet (siehe Tabelle 5.10).

$\boldsymbol{Alter_t/PJ_t}$	$\mathbf{0-0,067}$	$\mathbf{0,0671-0,161}$	$\mathbf{0,1611-0,301}$	$\mathbf{0,3011-1}$
$\mathbf{0-0,395}$	$1,000$	$4,063$	$0,000$	$0,000$
$\mathbf{0,3951-0,683}$	$2,311$	$14,300$	$13,000$	$45,067$
$\mathbf{0,6831-1}$	$3,939$	$14,083$	$19,500$	$0,000$

Tabelle 5.10: VFT: Risikoerhöhung bezogen auf Wertepaar $(0-0,395; 0-0,067)$ für transformierten realen DS

Nun werden die ermittelten Risikoerhöhungen in der Abbildung 5.9 grafisch dargestellt.

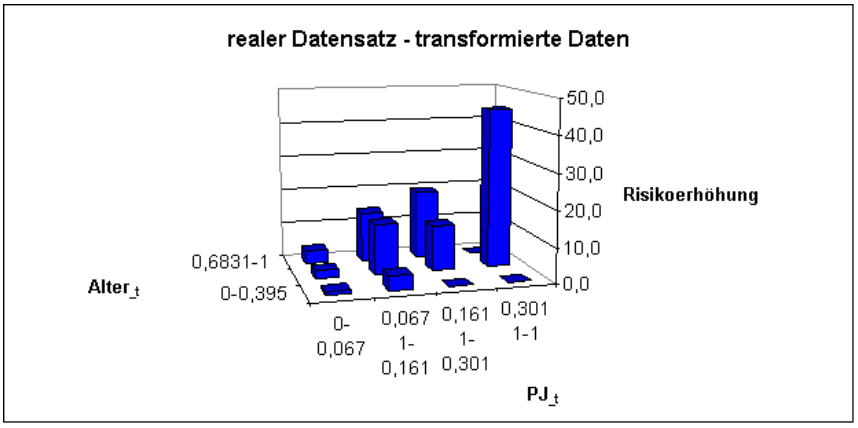

Abbildung 5.9: VFT: Risikoerhöhung bezogen auf Wertepaar $(0-0,395; 0-0,067)$ für transformierten realen DS

Die Berechnung der arithmetischen Mittelwerte für alle Intervalle:

$\boldsymbol{Alter_t}$	$\bar{\boldsymbol{X}}$	$\boldsymbol{PJ_t}$	$\bar{\boldsymbol{X}}$
$[0-0,395]$	$0,20$	$[0-0,067]$	$0,03$
$[0,3951-0,683]$	$0,55$	$[0,0671-0,161]$	$0,12$
$[0,6831-1]$	$0,85$	$[0,1611-0,301]$	$0,23$
		$[0,3011-1]$	$0,65$

Tabelle 5.11: Arithmetisches Mittel für transformierten realen DS

Es werden die Risikoerhöhungen mithilfe der MLR berechnet:

$\boldsymbol{Alter_t/PJ_t}$	$\mathbf{0,03}$	$\mathbf{0,12}$	$\mathbf{0,23}$	$\mathbf{0,65}$
$\mathbf{0,20}$	$1,000$	$1,729$	$0,000$	$0,000$
$\mathbf{0,55}$	$1,843$	$3,186$	$6,857$	$108,272$
$\mathbf{0,85}$	$3,135$	$5,421$	$11,667$	$0,000$

Tabelle 5.12: MLR: Risikoerhöhung bezogen auf Wertepaar $(0,20; 0,03)$ für transformierten realen DS

Es folgt die Berechnung der mittleren quadratischen Abweichung s^2:

$$s^2 = 4,412$$

Die Transformation der Daten hat eine Verbesserung des s^2-Wertes gebracht.

Es folgt das Oberflächendiagramm bezogen auf den Referenzpunkt $(0,20; 0,03)$:

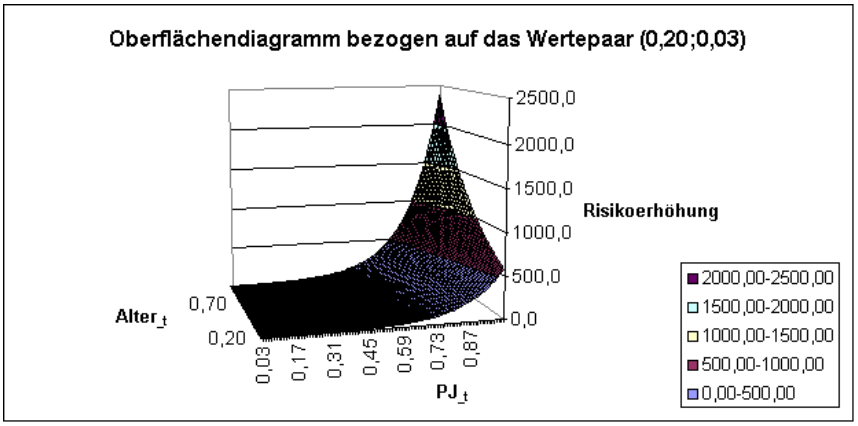

Abbildung 5.10: Oberflächendiagramm: Risikoerhöhung bezogen auf Wertepaar $(0,20; 0,03)$ für transformierten realen DS

