Logistische Regression mit R

Verfasst von Paul Ruppen, Versionsdatum: 18. September 2023.

Inhaltsverzeichnis

Vorbemerkung

Die Theorie und die Beispiele entstammen weitgehend dem Buch von Hosmer und Lemeshow (2000). Alle Datensätze auf

sozio.logik.ch/Statistiksupport PH Wallis/Materialien/Logistische Regression/Datensaetze.

Die R-Befehle des i-ten Kapitels findet man in den R-Skripten H&L_r_script_Kapitel_i.r. Man findet sie unter

sozio.logik.ch/Statistiksupport PH Wallis/Materialien/Logistische Regression/r-Skripten

Eine Beschreibung der Datensätze findet man in Hosmer und Lemeshow (2000). Datensätze werden eingelesen, indem man bei den angegebenen Befehlen „pfad“ durch den Pfad des Verzeichnisses ersetzt, in dem sich die entsprechende txt-Datei befinden. Die verwendeten Daten liegen auch im R-Paket LogisticDx (Dardis, 2015) vor. Lösungen der Übungen des Buches findet man im *Solutions Manual to Accompany Applied Logistic Regression* (E. D. Cook, 2001).

1 Einleitung

In der logistischen Regression geht es darum, den Zusammenhang zwischen mindestens einer erklärenden Variable und einer erklärten Variable zu modellieren. Sie ist damit mit der linearen Regression verwandt. Die logistische Regression hat mit der linearen Regression einiges gemeinsam, unterscheidet sich aber auch von dieser. Es ist fürs Verständnis nützlich, die Gemeinsamkeiten und Unterschiede zu analysieren. Es geht bei beiden Methoden darum, ein möglichst einfaches und doch passendes Modell zu finden. Dabei sind nicht nur statistische, sondern auch inhaltliche (theoretisch fachspezifische) Gesichtspunkte zu berücksichtigen. Während bei der linearen Regression die erklärte Variable metrisch und stetig skaliert ist, liegt bei der logistischen Regression eine dichotome Variable vor, d.h. die Variable weist nur zwei Werte auf. Diese werden üblicherweise in Hinblick auf die verwendeten Modelle mit 0 und 1 kodiert. Der Zusammenhang zwischen „erklärender“ und „erklärter Variable“ ist dabei nicht irgenwie kausal zu verstehen - wie auch die Benennungen „unabhängige“ und „abhängige Variable“ keinen solchen Zusammenhang postulieren. Es handelt sich vom statistischen Standpunkt her gesehen um praktische Benennungen, die nur bedeuten, dass durch die Berücksichtigung der erklärenden oder unabhängigen Variable bei der erklärten oder abhängigen Variable Streuung wegfällt, wodurch Schätzungen präziser werden.

Beispiel 1. *Mit den Daten von Hosmer und Lemeshow (2000, S. 3) möchte man den Zusammenhang zwischen Alter (AGE) und Herzproblemen (CHD) untersuchen. Das Alter wurde metrisch erfasst, Herzprobleme liegen vor (= 1) oder nicht (=0). Macht man ein Streudiagramm mit den Daten von Hosmer und Lemeshow (2000, S. 3), so erhält man mit*

```
ac=read.table("pfad/age_cor.txt",header=T)
plot(ac$AGE,ac$CHD,xlab="Alter",ylab="CHD",main="Alter und CHD")
```

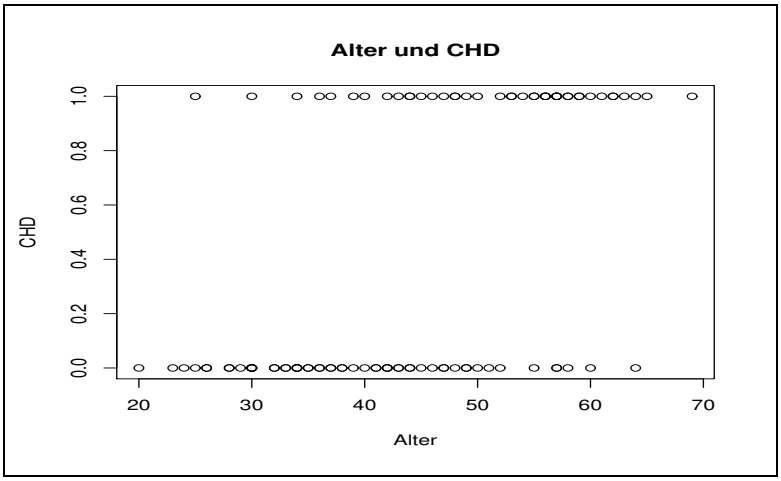

Abbildung 1: Punktediagramm der Alters- und CHD-Daten

Man erhält also die Punkte auf zwei parallelen Linien - auf 0 und auf 1. Die Punkte oben weisen eine Verschiebung nach rechts auf. Das Diagramm zeigt trotzdem nicht ein klares Bild des Zusammenhangs zwischen den beiden Variablen. Um ein deutlicheres Bild zu bekommen, kann man die Altersdaten klassifizieren - in den Daten bereits gemacht (Variable agrp, für die Klassengrenzen s. Tabelle 1). Man erhält bei dieser Klassifizierung mit

```
tab=table(ac$agrp)
names(tab)=c("[20,29]","]30-34]","]35-39]","]40-44]","]45-49]","]50-54]","]55-59]","]60-69]")
tab1=table(ac$agr,ac$CHD)
tab2=cbind(tab,tab1)
tab2=cbind(tab2,tab2[,3]/tab2[,1])
tab2=cbind(tab2,c(25,32.5,37.5,42.5,47.5,52.5,57.5,65))
colnames(tab2)=c("n","ohne CHD","mit CHD","Anteil CHD","Klassenmitten")
tab2
```

die Tabelle

Altersklassen	n	ohne CHD	mit CHD	Anteil CHD	Klassenmitten
[20, 30]	10	9	1	0.1	25
]30, 35]	15	13	2	0.1333333	32.5
]35, 40]	12	9	3	0.25	37.5
]40, 45]	15	10	5	0.3333333	42.5
]45, 50]	13	7	6	0.4615385	47.5
]50, 55]	8	3	5	0.625	52.5
]55, 60]	17	4	13	0.7647059	57.5
]60, 70]	10	2	8	0.8	65

Tabelle 1: Alters- und CHD-Daten klassifiziert

- wobei die Intervallangaben bei den obigen Befehlen als Namen der Zeilen eingeführt werden und entsprechend nicht die Beschriftung „Altersklassen“ aufweisen können und auch nicht eine Spalte der Tabelle darstellen (die erste Spalte ist also die Spalte mit der Überschrift „n“). Mit

```
plot(tab2[,5],tab2[,4],xlab="Alter",ylab="Anteile CHD",
main="Alter und Anteile CHD",ylim=c(0,1))
```

erhält man die Graphik:

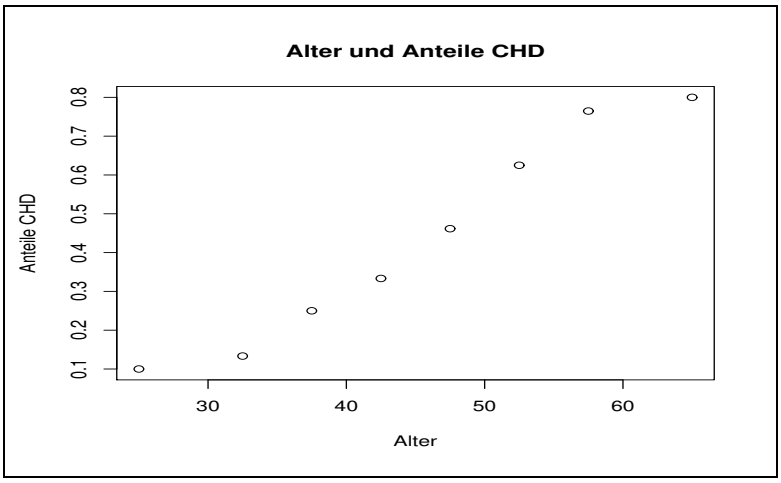

Abbildung 2: Punktediagramm der klassifizierten Alters- und CHD-Daten

Man erhält nun ein klareres Bild des Zusammenhangs der beiden Variablen: mit steigendem Alter nimmt die relative Häufigkeit der Personen mit CHD zu.

Bei allen Regressionsarten geht es darum, den Erwartungswert der erklärten Variable Y an der Stelle x der erklärenden Variable zu modellieren. Dies kann man ausdrücken mit

$$E(Y|x)$$

(Erwartungswert von Y bedingt auf x oder Erwartungswerte von Y gegeben x oder Erwartungswert von Y an der Stelle x). Man beachte, dass in der Tabelle 1 die relativen Häufigkeiten Mittelwerte der Werte der Zufallsvariablen der entsprechenden Klassen sind (man weise nach, dass bei mit 0-1-kodierten Variablen der Anteil der Einsen mit dem Mittelwert der Daten identisch ist). Diesbezüglich gibt es einen wesentlichen Unterschied zwischen der logistischen Regression und der linearen Regression. In der linearen Regression setzt man voraus, dass der Erwartungswert mit einer affinen Gleichung in x ausgedrückt werden kann, d.h.

$$E(Y|x) = \beta_0 + \beta_1 x.$$

Damit nimmt $E(Y|x)$ für $x \in \mathbb{R}$ Werte über den ganzen reellen Bereich an. Für dichotome Variablen muss allerdings

$$0 \leq E(Y|x) \leq 1$$

gelten - es werden die Werte 1 oder 0 angenommen. Die „Mittel“ müssen deshalb zwischen 0 und 1 liegen. Zudem ist die Steigung der Kurve nicht überall gleich. Wenn sich die Kurve 1 oder 0 nähert, wird die Steigung kleiner oder grösser. Betrachtet man z.B. die Wahrscheinlichkeit des Kaufs eines teuren Autos (Kauf ja oder nein) in Abhängigkeit vom Einkommen, so wird es einen Einkommensbereich geben, wo die Wahrscheinlichkeit des Kaufs durch eine Einheit zusätzliches Einkommen stark beeinflusst wird, während es Bereiche gibt, wo sich diese Wahrscheinlichkeit durch eine zusätzliche Einheit Einkommen kaum verändert - nämlich bei sehr tiefen und sehr hohen Einkommen: bei sehr tiefen

Einkommen genügt eine Einheit zusätzliches Einkommen nicht, um sich ein teures Auto kaufen zu können. Sehr hohe Einkommen haben sich bereits ein teures Auto geleistet, wenn sie sich ein solches wünschen. Die Situation entspricht einer S-förmigen Kurve und ähnelt manchen Verteilungsfunktionen wie der der Normalverteilung, wenn die Kurve steigend ist. Gewöhnlich verwendet man als Modell jedoch die logistische Funktion

$$f : \mathbb{R} \to \ ]0,1[$$
$$f(x) := \frac{\exp(x)}{1+\exp(x)},$$

deren graphische Darstellung der Verteilungsfunktion der Normalverteilung ähnelt. Die logistische Funktion wird verwendet, weil sie mathematisch gesehen flexibel sowie recht einfach zu verwenden ist, und weil sie eine einfache Interpretation erlaubt, wie wir sehen werden.

Nimmt eine Zufallsvariable Y an der Stelle x den Wert 0 oder 1 an - den Werte 1 mit einer Wahrscheinlichkeit von π, wobei π von x abhängt, so ist die Zufallsvariable Y bernoulli-verteilt und ihr Erwartungswert in x ist π. Um die Abhängigkeit von x auszudrücken, schreiben wir $\pi(x)$.

Um die Wahrscheinlichkeiten $\pi(x)$, dass die Zufallsvariable Y an der Stelle x den Wert 1 annimmt, mit Hilfe der logistischen Funktion zu modellieren, setzt man:

Definition 2.

$$\pi : \mathbb{R} \to \ ]0,1[$$
$$\pi(x) := \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$$

Praktisch gesehen ist die Definition eine Hypothese. Passt man ein solches Modell an die Daten an, so ist zu überprüfen, ob die Wahl des Modells auch geglückt ist. Ist $\beta_1 > 0$, ist die Funktion steigend, ist $\beta_1 < 0$, ist die Funktion sinkend. So nimmt die Wahrscheinlichkeit zu, dass bei grösserem Einkommen ein teures Auto gekauft wird. Die Wahrscheinlichkeit eine lange Ausbildung zu absolvieren, nimmt mit dem Alter ab. Es gilt

Theorem 3.

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0+\beta_1 x$$

Beweis.

$$\begin{aligned}
\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) &= \ln\left(\frac{\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}}{1-\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}}\right) \\
&= \ln\left(\frac{\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}}{\frac{1+\exp(\beta_0+\beta_1 x)-\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}}\right) \\
&= \ln\left(\exp(\beta_0+\beta_1 x)\right) \\
&= \beta_0+\beta_1 x
\end{aligned}$$

□

Man nennt $g(\pi(x)) := \ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$ „Logit von $\pi(x)$ in x“. Das Logit von $\pi(x)$ ist damit bei dieser Modellannahme eine affine Abbildung von x. g wird „Linkfunktion“ genannt. Sie verknüpft den Erwartungswert $\pi(x)$ mit dem Funktionswert von $\beta_0 + \beta_1 x$. Bei der linearen Regression ist die Linkfunktion die Identitätsfunktion $\mathbb{I}$, da $\mathbb{I}(E(Y|x)) = E(Y|x) = \beta_0 + \beta_1 x$ gilt.

Ein weiterer wichtiger Unterschied zwischen linearer und logistischer Regression besteht im Folgenden: in der linearen Regression stellt man jeden Wert y der erklärten Variable dar als

$$y = E(Y|x) + \varepsilon,$$

wobei ε der Fehlerterm (= Residuum) ist - die Abweichung der Beobachtung y vom Erwartungswert $E(Y|x)$ der Zufallsvariable Y an der Stelle x. Bei der linearen Regression setzt man voraus, dass die Fehlerterme ε Werte von normalverteilten Zufallsvariablen mit identischer Varianz und Erwartungswert 0 sind. Die Varianz des Fehlers $\varepsilon(x) = y - \pi(x)$ in der logistischen Regression hängt aber von der Lage von x ab und ist entsprechend nicht konstant. Es gilt nämlich

Theorem 4. *Sei Y eine bernoulli-verteilte Zufallsvariable mit dem Parameter $\pi(x)$. Für die Zufallsvariable $\mathcal{E}(x)$, welche an der Stelle x die Werte $\varepsilon(x) = y - \pi(x)$ annimmt, gilt: deren Erwartungswert $E(\mathcal{E}(x))$ ist mit 0 identisch und deren Varianz $V(\mathcal{E}(x))$ ist*

$$\pi(x)(1 - \pi(x)).$$

Beweis. (i) Der Erwartungswert von $\mathcal{E}(x)$ ist 0. Nimmt Y den Wert 1 an, ist $\varepsilon = 1 - \pi(x)$. Die Wahrscheinlichkeit, dass der Wert 1 angenommen wird, ist $\pi(x)$. Nimmt Y den Wert 0 an, ist $\varepsilon = 0 - \pi(x) = -\pi(x)$, wobei diese Wert mit einer Wahrscheinlichkeit von $1 - \pi(x)$ angenommen wird. Damit gilt gemäss Definition des Erwartungswertes: $E(\mathcal{E}(x)) = (1 - \pi(x))\pi(x) + -\pi(x)(1 - \pi(x)) = 0$.
(ii) Für die Varianz erhält man damit mit der Varianzdefinition (Von der dritten auf die vierte Zeile wird $\pi(x)(1 - \pi(x))$ ausgeklammert!). :

$$\begin{aligned} V(\mathcal{E}(x)) &= (1 - \pi(x) - 0)^2 \pi(x) + (-\pi(x) - 0)^2 (1 - \pi(x)) \\ &= (1 - \pi(x))^2 \pi(x) + (-\pi(x))^2 (1 - \pi(x)) \\ &= (1 - \pi(x))^2 \pi(x) + (\pi(x))^2 (1 - \pi(x)) \\ &= (1 - \pi(x))\pi(x)(1 - \pi(x) + \pi(x)) \\ &= (1 - \pi(x))\pi(x) \end{aligned}$$

□

Beispiel 5. *Sei $\pi(x) = 0.5$. Dann ist die Varianz von $\mathcal{E}(x)$ an der Stelle x: $0.5^2 = 0.25$. Ist $\pi(x) = 0.1$, ist die Varianz an der Stelle x: $0.1(1 - 0.1) = 0.09$*

Anpassung von logistischen Modellen

Um das Modell

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

an die Daten anzupassen, müssen die Parameter β_0 und β_1 bestimmt werden. In der linearen Regression werden die Parameter so gewählt, dass die Summe der quadrierten Abweichungen der y-Daten von der Regressionsgerade minimal wird. Diese Methode führt im Falle der linearen Regression zu Schätzern mit einer Anzahl günstiger Eigenschaften (Erwartungstreue Schätzer für β_1 sowie konsistente und effiziente Schätzer für β_i). Dies gilt aber im Falle der logistischen Regression nicht. Man verwendet deshalb eine allgemeinere Methode, die Maximum Likelihood-Methode (ML-Methode), die im Falle der linearen Regression bei normalverteilten Residuen und deren konstanter Varianz zu den Parameterschätzern führt, welche die Summe der quadrierten Abweichungen minimieren.

Die ML-Methode geht von der Wahrscheinlichkeits- oder Dichtefunktion f der Verteilung der erklärten Variable oder der Residuen aus. Die Dichte- oder Wahrscheinlichkeitsfunktion

$$f_{\boldsymbol{\beta}} : \mathbb{D} \to \mathbb{R}^+$$
$$f_{\boldsymbol{\beta}}(y_i) \in \mathbb{R}^+$$

ordnet jedem y_i einen Wert bei gegebenem Parametervektor $\boldsymbol{\beta} := (\beta_1, \beta_2)$ zu. Man geht von der gegenseitigen Unabhängigkeit der Variablen Y_i aus (diese sind zwar von x_i abhängig, nicht aber voneinander). Die gemeinsame Wahrscheinlichkeits- oder Dichtefunktion ist bei für alle Y_i identischem Parametervektor

$$\prod_{i=1}^{n} f_{\boldsymbol{\beta}}(y_i)$$

(n = Anzahl Daten y_i). Angesichts der gegebenen Daten y_i versucht man den Parametervektor $\boldsymbol{\beta}$ so zu bestimmen, dass die Erzeugung des gegebenen Datensets am plausibelsten ist. Variabel sind also die Parameter, während die y_i fix sind. Am plausibelsten sind die Daten, wenn die gemeinsame Wahrscheinlichkeits- oder Dichtefunktion bezüglich des Parametervektors $\boldsymbol{\beta}$ maximiert wird: die gemeinsame Funktion wird also so festgelegt, dass das Eintreffen der vorliegenden y_i bei dieser Funktion am wahrscheinlichsten ist. Die Funktion L, welche die Werte der Dichte- oder Wahrscheinlichkeitsfunktion in Abhängigkeit von den Parametern ausdrückt, wird Likelihood-Funktion genannt:

$$L : P \to \mathbb{R}$$
$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f_{\boldsymbol{\beta}}(y_i)$$

mit der Anzahl n der Daten und der Menge P der möglichen Parameter β oder Parametervektoren $\boldsymbol{\beta}$. Gewöhnlich wird die Likelihood-Funktion logarithmiert, da sie dann leichter

ableitbar ist. Wir nennen diese Funktion „Log-Likelihood-Funktion“ (LL-Funktion). Dies ist gerechtfertigt, da eine logarithmierte Funktion an derselben Stelle ihr Maximum annimmt wie die unlogarithmierte Funktion.

Beispiel 6. *Lineare Regression. Man geht davon aus, dass die Residuen (Fehler ε, s.o.) normalverteilt sind mit konstanter Varianz und dem Erwartungswert 0. Mit dem Modell $E(Y|x_i) = \mu_i = \beta_0 + \beta_1 x_i$, den Residuen $r_i := y_i - \mu_i$ und der Dichtefunktion*

$$f_{(\beta_0,\beta_1)}(r_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-0.5\left(\frac{r_i}{\sigma}\right)^2\right)$$

für die Residuen , erhält man

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-0.5\left(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma}\right)^2\right).$$

Man logarithmiert diese Funktion:

$$\begin{aligned} LL(\beta_0, \beta_1) &= \ln L(\beta_0, \beta_1) \\ &= \sum_{i=1}^{n} \left(\ln \frac{1}{\sigma\sqrt{2\pi}} + \ln \exp\left(-0.5\left(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma}\right)^2\right)\right) \end{aligned}$$

Der Term $\ln \frac{1}{\sigma\sqrt{2\pi}}$ verschieb die Funktion nur nach oben oder nach unten und beeinflusst die Lage des Maximums nicht. Er kann weggelassen werden. Man erhält:

$$\sum_{i=1}^{n} \left(-0.5\left(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma}\right)^2\right)$$

Der Ausdruck $\frac{0.5}{\sigma^2}$ verändert die Lage des Maximums nicht und kann weggelassen werden (beim Ableiten und Nullsetzen verschwindet der Term!). Man erhält damit die folgende Log-Likelihood-Funktion:

$$LL(\beta_0, \beta_1) := -\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{1}$$

Die Maximierung von $LL(\beta_0, \beta_1)$ erfolgt, wo $-LL(\beta_0, \beta_1)$ minimiert wird. Dies führt uns auf die Problemlage der Minimierung der Summe der quadrierten Abweichungen in der linearen Regression. Diese Methode ist also ein Spezialfall der ML-Methode. Man sieht durch die obige Herleitung auch, dass die Konstanz der Varianz σ der Residuen entscheidend ist. Im Modell ohne $\beta_1$ist mittels (1)

$$LL(\beta_0) := -\sum_{i=1}^{n} (y_i - \beta_0)^2$$

Durch Ableiten und Nullsetzen gilt:

$$\frac{d}{d\beta_0} - \sum_{i=1}^{n} \left(y_i - \beta_0\right)^2 = -\sum_{i=1}^{n} 2\left(y_i - \beta_0\right)(-1) = 0$$

und damit

$$\sum_{i=1}^{n} y_i - n\beta_0 = 0$$

$$\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Das arithmetische Mittel der y-Daten ist damit der Maximum-Likelihood-Schätzer von β_0 im Modell ohne β_1.

Im Falle der logistischen Regression ist die Wahrscheinlichkeitsfunktion (W-Funktion) für die einzelnen Zufallsvariable Y_i (Bernoulli-Verteilung):

$$P\left(Y_i = y_i\right) = \pi\left(x_i\right)^{y_i}\left(1 - \pi\left(x_i\right)\right)^{1-y_i}.$$

Wegen der vorausgesetzten Unabhängigkeit ist die Gemeinsame W-Funktion

$$\prod_{i=1}^{n} P\left(Y_i = y_i\right) = \prod_{i=1}^{n}\left(\pi\left(x_i\right)^{y_i}\left(1 - \pi\left(x_i\right)\right)^{1-y_i}\right).$$

Durch Logarithmieren erhält man die Log-Likelihoodfunktion (LL-Funktion)

$$\begin{aligned} LL\left(\beta_0, \beta_1\right) &= \sum_{i=1}^{n}\left(y_i \ln \pi\left(x_i\right) + \left(1 - y_i\right)\ln\left(1 - \pi\left(x_i\right)\right)\right) \\ &= \sum_{i=1}^{n}\left(y_i \ln \frac{\exp\left(\beta_0 + \beta_1 x_i\right)}{1 + \exp\left(\beta_0 + \beta_1 x_i\right)} + \left(1 - y_i\right)\left(1 - \ln \frac{\exp\left(\beta_0 + \beta_1 x_i\right)}{1 + \exp\left(\beta_0 + \beta_1 x_i\right)}\right)\right) \end{aligned} \tag{2}$$

An der Nullstelle $\hat{\boldsymbol{\beta}}$ der ersten Ableitungen der LL-Funktion liegt ein Maximum, sofern die LL-Funktion überall konkav ist und diese nicht überall monoton steigend ist.

Theorem 7. *Die Nullstelle $\hat{\boldsymbol{\beta}}$ der Ableitungen der LL-Funktion ist der Lösungsvektor der folgenden, sogenannten Likelihood-Gleichungen:*

$$\begin{aligned} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi\left(x_i\right) &= 0 \\ \sum_{i=1}^{n} x_i\left(y_i - \pi\left(x_i\right)\right) &= 0 \end{aligned} \tag{3}$$

Beweis.

$$\begin{aligned}\prod_{i=1}^{n}\left(\pi\left(x_i\right)^{y_i}\left(1-\pi\left(x_i\right)\right)^{1-y_i}\right) &= \prod_{i=1}^{n}\left(\pi\left(x_i\right)^{y_i}\left(1-\pi\left(x_i\right)\right)\left(1-\pi\left(x_i\right)\right)^{-y_i}\right)\\ &= \prod_{i=1}^{n}\left(\left(\frac{\pi\left(x_i\right)}{1-\pi\left(x_i\right)}\right)^{y_i}\left(1-\pi\left(x_i\right)\right)\right)\\ &= \prod_{i=1}^{n}\left(\exp\left(y_i\ln\left(\frac{\pi\left(x_i\right)}{1-\pi\left(x_i\right)}\right)\right)\left(1-\pi\left(x_i\right)\right)\right)\end{aligned}$$

Ersetzt man $\ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)$ durch $\beta_0+\beta_1 x_i$ und bildet man die Likelihood-Funktion bezüglich dieser Formulierung der W-Funktion, erhält man mit

$$\begin{aligned}1-\pi\left(x_i\right) &= 1-\frac{\exp\left(\beta_0+\beta_1 x_i\right)}{1+\exp\left(\beta_0+\beta_1 x_i\right)}\\ &= \frac{1+\exp\left(\beta_0+\beta_1 x_i\right)-\exp\left(\beta_0+\beta_1 x_i\right)}{1+\exp\left(\beta_0+\beta_1 x_i\right)}\\ &= \frac{1}{1+\exp\left(\beta_0+\beta_1 x_i\right)}\\ &= \left(1+\exp\left(\beta_0+\beta_1 x_i\right)\right)^{-1}:\end{aligned}$$

$$L\left(\boldsymbol{\beta}\right)=\prod_{i=1}^{n}\left(\exp\left(y_i\left(\beta_0+\beta_1 x_i\right)\right)\left(1+\exp\left(\beta_0+\beta_1 x_i\right)\right)^{-1}\right).$$

Durch Logarithmieren erhält man die folgende alternative Formulierung der LL-Funktion:

$$\begin{aligned}LL\left(\boldsymbol{\beta}\right):=\ln L\left(\boldsymbol{\beta}\right) &= \sum_{i=1}^{n}\left(\ln\exp\left(y_i\left(\beta_0+\beta_1 x_i\right)\right)+\ln\left(\left(1+\exp\left(\beta_0+\beta_1 x_i\right)\right)^{-1}\right)\right)\\ &= \sum_{i=1}^{n}y_i\left(\beta_0+\beta_1 x_i\right)-\sum_{i=1}^{n}\ln\left(\left(1+\exp\left(\beta_0+\beta_1 x_i\right)\right)\right)\end{aligned}$$

Damit gilt dann durch Ableiten

$$\begin{aligned}\frac{\partial}{\partial\beta_0}LL\left(\boldsymbol{\beta}\right) &= \sum_{i=1}^{n}y_i-\sum_{i=1}^{n}\frac{\exp\left(\beta_0+\beta_1 x_i\right)}{1+\exp\left(\beta_0+\beta_1 x_i\right)}\\ &= \sum_{i=1}^{n}y_i-\sum_{i=1}^{n}\pi\left(x_i\right)\end{aligned}$$

und

$$\begin{aligned}\frac{\partial}{\partial\beta_1}LL\left(\boldsymbol{\beta}\right) &= \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \frac{\exp\left(\beta_0+\beta_1 x_i\right) x_i}{1+\exp\left(\beta_0+\beta_1 x_i\right)} \\ &= \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \pi\left(x_i\right) x_i \\ &= \sum_{i=1}^{n} x_i \left(y_i - \pi\left(x_i\right)\right).\end{aligned}$$

Durch Nullsetzen erhält man die erwähnten Likelihood-Gleichungen:

$$\begin{aligned}\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi\left(x_i\right) &= 0 \\ \sum_{i=1}^{n} x_i \left(y_i - \pi\left(x_i\right)\right) &= 0\end{aligned}$$

□

Die zweiten Ableitungen der LL-Funktion sind:

$$\begin{aligned}\frac{\partial^2}{\partial\beta_0\beta_0}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{\exp\left(\beta_0+\beta_1 x_i\right)}{1+\exp\left(\beta_0+\beta_1 x_i\right)}\right) &= -\sum_{i=1}^{n} \pi\left(x_i\right)\left(1-\pi\left(\mathbf{x}_i\right)\right) \\ \frac{\partial^2}{\partial\beta_0\beta_1}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{\exp\left(\beta_0+\beta_1 x_i\right)}{1+\exp\left(\beta_0+\beta_1 x_i\right)}\right) &= -\sum_{i=1}^{n} \pi\left(\mathbf{x}_i\right)\left(1-\pi\left(\mathbf{x}_i\right)\right) x_i \\ \frac{\partial^2}{\partial\beta_1\beta_0}\left(\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \frac{\exp\left(\beta_0+\beta_1 x_i\right) x_i}{1+\exp\left(\beta_0+\beta_1 x_i\right)}\right) &= -\sum_{i=1}^{n} \pi\left(\mathbf{x}_i\right)\left(1-\pi\left(\mathbf{x}_i\right)\right) x_i \\ \frac{\partial^2}{\partial\beta_1\partial\beta_1}\left(\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \frac{\exp\left(\beta_0+\beta_1 x_i\right) x_i}{1+\exp\left(\beta_0+\beta_1 x_i\right)}\right) &= -\sum_{i=1}^{n} \pi\left(\mathbf{x}_i\right)\left(1-\pi\left(\mathbf{x}_i\right)\right) x_i^2\end{aligned}$$

Man kann zeigen, dass unter gewissen Bedingungen die LL-Funktion konkav und nicht überall monoton ist:

- Der Rang der Matrix $(\mathbf{1}, \mathbf{x})$ ist 2 (d.h. die beiden Vektoren $\mathbf{1}$ und $\mathbf{x}$ sind linear unabhängig).
- Die Daten dürfen nicht getrennt sein, d.h. es darf kein Datum geben, so dass unter diesem Datum für alle x_i die y-Daten denselben Wert aufweisen und von diesem x weg für alle x_i die y-Daten den anderen aber jeweils identischen Wert aufweisen.
- Die y-Daten dürfen nicht alle denselben Wert annehmen (alle 1 oder alle 0)

Gemäss (3) gilt im Maximum

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \pi\left(x_i\right)$$

d.h. die Summe der Werte y_i ist identisch mit der Summe der Wahrscheinlichkeiten, die den x_i zugeordnet sind. Während im Falle der linearen Regression die Ableitungen von $\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$ nach β_0 und β_1 und deren Nullsetzung auf einfache lineare Gleichungen führen, die leicht aufgelöst werden können, sind die Likelihood-Gleichungen nicht linear in β_1 und β_0. Ausgeschrieben werden sie ja mittels

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{\exp\left(\beta_0 + \beta_1 x_i\right)}{1 + \exp\left(\beta_0 + \beta_1 x_i\right)} = 0$$
$$\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \frac{\exp\left(\beta_0 + \beta_1 x_i\right) x_i}{1 + \exp\left(\beta_0 + \beta_1 x_i\right)} = 0$$

dargestellt. Die Lösung $\hat{\boldsymbol{\beta}}$ dieser Gleichungen wird mittels eines Näherungsverfahrens berechnet (s. Agresti (2013, S. 143 ff.)). $\hat{\boldsymbol{\beta}}$ ist der ML-Schätzer des tatsächlichen $\boldsymbol{\beta}$. Die aus der Schätzung von $\hat{\pi}\left(x_i\right) := \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)}$ gewonnen geschätzten Wahrscheinlichkeiten sind die ML-Schätzer für die tatsächlichen Wahrscheinlichkeiten $\pi\left(x_i\right)$. Im obigen Beispiel erhält man mit R - mit den Variablen AGE (erklärende) und CHD (erklärte)

```
ac=read.table("pfad/age_cor.txt",header=T)
m1=glm(CHD~AGE,data=ac,family="binomial"); summary(m1))
```

die folgenden Grössen :

Variable	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-5.30945	1.13365	-4.683	2.82e-06
AGE	0.11092	0.02406	4.610	4.02e-06

Tabelle 2: R-Ausgabe für die Alter-CHD-Daten

d.h. $\hat{\boldsymbol{\beta}} = (-5.30945, 0.11092)$. $\hat{\pi}\left(25\right) = \frac{\exp(-5.30945+0.11092\cdot 25)}{1+\exp(-5.30945+0.11092\cdot 25)} = 0.07334\,2$. Dies ist die geschätzte Wahrscheinlichkeit, als 25-jähriger CHD zu haben. Demgegenüber ist z.B. $\hat{\pi}\left(75\right) = \frac{\exp(-5.30945+0.11092\cdot 75)}{1+\exp(-5.30945+0.11092\cdot 75)} = 0.953$.

Die geschätzte Kurve ergibt in der obigen Abbildung 2:

```
plot(tab2[,5],tab2[,4],xlab="Alter",ylab="Anteile CHD",
main="Alter und Anteile CHD",ylim=c(0,1))
m1=glm(CHD~AGE,data=ac,family="binomial")
a=m1$coefficients["AGE"]; b=m1$coefficients["(Intercept)"]
f=function(x) exp(b+a*x)/(1+exp(b+a*x))
curve(f,add=T)
```

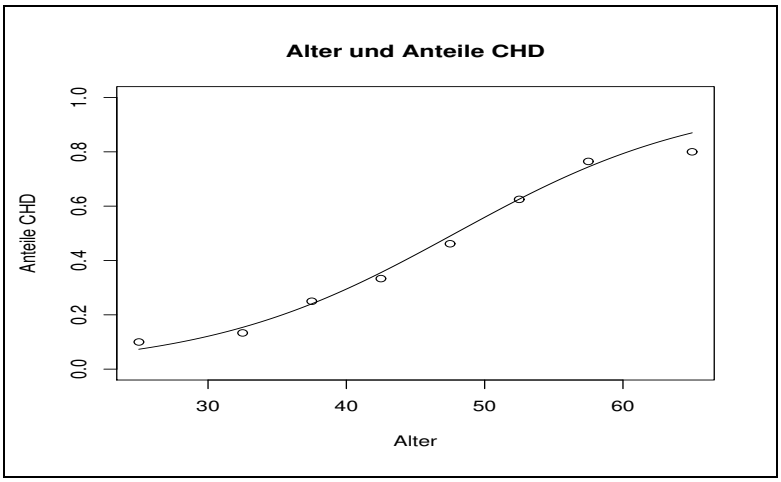

Abbildung 3: Punktediagramm der Alters- und CHD-Daten mit geschätzter logistischer Regressionskurve

Signifikanztests

Beim Testen geht es um die Frage, ob ein Modell mit der erklärenden Variable uns mehr über die erklärte Variable aussagt als ein Modell ohne die erklärende Variable. Die Antwort erhält man, indem man geeignete Kennzahlen der beiden Modelle vergleicht, wofür es verschiedene Möglichkeiten gibt.