Anschließend wird das Oberflächendiagramm bezogen auf den Referenzpunkt $(0; 0)$ dargestellt:

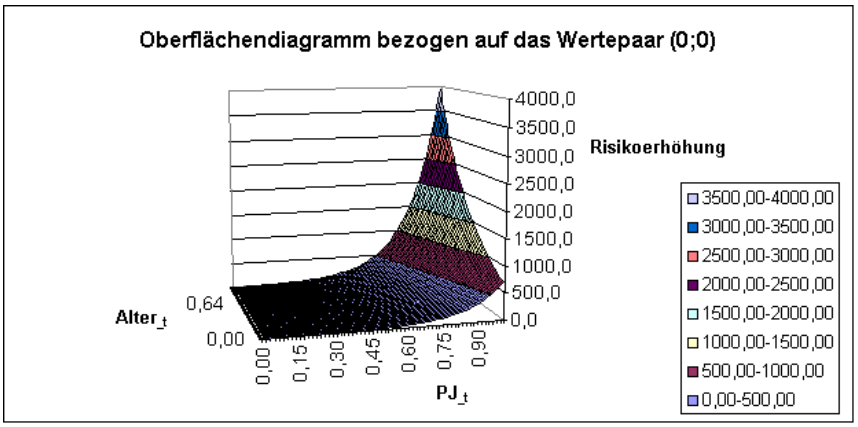

Abbildung 5.11: Oberflächendiagramm: Risikoerhöhung bezogen auf Wertepaar $(0; 0)$ für transformierten realen DS

Schritt 5

Rückrechnung: Hier wird für den Parameter Alter aufgrund der Betragsfunktion eine Fallunterscheidung vorgenommen.

$$\begin{aligned} \text{für } Alter \leq 53 : \text{Alter} &= 20 + 33 \cdot Alter_t \\ \text{für } Alter > 53 : \text{Alter} &= 86 - 33 \cdot Alter_t \\ PJ &= PJ_t \cdot 150 \end{aligned}$$

Es folgt das Oberflächendiagramm bezogen auf den Referenzpunkt $(20; 0)$:

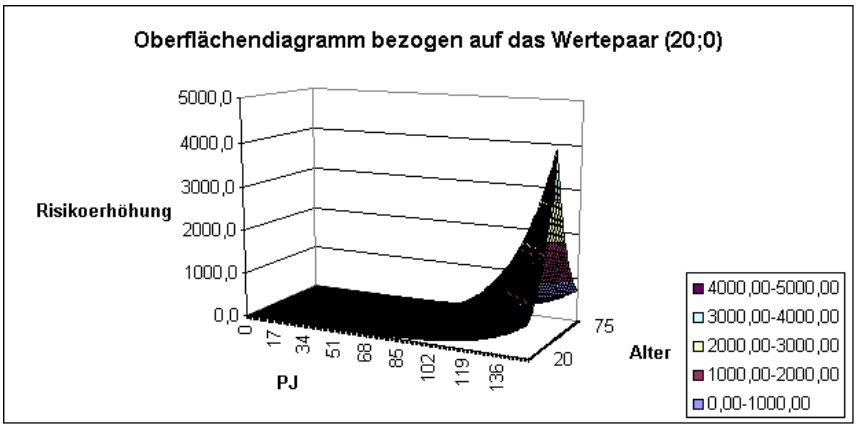

Abbildung 5.12: Oberflächendiagramm: Risikoerhöhung bezogen auf Wertepaar $(20; 0)$ für realen DS

Das Oberflächendiagramm stellt die Risikoerhöhungen in Abhängigkeit vom *Alter* und der PJ dar, wobei die Einflüsse der Parameter *Alter* und PJ erhalten bleiben. Die Risikoerhöhungen bezogen auf die PJ steigen mit zunehmender Anzahl an. Dagegen sorgt das *Alter* dafür, dass die Risikoerhöhungen erst ansteigen und mit zunehmendem *Alter* wieder fallen. Die risikohemmende Wirkung des Parameters *Alter* geht durch die Modellierung nicht verloren. Eine detailliertere Abbildung befindet sich im Anhang C.