1. Bei der linearen Regression wird die Varianzanalyse (ANOVA) durchgeführt. Man betrachtet die Summe
$$SQT := \sum_{i=1}^{n} (y_i - \bar{y})^2$$
der quadrierten Abweichungen der y-Daten vom Mittelwert $\bar{y}$ der erklärten Variable $\mathbf{y}$. SQT ist die mit $n-1$ multiplizierte Varianz der Daten der erklärten Variable. Diese Streuung teilt man auf in die Summe SQR der quadrierten Residuen und die Summe SQE der quadrierten Abweichungen der Werte auf der Regressionsgerade vom Mittelwert der y-Daten. Es gilt
$$\begin{array}{lcl} SQT & = & SQR + SQE \\ \sum_{i=1}^{n} (y_i - \bar{y})^2 & = & \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \end{array}$$
mit $\hat{y}_i = \hat{a} + \hat{b}x_i$. Wenn SQE in Bezug auf SQT gross ist (und damit SQR klein) ist die Aufnahme des Steigungskoeffizienten für die Reduktion der Streuung nützlich. Im Modell ohne erklärende Variable ist $\beta_0 = \bar{y}$. Bei der logistischen Regression vergleicht man hingegen Maxima der optimal an die Daten angepassten Log-Likelihoodfunktionen.

2. Man testet, ob die Koeffizienten von 0 verschieden sind (in der linearen Regression t-Tests, in der logistischen Regression Waldtests).

Diese beiden Verfahren werden nun bezüglich der logistischen Regression näher angeschaut. Man betrachtet die LL-Funktion und geht vom gesättigten Modell aus, das die Daten vollkommen vorausgesagt oder erklärt. Das gesättigte Modell hat so viele Parameter wie Daten (= n). Im Falle der logistischen Regression ist das gesättigte Likelihood-Modell für $y_i \in \{0, 1\}$ und $0^0 = 1$

$$L_{gesättigtes\ Modell} = \prod_{i=1}^{n} y_i^{y_i} \left(1 - y_i\right)^{1-y_i} = \prod_{i=1}^{n} 1 = 1$$

und entsprechend die logarithmierte Version

$$LL_{gesättigtes\ Modell} = 0.$$

Entsprechend ist das Maximum der LL-Funktion des gesättigten Modells 0, d.h.

$$\max\left(LL_{gesättigtes\ Modell}\right) = 0.$$

Dies vergleicht man mit dem Maximum der optimal an die Daten angepassten LL-Funktion des Modells mit den geschätzten Parametern $\hat{\beta}_0, \hat{\beta}_1$

$$\begin{aligned}
&\max\left(LL_{geschätztes\ Modell}\right) \\
&= LL\left(\hat{\beta}_0, \hat{\beta}_1\right) \\
&= \sum_{i=1}^{n} \left(y_i \ln \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)} + \left(1 - y_i\right) \left(1 - \ln \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)} \right) \right).
\end{aligned}$$

Es gilt

$$\max\left(LL_{geschätztes\ Modell}\right) < 0$$

Beweis. Wegen $0 < \pi\left(x_i\right) < 1$, ist $0 < P\left(Y_i = y_i\right) = \pi\left(x_i\right)^{y_i} \left(1 - \pi\left(x_i\right)^{1-y_i}\right) < 1$, und damit $0 < \prod_{i=1}^{n} P\left(Y_i = y_i\right) < 1$. Für $\ln \prod_{i=1}^{n} P\left(Y_i = y_i\right)$ gilt damit $\ln \prod_{i=1}^{n} P\left(Y_i = y_i\right) < 0$, womit auch $\max\left(LL_{geschätztes\ Modell}\right) < 0$ □

Je grösser $\max\left(LL_{geschätztes\ Modell}\right)$, desto näher liegt $\max\left(LL_{geschätztes\ Modell}\right)$ beim Maximum des gesättigten Modells. Entsprechend erklärt das Modell mit der erklärenden Variable die Daten umso besser, je grösser $\max\left(LL_{geschätztes\ Modell}\right)$ ist. Da das Modell mit der erklärenden Variable viel weniger Parameter hat als das gesättigte Modell, ist es ökonomischer. Die mit -2 multiplizierte Abweichung $\max\left(LL_{geschätztes\ Modell}\right)$ von 0

$$-2\max\left(LL_{geschätztes\ Modell}\right) = -2\left(\max\left(LL_{geschätztes\ Modell}\right) - \max\left(LL_{gesättigtes\ Modell}\right)\right)$$

wird „residuale oder restliche Abweichung“ (Residual Deviance) genannt. Sie kann von der Idee her mit SQR bei der linearen Regression verglichen werden. Die restliche Abweichung weist $n-p$ Freiheitsgrade auf (n ist die Anzahl der Daten, p ist die Anzahl der Parameter des Modells).

Üblicher Weise vergleicht man das Modell, das nur die Konstante aufweist (ohne_Var), mit dem Modell, dass eine erklärende Variable enthält (mit_Var). Für die Differenz der Maxima der entsprechenden Log-Likelihood-Funktionen gilt

$$\max\left(LL_{ohne\_Var}\right) - \max\left(LL_{mit\_Var}\right) < 0,$$

da das Modell ohne Variable schlechter passen wird als das mit Variable – d.h. $\max\left(LL_{ohne\_Var}\right) < \max\left(LL_{mit\_Var}\right) < 0$. Diese Differenz wird mit -2 multipliziert, um den Wert einer Statistik zu erhalten, deren Verteilung bekannt ist:

$$\begin{aligned} G &:= -2\left(\max\left(LL_{ohne\_Var}\right) - \max\left(LL_{mit\_Var}\right)\right) \\ &= -2\left(\max\left(LL_{ohne\_Var}\right) - \max\left(LL_{gesättigtes\ Modell}\right)\right) - \\ &\quad 2\left(\max\left(LL_{mit\_Var}\right) - \max\left(LL_{gesättigtes\ Modell}\right)\right) \end{aligned}$$

Wilks (1938) zeigte, dass die Statistik G, welche die Werte $\hat{G}$ annimmt, näherungsweise χ^2-verteilt mit $p-q$ Freiheitsgraden ist (p Anzahl der Freiheitsgrade der restlichen Abweichung des Modells mit Variable und q die Anzahl der Freiheitsgrade der restlichen Abweichung des Modells ohne Variable). G wird Devianz oder Abweichung (deviance) genannt.

Bemerkung 8.

$$\begin{aligned} G &= -2\left(\max\left(LL_{ohne\_Var}\right) - \max\left(LL_{mit\_Var}\right)\right) = -2\ln\left(\frac{\max\left(L_{ohne\_Var}\right)}{\max\left(L_{mit\_Var}\right)}\right) \\ &= -2\ln\left(\frac{\frac{\max(L_{ohne\_Var})}{\max\left(LL_{gesättigtes\ Modell}\right)}}{\frac{\max(L_{mit\_Var})}{\max\left(LL_{gesättigtes\ Modell}\right)}}\right) \end{aligned}$$

$\frac{\max(L_{ohne\_Var})}{\max(L_{mit\_Var})}$ *wird Likelihood-Verhältnis (Likelihood-Quotient, likelihood ratio) genannt.*

Theorem 9. *Mit* $n_1 := \sum_{i=1}^{n} y_i$ *gilt für den Fall der logistischen Regression*

$$\max\left(LL_{ohne\_Var}\right) = n_1 \ln\frac{n_1}{n} + (n-n_1)\ln\left(1-\frac{n_1}{n}\right)$$

Beweis. Ohne die erklärende Variable ist das Modell $\pi\left(x_i\right) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$ und es gilt $\ln\left(\frac{\pi(x_i)}{1+\pi(x_i)}\right) = \beta_0$. Man erhält mit (s. o.)

$$1-\pi\left(x\right) = 1 - \frac{\exp\left(\beta_0\right)}{1+\exp\left(\beta_0\right)} = \left(1+\exp\left(\beta_0\right)\right)^{-1}$$

$$\begin{aligned}\prod_{i=1}^{n}\left(\pi\left(x_i\right)^{y_i}\left(1-\pi\left(x_i\right)\right)^{1-y_i}\right) &= \prod_{i=1}^{n}\frac{\pi\left(x_i\right)^{y_i}}{1-\pi\left(x_i\right)^{y_i}}\left(1-\pi\left(x_i\right)\right)\\ &= \prod_{i=1}^{n}\exp\left(y_i\ln\frac{\pi\left(x_i\right)}{1-\pi\left(x_i\right)}\right)\left(1-\pi\left(x_i\right)\right)\\ &= \prod_{i=1}^{n}\exp\left(y_i\beta_0\right)\left(1+\exp\left(\beta_0\right)\right)^{-1}.\end{aligned}$$

Durch Logarithmieren erhält man:

$$LL\left(\beta_0\right) := \ln L\left(\beta_0\right) = \sum_{i=1}^{n} y_i\beta_0 - \sum_{i=1}^{n}\ln(1+\exp\left(\beta_0\right)$$

Ableiten nach β_0 ergibt

$$\begin{aligned}\frac{d}{d\beta_0}LL\left(\beta_0\right) &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\frac{\exp\beta_0}{1+\exp\left(\beta_0\right)}\\ &= \sum_{i=1}^{n} y_i - n\frac{\exp\beta_0}{1+\exp\left(\beta_0\right)}.\end{aligned}$$

Durch Nullsetzen und mit $n_1 = \sum_{i=1}^{n} y_i$ (Anzahl Einer in erklärter Variable) ergibt sich:

$$\frac{n_1}{n} = \frac{\exp\beta_0}{1+\exp\beta_0}$$

und damit

$$\begin{aligned}\frac{n_1}{n} + \frac{n_1}{n}\exp\beta_0 &= \exp\beta_0\\ n_1 + n_1\exp\beta_0 &= n\exp\beta_0\\ n_1 &= n\exp\beta_0 - n_1\exp\beta_0\\ n_1 &= \exp\beta_0\left(n-n_1\right)\\ \frac{n_1}{n-n_1} &= \exp\beta_0\\ \beta_0 &= \ln\frac{n_1}{n-n_1}.\end{aligned}$$

Damit erhält man (s. erste Zeile von 2 auf Seite 10)

$$\begin{aligned}\max\left(LL_{ohne\_Var}\right) &= \textstyle\sum_{i=1}^{n}\left[y_i\ln\pi\left(x_i\right) + \left(1-y_i\right)\ln\left(1-\pi\left(x_i\right)\right)\right]\\ &= \textstyle\sum_{i=1}^{n}\left(y_i\ln\frac{\exp\beta_0}{1+\exp\beta_0} + \left(1-y_i\right)\ln\left(1-\frac{\exp\beta_0}{1+\exp\beta_0}\right)\right)\\ &= \textstyle\sum_{i=1}^{n}\left(y_i\ln\frac{\frac{n_1}{n-n_1}}{1+\frac{n_1}{n-n_1}} + \left(1-y_i\right)\ln\left(1-\frac{\frac{n_1}{n-n_1}}{1+\frac{n_1}{n-n_1}}\right)\right).\end{aligned}$$

Mit

$$\frac{\frac{n_1}{n-n_1}}{1+\frac{n_1}{n-n_1}} = \frac{\frac{n_1}{n-n_1}}{\frac{n-n_1+n_1}{n-n_1}} = \frac{\frac{n_1}{n-n_1}}{\frac{n}{n-n_1}} = \frac{n_1}{n-n_1}\frac{n-n_1}{n} = \frac{n_1}{n}$$

erhält man

$$\begin{aligned}\max\left(LL_{ohne\_Var}\right) &= \sum_{i=1}^{n}\left(y_i \ln\frac{n_1}{n} + (1-y_i)\ln\left(1-\frac{n_1}{n}\right)\right)\\ &= n_1 \ln\frac{n_1}{n} + (n-n_1)\ln\left(1-\frac{n_1}{n}\right).\end{aligned}$$

□

Der vorausgesagte Wert π im Modell ohne die erklärende Variable ist damit $\frac{n_1}{n} = \frac{\exp\beta_0}{1+\exp\beta_0}$, das arithmetische Mittel der y-Daten.

Beispiel 10. *Im Alter-CHD-Beispiel ergibt sich mit den Parametern* $(\beta_0, \beta_1) = (-5.30945, 0.11092)$, *über denen das Maximum liegt:*

$$\begin{aligned}&\max\left(LL_{mit\_Var}\right)\\ &= \sum_{i=1}^{100}\left(y_i \ln\frac{\exp\left(-5.30945+0.11092x_i\right)}{1+\exp\left(-5.30945+0.11092x_i\right)} + (1-y_i)\ln\frac{\exp\left(-5.30945+0.11092x_i\right)}{1+\exp\left(-5.30945+0.11092x_i\right)}\right)\\ &= -53.67655.\end{aligned}$$

und damit die residuale Devianz $-2\cdot -53.67655 = 107.35$.
Mit R (für die Definition von m1 s. Tabelle 2 auf Seite 13):

```
b=m1$coefficients["(Intercept)"];a=m1$coefficients["AGE"]
MLL=ac$CHD*log(exp(b+a*ac$AGE)/ (1+exp(b+a*ac$AGE)))+
(1-ac$CHD)*log(1-(exp(b+a*ac$AGE)/(1+exp(b+a*ac$AGE))))
sum(MLL)
```

Fürs Modell ohne erklärende Variable erhält man gemäss Satz 9 mit $n_1 = 43$ und $n = 100$:

$$\begin{aligned}\max\left(\widehat{LL_{ohne\_Var}}\right) &= 43\ln\frac{43}{100} + (100-43)\ln\left(1-\frac{43}{100}\right)\\ &= -68.331.\end{aligned}$$

und damit die residuale Devianz $-2\cdot -68.331 = 136.66$. Wie zu erwarten, ist der Abstand des Maximums der LL-Funktion des Modells ohne erklärende Variable von 0 grösser als der Abstand des Maximums der LL-Funktion des Modells mit der Variable: das Modell mit Variable passt besser als das Modell ohne Variable. Zu testen bleibt, ob der Unterschied der beiden Modelle statistisch signifikant ist. Es ergibt sich die geschätzte Devianz

$$\hat{G} = 136.66 - 107.35 = -2\left(-68.331 - (-53.67655)\right) = 29.309.$$

Die Statistik G ist χ^2-verteilt mit $(100-1)-(100-2)=2-1=1$ Freiheitsgraden. Man erhält den p-Wert 6.17084E-08.
Mit R erfolgt durch (für die Definition von m1 s. S. 13)

```
anova(m1,test="Chisq")
```

die Ausgabe:

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			99	136.66	
AGE	1	29.31	98	107.35	6.168e-08 ***

Bei NULL stehen die Werte für das Modell, das nur die Konstante enthält. Unter Resid. Dev steht der Wert der restlichen Abweichung - der mit -2 multiplizierte Wert des Maximums der LL-Funktion des Modells ohne erklärende Variable ($-2 \cdot -68.331 = 136.66$) - und des Modells mit erklärender Variable ($-2 \cdot -53.67655 = 107.35$). Unter Deviance steht der Wert der Statistik G - also der Differenz der restlichen Abweichungen der beiden Modelle - und unter *df* deren Anzahl Freiheitsgrade (99 - 98=1). Es liegen 100 Daten vor. Die Freiheitsgrade von NULL sind also 100-1 (das Modell, das nur die Konstante enthält, weist einen Parameter auf) und die Freiheitsgrade von AGE sind 100-2 (das Modell, das die erklärende Variable enthält, weist zwei Parameter auf). Der p-Wert ist kleiner als 0.05. Die Variable Alter ist damit statistisch bedeutsam für die Erklärung von CHD.

Bemerkung 11. *SPSS: Unter „Omnibus-Tests der Modellkoeffizienten“ und bei „Modell“ steht die Devianz (29.310) und deren Freiheitsgrade (1). Die restliche Devianz des Modells mit Variable steht unter „Modellzusammenfassung“ und „-2 Log-Likelihood“ (107.353). Die übrigen Werte werden nicht geliefert, können aber aus diesen Angaben mit Hilfe der obigen Theorie leicht berechnet werden (die Restliche Abweichung des Modells ohne Variable sowie die Maxima der LL-Funktionen). Die Koeffizienten der logistischen Regression samt Waldstatistiken, p-Werten und ORs stehen in der Tabelle unter „Variablen in der Gleichung“.*

Bemerkung 12. *Im Falle der linearen Regression ist (s. Bemerkung 6):*

$$\begin{aligned} G &= -2\left(\max\left(LL_{ohne\_Var}\right) - \max\left(LL_{mit\_Var}\right)\right) \\ &= -2\left(-\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 - \left(-\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2\right)\right) \\ &= -2\left(-SQT-(-SQR)\right) \\ &= 2\left(SQT-SQR\right) \\ &= 2SQE \end{aligned}$$

Die Beobachtung mag die Berechnungen für den Fall der logistischen Regression motivieren: im Falle der linearen Regression würde man beim Testen von G die Abweichung der erklärten Streuung von 0 testen. Die durch das Modell mit Steigungskoeffizient erklärte Streuung ist die Differenz zwischen der restlichen Streuung ohne Steigungskoeffizient (= SQT) und der restlichen Streuung mit Steigungskoeffizient (= SQR).

Wir kommen nun zur zweiten Methode, die den t-Tests für die Koeffizienten in der linearen Regression entspricht. In der logistischen Regression wird die Wald-Statistik verwendet:

$$W_i = \frac{B_i}{SE\left(B_i\right)}$$

wobei $SE\left(B_i\right)$ die Standardabweichung (= Standardfehler) der Schätzstatistik B_i ist, welche die Werte $\hat{\beta}_i$ annimmt. Man testet, ob der Koeffizient von 0 verschieden ist. Diese Statistik ist näherungsweise standardnormalverteilt (Wald, 1943) und deren Quadrat näherungsweise chi-quadrat-verteilt mit einem Freiheitsgrad. Die geschätzte Standardabweichung wird in der R-Ausgabe (s. Tabelle 2, S. 13) unter „Std. Error“ angegeben. Damit ergibt sich fürs Beispiel

$$w_0 = \frac{-5.30945}{1.13365} = -4.683\,5$$
$$w_1 = \frac{0.11092}{0.02406} = 4.610\,1$$

Beide Abweichungen von 0 sind hochsignifikant (man vergleiche mit der R-Ausgabe von Tabelle 2, S. 13). Der Wald-Test wird in der Literatur nicht empfohlen, da er instabil sein kann. Es ist besser, den Likelihood-Verhältnis-Test oder beide Tests zu verwenden. Die Wald-Statistik erlaubt eine einfache Berechnung von Vertrauensintervallen, wie wir gleich sehen werden.

Neben den beiden diskutierten Testverfahren gibt es in der Literatur noch weitere Tests, z.B. den sogenannten Scoretest, der hier nicht diskutiert wird (s. Hosmer und Lemeshow (2000, S. 17) und Agresti (2013, S. 10)).

Schätzen von Vertrauensintervallen

Wie bei der linearen Regression kann man Vertrauensintervalle $VI\left(B_i\right)$ für die Zufallsvariablen B_i, welche die Werte $\hat{\beta}_i$ annehmen, sowie für die geschätzten Werte auf der Regressionskurve berechnen ($SE\left(B_i\right)$ für die Standardabweichung der Zufallsvariable B_i, $V\left(B_i\right)$ für die Varianz der Zufallsvariable B_i). Mit Hilfe der Wald-Statistik erhält man das α-Vertrauensintervall

$$VI\left(B_i\right) = B_i \pm \Phi^{-1}\left(\alpha + \frac{1-\alpha}{2}\right) SE\left(B_i\right)$$

mit der Wahrscheinlichkeit α, dass B_i ins Intervall

$$\left]B_i - \Phi^{-1}\left(\alpha + \frac{1-\alpha}{2}\right) SE\left(B_i\right), B_i + \Phi^{-1}\left(\alpha + \frac{1-\alpha}{2}\right) SE\left(B_i\right)\right[$$

fällt und mit der Verteilungsfunktion Φ der Standardnormalverteilung.

Die Varianz für die Statistiken $g(\Pi(x)) = B_0 + B_1 x$, mit den Statistiken $\Pi(x)$, welche die Wahrscheinlichkeiten $\hat{\pi}(x)$ als Werte annehmen, wird berechnet mittels

$$\begin{aligned} V(g(\Pi(x))) &= V(B_0 + B_1 x) \\ &= V(B_0) + V(xB_1) + 2KOV(B_0, xB_1) \\ &= V(B_0) + x^2 V(B_1) + 2xKOV(B_0, B_1) \end{aligned}$$

und damit das α-VI in Abhängigkeit von x

$$VI(g(\Pi(x))) = g(\Pi(x)) \pm \Phi^{-1}\left(\alpha + \frac{1-\alpha}{2}\right) SE(g(\Pi(x)))$$

Für $\Pi(x) = \frac{\exp(B_0+B_1 x)}{1+\exp(B_0+B_1 x)} = \frac{\exp(g(\Pi(x)))}{1+\exp(g(\Pi(x)))}$ erhält man die α-VIs:

$$VI(\Pi(x)) = \frac{\exp\left[g(\Pi(x)) \pm \Phi^{-1}\left(\alpha + \frac{1-\alpha}{2}\right) SE(g(\Pi(x)))\right]}{1 + \exp\left[g(\Pi(x)) \pm \Phi^{-1}\left(\alpha + \frac{1-\alpha}{2}\right) SE(g(\Pi(x)))\right]}$$

Beispiel 13. *Man erhält fürs obige Alter-CHD-Beispiel die Grenzen des 0.95-VIs mittels (s. für die Zahlenwerte Tabelle 2 auf Seite 13. Die Standardabweichungen können auch aus der gleich gelieferten Kovarianzmatrix (s. Tabelle 3) berechnet werden - man zieht die Wurzel aus den Varianzen, d.h. den Einträgen auf der Hauptdiagonalen).*

$$\begin{aligned} &-5.30945 \pm 1.96 \cdot 1.13365 \\ &0.11092 \pm 1.96 \cdot 0.02406 \end{aligned}$$

und damit die 0.95-VIs für $\hat{\beta}_0$ *und* $\hat{\beta}_1$ *:* $]-7.531\,4, -3.087\,5[$ *und* $]0.063\,762, 0.158\,08[$
Mit R kann man 0.95-VIs berechnen, die auf der sichereren Likelihood-Quotienten-Methode (Profile-Likelihood-VIs, s. Agresti (2013, S. 79)) basieren. Mit

```
confint(m1)
```

erhält man (für die Definition von m1 s. S. 13)

```
Waiting for profiling to be done...
                    2.5 %       97.5 %
(Intercept)     -7.72587162  -3.2461547
AGE              0.06693158   0.1620067
```

Alternative α-VIs durch

```
confint(m1,level=α)
```

.

Die geschätzte Kovarianzmatrix von B_0 und B_1 wird von R geliefert durch

```
vcov(m1)
```

.

Man erhält:Die geschätzte Varianz der Logits in x ist:

	(Intercept)	AGE
(Intercept)	1.28517059	-0.0266769747
AGE	-0.02667697	0.0005788748

Tabelle 3: Kovarianzmatrix des Modells Alter-CHD

$$\hat{V}\left(g\left(\hat{\pi}\left(x\right)\right)\right) = 1.28517059 + 0.0005788748x^2 - 2 \cdot 0.0266769747x$$

Im obigen Beispiel für $x = 50$ *gilt*

$$\hat{V}\left(g\left(\hat{\pi}\left(50\right)\right)\right) = 1.28517059 + 0.0005788748 \cdot 50^2 - 2 \cdot 0.0266769747 \cdot 50 = 0.064\,66$$

Mit

$$g\left(\hat{\pi}\left(50\right)\right) = -5.30945 + 0.11092 \cdot 50 = 0.23655$$

erhält man für das Logit in $x = 50$ *das 0.95-VI*

$$\widehat{VI}\left(g\left(\hat{\pi}\left(50\right)\right)\right) = 0.23655 \pm 1.96 \cdot \sqrt{0.064\,66}$$

und damit $[-0.261\,85,\ 0.734\,95]$. *Zudem kann man den Wert der an der Stelle* $x = 50$ *geschätzten Wahrscheinlichkeit berechnen:*

$$\hat{\pi}\left(50\right) = \frac{\exp 0.23655}{1 + \exp 0.23655} = 0.558\,86$$

samt dem 0.95-VI für die Wahrscheinlichkeit, im Alter von $x = 50$ *CHD zu haben:*

$$\begin{array}{ll} \textit{untere Grenze} & \frac{\exp(-0.261\,85)}{1+\exp(-0.261\,85)} = 0.43491 \\ \textit{obere Grenze} & \frac{\exp 0.734\,95}{1+\exp 0.734\,95} = 0.675\,89 \end{array}$$

2 Multiple logistische Regression

Die einfache logistische Regression mit einer erklärenden und einer erklärten Variable kann auf den Fall der multiplen logistischen Regression mit mehreren erklärenden Variablen und einer erklärten Variable erweitert werden. Dabei sind nicht nur metrisch skalierte, sondern auch nominalskalierte erklärende Variablen zulässig. Im Falle dichotomer Variablen sind diese mit 0 und 1 zu kodieren. Im Falle polytomer nominalskalierter erklärender Variablen sind diese zu dichotomisieren. Man führt bei k Ausprägungen $k-1$ sogenannte Hilfsvariablen, auch Dummy-, Design-, Schein- oder Stellvertretervariablen genannt, ein. Jede dieser Hilfsvariablen ist mit 0 oder 1 kodiert. Ein Objekt weist höchstens bei einer Hilfsvariable eine von 0 verschiedene Kodierung auf. Damit können alle Kategorien dargestellt werden. Z.B. für die Variable „Wohnort“ mit den Ausprägungen „Bergdorf“, „Aglo“ und „Stadt“ sind zwei Hilfsvariablen zu schaffen (z.B. „Bergdorf“ und „Aglo“. Wohnt die Person in einem Bergdorf, steht bei dieser Variable 1, bei der Variable „Aglo“ steht 0. Wohnt die Person in einer Agglomeration, steht bei „Aglo“ 1, bei „Bergdorf“ 0. Lebt die Person in der Stadt, steht bei beiden Hilfsvariablen 0). Statistikprogramme berechnen die Hilfsvariablen automatisch, wenn die Variablen als nominalskaliert gekennzeichnet sind. Dabei gibt es verschiedene Möglichkeiten der Darstellung der Ausprägungen mit Hilfsvariablen - statt „Aglo“ und „Bergdorf“, kann man z.B. auch „Aglo“ und „Stadt“ als Hilfsvariablen wählen. Die Wahl wirkt sich auch auf die Schätzer der Koeffizienten der übrigen Variablen aus!

Das Modell

Sei ein Vektor von erklärenden Variablen $\mathbf{x} = (x_1, ..., x_p)$ gegeben. Sei $P(Y = 1|\mathbf{x}) = \pi(\mathbf{x})$. Dann setzt wiederum

$$\pi(\mathbf{x}) := \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}$$

woraus sich das Modell

$$g(\pi(\mathbf{x})) = \ln \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p,$$

ergibt (zu oben analoge Herleitung, s. Satz 3 von S. 6).

Anpassen des Modells und Varianz der Koeffizientenschätzstatistiken

Es geht nun wiederum darum, mit ML-Methoden die Parameter $(\beta_0, \beta_1, ..., \beta_p)$ zu bestimmen. Die Herleitungen verlaufen wie oben, wobei für jeden Koeffizienten eine LL-Gleichungen hergeleitet wird:

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi\left(\mathbf{x}_i\right) = 0 \qquad \text{für } \beta_0$$
$$\sum_{i=1}^{n} x_{ij}\left(y_i - \pi\left(\mathbf{x}_i\right)\right) = 0 \quad \text{für } \beta_j,\ 1 \leq j \leq p$$

Sei $\boldsymbol{\beta} := (\beta_0, \beta_1, ..., \beta_p)$ die Stelle des Maximums der LL-Funktion (also die Lösung der Gleichungen). Die zweiten Ableitungen ergeben

$$\frac{\partial^2 LL\left(\boldsymbol{\beta}\right)}{\partial^2 \beta_j} = -\sum x_{ij}^2 \pi\left(\mathbf{x}_i\right)\left(1 - \pi\left(\mathbf{x}_i\right)\right)$$
$$\frac{\partial^2 LL\left(\boldsymbol{\beta}\right)}{\partial \beta_j \partial \beta_l} = -\sum x_{ij} x_{il} \pi\left(\mathbf{x}_i\right)\left(1 - \pi\left(\mathbf{x}_i\right)\right)$$

Die $(p+1) \times (p+1)$-Matrix $I\left(\hat{\boldsymbol{\beta}}\right)$, welche im konkreten Fall diese Terme mit -1 multipliziert enthält, wird beobachtete Informationsmatrix von $\hat{\boldsymbol{\beta}}$ genannt. Rao (1973) zeigte, dass für die Varianzmatrix $V\left(\mathbf{B}\right)$ gilt ($\mathbf{B} := (B_1, B_2, ..., B_p)$ mit den Zufallsvariblen B_i, welche die $\hat{\beta}_i$ als Werte annehmen):

$$V\left(\mathbf{B}\right) = \left(I\left(\mathbf{B}\right)\right)^{-1}.$$

Man kann $V\left(\mathbf{B}\right)$ darstellen als

$$V\left(\mathbf{B}\right) = \left(\mathbf{X}^\intercal \mathbf{V} \mathbf{X}\right)^{-1}$$

mit der Design-Matrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

- die Spalten weisen die Daten der Variablen $\mathbf{x}_j$ auf - und der Matrix

$$\mathbf{V} = \begin{pmatrix} \Pi\left(\mathbf{x}_1\right)\left(1 - \Pi\left(\mathbf{x}_1\right)\right) & 0 & \cdots & 0 \\ 0 & \Pi\left(\mathbf{x}_2\right)\left(1 - \Pi\left(\mathbf{x}_2\right)\right) & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \Pi\left(\mathbf{x}_n\right)\left(1 - \Pi\left(\mathbf{x}_n\right)\right) \end{pmatrix}.$$

Beispiel 14. *(bw=read.table("pfad/lowbwt.txt",header=T), Datensatz "lowbwt1"). Die Zielvariable drückt das Gewicht von Säuglingen aus (1 = zu tief, 0 = zureichend). Erklärende Variablen sind lwt (Gewicht der Frauen in Pfund), age (das Alter der Frauen), die nominale Variable „race", die mit 1 für white, 2 für black und 3 für other kodiert ist und die Variable ftv (Anzahl Arztbesuche im ersten Drittel der Schwangerschaft). Es wird hier nicht davon ausgegangen, dass „race" ein Ausdruck ist, der für Menschen angemessen ist. Hier wird die*

Bezeichnung einfach aus dem Buch übernommen. Gemeint ist wohl die soziale Schichtung, die hinter dem unangemessenen Begriff steckt, wobei die sozialen Schichtungen durch den Begriff wohl auch nicht gut eingefangen werden (z.B. Mittelschicht-Afroamerikaner und weisse „Unterschicht"). Anzahl Daten 189.
Die Variable race wird zuerst in einen Faktor, d.h. in eine nominalskalierte Variable verwandelt)

```
bw$raceF=as.factor(bw$race)
```

Mit

```
m36=glm(low~age+lwt+raceF+ftv,data=bw,family="binomial")
summary(m36)
```

erhält man

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.295366	1.071443	1.209	0.2267
age	-0.023823	0.033730	-0.706	0.4800
lwt	-0.014245	0.006541	-2.178	0.0294 *
raceF2	1.003898	0.497859	2.016	0.0438 *
raceF3	0.433108	0.362240	1.196	0.2318
ftv	-0.049308	0.167239	-0.295	0.7681

Tabelle 4: R-Ausgabe für die Daten Geburtsgewicht

Schnellinterpretation: je älter eine Frau ist, desto kleiner ist die Wahrscheinlichkeit, ein untergewichtiges Kind (= mit 1 kodierte Ausprägung der erklärten Variable) zu gebären (negatives Vorzeichen, nicht signifikant). Je schwerer eine Frau ist, desto kleiner ist die Wahrscheinlichkeit, ein untergewichtiges Kind zu gebären (signifikant). Schwarze Frauen gebären mit höherer Wahrscheinlichkeit untergewichtige Kinder als weisse Frauen (positives Vorzeichen; signifikant). Andere Frauen gebären mit höherer Wahrscheinlichkeit untergewichtige Kinder als weisse Frauen (nicht signifikant). Je mehr Arztbesuche, desto kleiner die Wahrscheinlichkeit, ein untergewichtiges Kind zu gebären (negatives Vorzeichen, nicht signifikant). Negative Koeffzienten bedeuten also, dass die Wahrscheinlichkeit des Auftretens der mit 1 kodierten Ausprägung der erklärten Variable bei steigenden Werten der entsprechenden erklärenden Variable sinkt, positive Koeffizienten bedeuten, dass die Wahrscheinlickeit des Auftretens der mit 1 kodierten Ausprägung der erklärten Variable bei steigenden Werten der erklärenden Variable steigt.

Zudem erhält man die Tabelle:

Null deviance:	234.67 on	188 degrees of freedom
Residual deviance:	222.57 on	183 degrees of freedom

Tabelle 5: R-Ausgabe für die Daten Geburtsgewicht

R und SPSS nehmen die mit der tiefsten Zahl kodierte Ausprägung der nominalskalierten Variable als Referenz (alle Ausprägungen der Hilfsvariablen auf 0 gesetzt), SAS die mit der höchsten Zahl kodierte Ausprägung. Mit R kann man die SAS-Version erhalten mit:

```
contr=contrasts(bw$raceF)=contr.SAS(3)
#die 3 gibt die Anzahl der Ausprägungen der  Variable an
m36c=glm(low~age+lwt+raceF+ftv,data=bw, contrasts=c(contr),
family= "binomial"  (link=logit))
summary(m36c)
```

oder mit

```
contr=contrasts(bw$raceF)=contr.treatment(3, base = 3)
m36d=glm(low~age+lwt+raceF+ftv,data=bw, contrasts=contr, family=
"binomial"(link=logit))
summary(m36d)
```

Mit dieser letzten Variante kann man beliebige Ausprägungen als Referenz wählen. Je nach der gewählten Referenz ergeben sich für alle Koeffizienten des Modells andere Ergebnisse. Werden mehrere nominalskalierte Variablen verwendet, kann für jede ein Kontrast definiert werden, der jeweils mit einem Namen zu versehen ist. Im glm-Befehl wird dann contrasts=c(...) eingegeben, wobei in c(...) die definierten Kontraste anzugeben sind - durch Komma abgetrennt.

Mit R erhält man die Ergebnisse der Voreinstellung, wenn man die Kontraste mit base=1 definiert. Lässt man dann contr ausgeben, erhält man:

	2	*3*
1	*0*	*0*
2	*1*	*0*
3	*0*	*1*

Werden die Kontraste für eine Variable in R alternativ zur Voreinstellung definiert, so wird diese Definition bei Modellen mit dieser Variable verwendet, selbst wenn man die Kontraste im glm-Befehl nicht mehr angibt. Will man die Berechnungen für die Voreinstellung vornehmen, müssen die Kontraste wieder entsprechend definiert werden.

Für die geschätzten Logits in $\mathbf{x}$ *berechnet man, wenn die Referenzkategorie die Ausprägung*

1 der Variable ist:

$$g\left(\hat{\pi}\left(\mathbf{x}\right)\right) = 1.295366 - 0.023823 \cdot age - 0.014245 \cdot lwt + 1.003898 \cdot raceF2 + \\ 0.433108 \cdot raceF3 - 0.049308 \cdot ftv$$

und für die Wahrscheinlichkeiten

$$\hat{\pi}\left(\mathbf{x}\right) = \frac{\exp\left(g\left(\hat{\pi}\left(\mathbf{x}\right)\right)\right)}{1 + \exp\left(g\left(\hat{\pi}\left(\mathbf{x}\right)\right)\right)}.$$

Für eine 38-jährige Frau mit 130 Pfund Gewicht, race=2 und drei Arztbesuchen im ersten Drittel der Schwangerschaft würde man damit erhalten.

$$\begin{aligned} g\left(\hat{\pi}\left(38, 65, 1, 0, 3\right)\right) &= 1.295366 - 0.023823 \cdot 38 - 0.014245 \cdot 130 + \\ &\quad 1.003898 \cdot 1 + 0.433108 \cdot 0 - 0.049308 \cdot 3 \\ &= -0.605\,78 \end{aligned}$$

Die geschätzte Wahrscheinlicheit, ein zu leichtgewichtiges Kind zu gebären, beträgt für diesen Datensatz

$$\hat{\pi}\left(38, 65, 1, 0, 3\right) = \frac{\exp\left(-0.605\,78\right)}{1 + \exp\left(-0.605\,78\right)} = 0.353\,02.$$

Signifikanztests

Man kann das Modell als Ganzes testen. Damit vergleicht man restliche Abweichung des vollen Models (mit allen Variablen) mit der restlichen Abweichung des Modells ohne erklärende Variablen (nur mit der Konstanten; Null deviance). Den Globaltest erhält man also mit der Statistik

$G = -2\left(\max LL_R - \max LL_0\right) =$ Null deviance - (Residual Deviance des vollen Modells)

Bemerkung 15. *Die maximale Log-Likelihood eines Modells m1 erhält man in R auch mit dem Befehl „logLik(m1)“*

Fürs Beispiel 14 erhält man also den Globaltest mit den Ergebnissen der Tabelle 5 mittels:

$$\begin{aligned} G &= \text{Null deviance - Residual Deviance des Modells „low ˜age+ltw+race2+race3+ftv“} \\ &= 234.67 - 222.57 = 12.1 \end{aligned}$$

mit $188 - 183 = 5$ Freiheitsgraden. Es hat 189 Daten. Im Modell ohne erklärende Variablen (aber mit der Konstanten; Null-Modell) geht ein Freiheitsgrad weg. Im Modell mit 6 Koeffizienten bleiben $189 - 6 = 183$ Freiheitsgrade. Man erhält mit

1-pchisq(12.1,5)

den p-Wert 0.033456146. Die Nullhypothese, dass keine der erklärenden Variablen einen Beitrag leistet, wird verworfen. Mindestens eine Variable leistet einen Erklärungsbeitrag.

Der Globaltest wird von R nicht angeboten - muss also selber nachgerechnet werden, z.B. für das Modell m36 mittels

```
a=summary(m36)
testw=a$null.deviance-a$deviance
freigr=a$df.null-a$df.residual
1-pchisq(testw,freigr)
```

Mit

```
anova(m36,test="Chisq")
```

rechnet R die Modelle

age
age+lwt
age+lwt+raceF
age+lwt+raceF+ftv

durch. Man erhält im Beispiel:

Analysis of Deviance Table
Model: binomial, link: logit
Response: low
Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			188	234.67	
age	1	2.7600	187	231.91	0.09665 .
lwt	1	4.7886	186	227.12	0.02865 *
raceF	2	4.4628	184	222.66	0.10738
ftv	1	0.0877	183	222.57	0.76708

Unter Resid.Dev wird jeweils die Restliche Abweichung (vom gesättigten Modell) angegeben. Unter „Deviance“ wird angegeben, um wieviel die Restliche Abweichungen vom Modell mit allen vorangehenden Variablen durch die Hinzunahme der entsprechenden Variable sinkt. Es handelt sich also um die durch die Hinzunahme einer Variablen zusätzlich erklärte Abweichung. Die Deviance bei age ist: $234.67 - 231.91 = 2.76$ und bei lwt: $231.91 - 227.12 = 4.79$.

Es gilt

$$222.57 + 2.7600 + 4.7886 + 4.4628 + 0.0877 = 234.67$$

und damit

$$\begin{aligned} 12.099 &= 2.7600 + 4.7886 + 4.4628 + 0.0877 \\ &= 234.67 - 222.57. \end{aligned}$$

Der Globaltest kann also auch aus den obigen Angaben berechnet werden. Die p-Werte sind Resultate des Vergleichs des Models ohne die entsprechende Variable mit dem Modell samt der entsprechenden Variable. Den Vergleich des Modells mit der Variable age und dem Nullmodell - erhält man mit den ungerundeten $234.67 - 231.96$ oder auch mit

```
m38=glm(low~age,data=bw,family="binomial"))
summary(m38)
anova(m38, test="Chisq")
```

Man erhält den p-Wert 0.09665 (s. obige „Analysis of Deviance Table"-Tabelle hinter „age"). Den Vergleich des Modells mit den Variablen age und lwt und des Modells mit der Variable age - erhält man auch, indem man

```
m38b=glm(low~age+lwt,data=bw,family="binomial")
anova(m38, m38b, test="Chisq")
```

rechnet (identischer p-Wert).

Mit den Tests sieht man also, ob ein Modell, das eine zusätzliche Variable aufweist, sich vom Modell ohne diese Variable signifikant unterscheidet. Im obigen Beispiel gibt es einen signifikanten Unterschied zwischen dem Modell age und dem Modell age+lwt. Die übrigen Modellvergleiche ergeben keine signifikanten Unterschiede.

Die Modellvergleiche kann man als LL-Verhältnis-Tests für die Koeffizienten selber betrachten. Man testet, ob diese von 0 verschieden sind.

Neben diesen LL-Quotienten-Tests können wiederum Wald-Tests durchgeführt werden. Mit den Abkürzungen von S. 24 gilt für die Teststatistik

$$W_j = \frac{B_j}{SE\,(B_j)},$$

dass diese näherungsweise standardnormalverteilt ist. Die entsprechenden R-Ergebnisse wurden schon geliefert (s. Tabelle 4, S. 25). Man kann diese nachrechnen: den Vektor der Standardabweichungen $\widehat{SE}\left(\hat{\beta}_j\right)$ erhält man mit R mittels

```
sqrt(diag(vcov(m36)))
```

.

Den Testwert W_0 für die Konstante erhält man entsprechend mittels

```
m36$coefficients[1]/sqrt(diag(vcov(m36)))[1]
```

Man erhält $W_0 = 1.208992$. Für den zweiseitigen Test erhält man entsprechend mit

```
2*min(pnorm(1.208992),1-pnorm(1.208992))
```

den p-Wert $0.2266659 > 0.05$. Analog für die anderen Koeffizienten.

Im Allgemeinen möchte man ein sparsames Modell haben, das die theoretisch relevanten Variablen und die statistisch signifikanten Variablen enthält. Im obigen Beispiel könnte

man eventuell auf die Variablen ftv und age verzichten. Man erhält mit (Im Beispiel geht man davon aus, dass vorgängig die Kontraste umdefiniert wurden. Deshalb wird hier die Voreinstellung wieder hergestellt).

```
contr=contrasts(bw$raceF)=contr.treatment(3, base = 1)
m38c=glm(low~lwt+raceF,data=bw,contrasts=contr,family="binomial")
summary(m38c)
```

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.8058	0.8452	0.95	0.3404
lwt	-0.0152	0.0064	-2.36	0.0181
raceF2	1.0811	0.4881	2.22	0.0268
raceF3	0.4806	0.3567	1.35	0.1778
Null deviance: 234.67 on 188 degrees of freedom				
Residual deviance: 223.26 on 185 degrees of freedom				

Den Modellvergleich der Modelle mit und ohne ftv sowie age, kann man mit der Differenz der residualen Abweichungen der beiden Modelle rechnen: $223.26 - 222.57 = 0.688$ oder mit

```
anova(m38c,m36,test="Chisq")
```

(Test mit $185 - 183 = 2$ Freiheitsgraden, p-Wert 0.7096). Die entsprechenden Variablen age und ftv sind also statistisch gesehen überflüssig.

Vertrauensintervalle

Die Methoden für die VIs der Koeffizienten, der Logits und der geschätzten Wahrscheinlichkeiten entsprechen im Wesentlichen den bereits für den zweidimensionalen Fall behandelten. Man kann VIs mit Hilfe der Wald-Test liefern oder mit der Profile-Likelihood-Methode (mit confint(modell)-Befehl). Für die Logits verläuft die Theorie analog, wobei die Formeln komplizierter werden, da die paarweisen Kovarianzen verwendet werden. Es gilt für die Varianz der Logits

$$\begin{aligned} V\left(g\left(\Pi\left(\mathbf{x}\right)\right)\right) &= V\left(B_0 + B_1 x_1 + ... + B_p x_p\right) \\ &= \sum_{j=0}^{p} x_j^2 V\left(B_j\right) + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2 x_j x_k KOV\left(B_j, B_k\right) \end{aligned}$$

Dies kann man übersichtlicher mit Hilfe von Matrizen darstellen mittels der Kovarianzmatrix $V\left(\mathbf{B}\right)$ der Koeffizientenstatistiken:

$$V\left(\mathbf{B}\right) = \left(\mathbf{X}^\intercal \mathbf{V} \mathbf{X}\right)^{-1}.$$

(für $\mathbf{X}$ und $\mathbf{V}$ s. Seite 24). Man erhält

$$V\left(g\left(\Pi\left(\mathbf{x}\right)\right)\right) = \mathbf{x}^\intercal V\left(\mathbf{B}\right) \mathbf{x} = \mathbf{x}^\intercal \left(\mathbf{X}^\intercal \mathbf{V} \mathbf{X}\right)^{-1} \mathbf{x}.$$

Beispiel 16. *Für eine Frau mit LWT=150 und race=1 erhält man für das Modell m38c (s. oben) die Schätzer (für race=1 ist raceF2=0 und raceF3=0)*

$$g\left(\hat{\pi}\left(150,0,0\right)\right)=0.80575-0.01522\cdot 150+1.08107\cdot 0+0.48060\cdot 0=-1.4773$$

und

$$\hat{\pi}\left(150,0,0\right)=\frac{\exp g\left(\hat{\pi}\left(150,0,0\right)\right)}{1+\exp g\left(\hat{\pi}\left(150,0,0\right)\right)}=\frac{\exp\left(-1.4773\right)}{1+\exp(-1.4773)}=0.185\,84.$$

und mit der Kovarianzmatrix (vcov(m38c)) und für Intercept=1, lwt=150, raceF2=0, raceF3=0 erhält man die Varianz

$$\begin{aligned}&\hat{V}\left(g\left(\hat{\pi}\left(150,0,0\right)\right)\right)\\&=\begin{pmatrix}1 & 150 & 0 & 0\end{pmatrix}\begin{pmatrix}0.714307 & -5.2137\cdot 10^{-3} & 0.0226026 & -0.1034970\\ -0.005214 & 4.1465\cdot 10^{-5} & -0.0006470 & 0.0003558\\ 0.0226036 & -6.4703\cdot 10^{-4} & 0.2381949 & 0.0532002\\ -0.103497 & 3.5585\cdot 10^{-4} & 0.0532002 & 0.1272161\end{pmatrix}\begin{pmatrix}1\\150\\0\\0\end{pmatrix}\\&=0.08316\,4\end{aligned}$$

und die Standardabweichung $\sqrt{0.08316\,4}=0.288\,38$, *also die 0.95-VIs*

$$\widehat{VI}\left(g\left(\hat{\pi}\left(150,0,0\right)\right)\right)=-1.4773\pm 1.96\cdot 0.288\,38=\ \ ]-2.042\,5,-0.912\,08[$$

und

$$\begin{aligned}\widehat{VI}\left(\hat{\pi}\left(150,0,0\right)\right)&=]\frac{\exp\left(-2.042\,5\right)}{1+\exp\left(-2.042\,5\right)},\frac{\exp\left(-0.912\,08\right)}{1+\exp\left(-0.912\,08\right)}[\\&=]0.114\,81,0.286\,57[.\end{aligned}$$

Mit R: Mittels

```
c(1,150,0,0) %*% m38c$coefficients
```

erhält man den Logit–Schätzung -1.477712 *und mit*

```
sch=c(1,150,0,0) %*% m38c$coefficients
exp(sch)/(1+exp(sch))
```

die Schätzung der Wahrscheinlichkeit 0.1857733. *Die VIs erhält man mit*

```
var=t(c(1,150,0,0)) %*% vcov(m38c) %*% c(1,150,0,0)
```

(var = 0.08316444*) und damit für das geschätzte Logit das 0.95-VI*

```
sch-qnorm(0.975)*sqrt(var)
sch+qnorm(0.975)*sqrt(var)
```

$]-2.042931,-0.9124927[$ ($\ni -1.477712$) *sowie für die geschätzte Wahrscheinlichkeit das 0.95-VI*

*u=sch-qnorm(0.975)*sqrt(var)* *o=sch+qnorm(0.975)*sqrt(var)*
exp(u)/(1+exp(u))
exp(o)/(1+exp(o))

$]0.1147686, 0.28649[$ ($\ni 0.1857733$)

3 Interpretation des geschätzten logistischen Regressionsmodells

Interpretation bei dichotomer erklärender Variable

Im Falle einer dichotomen erklärenden Variable, die mit 0 und 1 kodiert ist, erhält man mit dem Modell

$$g\left(\pi\left(x\right)\right) = \beta_0 + \beta_1 x$$

die Logitdifferenz

$$g\left(\pi\left(0+1\right)\right) - g\left(\pi\left(0\right)\right) = \beta_0 + \beta_1 \cdot 1 - \left(\beta_0 + \beta_1 \cdot 0\right) = \beta_1$$

und die OR

$$\begin{aligned}
\exp\left(g\left(\pi\left(0+1\right)\right) - g\left(\pi\left(0\right)\right)\right) &= \exp\left(\ln\frac{\pi\left(0+1\right)}{1-\pi\left(0+1\right)} - \ln\frac{\pi\left(0\right)}{1-\pi\left(0\right)}\right) \\
&= \exp\left(\ln\frac{\frac{\pi(0+1)}{1-\pi(0+1)}}{\frac{\pi(0)}{1-\pi(0)}}\right) \\
&= \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} = \exp\beta_1.
\end{aligned}$$

Dasselbe Resultat erhält man auch auf die folgende Weise: in einer Vierfelder-Kreuztabelle für die Wahrscheinlichkeiten, dass die erklärte Variable y einen spezifischen Wert annimmt, unter der Bedingung, dass die erklärende Variable x einen spezifischen Wert annimmt, ergibt sich

	y=1	y=0
x=1	$\pi\left(1\right) = \frac{\exp(\beta_0+\beta_1\cdot 1)}{1+\exp(\beta_0+\beta_1\cdot 1)}$	$1-\pi\left(1\right) = \frac{1}{1+\exp(\beta_0+\beta_1)}$
x=0	$\pi\left(0\right) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$	$1-\pi\left(0\right) = \frac{1}{1+\exp(\beta_0)}$

$\frac{\exp(\beta_0+\beta_1\cdot 1)}{1+\exp(\beta_0+\beta_1\cdot 1)}$ ist die Wahrscheinlichkeit, dass 1 angenommen wird unter der Bedingung, dass $x = 1$. $1-\pi\left(1\right)$ ist dann die Gegenwahrscheinlichkeit, nämlich dass 0 angenommen wird, unter der Bedingung, dass $x = 1$.

Damit ergibt sich wiederum

$$\begin{aligned}
\frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} &= \frac{\pi(1)(1-\pi(0))}{(1-\pi(1))\pi(0)} \\
&= \frac{\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}\frac{1}{1+\exp(\beta_0)}}{\frac{1}{1+\exp(\beta_0+\beta_1)}\frac{\exp(\beta_0)}{1+\exp(\beta_0)}} \\
&= \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}\frac{1}{1+\exp(\beta_0)}\frac{1+\exp(\beta_0+\beta_1)}{1}\frac{1+\exp(\beta_0)}{\exp(\beta_0)} \\
&= \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} \\
&= \exp(\beta_0+\beta_1-\beta_0) \\
&= \exp\beta_1
\end{aligned}$$

Der Zusammenhang zwischen dem Quotenverhältnis und dem Steigungs-Koeffizienten erleichtert die Interpretation der Koeffizienten und erklärt teilweise den Erfolg der logistischen Regression in der empirischen Forschung.

Beispiel 17. *Dichotomisiert man die Variable Alter im Alter-CHD-Beispiel mit dem Wert 55 und mittels R-Befehlen*

```
ac$agedich[ac$AGE>=55]=1
ac$agedich[ac$AGE<55]=0
```

,

so erhält man mit

```
table(ac$agedich,ac$CHD)
```

die folgende Kreuztabelle:

	$\geq 55:1$	$<55:0$	*Total*
present: 1	*21*	*22*	*43*
absent: 0	*6*	*51*	*57*
Total	*17*	*73*	*100*

Berechnet man die logistische Regression

```
m52=glm(CHD~agedich,data=ac,family="binomial"(link="logit"))
```

erhält man

Coefficients:	*(Intercept)*	*agedich*
	-0.8408	*2.0935*

Es gilt

$$\exp(2.0935) = 8.1133 = \frac{\frac{21}{22}}{\frac{6}{51}}.$$

Man erhält also das Quotenverhältnis mittels ML-Schätzung. Das Verhältnis CHD zu nicht-CHD ist bei den über 55-jährigen 8.11 mal grösser als dieses Verhältnis bei den unter 55-jährigen.
Mit den von R gelieferten Standardfehler (mit summary(m52) oder mit sqrt(diag(vcov(m52)))[2] erhält man 0.5285) kann man Wald-VIs für das Quotenverhältnis berechnen. Die 0.95-VIs für den Koeffizienten $\beta_1 = 2.0935$ *sind*

$$2.0935 \pm 1.96 \cdot 0.5285 =]1.\,057\,6, 3.\,129\,4[$$

und damit das 0.95-VI für das Quotenverhältnis

$$\begin{aligned}\exp\left(2.0935 \pm 1.96 \cdot 0.5285\right) &=]\exp\left(1.0576\right), \exp\left(3.129\,4\right)[\\ &=]2.879\,5, 22.86[\end{aligned}$$

Man kann zeigen, dass im 4-Felder-Fall gilt:

$$\hat{V}\left(\ln OR\right) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

und damit im Beispiel

$$\hat{V}\left(\ln OR\right) = \frac{1}{21} + \frac{1}{22} + \frac{1}{6} + \frac{1}{51} = 0.279\,35$$

und $\sqrt{0.279\,35} = 0.528\,54$.

Mit der Likelihood-Profile-Methode (confint(modell)-Befehl) erhält man für das Logit das 0.95-VI $]1.108356, 3.2068562[$ *und damit für das Quotenverhältnis die 0.95-VI-Grenzen* $\exp\left(1.108356\right) = 3.029\,4$ *und* $\exp\left(3.2068562\right) = 24.701$.

Bemerkung 18. *Statt 0 und 1 werden erklärende dichotome Variablen manchmal auch mit -1 und 1 kodiert (Abweichung vom Mittel-Kodierung). Damit erhält man z.B. fürs Geschlecht (1 Frauen, -1 Männer) eingesetzt ins Modell* $g\left(\pi\left(x\right)\right) = \beta_0 + \beta_1 x$

$$\begin{aligned}\ln\left(OR\left(Frauen, Männer\right)\right) &= g\left(\Pi\left(1\right) - g\left(\Pi\left(-1\right)\right)\right) \\ &= B_0 + B_1 \cdot 1 - \left(B_0 + B_1\left(-1\right)\right) \\ &= 2B_1\end{aligned}$$

und das geschätzte Quotenverhältnis $\exp 2\hat{\beta}_1$. *Mit*

$$SE\left(2B_1\right) = 2SE\left(B_1\right)$$

erhält man das α*-VI*

$$\exp\left(2B_1 \pm \Phi^{-1}\left(\alpha + \frac{1-\alpha}{2}\right) 2SE\left(B_1\right)\right)$$

Mit

```
ac$agedichm[ac$agedich==1]=1
ac$agedichm[ac$agedich==0] = -1
m55=glm(CHD~agedichm,data=ac,family="binomial" (link="logit"))
summary(m55)
```

erhält man

Coefficients:	*Estimate*	*Std. Error*	*z value*	*Pr(>|z|)*
agedichm	*1.0468*	*0.2643*	*3.961*	*7.46e-05 ****

Die Werte bei Referenzkodierung 0 *und* 1*, also die eigentlich interessierenden Werte bezüglich der OR, erhält man durch Multiplikation mit 2:*

$$\hat{\beta}_1 : 2 \cdot 1.0468 = 2.0936$$
$$\widehat{SE}\left(B_1\right) : 2 \cdot 0.2643 = 0.5286$$

Obwohl also die Koeffizienten verschieden sind, führt das Modell auf identische OR-Schätzer. Dies ist erforderlich - sonst wäre Identisches verschieden!

Interpretation bei polytomer erklärender Variable

Bei nominalskalierter, polytomer erklärender Variable gibt es ebenfalls verschiedene Kodiermethoden bei der Schaffung von Hilfsvariablen. Gewöhnlich wird eine Ausprägung bei allen Hilfsvariablen auf 0 gesetzt und als Referenz gewählt (Standardkodierung). z.B. bei vier Ausprägungen werden drei Hilfsvariablen geschaffen

	Hilfsvariablen		
Ausprägung	V1	V2	V3
r1	0	0	0
r2	1	0	0
r3	0	1	0
r4	0	0	1

und damit r1 als Referenz gewählt. Das entsprechende Modell ist (x_i für die Hilfsvariable Vi)

$$g\left(\pi\left(x_1, x_2, x_3\right)\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

Mit der logistischen Regression erhält man die logarithmierten Quotenverhältnisse der Ausprägungen 2,3,4 bezüglich der Ausprägung 1, denn mit

$$g\left(\Pi\left(x_1, x_2, x_3\right)\right) = \ln \frac{\pi\left(x_1, x_2, x_3\right)}{1 - \pi\left(x_1, x_2, x_3\right)}$$

ergibt sich für den Vergleich von r2 und r1 (bei r2 nehmen die drei Variablen 1, 0, und 0 an, bei r1 0, 0 und 0).

$$\begin{aligned}\ln\left(OR\left(r_2, r_1\right)\right) &= g\left(\pi\left(1,0,0\right)\right) - g\left(\pi\left(0,0,0\right)\right) \\ &= \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot 0 + \beta_3 \cdot 0 - \left(\beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 + \beta_3 \cdot 0\right) \\ &= \beta_1.\end{aligned} \tag{4}$$

$\exp\left(\beta_1\right)$ ist damit das Quotenverhältnis von r2 und r1. Auch hier kann man die Standardfehler von B_1 mit

$$SE\left(B_i\right) = \sqrt{\frac{1}{N_{11}} + \frac{1}{N_{1i}} + \frac{1}{N_{i1}} + \frac{1}{N_{ii}}}$$

berechnen (N_{ij} ist die Zufallsvariable, welche die Häufigkeit n_{ij} annimmt). Die α-VIs werden wie üblich durch

$$VI\left(B_i\right) = B_i \pm \Phi^{-1}\left(\alpha + \frac{1-\alpha}{2}\right) SE\left(B_i\right)$$

berechnet und für $OR\left(r_i, r_1\right)$ durch

$$\exp\left(B_i \pm \Phi^{-1}\left(\alpha + \frac{1-\alpha}{2}\right) SE\left(B_i\right)\right).$$

Beispiel 19. *Für die Tabelle (s. Seite 56 von H&L, Zielvariable ist CHD, erklärende, polytome Variable r mit den Ausprägungen r1, r2, r3 und r4)*

CHD\Farbe	*r1*	*r2*	*r3*	*r4*
Abwesend	*20*	*10*	*10*	*10*
Präsent	*5*	*20*	*15*	*10*

erhält man mit

```
r=c(rep(1,25),rep(2,30),rep(3,25),rep(4,20))
CHD=c(rep(0,20),rep(1,5),rep(0,10),rep(1,20),rep(0,10),rep(1,15),rep(0,10),rep(1,10))
ra_chd=cbind(r,CHD)
ra_chd=as.data.frame(ra_chd)
ra_chd$r=as.factor(ra_chd$r)
m57=glm(CHD~r,data=ra_chd,family="binomial" (link="logit" ))
summary(m57)
```

das Ergebnis

Coefficients:	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	-1.3863	0.5000	-2.773	0.00556 **
r_2	2.0794	0.6325	3.288	0.00101 **
r_3	1.7918	0.6455	2.776	0.00551 **
r_4	1.3863	0.6708	2.067	0.03878 *

Tabelle 6: Ergebnisse bei Standardkodierung der Hilfsvariablen und der Wahl von r1 als Referenz

wobei man 2.0794 *auch durch* $\ln\left(\frac{20\cdot 20}{5\cdot 10}\right) = 2.079\,4$ *(also das logarithmierte Quotenverhältnis von* r_2 *und* r_1 *bezüglich Anwesend-Abwesend berechnen kann. Ebenso erhält man* 1.7918 *durch* $\ln\left(\frac{20\cdot 15}{5\cdot 10}\right) = 1.791\,8$ *und* 1.3863 *durch* $\ln\left(\frac{20\cdot 10}{5\cdot 10}\right) = 1.386\,3$*. Die Quotenverhältnisse bezüglich CHD erhält für jedes* r_i *mit Referenz auf* r_1 *aus dem Ergebnis von R also durch* exp (*Koeffizient*)*. Das Verhältnis CHD zu nicht-CHD ist bei r2* $\exp(2.0794) = 7.999\,7$ *mal grösser als dieses Verhältnis bei r1.*
Die Standardfehler erhält man auch durch:

$$\sqrt{\frac{1}{20}+\frac{1}{10}+\frac{1}{5}+\frac{1}{20}} = 0.632\,46$$
$$\sqrt{\frac{1}{20}+\frac{1}{10}+\frac{1}{5}+\frac{1}{15}} = 0.645\,50$$
$$\sqrt{\frac{1}{20}+\frac{1}{10}+\frac{1}{5}+\frac{1}{10}} = 0.670\,82$$

Die folgenden Kodierung (Abweichung vom Mittel)

	Hilfsvariablen		
Ausprägung	V1	V2	V3
1	-1	-1	-1
2	1	0	0
3	0	1	0
4	0	0	1

drückt die Abweichung der Logits der Gruppen vom Gesamtmittel der Logits aus.

Beispiel 20. *Mit*

```
contrast=matrix(c(-1,-1,-1,1,0,0,0,1,0,0,0,1),byrow=T,ncol=3)
rownames(contrast)=c("1","2","3","4")
colnames(contrast)=c("2","3","4")
con=contrasts(ra_chd$r)=contrast
m59=glm(CHD~r,data=ra_chd,contrasts=con,family="binomial" (link="logit"))
summary(m59)
```

erhält man:

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.07192	0.21890	-0.329	0.7425
r2	0.76507	0.35059	2.182	0.0291 *
r3	0.47739	0.36228	1.318	0.1876
r4	0.07192	0.38460	0.187	0.8517

Tabelle 7: Abweichung-vom-Mittel-Kodierung bei Hilfsvariablen

Während die Abweichungen der r_i von r1 bei der Standardkodierung signifikant waren, ist bei der Mittel-Kodierung nur die Abweichung der Ausprägungen r2 vom Mittel signifikant. Die Abweichung der Ausprägung r1 vom Mittel wird nicht geliefert. Sie ist durch die übrigen drei Werte und das Mittel bestimmt. Man erhält für das Mittel der Logits (= Intercept = β_0):

$$\frac{1}{4}\left(\ln\frac{5}{20}+\ln\frac{20}{10}+\ln\frac{15}{10}+\ln\frac{10}{10}\right)=-0.07192\,1$$

Die Abweichung von r1 vom Mittel wäre also $\ln\frac{5}{20}-(-0.07192\,1)=-1.314\,4$. *Es gilt* $0.76507+0.47739+0.07192=1.314\,4$, *d.h. die Summe der Koeffizienten hebt sich auf. Für die Abweichung von r1 vom Mittel ist kein p-Wert zu liefern, da durch die übrigen Zahlen und das Mittel diese Abweichung bestimmt ist.*

$\beta_1=0.76507$ *z.B. erhält man bei gegebenem Mittel durch* $\ln\frac{20}{10}-(-0.07192\,1)=0.765\,07$ *(Abweichung des Logits der Ausprägung r2 vom Mittel).*

Die Interpretation bei der Mittel-Kodierung ist schwieriger, ausser das Mittel der Logits ergibt einen inhaltlichen Sinn. Die Werte, die man bei dieser Kodierung erhält, kann man verwenden, um die Werte bei der Standard-Kodierung zu erhalten.

Beispiel 21. *Mit der Abweichung-vom-Mittel-Kodierung und bei Betrachtung von r1 und r2 erhält man:*

$$\begin{aligned}\ln\left(OR\left(r2,r1\right)\right)&=g\left(\Pi\left(1,0,0\right)\right)-g\left(\Pi\left(-1,-1,-1\right)\right)\\&=B_0+B_1\cdot 1+B_2\cdot 0+B_3\cdot 0-\left(B_0+B_1\cdot -1+B_2\cdot -1+B_3\cdot -1\right)\\&=2B_1+B_2+B_3\end{aligned}$$

Dies ergibt: $2\cdot 0.76507+0.47739+0.07192=2.079\,5$, *also den Koeffizienten bei Standardkodierung von Tabelle 6 (*$\ln\left(OR\left(r2,r1\right)\right)$ *muss unabhängig von der Kodierung gleich bleiben!). Die Varianz von* $\ln\left(OR\left(r2,r1\right)\right)$ *ist*

$$\begin{aligned}V\left(\ln\left(OR\left(r2,r1\right)\right)\right)&=V\left(2B_1+B_2+B_3\right)\\&=V\left(2B_1\right)+V\left(B_2\right)+V\left(B_3\right)+2KOV\left(2B_1,B_2\right)\\&+2KOV\left(2B_1,B_3\right)+2KOV\left(B_2,B_3\right)\\&=4V\left(B_1\right)+V\left(B_2\right)+V\left(B_3\right)+4KOV\left(B_1,B_2\right)\\&+4KOV\left(B_1,B_3\right)+2KOV\left(B_2,B_3\right)\end{aligned}$$

Im Beispiel erhält man also mit der Kovarianzmatrix (vcov(m18))

	(Intercept)	*r2*	*r3*	*r4*
(Intercept)	*0.047916655*	*-0.01041667*	*-0.006249989*	*0.002083345*
r2	*-0.010416669*	*0.12291663*	*-0.031249997*	*-0.039583331*
r3	*-0.006249989*	*-0.03125000*	*0.131249988*	*-0.043750011*
r4	*0.002083345*	*-0.03958333*	*-0.043750011*	*0.147916655*

$V\left(\ln\left(\widehat{OR\left(r2,r1\right)}\right)\right) = 4 \cdot 0.12291663 + 0.131249988 + 0.147916655 +$
$4\left(-0.03125000\right) + 4\left(-0.03958333\right) + 2\left(-0.043750011\right) = 0.400\,00$
und dem geschätzten Standardfehler $\sqrt{0.400\,00} = 0.632\,46$ *von* $\ln\left(OR\left(r2,r1\right)\right)$ *(s. Tabelle 6)*

Interpretation bei stetiger erklärender Variable

Bei der bivariaten linearen Regression ist die entsprechende Interpretation einfach. Da

$$\begin{aligned} g\left(x+1\right) - g\left(x\right) &= \beta_0 + \beta_1(x+1) - \left(\beta_0 + \beta_1(x)\right) \\ &= \beta_0 + \beta_1 x + \beta_1 - \beta_0 - \beta_1(x) \\ &= \beta_1, \end{aligned}$$

drückt β_1 den Zuwachs der erklärten Variable aus, wenn auf der x-Achse eine Einheit hinzukommt. Zur linearen Regression analoge Aussagen können für die logistische Regression gemacht werden. Statt für die Werte der erklärten Variable gelten sie aber für die Logits. Für das Modell

$$g\left(\pi\left(x\right)\right) = \beta_0 + \beta_1 x$$

gilt

$$g\left(\pi\left(x+1\right)\right) - g\left(\pi\left(x\right)\right) = \beta_0 + \beta_1\left(x+1\right) - \left(\beta_0 + \beta_1 x\right) = \beta_1$$

Damit ist

$$g\left(\pi\left(x+1\right)\right) - g\left(\pi\left(x\right)\right) = \ln\frac{\pi\left(x+1\right)}{1-\pi\left(x+1\right)} - \ln\frac{\pi\left(x\right)}{1-\pi\left(x\right)} = \ln\left(\frac{\frac{\pi(x+1)}{1-\pi(x+1)}}{\frac{\pi(x)}{1-\pi(x)}}\right) = \beta_1$$

und somit ist das Quotenverhältnis $\left(\frac{\frac{\pi(x+1)}{1-\pi(x+1)}}{\frac{\pi(x)}{1-\pi(x)}}\right)$ mit $\exp\left(\beta_1\right)$ identisch.

Beispiel 22. *Im Age-CHD-Beispiel erhielt man* $\beta_1 = 0.11092$. *Es gilt* $\exp\left(0.11092\right) = 1.117\,3$. *Wird die Altersvariable in* x *um eine Einheit erhöht, so beträgt das Quotenverhältnis*

$$\frac{\frac{\pi(x+1)}{1-\pi(x+1)}}{\frac{\pi(x)}{1-\pi(x)}} = 1.01173,$$

d.h.

$$\frac{\pi(x+1)}{1-\pi(x+1)} = 1.01173\frac{\pi(x)}{1-\pi(x)}.$$

In $x+1$ ist das Verhältnis der Wahrscheinlichkeit CHD zu der von nicht-CHD um 1.01173 mal grösser als in x.