6 Zusammenfassung und Ausblick

6.1 Zusammenfassung

In den vorangegangenen Kapiteln wurde gezeigt, dass sich mithilfe einer Datentransformation eine Verbesserung der Modellierung durch die MLR erzielen lässt.

Es ist möglich Datensätze auf der Basis der MLR mit einem oder zwei Parametern zu simulieren. Dabei kann gewählt werden, ob es sich um einen modelladäquaten Datensatz handeln soll oder ob die Modelladäquatheit des Datensatzes erst nach der festgelegten nichtlinearen Transformation eintreten soll.
Für die anschließende Untersuchung der simulierten Datensätze wurde ein Verfahren bestehend aus fünf Schritten entwickelt. Dabei enthält der **erste Schritt** die Durchführung der MLR mithilfe eines Statistik Programms. Der **zweite Schritt** dient der Berechnung der VFTs. Dazu müssen die Parameter in geeignete Intervalle eingeteilt werden. So kann eine Aussage über die Plausibilität der MLR durch die VFTs getroffen werden. Um ein Maß für die Güte der Modellierung festzulegen, wird die mittlere quadratische Abweichung verwendet. Nach der Untersuchung des letzten Datensatzes (DS(4)) lässt sich einschätzen, in welcher Größenordnung die mittlere quadratische Abweichung liegen muss, damit von einer Verbesserung der Modelladäquatheit des Datensatzes gesprochen werden kann. Als Nächstes wird im **dritten Schritt** eine nichtlineare Transformation für den Datensatz gewählt. Für diesen wird dann mithilfe eines Statistik-Programms die MLR durchgeführt. Anschließend wird im **vierten Schritt** beschrieben, wie untersucht werden kann, ob durch die im Schritt drei festgelegte Transformation eine Verbesserung der mittleren quadratischen Abweichung erreicht werden kann oder nicht. Der letzte Schritt der Modellanalyse ist der **Schritt fünf**. Dieser dient der Rückrechnung der transformierten Parameter auf den Ausgangsdatensatz. Dabei ist es besonders wichtig, dass die ermittelten Risikoerhöhungen erhalten bleiben.

6.2 Ausblick

Nachdem das beschriebene Verfahren auf zwei Parameter anwendbar ist, lässt es sich auf beliebig viele Parameter übertragen. Nun gilt es, ein Klassifikationsverfahren für die Transformation der Einzelparameter zu entwickeln. In Kapitel 5 wurde anhand des Applikationsbeispiels gezeigt, dass die Einflüsse der einzelnen Parameter eine wichtige Rolle für die Entwicklung der gesamten Risikoerhöhung spielen. Die Parameter sollten durch eine Transformation so gut wie möglich an das Modell der MLR angepasst werden, sodass bei der Rücktransformation des Datensatzes die speziellen Einflüsse sichtbar werden. Es wird ein Verfahren für die Optimierung der Datentransformation benötigt. So lässt sich bei Datensätzen mit geringer mittlerer quadratischer Abweichung die Modellanalyse problemlos realisieren.

In der vorliegenden Arbeit war aufgrund der Verwendung der MLR die multiplikative Verknüpfung der Einzelrisiken festgelegt. Die verwendete Verknüpfung ist somit eine Form des Kompromisses. Eine andere Form wäre der Mittelwert der Auftretenswahrscheinlichkeiten. Zwei Extreme hinsichtlich des Kompromisses sind Optimismus und Pessimismus. Bei diesen ist jeweils die Auftretenswahrscheinlichkeit für die Nicht-Erkrankung bzw. Erkrankung an Lungenkrebs ausschlaggebend. Allerdings können diese Verknüpfungen nur auf Datensätze mit mindestens zwei Parametern angewendet werden. Hier kann untersucht werden, wie sich durch die beschriebenen Verknüpfungen die Simulation der Datensätze ändert.