Betrachtet man $g(\pi(x+c)) - g(\pi(x))$ erhält man

$$\begin{aligned}&\beta_0 + \beta_1(x+c) - (\beta_0 + \beta_1 x)\\ &= c\beta_i\end{aligned}$$

Entsprechend ist

$$\begin{aligned}g(\pi(x+c)) - g(\pi(x)) &= \ln\frac{\pi(x+c)}{1-\pi(x+c)} - \ln\frac{\pi(x)}{1-\pi(x)}\\ &= \ln\frac{\frac{\pi(x+c)}{1-\pi(x+c)}}{\frac{\pi(x)}{1-\pi(x)}} = c\beta_i\end{aligned}$$

und

$$\exp(g(\pi(x+c)) - g(\pi(x))) = \frac{\frac{\pi(x+c)}{1-\pi(x+c)}}{\frac{\pi(x)}{1-\pi(x)}} = \exp(c\beta_i).$$

und man erhält die Endpunkte des $\alpha-$VIs mittels

$$\exp\left(cB_1 \pm \Phi^{-1}\left(\alpha + \frac{1-\alpha}{2}\right)cSE(B_1)\right).$$

Beispiel 23. *Im CHD ˜Age-Beispiel erhielt man $\beta_1 = 0.11092$. Es gilt also für $c = 10$ − das Alters steigt um 10 Jahre:* $\exp(10 \cdot 0.11092) = 3.031\,9$. *d.h.*

$$\frac{\pi(x+10)}{1-\pi(x+10)} = 3.031\,9\frac{\pi(x)}{1-\pi(x)}.$$

In $x+10$ ist das Verhältnis der Wahrscheinlichkeit CHD zu der von nicht-CHD um $3.031\,9$ *mal grösser als in x.*

Drückt man dies statt mit einer Konstanten c durch zwei a, b auf der Achse aus, bezüglich der man das Quotenverhältnis ausrechnet, ergibt sich mit den obigen Überlegungen:

$$\exp(g(\pi(b)) - g(\pi(a))) = \frac{\frac{\pi(b)}{1-\pi(b)}}{\frac{\pi(a)}{1-\pi(a)}} = \exp((b-a)\beta_1)$$

Im multiplen Fall gilt für die lineare Regression: Lässt man die übrigen erklärenden Variablen konstant, erhält man

$$\begin{aligned}&g(x_1, ..., x_i + 1, ..., x_p) - g(x_1, ..., x_i, ..., x_p)\\ &= \beta_0 + \beta_1 x_1 + ... + \beta_i(x_i + 1) + ... + \beta_p x_p - (\beta_0 + \beta_1 x_1 + ... + \beta_i x_i + ... + \beta_p x_p)\\ &= \beta_i\end{aligned}$$

Jeder Koeffizient drückt aus, um wieviel der Zielwert steigt, wenn eine Einheit auf der x_i-Achse hinzukommt und die übrigen Variablen konstant gehalten werden. Für die multiple logistische Regression gilt analog:

$$\begin{aligned} & g\left(\pi\left(1+x_1+...+\left(x_i+1\right)+x_{i+1}+...+x_p\right)\right)-g\left(\pi\left(1+x_1+...+x_i+x_{i+1}+...+x_p\right)\right) \\ & =\beta_0+\beta_1 x_1+...+\beta_i\left(x_i+1\right)+...+\beta_p x_p-\left(\beta_0+\beta_1 x_1+...+\beta_i x_i+...+\beta_p x_p\right) \\ & =\beta_i \end{aligned}$$

d.h. der Koeffizient β_i drückt aus, um wieviel

$$\ln \frac{\pi\left(x_1, ..., x_i, ..., x_p\right)}{1-\pi\left(x_1, ..., x_i, ..., x_p\right)}$$

ansteigt, wenn auf der x_i-Achse eine Einheit hinzukommt und die übrigen Variablen konstant gehalten werden. Es gilt ja

$$\begin{aligned} & \ln \frac{\pi\left(x_1, ..., x_i+1, ..., x_p\right)}{1-\pi\left(x_1, ..., x_i+1, ..., x_p\right)}-\ln \frac{\pi\left(x_1, ..., x_i, ..., x_p\right)}{1-\pi\left(x_1, ..., x_i, ..., x_p\right)} \\ & =\ln \frac{\frac{\pi\left(x_1, ..., x_i+1, ..., x_p\right)}{1-\pi\left(x_1, ..., x_i+1, ..., x_p\right)}}{\frac{\pi\left(x_1, ..., x_i, ..., x_p\right)}{1-\pi\left(x_1, ..., x_i, ..., x_p\right)}}=\beta_i \end{aligned}$$

und damit

$$\frac{\frac{\pi\left(x_1, ..., x_i+1, ..., x_p\right)}{1-\pi\left(x_1, ..., x_i+1, ..., x_p\right)}}{\frac{\pi\left(x_1, ..., x_i, ..., x_p\right)}{1-\pi\left(x_1, ..., x_i, ..., x_p\right)}}=\exp\left(\beta_i\right).$$

Störvariablen

Mehrere bivariate Analysen im Falle mehrerer erklärender Variablen sind nicht angemessen, da Effekte von Drittvariablen dann unter Umständen der erklärenden Variable im Modell zugeschrieben werden. Wird der Effekt einer interessierenden erklärenden Variable durch eine dritte Variable beeinflusst, so spricht man von einer „Störfaktor“ (confounder, Mischfaktor). Der Ausdruck wird in der Epidemiologie verwendet, um Variablen (Kovariablen) zu kennzeichnen, welche zusätzlich zur eigentlich interessierenden erklärenden Variable einen Einfluss auf die Zielvariable haben. Man sagt, die erklärte Variable sei gestört (confounded). So hängt das Gewicht z.B. vom Alter ab. Gewichtsgruppenvergleiche ohne die Berücksichtigung eventueller Altersunterschiede können damit irreführend sein. Will man den Zusammenhang einer Variable mit einer anderen analysieren, geht es also darum, den Einfluss weiterer relevanter Variablen zu *kontrollieren*. Jeder Koeffizient in einer multiplen logistischen Regression drückt dann eine Schätzung der Log-ORs aus, unter Berücksichtigung des Einflusses der übrigen Variablen. Ist das Modell geschickt gewählt, erhält man so den Einfluss der untersuchten erklärenden Variable auf die Zielvariable.

Beispiel 24. *Lineare Regression: Man möchte den Einfluss von Ernährungsgewohnheiten auf das Gewicht von Knaben untersuchen, wobei die Ernährungsgewohnheiten regional verschieden sind. Man zieht je eine Stichprobe aus den zu vergleichenden Regionen. Da man weiss, dass das Alter einen Einfluss auf das Gewicht von Knaben hat, muss man die Variable Alter berücksichtigen, falls sich die Altersmittel der beiden Stichproben unterscheiden. Wir gehen zudem davon aus, dass sich die Gruppen bezüglich weiterer möglicher Einflussfaktoren auf das Gewicht nicht unterscheiden. Wenn sich die beiden Gruppen auch bezüglich Alter nicht unterschieden, würde eine bivariate Analyse ausreichen. Die Schätzer ergäben die korrekten Werte für die Gruppenmitteldifferenz bezüglich des Gewichts. Wenn eine Gruppe aber jünger ist, können wir die Gruppen derart nicht sinnvoll untersuchen. Ein Teil der Gewichtsdifferenz wird dem Alter zuzuschreiben sein und nicht der Ernährung. Um die Situation zu analysieren, gehen wir davon aus, dass es in jeder Gruppe denselben affinen Zusammenhang zwischen Alter und Gewicht hat, d.h. es gilt allgemein*

$$w(x_2) = \beta_0 + \beta_2 x_2$$

(x_2 für das Alter und $w(x_2)$ für das Gewicht in Abhängigkeit vom Alter x_2, β_0 ist die Konstante. β_2 ist die Steigung der Geraden: wächst das Alter um eine Einheit, so steigt das Gewicht um β_2). Ist nun ein regionaler Ernährungseffekt zu beobachten, so gibt es einen Gewichtsunterschied zwischen den Gruppen, der durch den Altersunterschied nicht erklärt wird. Wir erhalten durch Hinzufügen der Variable Gruppenzugehörigkeit mit den Ausprägungen 0 *(für „gehört zur Gruppe 1“) und* 1 *(für „gehört zur Gruppe 2“) das Modell*

$$w(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

(x_1 für die Variable „Gruppenzugehörigkeit“). Damit gilt:

$$\begin{aligned} w_0(x_2) &:= w(0, x_2) = \beta_0 + \beta_1 0 + \beta_2 x_2 = \beta_0 + \beta_2 x_2 \\ w_1(x_2) &:= w(1, x_2) = \beta_0 + \beta_1 1 + \beta_2 x_2 = \beta_0 + \beta_1 + \beta_2 x_2 \end{aligned}$$

Die Gerade $w_1(x_2)$ resultiert also als Verschiebung von $w_0(x_2)$ um die Konstante β_1. Die beiden Geraden verlaufen entsprechend parallel - sie haben denselben Steigungsparameter β_2 und unterscheiden sich nur durch die Konstante β_1. Wegen

$$w(1, x_2) - w(0, x_2) = \beta_0 + \beta_1 + \beta_2 x_2 - (\beta_0 + \beta_2 x_2) = \beta_1 \tag{5}$$

ist β_1 ist der Ernährungseffekt ohne den Einfluss des Alters. Diese sieht man noch deutlicher, wenn man statt x_2 z.B. a_2 − das Durchschnittsalter der zweiten Gruppe − einsetzt. Hätten beide Gruppen dasselbe Durchschnittsalter, betrüge der Unterschied der Gruppen β_1.

Vergleicht man mit Hilfe des Modells die Gewichtsunterschiede an den Altersmittelwerten a_1 und a_2, so erhält man

$$w(1, a_2) - w(0, a_1) = \beta_0 + \beta_1 + \beta_2 a_2 - (\beta_0 + \beta_2 a_1) = \beta_1 + \beta_2 (a_2 - a_1). \tag{6}$$

$\beta_2 (a_2 - a_1)$ ist der Gewichtsunterschied, der dem unterschiedlichen mittleren Alter der Gruppen zuzuschreiben ist. Zu beachten ist, dass $w(1, a_2)$ und $w(0, a_1)$ die tatsächlichen mittleren Gewichte pro Gruppe sind – Regressionsgeraden verlaufen ja durch Mittelwerte beider Variablen. Die folgende Graphik mit den beiden Funktionen $w_0(x_2)$ und $w_1(x_2)$ kann die Sachlage verdeutlichen:

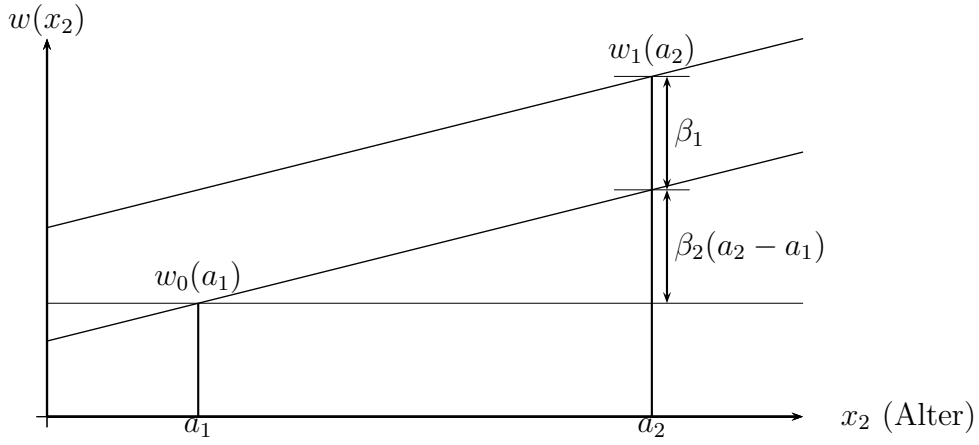

Abbildung 4: Regressionsgeraden durch zwei Gruppen; obere Gerade: $w_1(x) = \beta_0 + \beta_1 + \beta_2 x_2$; untere Gerade: $w_0(x) = \beta_0 + \beta_2 x_2$. Die im Beispiel konstante Differenz zwischen den zwei Gruppenregressionsgeraden beträgt β_1. a_i ist das mittlere Alter der Gruppe i. Die Gruppe 1 ist mit 0 kodiert, die Gruppe 2 mit 1.

β_2 ist als Steigung mit $\frac{x}{a_2 - a_2}$ identisch, wobei x die Distanz zwischen den Konstanten $w_0(a_2)$ und $w_0(a_1)$ ist. Entsprechend gilt für diese Distanz $x = \beta_2 (a_2 - a_1)$. Man sieht graphisch, dass sich der Alterseffekt und der Gruppeneffekt additiv zum Gesamtunterschied der Gruppen zusammensetzen (s. Gleichung 6). Ein Zahlenbeispiel zum eben diskutierten Beispiel kann die Sachlage zusätzlich erhellen. Mit den R-Befehlen

```
alter=rnorm(100,12,1)
altera=cbind(alter,rep(0,100))
alter2=rnorm(100,14,1)
alter2a=cbind(alter2,rep(1,100))
ag=rbind(altera,alter2a)
ag=as.data.frame(ag)
names(ag)=c("Alter","Gruppe")
ag$Gewicht= 15+2.5*ag$Alter+rnorm(200,0,1)
for (i in 1:200) if(ag$Gruppe[i]==1) ag$Gewicht[i]= ag$Gewicht[i]+2
```

erhält man einen Datensatz, welcher die Struktur des Beispiels hat (der Ernährung zuzuschreiben wären entsprechend ca. 2 kg, dem Altersunterschied der Gruppen ca. $2 \cdot 2.5$ kg – der durchschnittlich Altersunterschied zwischen den Gruppen beträgt ca. 2 Jahre). Zeichnet

man einen Streudiagramm der Daten, erhält man mit

```
plot(ag$Alter,ag$Gewicht)
malle=lm(ag$Gewicht~ag$Alter); abline(malle)
ag1=ag[ag$Gruppe==0,]   #daten Gruppe 1
mg1=lm(ag1$Gewicht~ag1$Alter); abline(mg1,lty=2)
ag2=ag[ag$Gruppe==1,]  #daten Gruppe 2
mg2=lm(ag2$Gewicht~ag2$Alter); abline(mg2,lty=3)
```

die Graphik

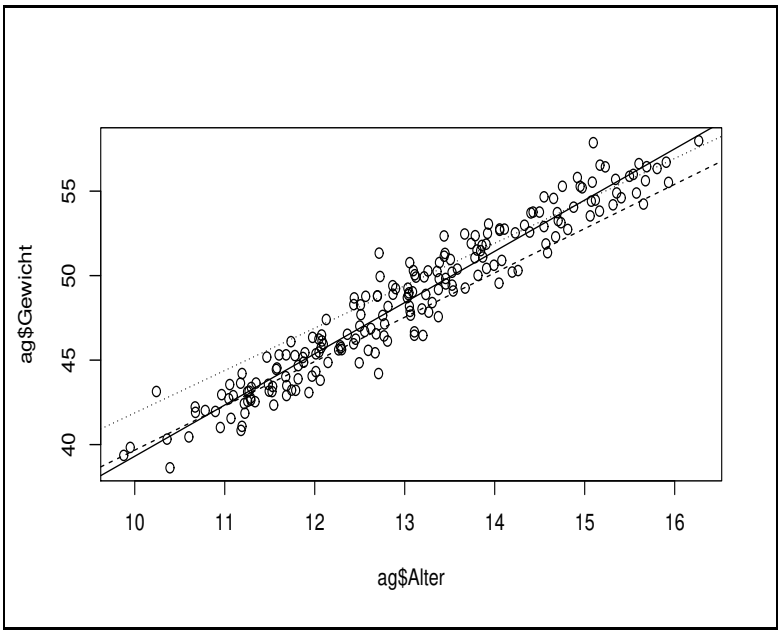

Abbildung 5: Durchgezogene Gerade: Regressionsgerade ohne Unterscheidung der Altergruppen; gepunktete Regressionsgerade: Gruppe 2; gestrichelte Regressionsgerade: Gruppe 1;

Rechnen wir eine bivariate Regression durch die Punktwolke, erhält man mit

```
(m2=lm(ag$Gewicht~ag$Gruppe))
```

die Grösse 7.163 *für den Gruppeneffekt (bei jedem Durchlauf wird eine andere Zahl produziert, wenn die Daten neu erzeugt werden. Die konkreten Zahlen des Beispiels sind zu finden in der Datei knaben.txt). Dieser Effekt ist offensichtlich zu gross, da ein Altersunterschied von* 2.065222 *vorliegt:*

```
mA1=mean(ag$Alter[ag$Gruppe==0])
mA2=mean(ag$Alter[ag$Gruppe==1])
mA1;mA2
mA2-mA1
```

mA1=11.98395, *mA2*=14.04917 *und mA2-mA1*= 2.065222. *Entsprechend wird der Ernährungseffekt durch die bivariate Regression überschätzt. Wir berechnen nun das Modell:*

lm(ag$Gewicht ~ag$Gruppe + ag$Alter)

und erhalten die Regressionsgerade

$$g(x_1, x_2) = 14.16143 + 1.87013x_1 + 2.56288x_2$$

Damit wird der Ernährungseffekt auf 1.87013 *reduziert. Es gilt:*

$$1.87013 + 2.56288\,(14.04917 - 11.98395) = 7.163$$

$(2.56288\,(14.04917 - 11.98395) = 5.\,292\,9$ *ist der Gewichtsunterschied, der dem unteschiedlichen Alter der beiden Gruppen zuzuschreiben ist). Der Gewichtsunterschied der Gruppen setzt sich additiv aus dem Gewichtsunterschied, der der Ernährung zuzuschreiben ist, und dem Gewichtsunterschied, der dem Altersunterschied zuzuschreiben ist, zusammen (s. Gleichung 6 von Seite 43).*

Im Falle der logistischen Regression, sind die Funktionswerte der Regressionsgerade Logits. Sonst ändert sich aber nicht viel: betrachtet man die (5) und (6) entsprechende Differenzen, ergeben sich auf der rechten Seite analoge Resultate, die aber bezüglich Quotenverhältnissen zu interpretieren sind. Sei die folgende Auswertungstabelle von Daten (Alter, Arztbesuch (ja, nein), 0.36 z.B. ist der Anteil der Personen mit Arztbesuch in der Gruppe 1) für zwei Gruppen von Männern gegeben:

	Gruppe 1		Gruppe 2	
Variable	Mittel	Std. Fehler	Mittel	Std. Fehler
Arztbesuch	0.36	0.485	0.80	0.404
Alter	39.65	5.272	47.34	5.259

Die univariaten Gruppenunterschiede ohne Berücksichtigung des Alters sind

$$\widehat{OR} = \frac{\frac{0.8}{0.2}}{\frac{0.36}{0.64}} = 7.1111$$

und

$$\ln \widehat{OR} = \ln\left(\frac{0.8}{0.2}\right) - \ln\left(\frac{0.36}{0.64}\right) = \ln(7.1111) = 1.\,961\,7$$

Es liegt ein beträchtlicher mittlerer Altersunterschiede vor. Deshalb und auf Grund inhaltlicher Überlegungen ist zu vermuten, dass ein Teil dieses ziemlich starken Unterschiedes der Gruppen dem Alter zuzuschreiben ist. Wir nehmen an, eine Auswertung der Daten würden die folgenden Ergebnisse zeigen:

Variable	Koeff.	St. Fehler	z	P>|z|
Gruppe	1.263	0.5361	2.36	0.018
Alter	0.107	0.0465	2.31	0.021
Konstante	-4.866	1.9020	-2.56	0.11

Das Quotenverhältnis der Gruppe 2 (mit 1 kodiert) bezüglich Gruppe 1 (mit 0 kodiert) ist nun

$$\exp(1.263) = 3.536,$$

also kleiner als 7.1111. Ein grosser Teil des Unterschiedes zwischen den Gruppen ist damit mit dem Altersunterschied erklärbar.

Bei der Interpretation ist im Auge zu behalten, dass diese davon abhängt, dass das Modell passt (Linearität der Logits). Abweichungen von dieser Voraussetzung können die Resultate nutzlos machen. Eine solche Abweichung besteht in der Interaktion.

Interaktion und Störung

Die bisherigen Methoden sind bei Störung geeignet, wenn es keine Interaktion hat: die Störvariable hat auf allen Ebenen der erklärenden Variable denselben Einfluss auf die erklärte Variable. Im Falle einer erklärenden Variable (Risikofaktor) mit zwei Ausprägungen (z.B. Gruppe) und einer stetigen Kovariable (z.B. Alter) liegt also keine Interaktion vor, wenn bei beiden Ausprägungen des Risikofaktors der Einfluss das Alters auf die erklärte Variable identisch ist. Graphisch bedeutet das, dass man in jeder Gruppe für die Logits zwei parallele Linien erhält (Logits in Abhängigkeit vom Alter pro Gruppe). Im allgemeinen liegt keine Interaktion vor, wenn das Modell keine Terme zweiter oder höherer Ordnung mit zwei oder mehr Variablen enthält (d.h. keine Produkte von Variablen; Berücksichtigung der Interaktion erfolgt rechnerisch durch ein Modell mit einer neuen, zusätzlichen Variable, welche das Produkt der Werte der eventuell interagierenden Variablen enthält). Bei Interaktion beeinflusst die Kovariable die Wirkung des Risikofaktors in Abhängigkeit von dessen Ausprägung (Stufe). Man spricht von einem „Effektmodifikator“ (effect modifier). Das einfachste und am häufigsten verwendete Modell für die Berücksichtigung von Interaktion ist damit eines, das linear in den Logits ist - aber mit unterschiedlichen Steigungen der Geraden pro Gruppe. Im folgenden Beispiel ist das Alter eine Störvariable, es liegt aber keine Interaktion vor:

Modell	Constant	Sex	Age	Sex×Age	resid. Devianz	G (Devianz)
1	0.060	1.981			419.816	
2	-3.374	1.356	0.082		407.780	12.036
3	-4.216	4.239	0.103	-0.062	406.392	1.388

(Testwert des Vergleichs der Modelle 2 und 3: 1.388 mit einem Freiheitsgrad: p-Wert 0.238743142). Bemerkenswert ist, dass durch die Zunahme der Interaktion der Koeffizient von „Sex“ stark verändert wird. Dies ist gemäss Hosmer und Lemeshow (2000, S. 73) häufig der Fall. Im folgenden Modell hingegen liegt Interaktion vor:

Modell	Constant	Sex	Age	Sex×Age	resid. Devianz	G (Devianz)
1	0.201	2.386			376.712	
2	-6.672	1.274	0.166		338.688	38.024
3	-4.825	-7.838	0.121	0.205	330.654	8.034

(Testwert des Vergleichs der Modell 2 und 3: 8.034 mit einem Freiheitsgrad: p-Wert: 0.004590735). Damit ist das Alter in diesem zweiten Beispiel sowohl Störvariable als auch Effektmodifikator. Hosmer und Lemeshow (2000, S. 74) vertreten die Meinung, dass es oft sinnvoll ist, Störvariablen im Modell zu halten, selbst wenn sich die Modelle mit und ohne die betreffende Variable nicht signifikant unterscheiden: jede klinisch bedeutsame Veränderung der Koeffizienten ist ein Hinweis darauf, dass Störung vorliegt. Bei Effektmodifikatoren vertreten sie demgegenüber die Ansicht, dass nur solche mit signifikantem Modellunterschied berücksichtigt werden sollten - wobei sie ein 10% Signifikanzniveau vorauszusetzen scheinen.

Schätzen von Quotenverhältnissen beim Vorliegen von Interaktion

Schreibt man bei Vorliegen von Interaktion eines dichotomen, eines Risikofaktors f, dessen Einfluss auf die Zielvariable man untersuchen will, und einer Störvariable x die Gleichungen für die Logits auf beiden Stufen des Risikofaktors hin und zählt diese voneinander ab, erhält man die korrekten Schätzer für die logarithmierten Quotenverhältnisse. Wir betrachten ein solches Modell: dichotomer Risikofaktor f, Störvariable x und Interaktion $f \times x$:

$$g\left(\pi\left(f,x\right)\right) = \beta_0 + \beta_1 f + \beta_2 x + \beta_3\left(f \times x\right)$$

Die Differenz der Gleichungen für die zwei Faktorstufen ($f_0 = 0$; $f_1 = 1$) ergibt:

$$\begin{aligned} g\left(\pi\left(f_1,x\right)\right) - g\left(\pi\left(f_0,x\right)\right) &= \beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3\left(f_1 \times x\right) - \left(\beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3\left(f_0 \times x\right)\right) \\ &= \beta_0 + \beta_1 + \beta_2 x + \beta_3\left(1 \times x\right) - \left(\beta_0 + \beta_2 x\right) \\ &= \beta_1 + \beta_3 x \end{aligned}$$

Man erhält das Quotenverhältnis mit $\exp\left(\beta_1 + \beta_3 x\right)$. Dieses ist bei Interaktion also von der Störvariable x abhängig.

Der Schätzer für die Varianz ist

$$V\left(B_1 + B_3 x\right) = V\left(B_1\right) + x^2 V\left(B_3\right) + 2x KOV\left(B_1, B_3\right)$$

Auch diese ist von der Störvariable x abhängig. Die Endpunkte des α-VIs sind

$$\left(B_1 + B_3 x\right) \pm \Phi^{-1}\left(\alpha + \frac{1-\alpha}{2}\right) SE\left(B_1 + B_3 x\right)$$

Beispiel 25. *Im folgenden Beispiel werden die Daten des Beispiels von Seite 24 wiederverwendet, wobei die Variable LWT (Gewicht von Frauen) dichotomisiert wird (= Variable LWD: 1 wenn < 110 Pfund, sonst 0). Es handelt sich um den dichotomen Risikofaktor (unabhängige Variable), dessen Einfluss auf die Variable low (Gewicht von Neugeborenen, 1 = ungenügend, 0 = genügend) untersucht wird. Man vermutet, dass die Variable Alter Störvariable ist oder mit LWD interagiert: Man erhält für die drei Modelle mit der Zielvariable Geburtsgewicht low in Abhängigkeit von*

(1) LWD
(2) LWD und Alter
(3) LWD, Alter, Interaktion von Alter und LWD
die Ergebnisse:

Model	*Constant*	*LWD*	*AGE*	*LWD×AGE*	*rest. Devianz*	*G*	*p*
0	*-0.790*				*234.68*		
1	*-1.054*	*1.054*			*226.24*	*8.44*	*0.004*
2	*-0.027*	*1.010*	*-0.44*		*224.28*	*1.96*	*0.160*
3	*0.774*	*-1.944*	*-0.080*	*0.132*	*221.14*	*3.14*	*0.076*

Hosmer und Lemeshow (2000, S. 77) nehmen die folgende Interpretation vor: Im Modell 1 erhält man das Quotenverhältnis $\exp(1.054) = 2.869\,1$. *Das Modell 2 weicht nicht stark vom Modell 1 ab: Von Modell 1 zu Modell 2 ändert sich der LWD-Koeffizient nur um ca. 4 Prozent* ($\Delta\beta_{LWD} = 100 - \frac{1.010}{1.054} \cdot 100 = 4.174\,6\%$). *Das Alter ist also nicht eine starke Störvariable. Durch die Hinzunahme der Interaktion verändern sich die Koeffizienten aber stark. Die Modelle 2 und 3 liegen weiter auseinander (p-Wert 0.076). Es liegt also Interaktion von LWD mit AGE vor.*

Das Quotenverhältnis bezüglich LWD für ein spezifisches Alter a erhält man (s. o.) mit

$$\ln\left(\widehat{OR}(a)\right) = \beta_1 + \beta_3 a = -1.944 + 0.132a$$

und

$$\widehat{OR}(a) = \exp(0.132\,a - 1.944).$$

Mit R erhält man die Ergebnisse für das Modell 3 (m77_1) mittels:

```
bw$lwd[bw$lwt>=110]=0
bw$lwd[bw$lwt<110]=1
m77_1=glm(low~lwd+age+lwd*age,data=bw,family="binomial")
anova(m77_1,test="Chisq")
```

Die Varianz des logarithmierten Quotenverhältnisses beim Alter a ist dann mit der Kovarianzmatrix des Modells (vcov(m77_1)):

	(Intercept)	*lwd*	*age*	*lwd:age*
(Intercept)	*0.82827928*	*-0.82827928*	*-0.035266546*	*0.035266546*
lwd	*-0.82827928*	*2.97495564*	*0.035266546*	*-0.127603824*
age	*-0.03526655*	*0.03526655*	*0.001570894*	*-0.001570894*
lwd:age	*0.03526655*	*-0.12760382*	*-0.001570894*	*0.005730240*

$$\begin{aligned} V\left(\ln\widehat{(OR}(a))\right) &= V(B_1 + B_3 x) \\ &= V(B_1) + x^2 V(B_3) + 2x KOV(B_1, B_3) \\ &= 2.97495564 + a^2 0.005730240 + 2a(-0.12760382) \end{aligned}$$

Man erhält für verschiedene ausgewählte Alter die Quotenverhältnisse (Beispiele) und 0.95-VIs

Alter	$\widehat{OR}$	$\widehat{VI}$*:Untere Grenze*	$\widehat{VI}$*: Obere Grenze*
15	$\exp(-1.944 + 0.132 \cdot 15) = 1.036\,7$	$0.284\,1$	$3.782\,6$
20	$\exp(-1.944 + 0.132 \cdot 20) = 2.005\,7$	$0.909\,31$	$4.424\,1$
25	$\exp(-1.944 + 0.132 \cdot 25) = 3.880\,6$	$1.704\,7$	$8.834\,3$
30	$\exp(-1.944 + 0.132 \cdot 30) = 7.508\,2$	$1.942\,2$	29.025
35	$\exp(-1.944 + 0.132 \cdot 35) = 14.527$	$1.927\,0$	109.51

Berechnungsbeispiel für VIs: Varianz für 15: $2.97495564+15^2 0.005730240+2\cdot 15\,(-0.12760382) = 0.436\,15$
Unter Grenze des VIs für $a = 15$*:* $\exp\left(-1.944 + 0.132 \cdot 15 - 1.96 * \sqrt{0.436\,15}\right) = 0.284\,1$
Obere Grenze des VIs für $a = 15$*:* $\exp\left(-1.944 + 0.132 \cdot 15 + 1.96 * \sqrt{0.436\,15}\right) = 3.782\,6$
Das Quotenverhältnis ist also ab dem Alter von 25 für tiefgewichtige Frauen signifikant von 1 verschieden. Für 30jährige Frauen ist das Verhältnis (ein Kind mit zu tiefem Gewicht zu einem genügend schweren Kind zu gebären) 7.5-mal höher, wenn Sie tiefgewichtig statt hochgewichtig sind. Zudem sieht, man dass die VIs immer grösser werden.

Bemerkung 26. *Man erhält beim Modell (a für "age")*

$$Logit\,(LWD, a) = \beta_0 + \beta_1 LWD + \beta_2 a$$

für $LWD = 1$ *und* $LWD = 0$ *je eine affine Funktionen, welche die Logits in Abhängigkeit vom Alter ausdrücken:*

$$\begin{aligned} Logit\,(1, a) &= \beta_0 + \beta_1 \cdot 1 + \beta_2 a + \beta_3 \cdot 1 \cdot a \\ &= \beta_0 + \beta_1 + \beta_2 a + \beta_3 a \\ &= 0.774 - 1.944 + (0.132 - 0.080)\, a \\ &= 0.052\, a - 1.17 \end{aligned}$$

und

$$\begin{aligned} Logit\,(0, a) &= \beta_0 + \beta_1 \cdot 0 + \beta_2 a + \beta_3 \cdot 0 \cdot a \\ &= \beta_0 + \beta_2 a \\ &= 0.774 - 0.080 a \end{aligned}$$

Während die erste Funktion steigend ist, ist die zweite sinkend. Bei tiefgewichtigen Frauen (LWD=1) steigt mit dem Alter das Quotenverhältnis, ein zu leichtes Kind zu gebären (low=1), bei hochgewichtigen Frauen (LWD = 0), sinkt mit dem Alter das Quotenverhältnis, ein zu leichtes Kind zu gebären.
Da das minimale Alter 14 und das maximale Alter 45 ist, kann man die folgende Graphik erhalten (s. Abbildung 6).

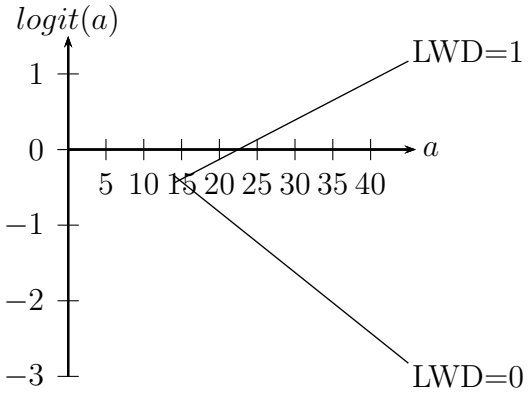

Abbildung 6: Geschätzte affine Funktionen (Logits auf den Stufen von LWD in Abhängigkeit vom Alter) bei Interaktion von Alter mit LWD

Wie oben bereits allgemein berechnet, erhält man durch Differenzbildung der beiden Logit-Funktionen $\ln\left(\widehat{OR}\,(a)\right)$:

$$\begin{aligned}\ln\left(\widehat{OR}\,(a)\right) &= Logit\,(1,a) - Logit\,(0,a)\\ &= 0.052a - 1.17 - (0.774 - 0.080a)\\ &= 0.132\,a - 1.\,944\end{aligned}$$

Vergleich der logistischen Regression mit der geschichteten Analyse von 2×2×p-Tabellen

Traditionell wurde eine in der Folge skizzierte geschichtete Analyse für 2×2×p-Tabellen vorgenommen, um Störung zu kontrollieren und Interaktion aufzudecken. Das wesentliche Ziel einer solchen Analyse ist die Überprüfung auf Konstanz der Quotenverhältnisse über die p Schichten hinweg. Bei konstanten Quotenverhältnissen wird ein geschichteter Quotenverhältnis-Schätzer wie der Mantel-Haenszel Schätzer (Mantel & Haenszel, 1959) oder ein gewichteter, logit-basierter Schätzer berechnet. Dieselbe Analyse kann man auch mit der logistischen Regressionsmodellierung erreichen. Die beiden Vorgehensweisen werden nun verglichen. Ein Beispiel aus den LWD-Daten kann die Gemeinsamkeiten und die Unterschiede illustrieren (low: Geburtsgewicht: $0 = \geq 2500g$. $1 = < 2500g$. Smoke: 0 = no Smoking during Pregnancy, 1 = yes).

```
table(bw$low,bw$smoke)
```

ergibt

		Smoke		
		1	0	Total
LOW	1	30	29	59
	0	44	86	130
	Total	74	115	189

Bei den Frauen, die normalgewichtige Kinder zur Welt bringen ($\geq$ 2500 g), rauchen viel weniger als bei den anderen. OR=$\frac{30 \cdot 86}{44 \cdot 29}$ = 2.021943574. Schichtet man nach Hautfarbe (p = 3), so sei

White		Smoke		
		1	0	Total
LOW	1	19	4	23
	0	33	40	73
	Total	52	44	96

OR:	$\frac{19 \cdot 40}{33 \cdot 4} = 5.7576$
$\ln(OR)$	$\ln(5.7576) = 1.7505$
$\mathrm{V}(\ln(OR))$	$\frac{1}{19} + \frac{1}{4} + \frac{1}{33} + \frac{1}{40} = 0.35793$
w	$\frac{1}{0.35793} = 2.7938$

black		Smoke		
		1	0	Total
LOW	1	6	5	11
	0	4	11	15
	Total	10	16	26

OR:	$\frac{6 \cdot 11}{4 \cdot 5} = 3.3$
$\ln(OR)$	$\ln(3.3) = 1.1939$
$\mathrm{V}(\ln(OR))$	$\frac{1}{6} + \frac{1}{5} + \frac{1}{4} + \frac{1}{11} = 0.70758$
w	$\frac{1}{0.70758} = 1.4133$

other		Smoke		
		1	0	Total
LOW	1	5	20	25
	0	7	35	42
	Total	12	55	67

OR:	$\frac{5 \cdot 35}{7 \cdot 20} = 1.25$
$\ln(OR)$	$\ln(1.25) = 0.22314$
$\mathrm{V}(\ln(OR))$	$\frac{1}{5} + \frac{1}{20} + \frac{1}{7} + \frac{1}{35} = 0.42143$
w	$\frac{1}{0.42143} = 2.3729$

Der Mantel-Haenszel-Schätzer des Quotenverhältnisses ist - mit

$$\begin{matrix} a_i & b_i \\ c_i & d_i \end{matrix}$$

und $N_i = a_i + b_i + c_i + d_i$ -

$$OR_{MH} = \frac{\sum \frac{a_i d_i}{N_i}}{\sum \frac{b_i c_i}{N_i}}$$

Im Beispiel ergibt sich:

$$\frac{\frac{19 \cdot 40}{96} + \frac{6 \cdot 11}{26} + \frac{5 \cdot 35}{67}}{\frac{33 \cdot 4}{96} + \frac{4 \cdot 5}{26} + \frac{7 \cdot 20}{67}} = 3.0864$$

(SPSS unter Analysieren, Kreuztabellen, Cochran-Mantel-Haenszel;
mit R: mantelhaen.test() in der offiziellen R-Distribution, im Beispiel

mantelhaen.test(bwlow,bwsmoke,bw$race); in R kann man auch dreidimensionale Kreuztabellen eingeben, s. für Beispiele Hilfe zum Test).
Die MH-Teststatistik ist (es wird getestet, ob die $\widehat{OR}$ von 1 verschieden ist):

$$X^2_{MH} = \frac{\left(\left|\sum\left(a_i - \frac{(a_i+b_i)(a_i+c_i)}{N_i}\right)\right| - 0.5\right)^2}{\sum\frac{(a_i+b_i)(a_i+c_i)(b_i+d_i)(c_i+d_i)}{N_i^3-N_i^2}}$$

(Im Beispiel gemäss R: Testwert 8.3779, df=1, p-Wert = 0.003798 (mit Excel nachvollzogen. Die Formel ergibt die Resultate des Programms; SPSS liefert 0.004; von SPSS wird noch ein Cochran-Wert geliefert, der grösser ist als der MH-Testwert, p-Wert ist entsprechend kleiner).