A Oberflächendiagramm für DS(3)

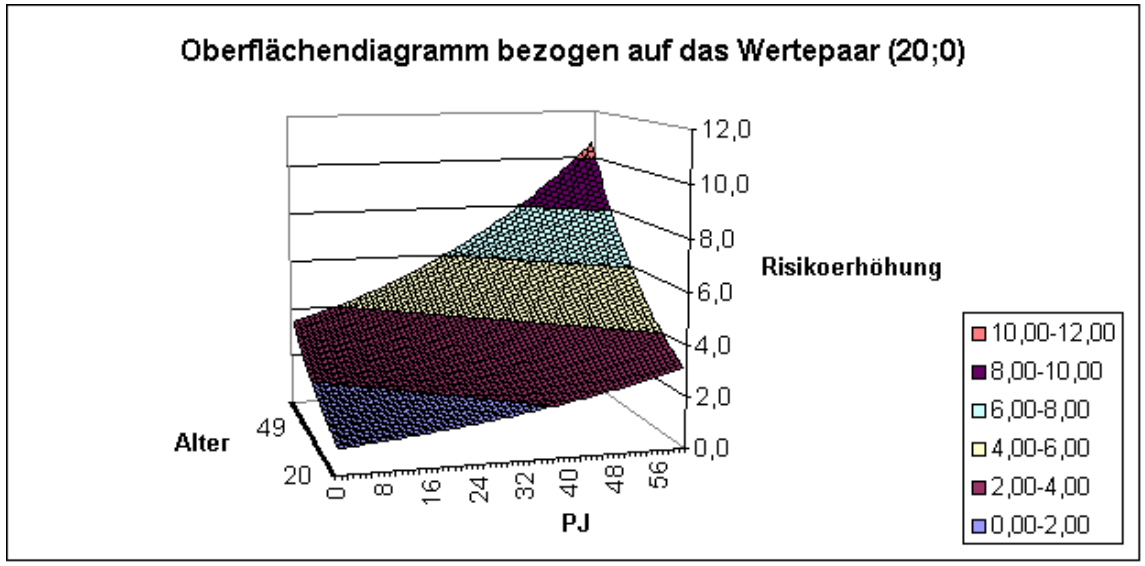

Abbildung A.1: Oberflächendiagramm: Risikoerhöhung bezogen auf Wertepaar $(20; 0)$ für DS(3)

B Oberflächendiagramm für DS(4)

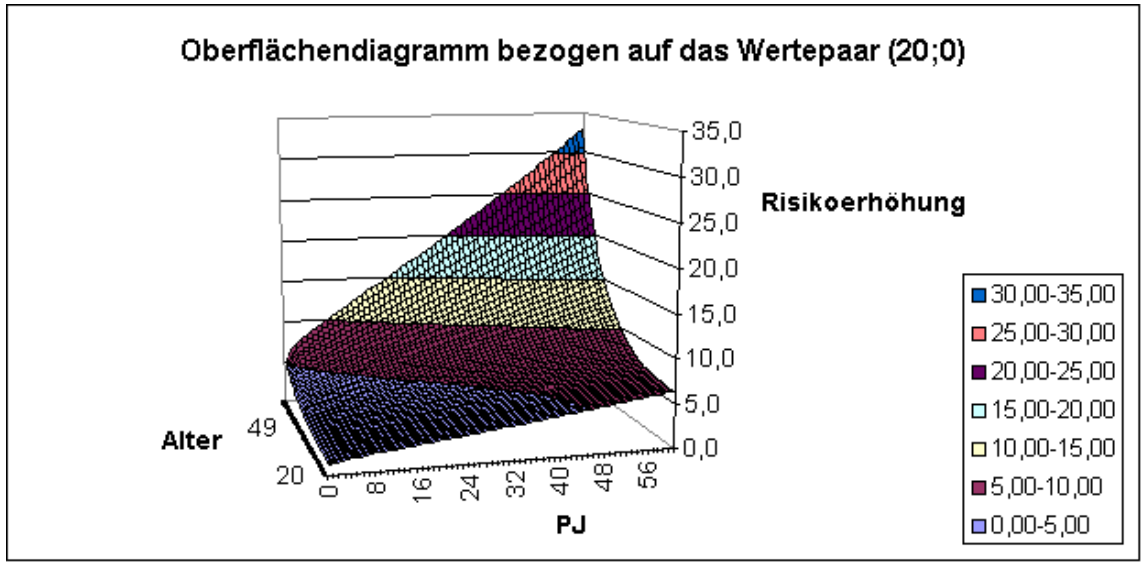

Abbildung B.1: Oberflächendiagramm: Risikoerhöhung bezogen auf Wertepaar $(20; 0)$ für DS(4)