Der logit-basierte Schätzer der OR ist das gewichtete Mittel der schichtspezifischen logarithmierten ORs, wobei die Gewichte der Kehrwert der Varianz der ORs in den Schichten sind :

$$OR_L = \exp\left(\frac{\sum w_i \ln(OR_i)}{\sum w_i}\right)$$

Im Beispiel erhält man:

$$\exp\left(\frac{2.7938\ln(5.7576) + 1.4133\ln(3.3) + 2.3729\ln(1.25)}{2.7938 + 1.4133 + 2.3729}\right) = 2.9452$$

Im Allgemeinen liegen der MH- und der logitbasierte Schätzer nahe beieinander, wenn es nicht zu wenige Daten in den Schichten hat. Der Vorteil des MH-Schätzers ist, dass einzelne Zellen leer sein können.

Es ist wichtig hervorzuheben, dass diese Schätzer nur korrekte Schätzungen des Einflusses des Risikofaktors liefern, wenn die ORs über die Schichten hinweg konstant sind. Die hohen Schwankungen der ORs über die Schichten hinweg lassen vorerst vermuten, dass mittels Hautfarbe Störung oder Effektmodifikation vorliegt. Deshalb muss getestet werden ob diese Abweichungen gut oder schlecht durch Zufall erklärbar sind. Eine klassische Teststatistik dafür ist die folgenden Summe quadrierter Abweichungen:

$$X^2_H = \sum\left(w_i\left(\ln(OR_i) - \ln(OR_L)\right)^2\right)$$

Die Statistik ist bei genügend vielen Daten pro Schicht näherungsweise chiquadratverteilt mit $p-1$ Freiheitsgraden (p = Anzahl Schichten). Im Beispiel ergibt sich:

$$2.7938\left(\ln(5.7576) - \ln 2.9452\right)^2 + 1.4133\left(\ln(3.3) - \ln 2.9452\right)^2 + 2.3729\left(\ln(1.25) - \ln 2.9452\right)^2 = 3.0166$$

Entsprechend gilt: $p = 0.221$ (mit 2 df). Der Test liefert keinen Hinweis auf die Verletzung der Annahme. Die doch ziemlich grossen Unterschiede liegen also im Bereich zufälliger Schwankungen. Ein weiterer Test ist der Breslow und Day-Test (N. E. Breslow und Day (2000) und N. Breslow (1996)); in R nicht in offizieller Distribution; in SPSS in Ausgabe

für den MH-Test). Sie erhalten mit OR_{MH} 3.10 für die Teststatistik (SPSS: 3.126; manche Software-Pakte rechnen gemäss Autoren mit etwas anderen Formeln). Eine Version des BD-Testes ist der von Tarrone (in SPSS geliefert; Testwert 3.125).

Dieselbe Analyse - Kontrolle von Störung und Aufdecken von Interaktion - kann mit Hilfe der logistischen Regression durchgeführt werden und es ergeben sich vergleichbare Resultate. Gibt man mit R

```
bw$raceF=as.factor(bw$race)
m84_3=glm(low~smoke+raceF+smoke*raceF,data=bw,family="binomial")
anova(m84_3,test="Chisq")
```

ein, erhält man:

Analysis of Deviance Table
Model: binomial, link: logit
Response: low
Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			188	234.67	
smoke	1	4.8674	187	229.8	0.027369 *
raceF	2	9.8299	185	219.97	0.007336 **
smoke:raceF	2	3.1569	183	216.82	0.206291

Modell 1 unterscheidet sich signifikant vom Null-Modell. Das Modell mit raceF ebenfalls vom Modell ohne diese Variable (Testwert: $G = 229.8 - 219.97 = 9.8299$ mit zwei Freiheitsgraden). Das Modell mit Interaktion unterscheidet sich nicht signifikant vom Modell ohne Interaktion ($G = 219.97 - 216.82 = 3.1569$ mit 2 Freiheitsgraden). Um die Koeffizienten für Smoke in den verschiedenen Modelle zu erhalten, müssen diese durchgerechnet werden:

```
m84_1=glm(low~smoke,data=bw,family="binomial"); summary(m84_1)
m84_2=glm(low~smoke+raceF,data=bw,family="binomial"); summary(m84_2)
m84_3=glm(low~smoke+raceF+smoke*raceF,data=bw,family="binomial");
summary(m84_3)
```

und man erhält die Koeffizienten:

	Koeffizient Smoke	Geschätzte OR
m84_1	0.704	$\exp(0.704) = 2.0218$
m84_2	1.116	$\exp(1.116) = 3.0526$
m84_3	1.751	$\exp(1.751) = 5.7604$

Der Wechsel von 2.0218 (OR von Smoke und LWD) zu 3.0526 (OR von Smoke+raceF und LWD) durch die Aufnahme der Variable raceF weist auf Störung hin. Die Beurteilung der

Homogenität der ORs durch die Schichten von raceF hindurch wird mit dem Testen des Vergleichs von Modell 2 und 3 geleistet - da Modell 3 sich nicht signifikant von Modell 2 unterscheidet, liegt keine Effektmodifikation vor. Die Devianz (Testwert) für den Vergleich von Modell smoke+raceF+smoke*raceF mit dem Modell smoke+raceF ergibt einen Wert von 3.1569 mit 2 Freiheitsgraden. Die übrigen disktutierten Verfahren fürs Testen auf Gleichheit der ORs über die Schichten hinweg führen auf vergleichbare Testwerte:

	Testwerte mit jeweils 2 Freiheitsgraden
X^2_H	3.0166
Breslow Day	3.10

Vergleicht man das Resultat 3.0526 (Interaktion ist ja zu vernachlässigen!) mit denen der anderen OR-Schätzverfahren, ergibt sich:

OR_{MH}	3.0864
OR_L	2.9452
$OR_{Log\text{Reg}}$	3.0526

Die OR-Schätzer für raceF im Modell smoke+raceF 3.0526 liegt also nahe beim MH-Schätzer (oder bei der logitbasierten Schätzung).

Vertrauensbänder für geschätzte Logits und Wahrscheinlichkeiten

Gewöhnlich interessiert man sich bei der logistischen Regression für die Quotenverhältnisse. Es kommt aber durchaus vor, dass man sich für die Logits oder die Wahrscheinlichkeiten in Abhängigkeit von den Werten einer erklärenden Variable interessiert. In diesem Fall kann man an der graphischen Darstellung der geschätzten Logits und eines entsprechenden Vertrauensbandes oder der geschätzten Wahrscheinlichkeiten und eines VI-Bandes interessiert sein. Für die Logits erhält man mit R

```
m85=glm(low~lwt+raceF,data=bw,family="binomial")

logit=function(x) m85$coefficients["(Intercept)"]+m85$coefficients["lwt"]*x

plot(logit,xlim=c(90,240),ylim=c(-4.25,0.5),xlab="lwt")
A=vcov(m85)
varg=function(x) c(1,x,0,0) %*% A %*% c(1,x,0,0)
upper=function(x) logit(x)+1.96*sqrt(varg(x))
xpoints=seq(95,250,1)
ypointso=vector(length=length(xpoints))
for (i in 1:156) ypointso[i]=upper(i+95)
points(xpoints,ypointso,type="l",col="red")
under=function(x) logit(x)-1.96*sqrt(varg(x))
ypointsu=vector(length=length(xpoints))
for (i in 1:156) ypointsu[i]=under(i+95)
points(xpoints,ypointsu,type="l",col="red")
```

die Graphik:

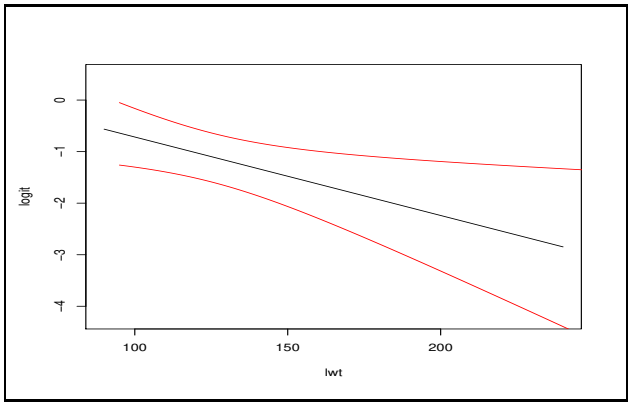

Abbildung 7: 0.95-Vertrauensband für die geschätzten Logits des Beispiels *Tiefes Geburtsgewicht* in Abhängigkeit vom *Gewicht bei letzter Menstruation* (lwt) unter Berücksichtigung der Variable *Race*. Zeichnung für Race=White

Ebenso kann man bezüglich der geschätzten Wahrscheinlichkeiten verfahren. Man erhält mit R mittels des eben definierten Funktion logit:

```
prob=function(x) exp(logit(x))/(1+exp(logit(x)))
plot(prob,xlim=c(95,240),ylim=c(0,0.6), xlab="lwt")
points(xpoints,exp(ypointso)/(1+exp(ypointso)),type="l",col="red")
points(xpoints,exp(ypointsu)/(1+exp(ypointsu)),type="l",col="red")
```

die Graphik:

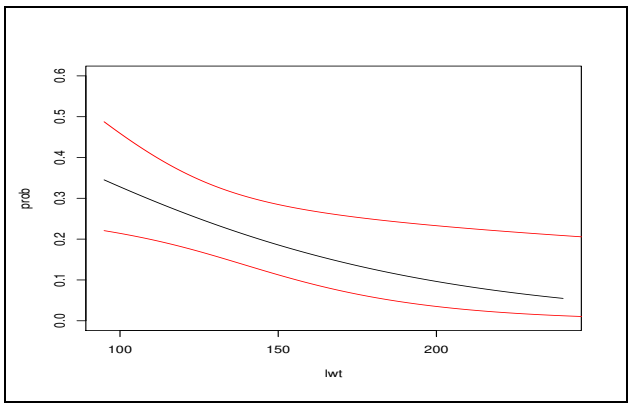

Abbildung 8: 0.95-Vertrauensband für die geschätzten Wahrscheinlichkeiten des Beispiels *Tiefes Geburtsgewicht* in Abhängigkeit vom *Gewicht bei letzter Menstruation* (lwt) unter Berücksichtigung der Variable *Race*. Zeichnung für Race=White

Derartige Graphiken kann man auch für raceF=black - dann ist $\mathbf{x} = (1, x, 1, 0)$ - und raceF = other - dann ist $\mathbf{x} = (1, x, 0, 1)$ - machen (passendes in der varg-Funktion ersetzen und die Logit-Regressions-Gerade mit den Koeffizienten berechnen). Die geschätzten

Wahrscheinlichkeiten kann man so deuten, dass der entsprechende Prozentsatz einer Population die Eigenschaft (= 1) der Zielvariable hat. Schätzungen sind nicht ausserhalb des Bereichs der vorliegenden Daten zu extrapolieren. Zudem sind die Schätzer nur gut, wenn das Modell passt. Die Anpassung wurde noch nicht überprüft.

4 Modellbildungsstrategien für die Logistische Regression

Liegen viele mögliche erklärende Variablen vor, so stellt sich die Frage, wie man optimale Modelle gewinnen kann. Das Ziel der Modellkonstruktion ist es, in einem wissenschaftlichem Zusammenhang das beste Modell zu finden. Um dieses Ziel zu erreichen, sollte man eine Vorgehensweise entwickeln, wie die Variablen fürs Modell auszuwählen sind, und Methoden, um die Adäquatheit des Modells zu überprüfen – bezüglich der einzelnen Variablen und bezüglich der Gesamtanpassung des Modells. Solche Methoden können dabei Sachverstand nicht ersetzen. Erfolgreiches Modellieren komplexer Variablenmengen ist teils Wissenschaft, teils Statistik, aber auch Erfahrung und gesunder Menschenverstand.

Kriterien für die Variablenselektion können von einem Modell zum anderen oder von einer wissenschaftlichen Disziplin zur anderen variieren. Der traditionelle Weg des Modellbildens sucht das sparsamste Modell, das die Daten erklärt. Der Grund für diese Vorgabe besteht darin, dass ein solches Modell eher numerisch stabil ist und einfacher verallgemeinerbar ist. Die Standardfehler werden kleiner und das Modell ist weniger von den Daten abhängig. Epidemiologen empfehlen demgegenüber, zusätzlich alle klinisch und intuitiv relevanten Variablen ins Modell aufzunehmen, unabhängig von allfälliger statistischer Signifikanz. Man möchte so weit als möglich alle potentiellen Störvariablen kontrollieren, dies auf dem Hintergrund der Erfahrung, dass einzelne Variablen wenig stören können, einige dieser Variablen zusammen jedoch ein beachtliches Störpotential aufweisen können. Der Nachteil einer solchen Vorgehensweise besteht in der Überanpassung des Modells an die Daten und die Erzeugung instabiler Schätzer. Überanpassung wird üblicherweise durch unrealistisch grosse Koeffizienten oder Standardfehler angezeigt. Dies kann besonders vorkommen, wenn die Anzahl der Variablen im Verhältnis zur Anzahl Daten oder zum Anteil von Personen mit einem spezifischen Antworteverhalten gross ist (s. für weitere nützliche Hinweise zur Modellkonstruktion Harrell, Lee und Mark (1996)). Der Selektionsprozess kann mit dem in der linearen Regression verglichen werden. Es erfolgt zuerst eine Übersicht über den Selektionsprozess in sechs Stufen

4.1 Erste Stufe des Selektionsprozesses: univariate Analysen

Der Selektionsprozess sollte mit einer sorgfältigen univariaten Analyse der Variablen starten (Häufigkeitsverteilungen). Sind bei manchen Variablen die Daten auf einer oder wenigen Ausprägungen massiert, so steigt die Gefahr von leeren Zellen bei zwei- oder mehrdimensionalen Kreuztabellen. Probleme können sich auch durch extrem schiefe Verteilungen ergeben.

4.2 Zweite Stufe des Selektionsprozesses: bivariate Analysen

Für nominale, ordinale und stetige Variablen mit wenigen Werten sollte man gemeinsame Häufigkeitsverteilungen mit der Zielvariable untersuchen (bivariate Analysen). Dabei kann

man den Likelihood-Verhältnistest für Kreuztabellen verwenden, der äquivalent ist mit den Ergebnissen des Likelihood-Verhältnistests der Logistischen Regression. So ergeben

```
b=table(bw$smoke,bw$lwd)
expected=chisq.test(b)$expected
2*sum((b*log(b/expected)))
```

und die Befehle

```
m76=glm(lwd~smoke,data=bw,family="binomial")
anova(m76,test="Chisq")
```

bei Deviance dasselbe Ergebnis (s. für die Beschreibung der Variablen Beispiel 14 S. 24, mit Dichotomisierung der Variable lwt in lwd gemäss Beispiel 25 von Seite 48).

Wegen der asymptotischen Äquivalenz des Likelihood-Verhältnistestes für Kreuztabellen mit dem Chiquadrat-Test, kann man auch letzteren verwenden. Bei mehreren Ausprägungen kann man zudem die einzelnen Quotenverhältnisse samt VIs betrachten, indem man eine der Ausprägungen als Referenz wählt.

Spezielle Beachtung erfordern Kreuztabellen mit leeren Zellen. Diese ergeben für eines oder mehrere Quotenverhältnisse entweder 0 oder Unendlich. Nimmt man eine solche Variable in ein logistisches Regressionsmodell auf, werden unerwünschte Resultate produziert (unrealistische Koeffizienten oder Standardfehler der Koeffizienten). Im Falle von leeren Zellen kann man inhaltlich sinnvoll Kategorien zusammenlegen, um die Nullzellen zu eliminieren. Bei ordinalskalierten Variablen schlagen Hosmer und Lemeshow (2000) für diesen Fall auch vor, diese als metrisch skaliert zu betrachten. Das kann man eventuell als fragwürdig betrachten oder man sieht es pragmatisch – interessante Ergebnisse sanktionieren eventuell fragwürdige Voraussetzungen.

Bei stetigen Variablen kann man eine einfache logistische Regression berechnen, und man begutachtet die geschätzten Koeffizienten, den geschätzten Standardfehler, die Wald-Statistik und den Likelihood-Verhältnis-Test. So kann man vorläufig den Einfluss der unabhängigen Variable auf die abhängige Variable beurteilen.

Die Anpassung des Modells der einfachen logistischen Regression an die Daten kann man mit Hilfe geglätteter Streudiagramme begutachten. Man kann dabei die Form der geglätteten Streudiagramme analysieren, um zu entscheiden, ob für die logistische Regression die Daten eventuell zu transformieren sind. Zudem kann man den Effekt extremer Werte einschätzen. Eine einfache Methode geglätteter Streudiagramme ist die Berechnung von relativen Häufigkeiten der klassifizierten Daten (s. Abbildung 2, S. 5). Es gibt komplexere Methoden, die sensibler auf die einzelnen Daten reagieren. Man kann z.B. für jedes Datum ein gewichtetes Mittel aller Daten berechnen. Das Gewicht jedes Datums ist eine sinkende Funktion der Distanz der Werte vom Datum, für das der geglättete Wert berechnet wird – gemäss Methoden, die in Cleveland (1993) dargestellt werden. In R kann man die R-Funktion loess („Local Polynomial Regression Fitting“ in der offiziellen R-Distribution im stat-Paket vorliegend) verwenden. Der Parameter „s“ (=span) kontrolliert den Glättungsgrad – bei kleineren Zahlen wird mehr geglättet, bei grösseren weniger.

Mit der folgenden R-Funktion (x = unabhängige Variable, y = abhängige Variable, s = Glättungsparameter der loess-Funktion)

```
plotSmooth_pi=function(x,y,s=0.75,main=NULL,xlab=NULL,ylab=NULL){
if (is.null(main)) main="Geglaettete Punkte und geschaetzte Regression-Kurve"
if(is.null(xlab)) xlab="x"
if(is.null(ylab)) ylab="y"
plot(y~x,main=main,xlab=xlab,ylab=ylab)
dat=cbind(x,predict(loess(y~x,span=s)))
dat=dat[order(dat[,1]),]
lines(dat[,1],dat[,2],col="red",lty=2)
m=glm(y~x,family="binomial")
b0=m$coefficients[1];b1=m$coefficients[2]
curve(exp(b0+b1*x)/(1+exp(b0+b1*x)),add=T)
if (b1>0) l="bottomright" else l= "topright"
legend(l,legend=c("Modell","geglaettet"),lty = c(1,2),col=c("black","red"))}
plotSmooth_pi(bw$age,bw$lwd,s=0.5,xlab="age",ylab="lwd"))
```

ergibt sich mit den eingezeichneten Daten, die auf 0 und 1 liegen, mit der geschätzten logistischen Funktion (Modell, durchgezogene Linie) und der gestrichelten geglätteten Kurve die Abbildung 9:

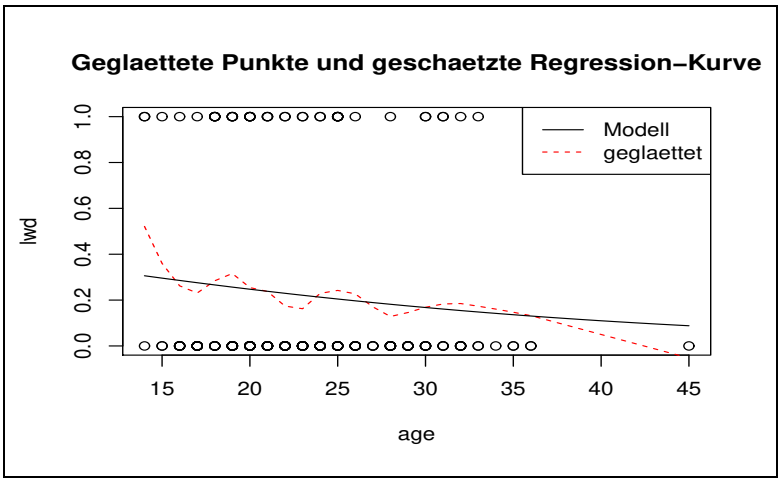

Abbildung 9: Mit der R-loess-Funktion geglätteter Punkteplot (gestrichelte Kurve) für die Daten, die auf 0 und 1 liegen, mit geschätzter logistischer Funktion (Modell, durchgezogene Fast-Gerade).

Man kann mit verschiedenen „s“ experimentieren. Die geglätteten Punkte können als grob geschätzte Wahrscheinlichkeiten betrachtet werden, dass in x der Wert 1 angenommen wird.

Die Ergebnisse der Funktion können auch auf die Logitskala umgerechnet werden, indem man für jeden oben berechneten geglätteten Wert y_s den Wert $\ln \frac{y_s}{1-y_s}$ rechnet und die Punkte mit einer Kurve verbindet. Mit diesem Plot kann man die Linearität der Logits überprüfen. Fürs Beispiel erhält man mit (es wird noch das geschätzte Modell, d.h. die Gerade $\hat{\beta}_0 + \hat{\beta}_1 x$ hinzugefügt, „l“ für Lage der Legende):

```
plotSmooth_logit= function(x,y,s=0.75,main=NULL,xlab=NULL,ylab=NULL,l=NULL) {
if (is.null(main)) main="Geglaettete Logits und geschaetzte Logit-Gerade"
if(is.null(xlab)) xlab="x"
if(is.null(ylab)) ylab="Logits"
logit = function(p) {log(p/(1-p))}
smooth = predict(loess(y~x,span=s))
p = pmax(pmin(smooth,0.9999),0.0001)
logits = logit(p)
dat=cbind(x,logits)
dat=dat[order(dat[,1]),]
m=glm(y~x,family="binomial")
b0=m$coefficients[1];b1=m$coefficients[2]
plot(dat[,1], dat[,2],lty=1,main=main,xlab=xlab,ylab=ylab)
lines(dat[,1], dat[,2],lty=2,col="red")
abline(b0,b1)
if(is.null(l)) {if (b1>0) l="bottomright" else l= "topright" }
legend(l,legend=c("Modell","geglaettet"),lty = 1,lty = c(1,2) , col=c("black","red" ))}
plotSmooth_logit(bw$age,bw$lwd,s=0.75,xlab="age",ylab="lwd",l="bottomleft")
```

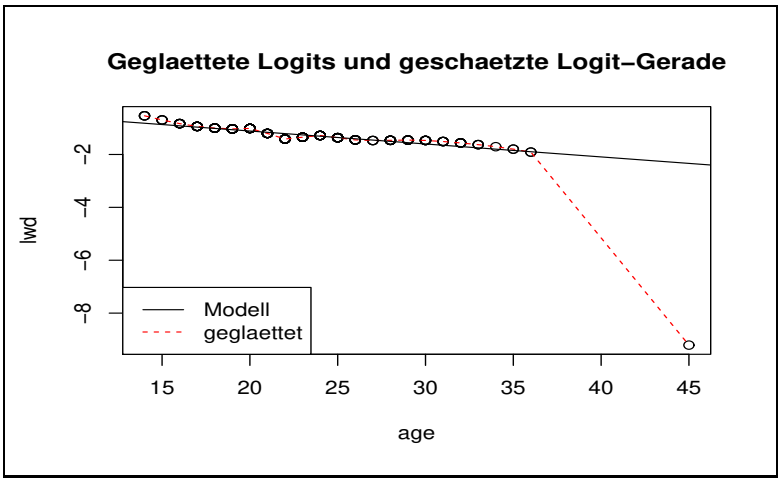

Abbildung 10: Geglätteter Punkteplot in der Logitskala: runde Punkte = $\ln \frac{y_x}{1-y_s}$; y_s = geglättetes Datum; gestrichelte Linie = Verbindung der geglätteten Punkte; Gerade = geschätzte Gerade $\hat{g}(\pi(x)) = \hat{\beta}_0 + \hat{\beta}_1 x$;

4.3 Dritte Stufe des Selektionsprozesses: Variablenselektion

Nach der Einzelanalyse der Variablen werden Variablen für die multivariate Analyse gewählt. Dazu gibt es verschiedene Verfahren: Das Schwellenverfahren legt eine Schwelle für die Aufnahme von Variablen fest. Ist der p-Wert in bivariaten Auswertungen kleiner als eine bestimmte Schwelle, wird die Variable ausgewählt. Weitere Variablenselektionsverfahren sind die schrittweise Ausweitung oder Einschränkung von Variablenmengen (Rückwärts- und Vorwärtselimination), sowie die Selektion der besten Teilmenge. Solche Verfahren wurden kritisiert, weil sie zur Aufnahme von inhaltlich irrelevanten Variablen führen können. Diese Gefahr besteht allerdings nicht, wenn man die Verfahren mit Umsicht anwendet: sie können inhaltliche Überlegungen nicht ersetzen. Gepaart mit inhaltlichem Sachverstand sind sie durchaus nützlich.

4.3.1 Das Schwellenverfahren

Jede Variable, deren zweidimensionales logistisches Modell für den Steigungskoeffizienten einen p-Werte z.B. kleiner als 0.25 ergibt, ist Kandidat für das multivariate Modell − zusätzlich zu allen Variablen mit bekannter klinischer oder inhaltlicher Relevanz. Man startet mit einem Modell mit all diesen Variablen. Ein 0.05-Niveau würde Variablen ausschliessen, die im Gesamtmodell wichtig sein können. Ein noch höheres Niveau würde Variablen berücksichtigen, die von zweifelhafter Relevanz sind. Ein Problem mit Einzeluntersuchung von Variablen ist, dass diese einzeln unbedeutend, im Paket mit anderen aber von etlicher Bedeutung sein können. Auch dies legt es nahe, nicht 0.05 bei der Variablenselektion zu wählen. Ein wesentlicher Gesichtspunkt ist auch die Anzahl der Daten. Je mehr Daten man hat (mit günstiger Verteilung der Antworten), desto eher kann man viele Variablen berücksichtigen. Liegen zu wenig Daten vor oder solche mit ungünstiger Verteilung der Antworten, können numerisch instabile Schätzer erzeugt werden. Die Waldstatistiken sind dann für die Variablenselektion nicht geeignet.

4.3.2 Stufenweise logistische Regression

Eine Zeitlang war die stufenweise Auswahl von Variablen verbreitet, geriet dann aber in Verruf. Mit Bedacht angewendet, können diese Methoden aber sehr nützlich sein − besonders wenn man ein Gebiet nicht gut kennt. Die Methoden ruhen auf einem statistischen Mass für Bedeutsamkeit. Es gibt verschiedene solcher Methoden: Rückwärts- und Vorwärtselimination sowie gemischte Methoden.

Die von H&L vorgestellte Vorwärtselemination − gemischt mit Rückwärtselimination − lässt sich wie folgt beschreiben:

1. Sie beginnt mit dem 0-Modell (nur Konstante). Dann wird für jede der p Variablen ein Modell $g(\pi(x_i) = \beta_0 + \beta_i x_i$, $i \in \{1, ..., p\}$, berechnet. Man berechnet für jedes der Modelle einen G-Test. Man vergleicht also das 0-Modell mit jedem der p Modelle. Die Variable, die zum kleinsten p-Wert führt, wird ins Modell aufgenommen, sofern von diesem p-Wert eine zu bestimmende Bedeutsamkeitsschwelle unterschritten

wird. Sie empfehlen ein Schwelle zwischen 0.15 und 0.2 − wobei auf Grund inhaltlicher Überlegungen auch eine höhere Schwelle, z.B. 0.25 ins Auge gefasst werden kann. Eine Schwelle von 0.05 würde zu viele Variablen ausschliessen − unter anderem Variablen, die nach der Aufnahme von weiteren Variablen signifikant werden könnten. Wir nehmen an, die Variable x_i sei gewählt worden.

2. Im Erweiterungsschritt wir für jede verbleibende Variable ein Modell $g(\pi(x_i, x_j) = \beta_0 + \beta_i x_i + \beta_j x_j$ berechnet, $j \in \{1, ..., p\} \backslash \{i\}$. Die Variable, die beim Modellvergleich $g(\pi(x_i))$ mit $g(\pi(x_i, x_j)$ zum kleinsten p-Wert führt, wird ins Modell aufgenommen, sofern die erwähnte Schwelle unterschritten wird.

3. Die bisher aufgenommen Variablen werden nach der Aufnahme der neuen Variable jeweils auf Wichtigkeit getestet (Rückwärts-Eliminationsschritt). Durch die Aufnahme einer Variablen kann nämlich der Beitrag einer bereits aufgenommenen Variable verschwinden oder bedeutungslos werden. Man vergleicht mit dem G-Test das Modell ohne die zu untersuchende Variable mit dem Modell, das im Erweiterungsschritt gewonnen wurde. Dazu wird eine zweite Schwelle definiert, die zur Elimination oder Beibehaltung von Variablen festzulegen ist − diese Eliminations-Schwelle muss grösser als die Aufnahme-Schwelle sein − sonst würden Variablen in einer Endlosschlaufe eliminiert und dann wieder aufgenommen.

4. Man geht wieder zum Erweiterungsschritt.

Das Verfahren stoppt, wenn keine zusätzlichen Variablen die Schwelle unterschreiten oder bei der Rückwärtselimination ausgeschieden werden.

Am Schluss kann man eventuell Variablen ausscheiden, welche die eigentliche Signifikanzschwelle nicht erreichen. Dazu gibt es zwei Methoden (q sei die Anzahl der Variablen im Modell, das durch die obige Vorwärtselimination (VE-Modell) berechnet wurde):

Die *erste Methode* überprüft die Variablen in der Ordnung ihrer Aufnahme ins VE-Modell bezüglich eines spezifischen Signifikanzniveaus (z.B. 0.05). Man vergleicht jeweils das Modell mit den ersten k-1 Variablen mit dem Modell mit den ersten k Variablen. Man bricht ab, wenn das k-Modell sich nicht signifikant vom k-1-Modell unterscheidet. Bei dieser Methode ergeben sich nicht unbedingt die G-Tests, die bei der oben beschriebenen Vorwärtselimination berechnet werden, da beim VE-Modell nicht dieselben Variablen vorkommen müssen, die bei diesen Tests auftauchten. Durch die Rückwärtselimination können Variablen eliminiert worden sein.

Bei der *zweiten Methode* werden die Modelle mit den ersten k Variablen − in der Ordnung ihrer Aufnahme ins VE-Modell − mit dem VE-Modell verglichen. Es wird also getestet, ob die Koeffizienten k+1 ... q von 0 verschiedene sind oder nicht. Man stoppt, wenn der erste solche Test nicht signifikant ist.

Eine Variante des eben beschriebenen Verfahrens besteht darin, mit einem Modell zu starten, das alle als wesentlich betrachteten Variablen enthält. In der Folge wird mit dem stufenweisen Vorgehen untersucht, ob weitere Variablen aufzunehmen sind.

Beispiel 27. *Als Beispiel liefern sie die UIS-Daten (s. Beispiel 28, S. 68), wobei sie für das Verfahren das Programm BMDPLR verwenden. In R gibt es zwei Verfahren: help.search("stepwise") ergibt (1) MASS::stepAIC − Choose a model by AIC in a Stepwise Algorithm und (2) stats::step − Choose a model by AIC in a Stepwise Algorithm). Es wird also nach dem AIC-Kriterium (Devianz+k·df, k = Strafe pro Parameter − gewöhnlich ist k = 2; df = Freiheitsgrade des Modells; die Variable mit dem tiefsten AIC wird bei der Rückwärtselimination weggelassen) gerechnet. Man kann Vorwärts- oder Rückwärtselimination oder beide (direction="both") anwenden (Voreinstellung ist "backwards"). z.B.*

```
m105=glm(dfree~age+beck+ndrgtx+ivhx+race+treat+site, data=uis,family="binomial")
m105st=step(m105)
summary(m105st)
```

Im letzten Durchgang erhält man:

Step: AIC=632.59
dfree ~age + ndrgtx + ivhx+ treat

	Df	*Deviance*	*AIC*
<none>		*620.59*	*632.59*
- treat	*1*	*625.80*	*635.80*
- ndrgtx	*1*	*628.07*	*638.07*
- age	*1*	*630.05*	*640.05*
- ivhx	*2*	*632.44*	*640.44*

*Man erhält das gleiche Modell wie im Buch (s. 123), aber nicht mit derselben Reihenfolge der Variablen (bei R wird die aufsteigende Reihenfolge der AICs übernommen; Es hat 5 Freiheitsgrade. 2*5=10; Es wird also jeweils 10 hinzugezählt, bei ivhx 8 − wobei nicht klar wird, wieso dort 8 und nicht 10 steht). Je nach Verfahren werden andere Variablenmengen ausgewählt.*
Mit library(MASS); stepAIC(m105) werden die nominalen Variablen nicht nach Hilfsvariablen aufgeteilt dargestellt. Mit SPSS: es werden 7 Verfahren offeriert (bei Methode; Erläuterungen unter Hilfe).

Das Verfahren kann auch für Interaktion angewendet werden. Man berechnet ein Modell mit allen Interaktionen und wendet dann den step-Befehl an.

4.3.3 Beste Teilmenge

Beste-Teilmengen-Modelle werden für die lineare Regression von verschiedenen Softwarepaketen angeboten. Lawless und Singhal (1987b) und Lawless und Singhal (1987a) erweitern die Methoden auf GLMs. Ihre Methode besteht in einer linearen Approximation der Matrix der Kreuzprodukte der Summenquadrate (Furnival-Wilson Algorithmus). Dadurch erhält man eine Approximation der Maximum-Likelihood-Schätzer. Die Modelle werden dann mit dem Modell verglichen, das alle Variablen enthält (MLL-Verhältnis-Tests). H&L

stellen ein Verfahren vor, wie man im Falle der logistischen Regression mit den Routinen für die lineare Regression vorgehen kann (S. 128 ff.).

Für R gibt es das Paket bestglm von McLeod und Xu (2014): Best Subset GLM. Mit

```
bestglm(uis2,family=binomial,IC="AIC")
```

erhält man

Morgan-Tatar search since family is non-gaussian.
Note: factors present with more than 2 levels.
AIC

Best Model:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age	1	0.27	0.2676	1.471	0.2256
ivhx	2	3.63	1.8127	9.967	5.57e-05 ***
ndrgtx	1	1.08	1.0777	5.925	0.0152 *
treat	1	0.96	0.9582	5.268	0.0221 *
Residuals	569	103.49	0.1819		

Man beachte die Beschriftung (nicht Abweichungen oder Abweichungsdifferenzen sondern Summenquadrate). Wählt man für IC (Informationskriterium) andere Kriterien, so erhält man nur ndrgtx. Die MLL-Verhältnistests werden nicht als Informationskriterium angeboten.

Hat man ein Modell angepasst, muss die Bedeutung jeder Variable im Modell überprüft werden. Dies erfolgt mit Hilfe der Waldstatistiken und den Likelihood-Verhältnistests (Modellvergleiche Modell mit und ohne diese Variable). Unterscheiden sich die Modelle nicht signifikant, verzichtet man vermutlich auf die Variable. Vorher sollte man noch die verbleibenden Koeffizienten mit denen des Modells mit der wegfallenden Variable vergleichen. Ergeben sich starke Veränderungen in der Schätzung der Koeffizienten? Solche Schwankungen zeigen an, dass die ausgeschlossene Variable einen wichtigen Einfluss bezüglich der Effekte verbleibender Variablen hat (Störung und Interaktion). Dies könnte dafür sprechen, die Variable doch nicht auszuschliessen. Dieser Prozess wird solange fortgeführt, bis man ein Modell mit allen wichtigen Variablen hat. Am Schluss sollte man jede ausgeschlossene Variable ins verbleibende Modell einschliessen, um die Effekte auf die verbleibenden Variablen zu testen. Am Ende dieses dritten Schrittes gelangt man auf ein Modell, dass sie „Vorläufiges Haupteffekt-Modell“ nennen.

4.4 Vierte Stufe des Selektionsprozesses: Modellverfeinerung

Für stetige Variablen des Vorläufigen Haupteffekt-Modells sollte man überprüfen, ob die Annahme der Linearität in den Logits haltbar ist. Eventuell sind die erklärenden, stetigen Variablen durch geeignete Transformationen anzupassen (Modellverfeinerung). Das Modell, dass man damit erhält, nennen sie „Haupteffekt-Modell“.

Die Methoden der Glättung von Streudiagrammen kann man schlecht auf multiple Modelle anwenden. Eine Ausnahme ist die Verwendung von klassifizierten Daten. Man wählt

die Quartile der Variablen als Klassengrenzen. Dann kreiert man eine kategoriale Variable mit 4 Ausprägungen. Man passt ein multiples Modell mit der neuen Variable an, welche die alte ersetzt. Man schafft also 3 Hilfsvariablen und wählt die tiefste Klasse als Referenzgruppe. Dann zeichnet man die geschätzten Koeffizienten an den Mitten zwischen den Quartilen. Für die erste Klasse wählt man den Punkt mit den Koordinaten (Mitte zwischen dem ersten Datum und dem ersten Quartil, 0). Die vier Punkte werden dann mit Linien verbunden. Diese sollten monoton wachsend oder sinkend sein, da die β_i die LogORs sind (s. 4 auf S. 37) und diese sollten bei Linearität der Logits mindestens monoton sein − für die konkrete Durchführung s. Beispiel 28 auf Seite 68.

Man kann auf diesem Hintergrund ein alternatives Modell wählen, wenn das lineare Modell nicht angemessen erscheint. Wählt man ein alternatives Modell, so testet man dieses − ist es signifikant vom alten verschieden und macht das Modell klinisch oder inhaltlich Sinn. Sind mehrere Modelle möglich, so wählt man dasjenige, das inhaltlich am meisten Sinn ergibt.

Eine diesbezügliche Methode besteht in der Verwendung von Fraktionalen Polynomen. Man möchte bestimmen, welcher Wert p in x^p die beste Wahl für eine erklärende Variable ist. Theoretisch könnt man p als weiteren zu schätzenden Parameter in die Modellschätzung einbauen. Dies verschärft die Komplexität der Schätzprozesse jedoch erheblich. Es ist besser, p aus einer kleinen Menge von möglichen Werte zu wählen und die entsprechenden Modelle durchzutesten. Das Modell, das linear in den Logits ist, ist

$$g\left(x,\boldsymbol{\beta}\right)=\beta_0+x\beta_1.$$

Diese Funktion kann man wie folgt verallgemeinern:

$$g\left(x,\boldsymbol{\beta}\right)=\beta_0+\sum_{j=1}^{J}F_j\left(x\right)\beta_j$$

wobei $F_j\left(x\right)$ Potenzfunktionen eines definierten Typs sind, nämlich

$$F_1\left(x\right)=\begin{cases}x^{p_1} & \text{für } p_1\in\{-2,-1,-0.5,0.5,1,2,3\}\\ \ln\left(x\right), & \text{für } p_1=0\end{cases}$$

Die verbleibenden Funktionen $j=2$ bis J werden definiert als

$$F_j\left(x\right)=\begin{cases}x^{p_j}, & \text{für } p_j\neq p_{j-1}\\ F_{j-1}\left(x\right)\ln\left(x\right), & \text{für } p_j=p_{j-1}\end{cases}$$

für $j=2,...,J$. Ist zum Beispiel $J=2$ und $p_1=0$, und $p_2=-0.5$ erhält man

$$g\left(x,\boldsymbol{\beta}\right)=\beta_0+\beta_1\ln\left(x\right)+\beta_2\frac{1}{\sqrt{x}}$$

oder ist $J=2$ und $p_1=2=p_2$ erhält man

$$g\left(x,\boldsymbol{\beta}\right)=\beta_0+\beta_1x^2+\beta_2x^2\ln\left(x\right)$$

Das Modell ist quadratisch mit $J = 2$ mit $p_1 = 1$ und $p_2 = 2$, d.h.

$$g(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \beta_2 x^2$$

Ist $J = 1$ und $p_1 = 2$ erhält man

$$g(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 x^2.$$

Die Methode erfordert, dass man bei $J = 1$ acht Modelle durchrechnet, bei $J = 2$ sechsunddreissig (Ziehen ohne Reihenfolge ohne Zurücklegen $\binom{8}{2} = 28$, zusätzlich 8 Möglichkeiten mit $p_j = p_{j-1}$). Es wurde vermutet und mit Hilfe von Simulationen bestätigt, dass jeder Term des fraktionalen Polynoms näherungsweise zwei Freiheitsgrade aufweist (einen für den Koeffizienten und den anderen für die Hochzahl). Entsprechend weist der Test, welcher das lineare Modell mit dem besten J=1-Modell vergleicht, einen Freiheitsgrad auf (unter der Nullhypothese gilt das lineare Modell; das J=1-Modell weist 2 Freiheitsgrade auf, das lineare Modell 1. Die Differenz ergibt 1). Der Test, welcher das beste J=1-Modell mit dem besten J=2-Modell vergleicht, weist zwei Freiheitsgrade auf (unter der Nullhypothese gilt das beste J=1-Modell; das J=2-Modell weist 4 Freiheitsgrade auf, das J=1-Modell 2. Die Differenz ergibt 2). Der Vergleich des besten J=2-Modells mit dem linearen Modell weist 3 Freiheitsgrade auf.

Sie schlagen dann die folgende Vorgehensweise vor: Fraktionale Polynome werden nicht benutzt, um Variablen auszuwählen, sondern nur um diese besser anzupassen. Die Menge der Variablen bleibt also bis auf die durch die Methode der Fraktionalen Polynome transformierte Variable unverändert. Man ordnet die Variablen gemäss Wald-Test − Variablen mit dem kleinsten p-Wert kommen zuerst. Für jede Variable, bei denen die Linearität in den Logits fragwürdig ist, wird von oben nach unten wie folgt zweistufig verfahren: Zuerst testet man das beste J=2-Modell gegen das lineare Modell. Ist der Unterschied signifikant, testet man das beste J=1-Modell gegen das beste J=2-Modell, sonst stoppt man und belässt die ursprüngliche Variable. Ist der Unterschied des besten J=1-Modell gegen das beste J=2-Modell signifikant, wählt man die J=2-Modell-Version, sonst die J=1-Modell-Version (Hosmer und Lemeshow (2000, S. 102), s. auch Sauerbrei, Meier-Hirmer, Benner und Roystond (2006)). Hat man die Variablen so angepasst, überprüft man in einem zweiten Gang, wie die veränderten Variablen die Form der übrigen Variablen beeinflusst (immer noch linear, oder muss für manche Variablen eine (andere) Transformation gewählt werden?). Dies wird solange verfolgt, bis alles stabil ist. Wichtig ist, dass man dabei inhaltliche Gesichtspunkte berücksichtigt. Sie diskutieren dann noch weitere Möglichkeiten (S. 103).

Besonders erwähnenswert ist ein manchmal auftretender Spezialfall: werden Personen befragt, wie viele Zigaretten sie pro Tag rauchen, so gibt es wegen den Nichtrauchern mehr Nullen als gemäss einem logistischen Modells zu erwarten ist. Zudem ist die Verteilung rechtsschief. Es wurde vorgeschlagen, die Situation mit

$$g(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 d + \beta_2 x$$

zu modellieren (mit der dichotomen Variable d − Raucher oder nicht − und der Variable x Anzahl Zigaretten. $d = 0$ wenn $x = 0$ und $d = 1$ wenn $x > 0$, s. S. 104).

Beispiel 28. *(UIS-Datensatz; UMARU IMPACT Study, S. 28) Man möchte den Einfluss der Variable „Treatment Randomization Assignment“ (0 = Short, 1 = Long) (treat) auf die Variable DFEE (Remained Drug Free = 1, otherwise = 0) untersuchen − unter Kontrolle von Störvariablen. Als mögliche Kandidaten werden die Variablen „age of enrollment“ (age), „Beck Depression Score“ (pseudometrisch) (beck), Drug Use History (1 = never, 2 = previous, 3 = recent) (IVHX), „Number of Prior Drug Treatments“ (ndrgtx), race (0 =white, 1 = other), und "Treatment Site (0 = A und 1 = B) (site) betrachtet.*
Schritt 1: *Modelle mit jeweils einer Variable:*

Variable	*Estimate*	*Std. Error*	*z value*	*Pr(>|z|)*	*G*	*Pr(>G)*
age	*0.01817*	*0.01534*	*1.184*	*0.23628*	*1.398*	*0.2371*
beck	*-0.008225*	*0.010343*	*-0.795*	*0.426*	*0.63649*	*0.425*
ndrgtx	*-0.07496*	*0.02468*	*-3.037*	*0.00239 ***	*11.839*	*0.00058 ****
ivhx2	*-0.4810*	*0.2657*	*-1.810*	*0.070242 .*		
ivhx3	*-0.7748*	*0.2166*	*-3.578*	*0.000347 ****	*13.352*	*0.001261 ***
race	*0.4591*	*0.2110*	*2.176*	*0.0295 **	*4.6235*	*0.03154 **
treat	*0.4372*	*0.1931*	*2.264*	*0.0236 **	*5.1782*	*0.02287 **
site	*0.2642*	*0.2034*	*1.299*	*0.194*	*1.6659*	*0.1968*

Man könnte auch die Quotenverhältnisse angeben ($\exp(Estimate)$*, z.B.* $\exp(0.4591) = 1.582\,6$ *oder Quotenverhältnisse für längere Intervalle, z.B. für zehn Jahre* $\exp(10 \cdot 0.01817) = 1.199\,3$ *oder für beck (depression score) 5 Punkte-Schritte:* $\exp(5 \cdot -0.008225) = 0.959\,71$. $\exp(10 \cdot -0.008225) = 0.921\,04$ *etc.*
Man kann zudem die VIs angeben, z.B. für
race: $\exp(0.4591 \pm 1.96 \cdot 0.2110) :]1.046\,6, 2.393\,3[$
age: $\exp(10 \cdot 0.01817 \pm 10 \cdot 1.96 \cdot 0.01534)$ *und*
beck: $\exp(5 \cdot (-0.008225) \pm 5 \cdot 1.96 \cdot 0.010343)$
oder auch mit der Profile-Likelihood-Methode.

Würde man als Kriterium für die Aufnahme ins Modell eine Signifikanzniveau von 0.05 wählen, würden alle Variablen ausser beck, race und site ins Modell aufzunehmen sein. Inhaltliche Überlegungen (Diskussionen mit Fachpersonen) veranlassen sie jedoch, age und race ins Modell aufzunehmen. site wird berücksichtigt, weil die Stichproben innerhalb von Standorten erhoben wurden (Kontrolle von Unabhängigkeit?).

Schritt 2*: Man überprüft die Logit-Linearität der stetigen Variable age bezüglich der Zielvariable dfree. Mit der oben gelieferten R-Funktion*

```
plotSmooth_pi(uis$age,uis$dfree,s=0.75,xlab="age", ylab="dfree")
```

erhält man:

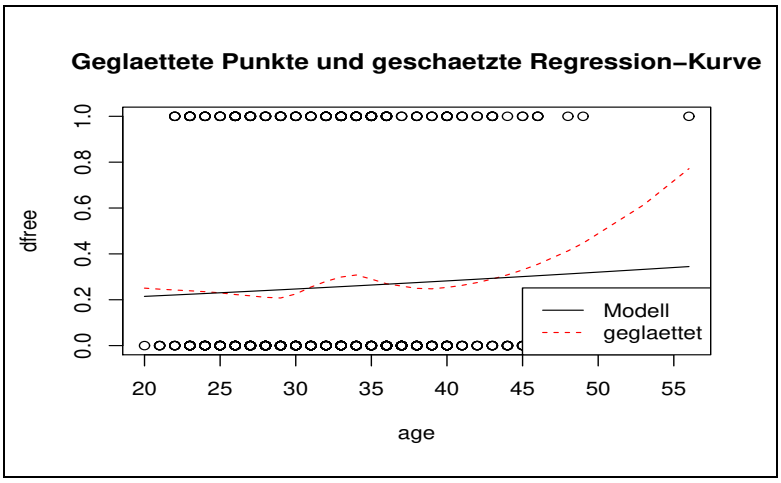

Abbildung 11: Geglätteter Punkteplot für die Daten mit geschätzter logistischer Funktion

Mit der R-Funktion (s. S.

plotSmooth_logit(uisage,uisdfree,s=1,xlab="age")

erhält man

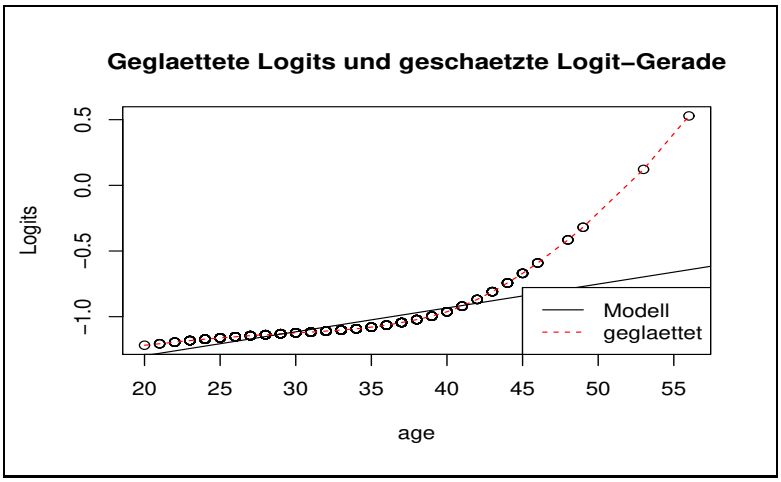

Abbildung 12: Geglätteter Punkteplot für die Daten auf der Logit-Skala mit der Logit-Geraden

Man sieht, dass die Logits bei etwas unter 40 die Richtung ändern. H&L meinen, man könne trotzdem davon ausgehen, dass die Linearität der Logits gegeben ist.

*Mit der zweiten Methode (stetige Variable in Quartile unterteilt und das Model mit den neuen Variablen durchgerechnet; Graphische Darstellung Mitten der Quartile * Koeffizienten) erhält man mit der R-Funktion zur Klassifikation von Daten:*

uis$ageklass=cut(uis$age,quantile(uis$age),include.lowest=T)

die neue Variable uis$ageklass . Damit wird eine neues Modell gerechnet:

```
m555=glm(dfree~ageklass+ndrgtx+ivhx+race+treat+site, data=uis,family="binomial")
```

(die klassifizierte Variable muss an erste Stelle im Modell erscheinen) und dann erhält man mit der folgenden R-Funktion (xklass = mit Quantilen klassifizierte Variable (im Beispiel uis$ageklass), modell = Resultat des glm-Befehls (im Beispiel m555), plotvi=T zeichnet Vertrauensbänder):

```
plotKoeff=function(xklass, modell, xlab=NULL,main=NULL,plotvi=F,pos="topleft"){
intervalle=levels(xklass); a=sub("\\[", "\\1", levels(xklass))
a=sub("\\(", "\\1", a); a=sub("\\]", "\\1", a); a=strsplit(a,",")
quant=rep(NA,5)
for (i in (1:4)) quant[i]=a[[i]][1]; quant[5]=a[[4]][2]; a=as.numeric(quant)
mitten=rep(NA,4); for (i in 1:4) mitten[i]=a[i]+(a[i+1]-a[i])/2
tabelle=mitten; tabelle=rbind(tabelle, as.integer(table(xklass)))
tabelle=rbind(tabelle,modell$coefficients[1:4])
tabelle[3,1]=0; vi=confint(modell); tabelle=rbind(tabelle,c(0,vi[2:4,2]))
tabelle=rbind(tabelle,c(0,vi[2:4,1]))
rownames(tabelle)=c("Intervallmitten","Anzahl Daten", "Koeffizienten",
   "vi_ober_Grenz","vi_unter-Grenz")
colnames(tabelle)=c("1. Intervall", "2. Intervall", "3. Intervall","4. Intervall")
plot(mitten[1:4],c(0,modell$coefficients[2:4]),type="b",
   ylim=c(min(tabelle[5,]), max(tabelle[4,])),
   ylab="Koeffizienten",xlab=xlab,main=main)
if (plotvi==T) { lines(x=mitten[2:4],y=tabelle[5,2:4],col=3,lty=2)
   lines(x=mitten[2:4],y=tabelle[4,2:4],col=2,lty=2)
legend(pos,title="VIs",legend=c("og","ug"),lty=2,col=c(2,3))}
abline(0,0); tabelle}
```

und dem Befehl

```
plotKoeff(uis$ageklass,m555,plotvi=T)
```

die graphische Darstellung

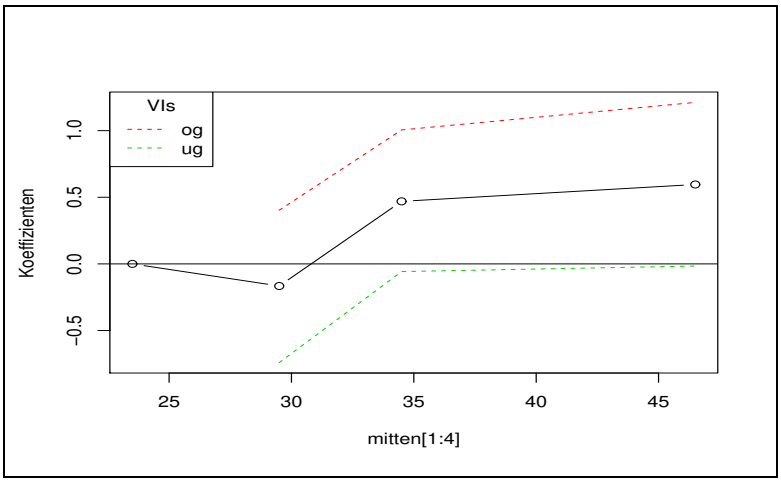

Abbildung 13: Mitten der Quartile der Variable *age* und Koeffizienten der entsprechend klassifizierten Daten der Variable *age* samt den geschätzten Koeffizienten – Referenzfaktorausprägung ist das erste Quartil

Zudem erhält man die Tabelle:

	1. Quartil	*2. Quartil*	*3. Quartil*	*4. Quartil*
Mitten	*23.5*	*29.5*	*34.5*	*46.5*
Anzahl	*148*	*144*	*166*	*117*
Koeffizienten	*0.0*	*-0.1658640*	*0.46933987*	*0.59577100*
vi_ober_Grenz	*0.0*	*0.4026872*	*1.00538653*	*1.21182679*
vi_unter-Grenz	*0.0*	*-0.7409699*	*-0.05773024*	*-0.01609988*

Die VI-Grenzen wurden im Gegensatz zur Tabelle auf Seite 107 des Buches von H&L mit der Profile-Likelihood-Methode berechnet; wie man auf die Klassenmitten im Buch kommt, ist nicht klar.

Die Koeffizienten sind zwar nicht durchgehend steigend oder durchgehend sinkend. Allerdings enthalten alle VIs der Koeffizienten 0, so dass die These eines linearen Trends vertretbar ist. L&H weisen darauf hin, dass diese Art von Untersuchung nicht ein besonders kräftiges Analyseinstrument ist. In der Tat sieht das Resultat z.B. ziemlich anders aus, wenn man die Quartilsintervallgrenzen anders definiert (statt]] [[). Die Ergebnisse seien den statistischen Laien aber relativ einfach vermittelbar.

Ist die Linearität nicht gegeben, wird man nun versuchen, die Variablen mit Hilfe der Methode der fraktionalen Polynome zu transformieren. Das R-Paket mfp führt die Analyse für alle Variablen auf einmal durch – mit dem Befehl:

```
mfp(dfree~age+ndrgtx+ivhx+race+treat+site, data=uis,family="binomial")
```

Es werden im Beispiel keine alternative Transformationen von Variablen vorgeschlagen. Es erfolgt allerdings kein detaillierter Output bezüglich der verschiedenen Tests.

Als nächstes wird die Variable ndrgtx nach den drei Methoden untersucht. Mit

```
plotSmooth_logit(uis$ndrgtx,uis$dfree,xlab="ndrgtx", s=0.5)
```

erhält man

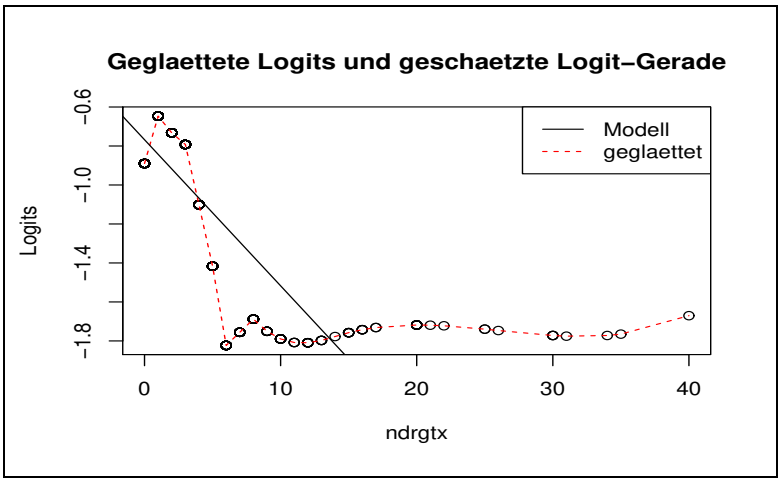

Abbildung 14: Plot der geglätteten Logits - stetige Variable ndrgtx gegen dichotome Variable dfree

Die Linearität ist nicht wirklich gegeben, besonders wenn man die geglättete Kurve mit dem geschätzten Modell vergleicht. Das Resultat stimmt im Detail relativ schlecht mit dem Ergebnis im Buch (S. 110) überein. Übereinstimmung besteht im Ansteigen zu Beginn. Die Kurve geht dann aber schneller nach unten als im Buch. Eine ähnlich Tendenz wie im Buch erhält man bis 15, wenn man $s = 0.8$ eingibt. Im Gegensatz zum Buch steigt dann die Kurve aber nachher ziemlich stark und monoton an. Die Abflachung nach unten ergibt sich nur bei kleineren s, bei denen aber die Kurve vor 15 sich anders verhält. Besonders interessant ist die Frage, ob die anfängliche Steigung ein Artefakt ist oder inhaltlich sinnvoll interpretiert werden kann.

In einem zweiten Schritt verwenden sie die Methode mit den Hilfsvariablen an, allerdings nicht mit Quartilen, sondern mit Quantilen, die sie auf Grund der obigen Graphik festlegen: sie klassifizieren nach 0, 1 - 2, 3 - 15, 16 - 40 (rechtsoffene Intervalle). Man erhält mit

```
uis$ndrgtxklass=cut(uis$ndrgtx,c(0,0.99,2,16,max(uis$ndrgtx),include.lowest=T)
m601=glm(dfree~ndrgtxklass+age+ivhx+race+treat+site,data=uis,family="binomial")
```

und mit dem Befehl plotKoeff

```
plotKoeff(uis$ndrgtxklass,m601,main="NDRGtX*Koeffizienten",plotvi=T,pos=
"bottomleft")
```

die Abbildung

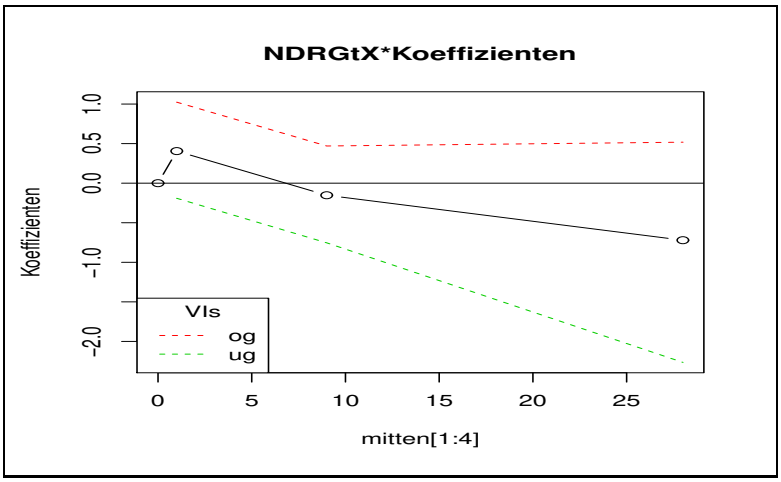

Abbildung 15: Plot der Koeffizienten der klassifizierten Variable ndrgtx

Man erhält also bis auf den letzten Koeffizienten ein ähnliches Muster wie beim Logit-Plot. Die Wald-VIs der Hilfsvariablen enthalten alle 0, obwohl die Variable ndrgtx einen signifikanten Beitrag zum Modell leistet. Die Zahlen, mit denen die Graphik erstellt wurde, sind:

	Estimate	Std. Error	z value	Pr($>\|z\|$)
ndrgtxklass2	0.40601	0.30902	1.314	0.18890
ndrgtxklass3	-0.15369	0.31168	-0.493	0.62193
ndrgtxklass4	-0.58528	0.62057	-0.943	0.34561

Mit der Methode der Fraktionalen Polynome erhält man mit dem mfp-R-Paket kein signifikantes Resultat (selbst auf dem 0.1-Niveau). H&L kommen hingegen auf das Modell J = 2 mit unten stehender Transformation (p-Wert 0.06). Sie befürworten entsprechend das Ersetzen der Variable ndrgtx durch die zwei Variablen

```
ndrgtx1=10/(ndrgtx+1)
ndrgtx2=ndrgtx1*ln (ndrgtx1^-1)
```

Es wurde eine R-Routine implementiert, welche die Methode von H&L durchführt, zu finden in der Datei „fractional_polynomial.r“. Man kann sie einbinden durch

```
source("pfad\\fractional_polynomial.r")
```

Verfolgt man ihre Schritte, so sieht man, dass sie beim Vergleich des besten Modell der Gruppe II (zwei identische Variablen zu verschiedenen Potenzen) mit dem linearen Modell

ein Signifikanzniveau von 0.1 wählen (G = 6.301555 und p-Wert 0.09782595). Im Vergleich des besten Modell der Gruppe II mit dem besten Modell der Gruppe I ergibt G = 5.871182 und p-Wert = 0.05309932.

H&L überprüften, ob das Modell klinisch Sinn macht. Die Experten bejahten (die Rückfallquote ist nach erster Behandlung kleiner, und steigt bei zwei bis drei Behandlungen an, bis sie wieder sinkt). Mit

```
uis$ndrgtx1=10/(uis$ndrgtx+1)
uis$ndrgtx2= uis$ndrgtx1*log((uis$ndrgtx+1)/10)
m601b=glm(dfree~age+ndrgtx1+ ndrgtx2+ivhx+race+treat+site, data=uis,
family="binomial")
summary(m601b)
```

erhält man die Resultate des Buches (die Konstante ist im Buch allerdings -2.928; S. 113, Sie verwenden die Variablennamen NDFRGPF1 statt ndrgtx1 und NDFRGPF2 statt ndrgtx2):

Coefficients:	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-4.31381	0.79246	-5.444	5.22e-08 ***
age	0.05445	0.01749	3.113	0.001850 **
ndrgtx1	0.98145	0.28885	3.398	0.000679 ***
ndrgtx2	0.36113	0.10986	3.287	0.001012 **
ivhx2	-0.60883	0.29111	-2.091	0.036490 *
ivhx3	-0.72381	0.25556	-2.832	0.004623 **
race	0.24770	0.22422	1.105	0.269267
treat	0.42237	0.20037	2.108	0.035033 *
site	0.17321	0.22098	0.784	0.433123

und mit: anova(m601b,test="Chisq")

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>G)
NULL			574	653.73	
age	1	1.40	573	652.33	0.2371
ndrgtx1	1	7.16	572	645.17	0.0074
ndrgtx2	1	14.00	571	631.16	0.0002
ivhx	2	11.40	569	619.77	0.0034
race	1	1.33	568	618.43	0.2481
treat	1	4.37	567	614.06	0.0365
site	1	0.61	566	613.45	0.4346

Die beiden neuen Variablen verkleinern die residuale Devianz um 7.1632 und 14.0035 – neben ivhx mit einer Devianz von 11.3956 die weitaus grössten Verkleinerungen der residualen Devianz. Die untransformierte Variable ndrgtx hatte demgegenüber nur eine

Devianz von 11.839. Man gewinnt also eine Reduktion der residualen Devianz von 7.1632 + 14.0035 − 11.839 = 9. 327 7.

Als nächstes geht es darum, ob die vorliegenden Variablen beim Hinzufügen von ausgeschiedenen Variablen stabil sind (sind diese Störvariablen oder sind sie nunmehr signifikant?). Nach dieser Abklärung hat man das Haupt-Effekte-Modell erlangt.

4.5 Fünfte Stufe des Selektionsprozesses: Interaktionen

Überprüfung auf Interaktionen. Man listet Paare von Variablen auf, bei denen auf Grund inhaltlicher Überlegungen Interaktion realistisch ist. Sind solche Interaktionen statistisch signifikant, werden sie ins Modell aufgenommen. Diese Modell nennen sie „Vorläufiges Endmodell“. In der Folge wird das Vorliegen von Interaktionen überprüft. Sie empfehlen, dabei von inhaltlich sinnvollen Interaktionen auszugehen und finden, dass im vorliegenden Fall die Interkation aller Variablen inhaltlich Sinn macht. Sie fügen entsprechend jeweils eine Interaktion zu den übrigen Variablen hinzu und berechnen das Modell. Zu berechnen sind also 15 Modelle (6 erklärende Variablen ergeben $\frac{5(5+1)}{2} = 15$ Varianten). Bei ndrgtx1 und ndrgtx2 nehmen sie dabei beide Interaktionen und zählen die Devianzen zusammen. Mit

```
m601c=glm(dfree~age+age*ndrgtx1+ndrgtx1+ age*ndrgtx2+ivhx+race+treat+site,
data=uis,family="binomial")
anova(m601c,test="Chisq")
```

erhält man für diesen Fall für die beiden Interaktionen

	df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
age:ndrgtx1	1	7.7029	565	605.75	0.0055134 **
age:ndrgtx2	1	0.0855	564	605.66	0.7699979

und damit die Gesamt-Abweichung 7.7029 + 0.0855 = 7. 788 4 mit 2 Freiheitsgraden (p-Wert: 0.020359656). Die 15 Paare, die sich ergeben, werden auf der Seite 352 geliefert. Man erhält bei dieser Analyse zwei signifikante Interaktionen − die eben gelieferte und race*site mit dem p-Wert 0.004. Auf dem 0.1-Signifikanzniveau würde sich zusätzlich age*treat mit dem p-Wert 0.096 ergeben. Sie fügen dann alle diese drei Interaktionen zum Haupt-Effekte-Modell. Dieses Modell ergibt Wald-Statistiken für die beiden Koeffizienten für age*ndrgtx, die nicht signifikant sind von 0 verschieden sind. Der Zwei-Freiheitsgrade-Test bei der Entfernung der beiden Interaktionen ist aber signifikant (p = 0.026, die Devianzen für age:ndrgtx1 und age:ndrgtx2 verändern sich bei Berücksichtigung der zwei zusätzlichen Interaktionen age*treat und race*site nicht). Die unterschiedlichen Testresultate sind darauf zurückzuführen, dass es eine hohe Korrelation zwischen den beiden ndrgtx-Variablen gibt. Im Modell, dass neben race*site die beiden Interaktionen age*ndrgtx1 und age*ndrgtx2 enthält, verliert age*treat die Signifikanz (p-Wert 0.1068267). Da die Interaktion age:ndrgtx2 nicht bedeutend ist, lässt man diese weg. Auch bei diesem Modell ist

age*treat nicht signifikant (p-Wert 0.11099) und wird deshalb weggelassen. Man erhält also das Modell

```
m601d=glm(dfree~age+age*ndrgtx1+ndrgtx2+ndrgtx1+ivhx+race*site+race+treat+site,
data=uis,family="binomial")
```

und mit

```
summary(m601d)
```

die Ausgabe

Coefficients:	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-6.843864	1.219290	-5.613	1.99e-08 ***
age	0.116638	0.028874	4.040	5.36e-05 ***
ndrgtx1	1.669035	0.407144	4.099	4.14e-05 ***
ndrgtx2	0.433688	0.116903	3.710	0.000207 ***
ivhx2	-0.634631	0.298717	-2.125	0.033627 *
ivhx3	-0.704948	0.261578	-2.695	0.007039 **
race	0.684107	0.264134	2.590	0.009598 **
site	0.516201	0.254886	2.025	0.042845 *
treat	0.434925	0.203758	2.135	0.032800 *
age:ndrgtx1	-0.015270	0.006027	-2.534	0.011287 *
race:site	-1.429457	0.529778	-2.698	0.006971 **

Bemerkenswert ist, dass *site* und *race* durch die Berücksichtigung von Interaktion signifikant geworden sind (Waldstatistiken, im Buch Tabelle auf Seite 115). Die Devianzen von age, race und site sind nicht signifikant von 0 verschieden. Die Variablen tauchen aber in Interaktionen auf, deren Devianzen signifikant sind:

	Df	Deviance (G)	Resid. Df	Resid. Dev	Pr(> G)
NULL			574	653.73	
age	1	1.40	573	652.33	0.2371
ndrgtx1	1	7.16	572	645.17	0.0074
ndrgtx2	1	14.00	571	631.16	0.0002
ivhx	2	11.40	569	619.77	0.0034
race	1	1.33	568	618.43	0.2481
site	1	0.50	567	617.93	0.4791
treat	1	4.48	566	613.45	0.0342
age:ndrgtx1	1	7.70	565	605.75	0.0055
race:site	1	7.79	564	597.96	0.0053

Das neue Modell ist das Vorläufig Endmodell.

4.6 Sechste Stufe des Selektionsprozesses: Güte der Anpassung

Das Vorläufige Endmodell wird auf Güte der Anpassung untersucht (s. nächstes Kapitel). Ist diese gegeben, erhält man das Endmodell.

Numerische Probleme

Bestimmte Muster in den Daten können zu numerischen Problemen führen. Das einfachste Problem besteht darin, dass man in der Häufigkeitstabelle Nullen hat. Die geschätzten Quotenverhältnisse werden dann 0 oder Unendlich.