C Oberflächendiagramm für realen DS

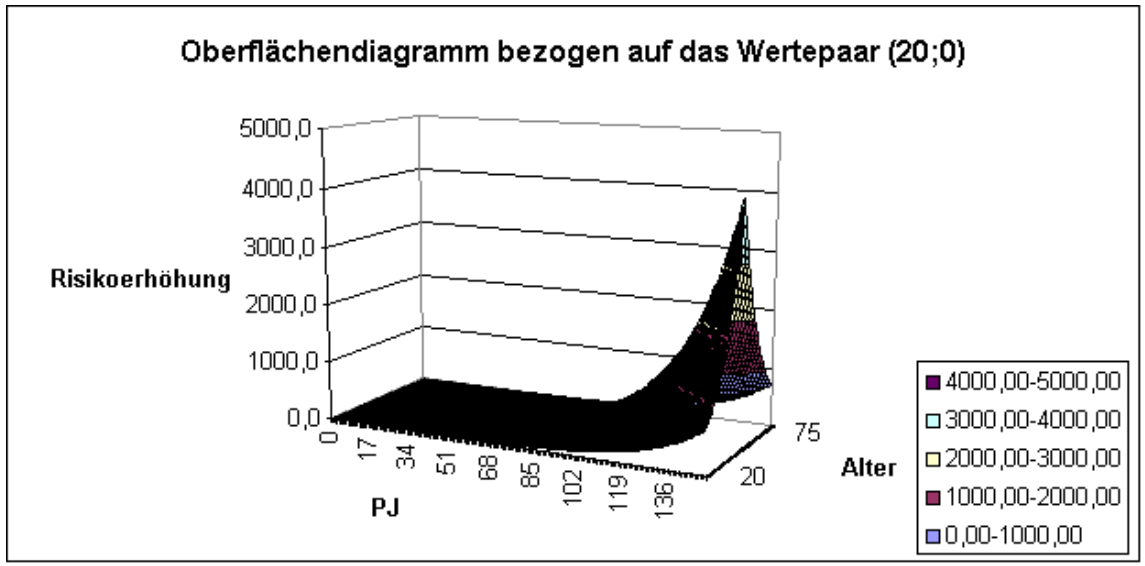

Abbildung C.1: Oberflächendiagramm: Risikoerhöhung bezogen auf Wertepaar $(20; 0)$ für realen DS

Literaturverzeichnis

[1] N. Bitterlich. OR-Analyse. Manuskript, Medizin und Service GmbH, 2007.

[2] L. Sachs; J. Hedderich. *Angewandte Statistik, Methodensammlung mit R.* Springer-Verlag Berlin-Heidelberg, 2006.

[3] E. Kvas. Basics in Statistik: Teil 1: Kennzahlen der Epidemiologie - Relatives Risiko und Chancenverhältnis (=Odds Ratio). `http://www.kup.at/kup/pdf/5244.pdf`, 2005. [Online; Stand 15. November 2009].

[4] R. Gross; M. Löffler. *Prinzipien der Medizin.* Springer-Verlag Berlin-Heidelberg, 1997.

[5] Dr. med. Dirk Mosshammer; Professor Dr. med. Gernot Lorenz. Odds Ratio. `http://www.medizin.uni-tuebingen.de/lehre/Hp_Allgemeinmedizin/Mosshammer/Odds%20Ratio.doc`, 2009. [Online; Stand 15. November 2009].

[6] N. Bitterlich; J. Schneider. *Datenaufbereitung mittels nichtlinearer Klassifikationsverfahren zur Erhöhung der statistischen Aussagekraft von Odds-Ratio-Analysen.* Arbeitsmed. Sozialmed. Umweltmed. Vol. 42, 2007.

[7] N. Bitterlich; J. Schneider. *Odds-Ratio-Analyse in Fall-Kontroll-Studien zur Risikoabschätzung bei Studien zu Genpolymorphismen.* Arbeitsmed. Sozialmed. Umweltmed. Vol. 43, 2008.

[8] Karl Ernst v. Mühlendahl. Odds Ratio (OR) und Relatives Risiko (RR). `http://www.ecomed-medizin.de/sj/ufp/Pdf/aId/1813`, 1998. [Online; Stand 15. November 2009].

[9] J. Schneider; U. Bernges; M. Philipp; HJ. Woitowitz. *GSTM1, GSTT1 and GSTP1 polymorphism and lung cancer risk in relation to tobacco smoking.* Cancer Letters, 2004.

[10] www.klinken.de. Packungsjahr. `http://www.gmdn.de/lexikon/Medizin/Diagnostik/Packungsjahr.html`, 1997-2007. [Online; Stand 15. November 2009].

[11] S. Lange; R. Bender; A. Ziegler. Logistische Regression -Artikel Nr.14 der Statistik-Serie in der DMW-. `http://www.rbsd.de/PDF/DMW/DMW-2007-S1-14.pdf`, 2002. [Online; Stand 15. November 2009].

Selbstständigkeitserklärung

Hiermit erkläre ich, Susanne Nicklisch, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Bearbeitungsort, Datum

Unterschrift