Beispiel 29. *Mit der Referenzgruppe 1 ergibt sich*

Ergebnis\x	*1*	*2*	*3*	*Total*
1	*7*	*12*	*20*	*39*
0	*13*	*8*	*0*	*21*
Total	*20*	*20*	*20*	*60*
OR		$\frac{13\cdot 12}{7\cdot 8} = 2.785\,7$	$\frac{20\cdot 13}{7\cdot 0} = \infty$	
$\ln(OR)$		$\ln\left(\frac{13\cdot 12}{7\cdot 8}\right) = 1.0245$	$\ln\left(\frac{20\cdot 13}{7\cdot 0}\right) = \infty$	
$\hat{\beta}$	*-0.62*	*1.03*	*11.7*	
$\widehat{SE}$	*0.47*	*0.65*	*34.9*	

Für die Berechnung wird diese Kreuztabelle in eine Urliste von Daten verwandelt (ein Spalte entspricht einer Variable, eine Zeile entspricht einem Untersuchungsobjekt):

```
dframe= c(rep(1,39),rep(0,21))
dframe=cbind(dframe,c(rep(1,7), rep(2,12),rep(3,20),rep(1,13),rep(2,8)))
colnames(dframe)=c("ab","unab")
dframe=as.data.frame(dframe)
dframe$unab=as.factor(dframe$unab)
```

Mit

```
mod137=glm(ab~unab,data=dframe,family="binomial")
summary(mod137)
```

erhält man

Coefficients:	*Estimate*	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	*-0.6190*	*0.4688*	*-1.320*	*0.187*
unab2	*1.0245*	*0.6543*	*1.566*	*0.117*
unab3	*20.1851*	*2404.6704*	*0.008*	*0.993*

Der im Verhältnis zum Schätzer absurd hohe Standardfehler ist ein Hinweis darauf, das ein Problem vorliegt. Im Buch steht (S. 137) bezüglich unab3 für den Schätzer 11.7 und für den Standardfehler 34.9. Unterschiede zwischen den Programmen sind ebenfalls ein deutlicher Hinweis auf numerische Probleme. Es gibt Programme, die eine Fehlermeldung liefern.

Eine gängige Praxis besteht darin, in der Kreuztabelle in allen Zellen 0.5 zu addieren. Dies funktioniert bei der Analyse von Kreuztabellen mit Hilfe von Quotenverhältnissen, nicht aber bei mehr Variablen.

Beim Testen auf Interaktion muss man dreidimensionale Tabellen begutachten.

Beispiel 30.

Schichten	*1*		*2*		*3*	
Ziel/x	*1*	*0*	*1*	*0*	*1*	*0*
1	*5*	*2*	*10*	*2*	*15*	*1*
0	*5*	*8*	*2*	*6*	*0*	*4*
Total	*10*	*10*	*12*	*8*	*15*	*5*
$\widehat{OR}$	*4*		*15*		∞	

Eine Zelle ist leer. Zuerst wird die Tabelle wiederum in eine Urliste von Daten verwandelt: mit

```
V1=c(rep(1,5), rep(0,5),rep(1,2),rep(0,8),rep(1,10),rep(0,2),rep(1,2),rep(0,6),
rep(1,15),1,rep(0,4))
V2=c(rep(1,10),rep(0,10),rep(1,12),rep(0,8),rep(1,15),rep(0,5))
V3=c(rep(1,20),rep(2,20),rep(3,20))
tab137c=as.data.frame(cbind(V1,V2,V3))
tab137c$V3=as.factor(tab137c$V3)
```

Ohne Berücksichtigung von Interaktion ergibt sich mit

```
m137c=glm(V1~ V2+V3,data=tab137c,family="binomial" )
summary(m137c)
```

die Tabelle

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.3189	0.7727	-3.00	0.0027
V2	2.7681	0.7156	3.87	0.0001
V32	1.1888	0.8118	1.46	0.1431
V33	2.0382	0.8889	2.29	0.0219

Bei Interaktion ergibt sich mit

```
m137cint=glm(V1~V2+V3+V2*V3,data=tab137c,family="binomial" )
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.3863	0.7906	-1.75	0.0795
V2	1.3863	1.0124	1.37	0.1709
V32	0.2877	1.1365	0.25	0.8002
V33	-0.0000	1.3693	-0.00	1.0000
V2:V32	1.3218	1.5138	0.87	0.3826
V2:V33	18.5661	1684.1387	0.01	0.9912

Man erkennt wiederum (1) den extrem grossen Koeffizienten und die sehr grosse Standardabweichung und (2) den Unterschied zu anderen Programmen (im Buch: 11.54 und 50.22). H&L bemerken, dass die Koeffizienten weniger sensibel auf das Problem reagieren als die Standardabweichungen. Entsprechend sollt man also vor allem letztere im Auge behalten.

Eine zweite Art von numerischen Problemen ergibt sich, wenn die Prädikatoren die Zielvariable sauber klassifizieren. Wenn z.B. aller Personen, die älter als 55 sind eine Eigenschaft aufweisen, alle darunter nicht. In diesem Falle existieren die ML-Schätzer nicht, da die LL-Funktion in diesem Fall monoton ist und damit keine Maximum existiert (Albert & Anderson, 1984). H&L liefern das folgende

Beispiel 31.

x	y
1	*0*
2	*0*
3	*0*
4	*0*
5	*0*
x_6	*0*
6	*1*
7	*1*
8	*1*
9	*1*
10	*1*
11	*1*

mit $x_6 = 5.5$ *oder* 6 *oder* 6.5 *oder* 6.1 *oder* 6.2 *oder* 8. *Mit* 5.5 *erhält man bei R mit*

```
dat138=as.data.frame(matrix(c(1,0,2,0,3,0,4,0,5,0,5.5,
0,6,1,7,1,8,1,9,1,10,1,11,1),ncol=2,byrow=T))
m138a=glm(V2~ V1,data=dat138,family="binomial" )
```

die folgende Fehlermeldung:

```
Warning messages:
1: glm.fit: Algorithmus konvergierte nicht
2: glm.fit: Angepasste Wahrscheinlichkeiten mit numerischem Wert 0
oder 1 aufgetreten
```

Auch bei $x_6 = 6 -$ berechnet mit

```
dat138b=dat138
dat138b[6,1]=6
m138b=glm(V2~ V1,data=dat138b,family="binomial" )
summary(m138b)
```

- taucht eine Fehlermeldung auf (die 2: von eben). Analog berechnet man die übrigen Modelle durch und erhält:

x_6	*(Intercept)*	*SE(Intercept)*	β_1	$SE(\beta_1)$
6.05	*-26.175*	*36.728*	*4.345*	*6.088*
6.1	*-21.978*	*25.393*	*3.636*	*4.184*
6.2	*-17.803*	*17.298*	*2.927*	*2.818*
8	*-6.1240*	*3.5853*	*0.9665*	*0.5312*

Es gibt Programme, welche bei 5.5 und 6 Parameter ausspucken – sie liefern einfach die letzten Ergebnisse vor dem Abbruch. Man sieht, dass die Koeffizienten und die Standardfehler kleiner werden, wenn das die Separierung vermeidende Datum weiter rechts liegt. Bei wenigen Daten und vielen Variablen ist die Chance gross, dass völlige Trennung vorliegt. Eine Möglichkeit, bei vielen Variablen Separierbarkeit zu entdecken, besteht in der Begutachtung der geschätzten SEs über die verschiedenen Approximationsrunden hinweg. Nehmen die SEs immer stärker zu, liegt vermutlich Separierbarkeit vor. Bei R kann man die Entwicklung der Devianz verfolgen, indem man eingibt:

```
glm(V2~V1,data=dat138f,family="binomial",epsilon=10^(-20), maxit = 25,trace=T)
```

Drittens ist der Schätzprozess sensibel auf das Vorliegen von Kollinearitäten der erklärenden Variablen. Die Koeffizienten und die SEs tendieren dann wiederum dazu, absurde Werte anzunehmen.

Beispiel 32. *Für die Tabelle 8*

ID	x1	x2	x3	y
1	0.22	0.23	1.03	0.00
2	0.49	0.49	1.02	1.00
3	-1.08	1.07	1.07	0.00
4	-0.87	-0.87	1.09	0.00
5	-0.58	-0.57	1.09	0.00
6	-0.64	-0.64	1.01	0.00
7	1.61	1.62	1.09	0.00
8	0.35	0.35	1.09	1.00
9	-1.02	-1.02	1.01	0.00
10	0.93	0.94	1.06	1.00

Tabelle 8: Künstliches Beispiel mit Daten

erhält man mit den Definitionen der drei Modelle m140, m141 und m142:

```
m140=glm(y~x1,data=dat140,family="binomial");summary(m140)
m141=glm(y~x1+x2,data=dat140,family="binomial");summary(m141)
m142=glm(y~x1+x2+x3,data=dat140,family="binomial");summary(m142)
```

das Ergebnis

m140	*Intercept*	*SE(Intercept)*	β_i	*SE*(β_i)
x_1	*-1.0017*	*0.8294*	*1.3803*	*1.0053*
m141				
x_1	*-0.4159*	*1.3689*	*124.4154*	*287.7183*
x_2			*-123.0143*	*287.2792*

Für das Modell m143 erfolgt eine Fehlermeldung (die 2. oben). In der Tat sind die ersten zwei Variablen korreliert (cor(dat140$x1,dat140$x2) ergibt: 0.7220211). Die Variablen wurden wie folgt geschaffen: $X_1 \sim N(0,1)$; $X_2 = X_1 + U(0, 0.1)$, $X_3 = 1 + U(0, 0.01)$. *Die dritte Variable ist also bezüglich der übrigen Variablen fast konstant. Auch hier zeigt sich, dass die Koeffizienten und Standardfehler absurd werden und zudem mit anderen Programmen nicht übereinstimmen (s. S. 141).*

5 Überprüfung der Güte der Anpassung des Modells

Nach der Modellanpassung, wie in den letzten Kapiteln beschrieben, muss in der Folge die Güte der Anpassung des Modells an die Daten überprüft werden. Sind die Daten $\mathbf{y}$ und die geschätzten Werte $\hat{\mathbf{y}}$ gegeben, ist die Anpassung gut, wenn (1) die Distanz zwischen diesen beiden Vektoren klein ist und (2) wenn der Beitrag der Distanzen zwischen den einzelnen Datenpaaren zur Gesamtdistanz unsystematisch und relativ klein ist. Um diese zwei Erfordernisse zu überprüfen, wurden (1) summative Kennzahlen der Güte der Anpassung, (2) Kennzahlen, welche dem Bestimmtheitsmass in der linearen Regression entsprechen sollen und (3) diagnostische Verfahren, um den Beitrag einzelner Daten oder Datenmuster zu den zusammenfassenden Statistiken zu bestimmen, entwickelt.

Zusammenfassende Kennzahlen der Güte der Anpassung

Summative Kennzahlen geben keine Auskunft über den Beitrag einzelner Daten zum Modell, können aber Hinweise auf ernsthafte Probleme des Modells liefern.

Bei den Güte-der-Anpassung-Tests spielen die Freiheitsgrade eine Rolle, die durch die Anzahl der „Kovariablenmuster“ bestimmt sind. Es handelt sich um die Muster, die durch die Kombinationen von Werten der erklärenden Variablen erzeugt werden. Bei drei dichotomen Kovariablen gibt es $2^3 = 8$ mögliche Muster, liegt eine stetige Variable vor, so kann es so viele Muster wie Daten haben. Sei n die Anzahl der Daten und J die Anzahl der Muster. Es gilt also $J \leq n$. Sei m_j die Anzahl Daten pro Muster j. Wenn die Anzahl der Kovariablenmuster mit n steigt, sind die m_j klein. Gibt es wenig Kovariablenmuster, deren Anzahl fix ist, so steigt m_j mit n. Verteilungsresultate, die nur auf steigendem n ruhen, nennt man n-asymptotisch, Verteilungsresultate, die auf steigenden m_j für alle j beruhen, m-asymptotisch (bei fixem J steigt m_j wenn n steigt - ist also m-asymptotisch). Sie starten mit dem Fall $J \approx n$, der in der Praxis auf häufigsten vorkomme (mindestens eine stetige Kovariable) und der am schwierigsten zu behandeln sei.

Pearson Chi-Quadrat-Statistik und Abweichung (Devianz) Während man in der linearen Regression die Residuen $y - \hat{y}$ berücksichtigt, um die Anpassung summativ zu begutachten, gibt es in der logistischen Regression verschiedene Möglichkeiten, um die Differenz zwischen faktischen und vorausgesagten Werten zu messen. Um zu berücksichtigen, dass der geschätzte Wert bei der logistischen Regression für jedes Kovariablenmuster berechnet wird und von der geschätzten Wahrscheinlichkeit bezüglich dieses Kovariablenmusters abhängt, definiert man:

$$\hat{y}_i = m_j \hat{\pi}_j = m_j \frac{e^{\hat{g}(\mathbf{x}_j)}}{1 + e^{\hat{g}(\mathbf{x}_j)}}$$

Es handelt sich um die Erwartungswert der entsprechenden binomialverteilten Zufallsvariablen pro Kovariablenmuster. Eine erste Kennzahl ruht auf den Pearson-Residuen

$$r\left(y_j, \hat{\pi}_j\right) = \frac{y_j - m_j \hat{\pi}_j}{\sqrt{m_j \hat{\pi}_j \left(1 - \hat{\pi}_j\right)}}, \tag{7}$$

wobei y_j die Anzahl der ausgezeichneten Beobachtungen der Klasse j sind. Die Pearson-Residuen sind also die standardisierten y_j-Daten gemäss Formel

$$\frac{Y_j - E(Y_j)}{\sqrt{Var(Y_j)}}$$

für die binomialverteilte Zufallsvariable Y_j.

Summiert man über die entsprechenden quadrierten Werte, erhält man die Pearson-Chi-Quadrat-Statistik (Summe quadrierter standardisierter Abweichungen)

$$X^2 = \sum_{j=1}^{J} r(y_j, \hat{\pi}_j)^2 = \sum_{j=1}^{J} \frac{(y_j - m_j\hat{\pi}_j)^2}{m_j\hat{\pi}_j(1-\hat{\pi}_j)}$$

Diesen letzten Wert kann man z.B. mit dem R-Paket LogisticDx von Dardis (2015) berechnen. Mit (dx für Diagnose des Modells x) erhält man fürs Modell für die uis-Daten (s. Seite 68)

```
m601d=glm(dfree~age+age*ndrgtx1+ndrgtx2+ndrgtx1+ivhx+race*site+race+treat+site,
data=uis,family="binomial")
```

```
dx(m601d,byCov=T)
```

etliche Kennzahlen pro Kovariablenmuster. Man berechnet das oben definierte X^2 mit

```
dx1=dx(m601d,byCov=T)
sum((dx1$y-dx1$n*dx1$P)^2/(dx1$n*dx1$P*(1-dx1$P)))
```

oder mit

```
sum(dx1$Pr^2)
```

Man erhält die Zahl 511.7807.

Eine zweite Möglichkeit, die Distanz der beobachteten von den erwarteten Werten zu messen, besteht in der Summation über die Abweichungsresiduen (Devianz).

$$d(y_j, \hat{\pi}_j) = \pm\sqrt{2\left(y_j \ln\frac{y_j}{m_j\hat{\pi}_j} + (m_j - y_j)\ln\frac{m_j - y_j}{m_j(1-\hat{\pi}_j)}\right)} \tag{8}$$

wobei $\pm$ das Vorzeichen von $y_j - m_j\hat{\pi}_j$ ist. $y_j \ln\frac{y_j}{m_j\hat{\pi}_j}$ ist die Differenz des logarithmierten Wertes der Statistik Y_j (Anzahl Erfolge) und deren Erwartungswert, gewichtet mit dem Wert der Statistik Y_j. Der zweite Ausdruck ist dieselbe Differenz bezüglich der Anzahl Misserfolge. Mit diesen Residuen erhält man die Statistik (Summe der quadrierten Residuen):

$$D = \sum_{j=1}^{J} d(y_j, \hat{\pi}_j)^2 = 2\sum_{j=1}^{J}\left(y_j \ln\frac{y_j}{m_j\hat{\pi}_j} + (m_j - y_j)\ln\frac{m_j - y_j}{m_j(1-\hat{\pi}_j)}\right) \tag{9}$$

Mit dem erwähnten Paket erhält man mit

```
sum(dx1$dr^2)
```

die Zahl 530.7412. Die Statistik ist näherungsweise chi-quadrat-verteilt mit $J-(p+1)$ Freiheitsgraden (p Anzahl Variablen im Modell), sofern die Anzahl der Kovariablenmuster konstant ist (m-Asymptotik: d.h. mit n nimmt die Anzahl der Daten pro Klasse m_j zu).

Nimmt die Anzahl der Kovariablenmuster mit n zu, kann nicht von einer Chi-Quadrat-Verteilung ausgegangen werden. In diesem Fall kann man versuchen, die Daten zu klassifizieren, um die m-Asymptotik anwenden zu können. Ein Beispiel dafür ist der Hosmer-Lemeshow-Test. H&L schlugen dafür zwei Klassifikationsstrategien vor. Man kann die Daten nach Perzentilen der nach geschätzten Wahrscheinlichkeiten geordneten Daten klassifizieren oder nach fixen Werten der geschätzten Wahrscheinlichkeiten. Sie wählen bei der ersten Methode z.B. $g=10$ Gruppen. Die erste Gruppe enthält dann $\frac{n}{g}$ Daten - die Daten mit der kleinsten Wahrscheinlichkeit, etc. So enthält die erste Klasse z.B. alle Daten mit einer geschätzten Wahrscheinlichkeit von kleiner als $\frac{1}{g}=0.1$. Bei der zweiten Methode enthalten die Klassen nicht alle gleichviele Daten.

H&L schlagen die folgende Teststatistik vor:

$$C=\sum_{k=1}^{g}\frac{(o_k-n_k\bar{\pi}_k)^2}{n_k\bar{\pi}_k\left(1-\bar{\pi}_k\right)}$$

mit

$$o_k=\sum_{j=1}^{c_k}y_j,$$

wobei c_k die Anzahl der Kovariablenmuster der Klasse k ist. o_k ist also die Anzahl der Einsen der Klasse k. n_k ist die Anzahl der Daten pro Klasse k. y_i ist die Anzahl der Einsen des Kovariablenmusters i der Klasse k. Zuletzt ist

$$\bar{\pi}_k=\sum_{j=1}^{c_k}\frac{m_j\hat{\pi}_j}{n_k}$$

die mittlere geschätzte Wahrscheinlichkeit der Klasse k. H&L haben gezeigt, dass für $J=n$ und einem passenden logistischen Modell die Verteilung von C näherungsweise chi-quadratisch ist (g-2 df). Sie weisen darauf hin, dass es besser ist, die erste der beiden Methoden zu verwenden. Für das Modell m601d (s. oben) erhält man mit:

```
library(LogisticDx)
gof(m601d)
```

(Parameter g=10 Voreinstellung) zusammen mit etlichen anderen Tests

	val	df	pVal
HL chiSq	4.394114	8	0.819931

Mit

```
go=gof(m601d)
go$ctHL
```

erhält man die Resultate der Tabelle S. 150. Zu überprüfen ist, ob die erwarteten Zellhäufigkeiten nicht in mehr als 20% der Zellen kleiner als 5 sind. Unter diesem Gesichtspunkt ist der Test im Beispiel unproblematisch.

Wenn $J \neq n$, ist es möglich, dass die Dezentile an Stellen mit mehreren identischen Mustern vorliegen. Je nach Programmierung ergeben sich dann unterschiedliche Resultate. g sollte nicht zu klein sein, da der Test sonst Güte signalisiert, wo keine vorliegt: wenn $g < 6$ würde der Test fast immer Güte anzeigen. Der Vorteil des Tests ist, dass eine einzige Grösse die Güte der Anpassung misst. Der Nachteil ist, dass durch Klassifizierung wichtige Details verschwinden. Entsprechend wichtig ist es, die einzelnen Residuen anzuschauen und die entsprechenden Diagnostiken zu nutzen. Es lohnt sich, den beobachteten und erwarteten Häufigkeiten der erwähnten Tabelle einzusehen. Man kann unter Umständen Regionen entdecken, wo die Anpassung nicht optimal ist.

Es gibt weitere Gruppierungsvorschläge, wie z.B. der von Tsiatis (1980). Mit Hilfe inhaltlicher Gesichtspunkte oder mechanisch, z.B. mit Dezentilen, wird der Raum der Kovariablen in Untergruppen aufgeteilt. Dann wird eine Variable definiert, welche diese Aufteilung repräsentiert und ein Score-Test oder ein MLL-Test für die Koeffizienten dieser Variable durchgeführt. Tsiatis (1980) zeigt, dass durch den Score-Test die erwarteten mit den faktischen Häufigkeiten verglichen werden. Auf der Seite 152 ff wird eine Fülle von weiteren Testverfahren erwähnt. Für den Test von Osius und Rojek (1992) stellen Sie dar, wie man ihn mit gewichteter linearer Regression nachrechnen kann. Er wird vom LogisticDx-Paket mit dem Befehl gof(modell) geliefert.

Hosmer, Hosmer, Le Cessie und Lemeshow (1997) zeigten mit Simulationen, dass all diese Tests für Stichroben kleiner als 400 nicht besonders mächtig sind. Im Beispiel ist die H&L-Statistik von der Wahl von g wenig abhängig. In anderen Beispielen ist sie aber sehr volatil - die Verwendung der H&L-Statistik wird deshalb nicht empfohlen.

Klassifikationstabellen Eine weitere, etwas grobe Weise, die Güte der Anpassung zu überprüfen, besteht in sogenannten Klassifikationstabellen. Man legt einen Trennpunkt fest (z.B. 0.5). Allen Datensätzen, die gemäss berechnetem Modell eine Wahrscheinlichkeit grösser-gleich 0.5 aufweisen, dass der ausgezeichnete Wert angenommen wird, wird 1 zugeordnet, den anderen 0. Es wird eine 2×2-Tabelle erstellt mit den faktischen Ergebnissen versus den so berechneten Ergebnissen. So erhält man für die uis-Daten mittels

```
table(m601d$y,fitted(m601d)>=0.5)
```

das Ergebnis

		erwartete Werte		
		Daten mit $\hat{P}(Y=1)<0.5$	Daten mit $\hat{P}(Y=1)\geq 0.5$	Summen
faktische	0	417	11	428
	1	131	16	147
	Summen	548	27	575

Definition 33. *Die Sensitivität eines logistischen Regressionsmodells bezüglich des Trennwertes a ist (n = Anzahl Daten, n_1 = Anzahl Einsen der Zielvariable)*

$$Sensitivität(a) = \frac{Anzahl\ Daten\ mit\ \left(\hat{P}(Y=1)\geq a\right)\ und\ y=1}{n_1}$$

Die Sensitivität bezüglich des Trennwertes a ist also der Anteil der Untersuchungsobjekte, die den ausgezeichneten Wert mit einer Wahrscheinlichkeit von mindestens a annehmen und die diesen faktisch annehmen, an den Untersuchungsobjekten, die den ausgezeichneten Wert faktisch annehmen. Anders ausgedrückt: es handelt sich um den Anteil der gemäss Modell und Trennpunkt korrekt als ausgezeichnet klassifizierten Untersuchungsobjekte an den ausgezeichneten Untersuchungsobjekten.

Definition 34. *Die Spezifizität eines logistischen Regressionsmodelle bezüglich des Trennwertes a ist*

$$Spezifizität(a) = \frac{Anzahl\ Daten\ mit\ \left(\hat{P}(Y=1)<a\right)\ und\ y=0}{n-n_1}$$

Die Spezifizität bezüglich des Trennwerte a ist also der Anteil der Untersuchungsobjekte, die den ausgezeichneten Wert mit weniger als einer Wahrscheinlichkeit von a annehmen und die den ausgezeichneten Wert faktisch nicht annehmen, an den Untersuchungsobjekten, die den ausgezeichneten Wert faktisch nicht annehmen. Anders ausgedrückt: es handelt sich um den Anteil der gemäss Modell und Trennpunkt korrekt als nicht ausgezeichnet klassifizierten Untersuchungsobjekte an den nicht ausgezeichneten Untersuchungsobjekten.

In der Folge spielt die Grösse $1 - Spezifizität$ eine Rolle. Es gilt

$$1 - Spezifizität\,(a) = 1 - \frac{\text{Anzahl Daten mit } \left(\hat{P}(X=1)<a\right) \text{ und } y=0}{n-n_1}$$
$$= \frac{n-n_1 - \text{Anzahl Daten mit } \left(\hat{P}(X=1)<a\right) \text{ und } y=0}{n-n_1}$$

Zählt man von den Untersuchungsobjekten, die nicht den ausgezeichneten Wert annehmen, die Untersuchungsobjekte ab, die sowohl mit einer Wahrscheinlichkeit von weniger als a den ausgezeichnete Wert annehmen und die faktisch den Wert 0 annehmen, erhält man

die Personen, die mit einer Wahrscheinlichkeit von mindestens a den Wert 1 annehmen, faktisch aber nicht den ausgezeichneten Wert annehmen. $1 - Spezifizität\,(a)$ ist also der Anteil Untersuchungsobjekt, die den ausgezeichneten Wert mit einer Wahrscheinlichkeit von mindestens a annehmen, die aber faktisch den nicht ausgezeichneten Wert annehmen, an den Personen, die den ausgezeichneten Wert nicht annehmen. Anders ausgedrückt: es handelt sich um den Anteil der gemäss Modell und Trennpunkt falsch als ausgezeichnet klassifizierten Untersuchungsobjekte an den nicht ausgezeichneten Untersuchungsobjekten.

Im Beispiel gilt also: Die $Sensitivität(0.5)$ ist der Anteil der Personen, die vom Modell bezüglich Trennpunkt 0.5 korrekt als drogenfrei klassifiziert werden, an den drogenfreien Personen:

$$\frac{16}{131+16} = \frac{16}{147} = 0.108\,84$$

Die Sensitivität bezüglich 0.5 ist also nicht sehr hoch.

Die $Spezifizität(0.5)$ ist der Anteil der Personen, die vom Modell bezüglich Trennpunkt 0.5 korrekt als nicht drogenfrei klassifiziert werden, an den nicht drogenfreien Personen:

$$\frac{417}{417+11} = \frac{417}{428} = 0.974\,30$$

Die Spezifizität bezüglich 0.5 ist recht hoch.

$1 - Spezifizität(0.5)$ ist der Anteil der Personen, die vom Modell bezüglich Trennpunkt 0.5 falsch als drogenfrei klassifiziert werden, an den nicht drogenfreien Personen:

$$\frac{11}{428}$$

oder auch

$$1 - \frac{417}{428} = \frac{428-417}{428} = \frac{11}{428} = 0.0257$$

Bemerkung 35. *Sensitivität und Spezifizität verhalten sich bei steigenden Trennpunkten gegenläufig: je kleiner der Trennpunkt, desto grösser die Sensitivität und desto kleiner die Spezifizität. Entsprechend verhalten sich Sensitivität und 1-Spezifizität gleichläufig (s. fürs Beispiel die Tabelle 5.6 von S. 161 in H&L).*

Bemerkung 36. *Bei medizinischen Tests: Sensitivität des Tests ist der Anteil der kranken Personen, die mit dem Test als solche erkannt werden, an den kranken Personen; Spezifizität ist der Anteil der gesunden Personen, die vom Test nicht als krank spezifiziert werden, an den gesunden Personen. 1- Spezifizität ist also der Anteil der gesunden Personen, die vom Test als krank spezifiziert werden, an den gesunden Personen.*

Klassifikationstabellen weisen gewisse Nachteile auf: sie sind erstens sensibel bezüglich der relativen Grösse der zwei Gruppen und begünstigt immer die Klassifikation in die grössere Gruppe - eine Tatsache, die mit der Güte der Anpassung nichts zu tun hat. Geht man z.B. von einer Welt aus, in der nur ca. 5% der Personen, die mit einer Wahrscheinlichkeit von weniger als 0.5 drogenfrei sind, faktisch drogenfrei sind, und nur ca. 5% der

Personen, die mit einer Wahrscheinlichkeit von mehr als 0.5 drogenfrei sind, faktisch Drogen konsumieren, dann hätte man bei den obigen Spaltenrandsummen z.B. (1 = drogenfrei; $\frac{27}{548} = 0.049\,27$; $\frac{1}{27} = 0.037037$):

		erwartete Werte		
		Anzahl mit $\hat{P}(X=1) < 0.5$	Anzahl mit $\hat{P}(X=1) \geq 0.5$	Summen
faktisch	0	521	1	522
	1	27	26	53
	Summen	548	27	575

Die Sensitivität bezüglich des Trennwertes 0.5 beträgt - trotz der tiefen Fehlklassifikationsraten - $\frac{26}{53} = 0.490\,57$. Die Spezifizität ist dagegen mit $\frac{521}{522} = 0.998\,08$ sehr hoch, ein Resultat, das sich wegen der sehr unterschiedlichen Grösse der Gruppen ergibt. Es zeigt, dass man sich nicht auf die eingeführten Kennzahlen, die mit Hilfe von Klassifikationstabellen berechnet werden, verlassen kann, um die Güter der Anpassung zu beurteilen.

Man kann zweitens Daten konstruieren, so dass die logistische Regression passt, die Klassifizierung jedoch nicht besonders gut ist (s. Beispiel Buch S. 156 f.). H&L raten entsprechend davon ab, sich bei der Überprüfung der Anpassungsgüte nur auf solche Tabellen zu stützen - obwohl deren Inspektion durchaus nützlich sein könne.

Betrachtet man die obige Tabelle und die geschätzten Wahrscheinlichkeiten des Modells m601d, ergibt sich

Befehl	Ergebnis
mean(fitted(m601d)[m601d$y==1 & fitted(m601d)>=0.5])	0.5673325
mean(fitted(m601d)[m601d$y==0 & fitted(m601d)<0.5])	0.2234832
max(fitted(m601d)[m601d$y==1 & fitted(m601d)>=0.5])	0.7283415
min(fitted(m601d)[m601d$y==1 & fitted(m601d)>=0.5])	0.5025006
max(fitted(m601d)[m601d$y==0 & fitted(m601d)<0.5])	0.4966168
min(fitted(m601d)[m601d$y==0 & fitted(m601d)<0.5])	0.02857757

Das Mittel der durch das logistische Modell geschätzten Wahrscheinlichkeiten der drogenfreien Personen, die mit einer Wahrscheinlichkeit von mindestens 0.5 drogenfrei sind, beträgt 0.56733. Der höchste Wert ist 0.7283415 und der tiefste 0.5025006. Das Mittel der Wahrscheinlichkeit der Drogen konsumierenden Personen, die mit einer Wahrscheinlichkeit von weniger als 0.5 drogenfrei sind, beträgt 0.2234832, der tiefste Wert ist 0.028, der höchste 0.49666 (es ergeben sich manchmal leicht andere Zahlen als im Buch - wohl Rundungen zuzuschreiben). Wenn viele Personen Wahrscheinlichkeiten in der Nähe des Trennpunktes haben, ist zu erwarten, dass es viele falsche Klassifizierungen gibt. Diese Überlegungen zeigen, dass man Klassifikationstabellen dann brauchen soll, wenn es um die Klassifikation geht. Sonst sind sie nur als Zusatzuntersuchung zu betrachten, die strengere Methoden nicht ersetzen kann.

Die Fläche unter der ROC-Kurve Während die Sensitivität und die Spezifizität bezüglich eines einzelnen Trennpunktes berechnet wird, verwendet man bei der ROC-Kurve (Receiver Operating Characteristic) viele Trennpunkte.

Zuerst eine Vorüberlegung:

Bemerkung 37. *Man kann die Sensitivität und die Spezifizität, für viele Trennpunkte oder für die durch die geschätzten Wahrscheinlichkeiten gegebenen Trennpunkte berechnen und die entsprechenden Punkte (Trennpunkte, Sensitiviät) beziehungsweise (Trennpunkte, Spezifizität) mit Strecken verbinden. Man erhält mit den beiden folgenden Funktionen (Berechnung der Sensitivität beziehungsweise der Spezifizität für einen spezifischen Trennpunkt)*

```
sens=function(modell, Trennpunkt){ tab=
table(m601d$y, fitted(m601d)>=Trennpunkt)
if (dim(tab)[1]==2 & dim(tab)[2]==2)
sensi=tab[2,2]/sum(tab[2,]) else sensi=0; sensi }
```

und

```
spez=function(modell, Trennpunkt){ tab=
table(m601d$y, fitted(m601d)>=Trennpunkt)
if (dim(tab)[1]==2 & dim(tab)[2]==2)
spezi=tab[1,1]/sum(tab[1,]) else spezi=0; spezi }
```

und der Funktion

```
sens_spez_plot=function(modell,Trennpunkte=NULL){
if (is.null(Trennpunkte)) Trennpunkte= sort(fitted(modell))
n=length(Trennpunkte)
punktesens=rep(NA,n)
for (i in 1:n) punktesens[i]=sens(modell,Trennpunkte[i])
m=1
while(punktesens[m]==0){punktesens[m]=1;m=m+1}
plot(1, type="n", xlim=c(0,1),ylim=c(0,1), xlab= "Trennpunkte",
ylab= "Sensitivitaet/Spezifizitaet")
lines(c(Trennpunkte,1),c(punktesens,0), lty=1)
punktespez=rep(NA,n)
for (i in 1:n) punktespez[i]=spez(modell,Trennpunkte[i])
lines(c(Trennpunkte,1),c(punktespez,1),lty=2,col="red")
legend("right",legend=c("Sensitivitaet","Spezifizitaet"),lty = c(1,2),
col=c("black", "red" )) }
sens_spez_plot(m601d)
```

für das Modell m601d die Graphik 16

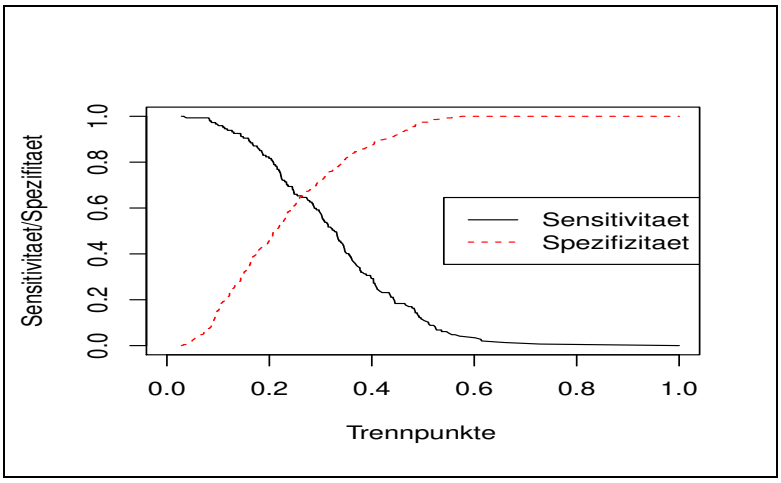

Abbildung 16: Graphische Darstellung der Sensitivität und der Spezifizität an den geschätzten Wahrscheinlichkeiten der UIS-Daten

Die optimale Klassifikationstabelle (mit möglichst wenig Fehlklassifikationen) würde man im Schnittpunkt der beiden Kurven erhalten.

Nun zur ROC-Kurve. Diese wurde in der Signaltheorie entwickelt. Sie vergleicht die Wahrscheinlichkeiten, ein korrektes Signal bei Rauschen korrekt zu erfassen (Sensitivität) mit den Wahrscheinlichkeiten, ein falsches Signal als korrektes aufzufassen (1- Spezifizität). Man zeichnet für alle geschätzten Wahrscheinlichkeiten (= Trennpunkte) die Punkte (1-Spezifizität, Sensitivität). Die Punkte werden dann mittels Strecken verbunden (Gerade zwischen Punkten). Ist die Kurve konkav, ist die Sensitivität jeweils grösser als der Anteil der als korrekt klassifizierten falschen Signale an den falschen Signalen. Je konkaver die Kurve, desto besser werden aus dem Rauschen die korrekten Signale herausgefiltert. Die Fläche unter der ROC-Kurve wird mittels Trapezformel berechnet. Je grösser die Fläche, desto konkaver ist die Kurve und desto besser wird Information herausgefiltert.

In der Anwendung auf die vorliegende Problemlage werden für jede geschätzte Wahrscheinlichkeit (= Trennpunkte) gemäss Definitionen 33 und 34 ebenfalls die Punkte

(1-Spezifizität, Sensitivität)

abgetragen und mit Strecken verbunden. Man erhält mit:

```
roc=function(modell,Trennpunkte=NULL){
if (is.null(Trennpunkte)) Trennpunkte= sort(fitted(modell))
Trennpunkte=unique(Trennpunkte)
n=length(Trennpunkte)
punktesens=rep(NA,n)
for (i in 1:n) punktesens[i]=sens(modell,Trennpunkte[i])
m=1
while(punktesens[m]==0){punktesens[m]=1;m=m+1}
punktespez=rep(NA,n)
for (i in 1:n) punktespez[i]=spez(modell,Trennpunkte[i])
plot(1, type="n", xlim=c(0,1),ylim=c(0,1), xlab="1-Spezifizitaet",
ylab="Sensitivitaet",main="ROC-Kurve")
lines(c(1-punktespez,0),c(punktesens,0),lty=1);fl=0
for (i in 1:(n-1)) fl=(1-punktespez[i]-(1-punktespez[i+1]))*
(punktesens[i]+punktesens [i+1])*0.5+fl; fl}
roc(m601d)
```

die Graphik 17

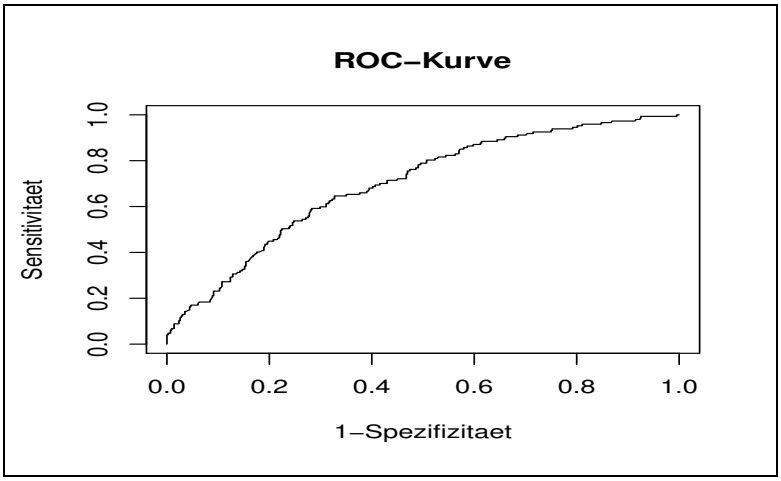

Abbildung 17: Graphische Darstellung der ROC-Kurve der UIS-Daten

und die Kennzahl ROC (Fläche unter der Kurve): 0.6989081

SPSS: Analysieren, ROC-Kurve, Testvariable (= Trennpunkte, z.B. die geschätzten Wahrscheinlichkeiten der logistischen Regression für spezifische Daten oder Datengruppen), Zustandsvariable (= Zielvariable, dichotome Variable) ergibt die Graphik und die Kennzahl ROC 0.698908.

Die ROC-Kurve samt Kennzahl wird auch vom Paket LogisticDx ausgespuckt (gof(m601d,

plotRoc=T)). Die Kennzahl samt Vertrauensintervall erhält man mit

```
go=gof(Modell)
go$au
```

Eine alternative Berechnungsweise der ROC, welche eine eventuell intuitivere Interpretation der Kennzahl erlaubt, besteht in der Berechnung der Anzahl Paare U_1, so dass die Einsen eine höhere geschätzte Wahrscheinlichkeit haben als die Nullen. Diese Zahl wird an der Anzahl der Paare gemessen. Man erhält mit 147 Einsen und 428 Nullen $147 \cdot 428 = 62916$ mögliche Paare. Die Anzahl der Paare U_1, so dass die Einsen eine höhere Wahrscheinlichkeit haben als die Nullen, kann man mittels Rängen berechnen. Man berechnet die Summe der Ränge R_2 der Nullen mit

```
rang=rank(fitted(m601d))
sum(rang[m601d$y==0]),
```

was 110749.50 ergibt (bei gleichen Rängen werden Rangmittel verwendet). Damit erhält man mit der hier nicht zu beweisenden Formel

$$U_1 = n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} - R_2$$

- mit n_1 der Anzahl der Einsen und n_2 der Anzahl der Nullen - das Ergebnis

$$U_1 = 147 \cdot 428 + \frac{428 (428 + 1)}{2} - 110749.5 = 43972.5.$$

Schliesslich ergibt sich:

$$\frac{43972.5}{62916} = 0.698\,91.$$

Alternativ erhält man die benötigten Zahlen mittels des Mann-Whitney-Rang-Testes. Mit

```
wilcox.test(fitted(m601d)~m601d$y)
```

erhält man $W = U_2 = 18944$ (Wobei $62916 - U_1 = U_2 = 18943.5$ - der Wilcoxon-Test-Befehl von R liefert nur ganze Ränge.). Man erhält also U_1 mittels

$$n_1 n_2 - W,$$

im Beispiel

$$147 \cdot 428 - 18944 = 43\,972$$

und die obige Kennzahl mittels

$$\frac{43\,972}{147 \cdot 428} = 0.698\,9$$

Bei der Interpretation von ROC beobachtet man oft die folgende pragmatische Regel:

$ROC = 0.5$	Keine Unterscheidungsfähigkeit (Diskrimination)
$0.7 \leq ROC < 0.8$	Akzeptable Unterscheidungsfähigkeit
$0.8 \leq ROC < 0.9$	Ausgezeichnete Unterscheidungsfähigkeit
$0.9 \leq ROC$	Aussergewöhnliche Unterscheidungsfähigkeit

Ein Wert von über 0.9 ist sehr unwahrscheinlich. Bei kompletter Trennung (alle Kovariablenmuster haben denselben Wert in der Zielvariable) ist die logistische Regression gar nicht berechenbar und Werte über 0.9 erfordern eine beinahe völlige Trennung.

Ein schlecht passende Modell kann durchaus eine gute Unterscheidungsgüte haben. Zählt man z.B. in einem gut passenden logistischen Modell mit guter Unterscheidungsfähigkeit zu jedem geschätzten Wert 0.25 hinzu, würde das Modell schlecht passen, aber die Unterscheidungsfähigkeit ändert sich nicht. Entsprechend genügt es nicht, die Unterscheidungsfähigkeit alleine zu betrachten.

Weitere summative Kennzahlen Aus Gründen der Vollständigkeit diskutieren H&L kurz R^2-Kennzahlen, die für logistische Modelle vorgeschlagen wurden. Diese Kennzahlen ruhen im Allgemeinen auf verschiedenen Vergleichen der vorausgesagten Werte des angepassten Modells mit denen des Nullmodells (nur Konstante). H&L vertreten die Ansicht, dass eine Kennzahl für die Güte der Anpassung die vorausgesagten mit den faktischen Werten vergleichen muss. Es gebe aber Umstände, wo R^2-Kennzahlen nützliche Statistiken für den Vergleich konkurrierender Modelle bezüglich spezifischer Daten bieten könnten. Mittlböck und Schemper (1996) haben die Eigenschaften von 12 Kennzahlen untersucht, indem sie folgende 3 Kriterien anwendeten:

1. die Kennzahl hat eine einfache Interpretation
2. $0 \leq R^2 \leq 1$
3. die Kennzahl wird nicht durch lineare Transformationen der Modell-Kovariablen verändert.

Sie empfehlen zwei Kennzahlen (die anderen verletzen mindestens eines der Kriterien):

1. die quadrierte Pearson-Korrelation zwischen beobachteten Werten und vorausgesagten Wahrscheinlichkeiten;
2. eine mit der linearen Regression vergleichbaren Quadratsumme R^2, d.h. eine Kennzahl, welche die erklärte Streuung ausdrückt,

wobei 1. und 2. für $J = n$ äquivalent sind.
Die Quadrierte Pearson-Korrelation ist bei n Kovariablenmustern (d.h. $J = n$)

$$r^2 = \frac{\left(\sum_{i=1}^{n} (y_i - \bar{y})(\hat{\pi}_i - \bar{\pi})\right)^2}{\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n} (\hat{\pi}_i - \bar{\pi})^2\right)}$$

mit $\bar{y} = \bar{\pi} = \frac{n_1}{n}$. Man erhält fürs UIS-Beispiel mit

```
cor(fitted(m601d),m601d$y)^2
```

die Kennzahl 0.09464774.

Bei $J < n$ ergibt sich die Formel.

$$r_c^2 = \frac{\left(\sum_{j=1}^{J}\left(y_j - m_j\bar{y}\right)\left(m_j\hat{\pi}_i - m_j\bar{\pi}\right)\right)^2}{\left(\sum_{j=1}^{J}\left(y_i - m_j\bar{y}\right)^2\right)\left(\sum_{i=1}^{n}\left(m_j\hat{\pi}_i - m_j\bar{\pi}\right)^2\right)}$$

H&L erhalten $r_c^2 = 0.3564$. Die Zahl erhält man mit dem R-Paket LogisticDx für das GLM-Modell m601d mit

```
library(LogisticDx)
dx1=dx(m601d,byCov=T)
cor(dx1$y,dx1$yhat)
```

Die starke Erhöhung ist dadurch zu erklären, dass $m_j y_j$ grössere Werte annehmen kann als y_i (y_i nur 0 oder 1).

Die Lineare Regressions-Kennzahl R_{SS}^2 (Bestimmtheitsmass) ist

$$R_{SS}^2 = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 - \sum_{i=1}^{n}\left(y_i - \bar{\pi}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \bar{\pi}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

Mit

```
summary(lm(m601d$y~fitted(m601d)))
```

erhält man unter Multiple R-squared: 0.09465. Zudem kann man rechnen:

```
anova(lm(m601d$y~fitted(m601d)))
```

und erhält

	Df	Sum Sq
fitted(m601d)	1	10.356
Residuals	573	99.063

womit man mit $10.356 + 99.063 = 109.\,42$ dasselbe Ergebnis erhält:

$$R_{SS}^2 = 1 - \frac{99.063}{109.\,42} = 0.09465\,4.$$

Die entsprechende Kennzahl bei Berücksichtigung identischer Kovariablenmuster ist

$$R_{SS_c}^2 = 1 - \frac{\sum_{j=1}^{J}\left(y_j - m_j\hat{\pi}_j\right)^2}{\sum_{j=1}^{J}\left(y_j - m_j\bar{y}\right)^2}$$

und man erhält $R^2_{SS_c} = 0.0997$ (Berechnung mit R s. unten).

Man erhält ein MLL-Analogon zu R^2_{SS}, wenn man die MLL (maximale Log-Likelihood) statt der Quadratsummen verwendet (in Programmen oder der Literatur u.a. „Pseudo-R^2“ genannt).

$$R^2_L = \frac{MLL_0 - MLL_p}{MLL_0} = 1 - \frac{MLL_p}{MLL_0}$$

(McFadden R^2, MLL_0: Maximum der Loglikelihoodfunktion der Nussmodells, MLL_p: Maximum der Loglikelihoodfunktion der Modells mit p erklärenden Variablen). Im UIS-Beispiel erhält man mit den R-Ergebnissen

```
summary(m601d)
```

```
Null deviance: 653.73 on 574 degrees of freedom
Residual deviance: 597.96 on 564 degrees of freedom
```

das Resultat:

$$1 - \frac{\frac{597.96}{-2}}{\frac{653.73}{-2}} = 1 - \frac{597.96}{653.73} = 0.085\,31$$

Die maximale Zahl für R^2_L erreicht man, wenn man das gesättigte Modell anpasst. Wenn $J = n$, dann ist $\text{MLL}_S = 0$ (Maximum der Log-Likelihoodfunktion des gesättigten Modells) und R^2_L nimmt den Wert

$$\frac{MLL_0 - MLL_S}{MLL_0} = \frac{MLL_0}{MLL_0} = 1$$

an. Beim Vorliegen identischer Kovariablenmuster wird 1 aber nicht erreicht. Die folgenden Statistik kann 1 erreichen:

$$R^2_{LS} = \frac{MLL_0 - MLL_p}{MLL_0 - MLL_S}$$

wobei man $\text{MLL}_S = MLL_p + 0.5D$ rechnet (D für die Summe quadrierter Residuen gemäss Formel 9 auf S. 83 - H&L liefern die Referenz auf (5.2) statt auf (5.4)). Man erhält also im UIS-Beispiel mit $D = 530.7412$

$$\frac{597.96}{-2} + 0.5\,(530.7412) = -33.\,609$$

und damit

$$\widehat{R^2_{LS}} = \frac{\frac{653.73}{-2} - \frac{597.96}{-2}}{\frac{653.73}{-2} - (-33.\,609)} = 0.095\,08\,8$$

oder

$$\widehat{R^2_{LS}} = \frac{653.73 - 597.96}{653.73 + 2\,(-33.\,609)} = 0.095\,08\,8$$

Mit dem Paket LogisticDx erhält man mit

```
go=gof(m601d)
go$R2
```

das Ergebnis

method	R2	
1: ssI	0.09464760	für R^2_{SS}
2: ssG	0.09959974	für $R^2_{SS_c}$
3: llI	0.08530447	für R^2_{L}
4: llG	0.09508152	für R^2_{LS}

Alle R^2-Werte sind im Vergleich zu üblichen R^2-Werten im Rahmen der linearen Regression klein. Dies ist bei logistischen Modellen normal. Rapportiert man Werte einem Publikum, das den diesbezüglichen Unterschied nicht kennt, kann dies Interpretationsprobleme schaffen. Sie empfehlen deshalb, die eingeführten R^2 nicht routinemässig zu publizieren. Sie seien aber nützlich, um bei der Modellbildung konkurrierende Modelle zu vergleichen.

Bemerkung 38. *Von SPSS werden das Maddala/Cox-und-Snell-R^2 (Cox & Snell, 1989) sowie das R^2 von Nagelkerke (1991) geliefert. Das erste wird wie folgt berechnet:*

$$R^2_{MCS} = 1 - \left(\frac{ML_0}{ML_p}\right)^{\frac{2}{n}} \in [0,1]$$

mit der Anzahl Daten n. Im Beispiel $1 - \left(\frac{\exp\frac{597.96}{2}}{\exp\frac{653.73}{2}}\right)^{\frac{2}{575}} = 0.092436$. *Das zweite wird berechnet mit:*

$$R^2_N = \frac{1 - \left(\frac{ML_0}{ML_p}\right)^{\frac{2}{n}}}{1 - (ML_0)^{\frac{2}{n}}} \in [0,1]$$

Nagelkerkes Pseudo-R^2 erweitert Maddalas/Cox und Snells Pseudo-R^2, sodass durch eine Reskalierung ein möglicher Wert von 1 erreicht werden kann, wenn das vollständige Modell eine perfekte Vorhersage mit einer Wahrscheinlichkeit von 1 trifft. Der Nenner des folgenden Ausdrucks ist nämlich das Maximum, das der Zähler annehmen kann. Im Beispiel erhält man: $\frac{1-\left(\frac{\exp\frac{597.962916}{2}}{\exp\frac{653.73}{2}}\right)^{\frac{2}{575}}}{1-\left(\exp\left(-\frac{597.962916}{2}\right)\right)^{\frac{2}{575}}} = 0.142\,97$. *Vermutlich wird bei SPSS noch irgendeine Korrektur vorgenommen. Mit:*

```
LOGISTIC REGRESSION VARIABLES dfree
/METHOD=ENTER age ivhx race treat site ndrgtx1 ndrgtx2 ageXndrgtx1 raceXsite
/CONTRAST (ivhx)=Indicator(1)
/PRINT=GOODFIT
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

erhält man

	Modellübersicht		
Schritt	*-2Log-Likelihood*	*R-Quadrat nach Cox & Snell*	*R-Quadrat nach Nagelkerke*
1	*597.963[a]*	*0.092*	*0.136*
a. Schätzung wurde bei Interation Nummer 5 beendet da Parameterschätzungen sich um weniger als 0.001 geändert haben.			

Bemerkung 39. *In R können die folgenden Funktionen verwendet werden, um die Ergebnisse des glm-Befehls zu nutzen:*

- *residuals oder resid, für die Abweichungsresiduen (Devianzresiduen)*
- *fitted oder fitted.values, für die geschätzten Wahrscheinlichkeiten*
- *predict, für den linearen Prädiktor (geschätzte Logits)*
- *coef oder coefficients, für die Koeffizienten, und*
- *deviance, für die Abweichung.*

Manche dieser Funktionen haben Optionen; so kann man z.B. 5 verschiedene Arten von Residuen extrahieren, genannt "deviance", "pearson", "response"(response - fitted value), "working"(the working dependent variable in the IRLS algorithm - linear predictor), und "partial"(a matrix of working residuals formed by omitting each term in the model). z.B. residuals(m601d,type="pearson"). drop1(glm-modell,test="Chisq") lässt jeweils eine Variable weg (s. Hilfe).

Diagnostik der logistischen Regression

Die summativen Statistiken wie die Summe der quadrierten Pearson-Residuen erlauben es mit einer Zahl, die geschätzten und die faktischen Werte zu vergleichen. Dieser Vorteil ist aber auch ein Nachteil. Es wird viel Information in eine einzige Zahl verarbeitet, was zu entsprechenden Informationsverlusten führt. Man sollte zusätzlich überprüfen, ob die Anpassung über alle Kovariablenmuster gut ist. Dies wird durch Verfahren bewerkstelligt, welche allgemein unter dem Titel *Diagnostik* abgehandelt werden. Die wichtigen Komponenten der Diagnostik sind wie in der linearen Regression die Komponenten der residualen Summenquadrate. In der linearen Regression ist die grundlegende Annahme die, dass die Fehlervarianz nicht von den erklärenden Variablen abhängt. Dies ist in der logistischen Regression anders (binomialverteile Fehler pro Kovariablenmuster), d.h. die Varianz ist

$$Var\left(y_j|\mathbf{x}_j\right) = m_j\pi\left(\mathbf{x}_j\right)\left(1-\pi\left(\mathbf{x}_j\right)\right).$$

Man beginnt die Analyse also mit Residuen, die durch die geschätzte Varianz dividiert werden (Pearson-Residuen r_j oder Abweichungs-Residuen d_j, s. die Definitionen 7 und 8 auf S. 82). Nachdem die Residuen durch die geschätzte Varianz dividiert wurden, ist zu erwarten, dass die Standardabweichung der Ergebnisse näherungsweise 1 ist und der Erwartungswert 0. Wie in der linearen Regression kann man diese aber weiter standardisieren, um dies besser zu erreichen (s. Bemerkung 40 von Seite 102)

Bei der Nutzung von Software ist es wichtig zu wissen, ob die Diagnostik-Statistiken für jeden Datensatz oder für Kovariablenmuster berechnet werden. Sie empfehlen, Statistiken zu verwenden, die auf Kovariablenmuster ruhen. Dies ist vor allem wichtig, wenn J wesentlich kleiner als n ist und wenn es Kovariablenmuster mit mehr als 5 Datensätzen hat. Im UIS-Beispiel ist $n = 575$ und $J = 521$. In diesem Falle sollte man auf alle Fälle mit Kovariablenmustern arbeiten. Ist J wesentlich kleiner als n und verzichtet man auf die Analyse der Kovariablenmuster, läuft man Gefahr, einflussreiche Kovariablenmuster nicht zu erfassen. Betrachten man jeden Datensatz individuell, ergibt sich mit $y_i \in \{0, 1\}$

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i \left(1 - \hat{\pi}_i\right)}}$$

sonst jedoch mit $y_i \in \{0, 1, ..., m_i\}$

$$r_j = \frac{y_i - m_j \hat{\pi}_i}{\sqrt{m_j \hat{\pi}_i \left(1 - \hat{\pi}_i\right)}}. \tag{10}$$

Der letzte Wert steigt mit wachsendem m_j. Wenn $\hat{\pi}_i = 0.5$ und $y_i = 0$, ist das Residuum $\frac{0-0.5}{\sqrt{0.5(1-0.5)}} = -1$, was nicht einen grossen Wert darstellt. Ist hingegen $m_j = 16$, erhält man $\frac{0-16 \cdot 0.5}{\sqrt{16 \cdot 0.5(1-0.5)}} = -4$, was eine starke Abweichung darstellt (STATA liefert -4 für alle 16 Individuen, SAS -1 für alle 16 Individuen, der dx-Befehl des R-LogisticDx-Pakets liefert mit der Voreinstellung -4 pro Variablenmuster, bei byCov=F -1).

Hutmatrix und die Hebelwerte Wenn einzelne Daten einen starken Einfluss auf die Ergebnisse einer Regression ausüben, sind entsprechende Resultate mit Vorsicht zu geniessen. Eine Methode, den Einfluss einzelner Daten zu überprüfen besteht in der Analyse der sogenannten Hebelwerte. Wir betrachten zuerst das linearen Regressionsmodell

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

mit

- $\mathbf{X} \in \mathbb{R}^{n \times p}$, der Designmatrix ($p$ = Anzahl erklärende Variablen und der ersten Spalte $\mathbf{1}$),
- $\boldsymbol{\beta} \in \mathbb{R}^p$, dem Parametervektor des linearen Modells,
- $\mathbf{y} = (Y_i, ..., Y_n)$, dem Zufallsvektor der Zufallsvariablen,

- $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n) \in \mathbb{R}^n$ dem Vektor der Fehlerzufallsvariablen.

Die Hutmatrix und die Hebelwerte, die aus dieser abgeleitete werden, spielen für die Diagnostik der linearen Regression eine wesentliche Rolle (Hoaglin & Welsch, 1978) und (Dümbgen, 2009). Die Hutmatrix $\mathbf{H} := \mathbf{X}\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal \in \mathbb{R}^{n\times n}$ produziert die geschätzten Werte $\hat{\mathbf{y}}$ als orthogonale Projektionen der Zielvariablen in den durch die Kovariablen aufgespannten Raum, d.h.

$$\mathbf{H}\mathbf{y} = \hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

da $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal\mathbf{y}$ und damit $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal\mathbf{y}$. $\hat{\mathbf{y}}$ ist eine Linearkombination der Spalten von $\mathbf{X}$ und der Schätzungen $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ von $\boldsymbol{\beta}$. Da $\hat{\boldsymbol{\beta}}$ so bestimmt wird, dass die Distanz zwischen $\mathbf{y}$ und $\hat{\mathbf{y}}$ minimal wird, ist die Projektion orthogonal.

Da $\mathbf{H}$ eine Projektionsmatrix ist, ist $spur\left(\mathbf{H}\right) = \dim\left(\mathbf{H}\right)$, und wenn $\dim\left(\mathbf{X}\right) = p$, ist $spur\left(\mathbf{H}\right) = p$. Der Zufallsvektor „geschätzte Residuen“ $\hat{\boldsymbol{\epsilon}} \in \mathbb{R}^n$ der linearen Regression kann mit Hilfe der Hutmatrix ausgedrückt werden durch

$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{I}\mathbf{y} - \mathbf{H}\mathbf{y} = \left(\mathbf{I} - \mathbf{H}\right)\mathbf{y}$$

mit der Einheitsmatrix $\mathbf{I}$. Zudem ist

$$\hat{\boldsymbol{\epsilon}} = \left(\mathbf{I} - \mathbf{H}\right)\boldsymbol{\epsilon}$$

denn mit $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ist

$$\begin{aligned}\hat{\boldsymbol{\epsilon}} &= \left(\mathbf{I} - \mathbf{H}\right)\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\right) = \mathbf{I}\mathbf{X}\,\boldsymbol{\beta} + \mathbf{I}\boldsymbol{\epsilon} - \mathbf{H}\mathbf{X}\,\boldsymbol{\beta} - \mathbf{H}\epsilon \\ &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} - \mathbf{H}\mathbf{X}\,\boldsymbol{\beta} - \mathbf{H}\epsilon \\ &= \left(\mathbf{I} - \mathbf{H}\right)\boldsymbol{\epsilon} + \boldsymbol{\beta}\left(\mathbf{I} - \mathbf{H}\right)\mathbf{X}\end{aligned}$$

wobei $\left(\mathbf{I} - \mathbf{H}\right)\mathbf{X} = \mathbf{0}$, da $\left(\mathbf{I} - \mathbf{H}\right)\mathbf{X} = \mathbf{X} - \mathbf{X}\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal\mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0}$.

Es gilt:

- $\mathbf{H}$ ist idempotent: $\mathbf{H}^2 = \mathbf{X}\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal\mathbf{X}\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal = \mathbf{X}\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal = \mathbf{H}$.
- $\mathbf{H}$ ist symmetrisch: $\mathbf{H}^\intercal = \left(\mathbf{X}\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal\right)^\intercal = \mathbf{X}^{\intercal\intercal}\left(\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\right)^\intercal\mathbf{X}^\intercal = \mathbf{X}\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal$.
- $\mathbf{H}$ ist nicht invertierbar, da bei $\dim\left(\mathbf{X}\right) = p$ gilt: $\dim\mathbf{H} = p < n$.
- $$0 \leq h_{ii} \leq 1$$

  denn wegen der Idempotenz gilt

  $$h_{ii} = \sum_{i=1}^{n} h_{ij}^2 = h_{ii}^2 + \sum_{i\neq j}^{n} h_{ij}^2$$

  Wäre $h_{ii} > 1$, könnte $h_{ii}^2 + \sum_{i\neq j}^{n} h_{ij}^2 \geq 0$ nicht gleich h_{ii} sein, da dann bereits $h_{ii}^2 > h_{ii}$ wäre.

Ist $V(\boldsymbol{\epsilon}) = \sigma\mathbf{I}$, so ist die Kovarianzmatrix von $\hat{\boldsymbol{\epsilon}}$

$$V(\hat{\boldsymbol{\epsilon}}) = V((\mathbf{I}-\mathbf{H})\boldsymbol{\epsilon}) = (\mathbf{I}-\mathbf{H})^{\intercal}(\mathbf{I}-\mathbf{H})V(\epsilon) = (\mathbf{I}-\mathbf{H})\sigma\mathbf{I} = (\mathbf{I}-\mathbf{H})\sigma$$

denn

$$\begin{aligned}(\mathbf{I}-\mathbf{H})^{\intercal}(\mathbf{I}-\mathbf{H}) &= (\mathbf{I}^{\intercal}-\mathbf{H}^{\intercal})(\mathbf{I}-\mathbf{H}) \\ &= \mathbf{I}^{\intercal}\mathbf{I} - \mathbf{I}^{\intercal}\mathbf{H} - \mathbf{H}^{\intercal}\mathbf{I} + \mathbf{H}^{\intercal}\mathbf{H} \\ &= \mathbf{I} - \mathbf{H} - \mathbf{H} + \mathbf{H} = \mathbf{I} - \mathbf{H}\end{aligned}$$

Bezüglich der Varianzen von $\hat{\epsilon}_i$ gilt damit: $V(\hat{\epsilon}_i) = (1-h_{ii})\sigma$. Je grösser h_{ii}, desto mehr wird die Varianz $V(\hat{\epsilon}_i)$ durch das entsprechende Datum beeinflusst. Die Grösse von h_{ii} hängt dabei nicht vom y_i-Wert ab, sondern von $\mathbf{x}_i$ - in $\mathbf{X}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}$ kommt $\mathbf{y}$ ja nicht vor. Betrachtet man die einfache Regression, erhält man für h_{ii} den Wert:

$$h_{ii} = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}.$$

h_{ii} ist also gross, wenn x_i weit weg von $\bar{x}$ liegt und die Summe der quadrierten Abweichungen der x_j von $\bar{x}$ eher klein ist. Ist h_{ii} gross, so werden die Ergebnisse der Regression stark vom entsprechenden Datum beeinflusst. Es gilt übrigens $1 \geq \max h_{ii} \geq \frac{p}{n}$. Das maximale h_{ii} kann also nur klein sein, wenn n im Verhältnis zu p gross ist. Es existieren verschiedene Formeln zur Berechnung, ab wann ein Hebelwert groß genug ist, um als Ausreißer klassifiziert zu werden. Viele davon richten sich nach der Anzahl der Prädiktoren p und der Anzahl der Fälle n. Huber (1981) empfiehlt eine kritische Grenze von 0.2. Igo (2010) empfiehlt die Formel $\frac{2p}{n}$ für einigermaßen große Datensätze von $n-p>50$. Velleman und Welsch (1981) empfehlen $\frac{3p}{n}$ für $p>6$ und $n-p>12$.

Wir wenden uns nun der logistischen Regression zu. Pregibon (1981) zeigte, dass man mit $\mathbf{X},\mathbf{H} \in \mathbb{R}^{J\times J}$ (J Anzahl der Kovariablenmuster) eine approximatives $\mathbf{H}$ für die logistische Regression berechnen kann durch

$$\mathbf{H} = \mathbf{V}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^{\intercal}\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbf{V}^{\frac{1}{2}},$$

mit der Diagonalmatrix $\mathbf{V} = diag\left(m_j\hat{\pi}(\mathbf{x}_j)(1-\hat{\pi}(\mathbf{x}_j))\right)$. Sei h_j das j-te Diagonalelement der Matrix $\mathbf{H}$. Man kann zeigen, dass

$$h_j = m_j\hat{\pi}(\mathbf{x}_j)(1-\hat{\pi}(\mathbf{x}_j))\mathbf{x}_j^{\intercal}(\mathbf{X}^{\intercal}\mathbf{V}\mathbf{X})^{-1}\mathbf{x}_j^{\intercal} = v_j b_j$$

mit

$$b_j := \mathbf{x}_j^{\intercal}(\mathbf{X}^{\intercal}\mathbf{V}\mathbf{X})^{-1}\mathbf{x}_j^{\intercal}$$

für das $j-$ te Kovariablenmuster mit dem Repräsentanten $\mathbf{x}_j$. Es gilt wie in der linearen Regression

$$\sum_{j=1}^{J} h_j = p$$

mit der Anzahl p der erklärenden Variablen und der Spalte **1**. Die obere Grenze für h_j ist $\frac{1}{m_j}$ mit der Anzahl m_j der Objekte des Kovarianzmusters j.

In der linearen Regression wächst h_{ii} mit der Distanz der x-Werte einer erklärenden Variable von ihrem Mittelwert. Dies ist in der logistischen Regression für h_j nicht auf dem ganzen Bereich der Fall - v_j ist nicht überall gleich gross und der Einfluss von v_j kann nicht überall vernachlässigt werden. So erhält man für die Werte von 100 Zufallsvariablen $X \sim N(0, 9)$ und

$$\pi(x) = \frac{\exp(g(x))}{1 - \exp(g(x))}$$

für $g(x) = 0.8x$ mit

```
dat1=rnorm(100,0,9)
dat2=cbind(dat1, exp(0.8*dat1)/(1+exp(0.8*dat1)))
dat3=cbind(dat2,rep(NA,100))
for (i in 1:100) dat3[i,3]=rbinom(n=1,size=1,prob=dat2[i,2])
dat4=as.data.frame(dat3)
names(dat4)=c("V1","V2","V3")
m171=glm(V3~V1,data=dat4,family="binomial")
library(logisticDx)
dx171=dx(m171,byCov=T)
plot(dx171$P,dx171$h)
```

die folgende Streudiagramm 18 der geschätzten Wahrscheinlichkeiten und der Hebel (die Ergebnisse sind bei den verschiedenen Zufallsrealisationen recht volatil):

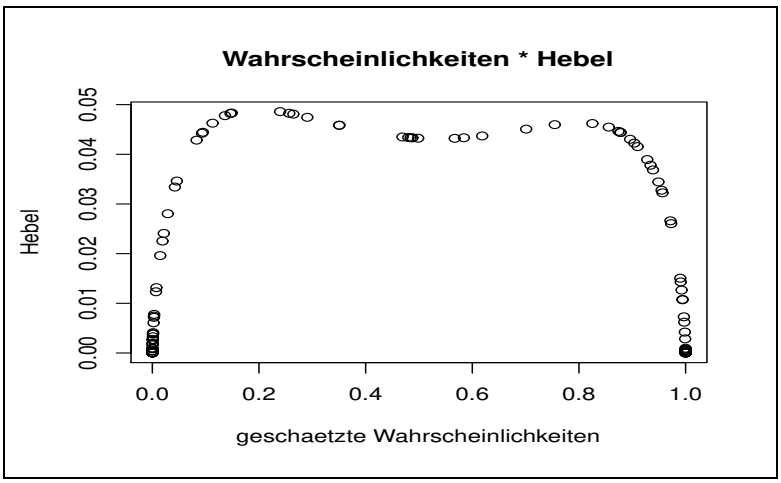

Abbildung 18: Graphische Darstellung der Hebel bezüglich geschätzter Wahrscheinlichkeiten für künstliches Beispiel

Das Mittel der $x-$Daten liegt ungefähr bei 0 und

$$\frac{\exp(0 \cdot 0.8)}{1 + \exp(0 \cdot 0.8)} = 0.5$$

Zeichnet man die Hebel bezüglich der x-Werte, erhält man die Abbildung 19:

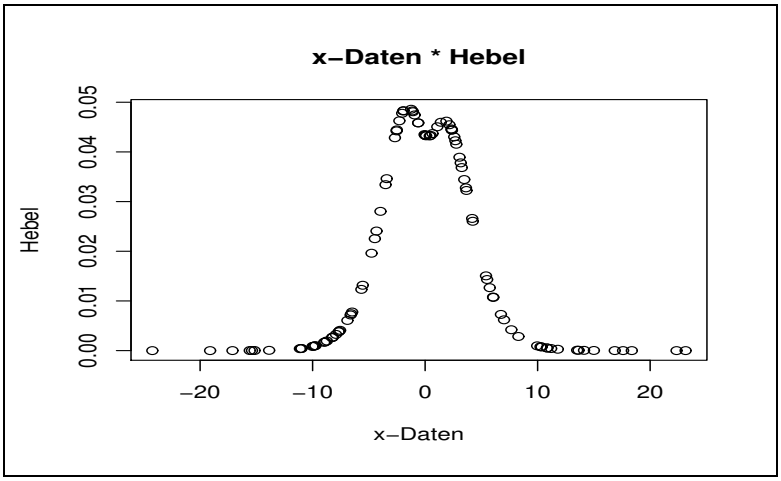

Abbildung 19: Graphische Darstellung der Hebel bezüglich der x-Werte fürs künstliche Beispiel

Man sieht, dass die Hebel mit sinkendem oder steigendem x zuerst ansteigen und dann sinken - im Gegensatz zur linearen Regression, in der diese mit der Entfernung der x vom Mittel ansteigen. Offenbar steigen die Hebel mit der Entfernung der geschätzten Wahrscheinlichkeiten von 0.5 im Intervall zwischen 0.1 und 0.9. Im Intervall [0.1, 0.9] steigt also der Hebel mit der Distanz der Wahrscheinlichkeiten von 0.5, in den Intervallen [0,0.1[und]0.95, 0] sinkt er. Im einen konkreten Hebel einschätzen zu können, muss man entsprechend die Wahrscheinlichkeit des entsprechenden Datenmusters berücksichtigen.

Bemerkung 40. *Verwendet man die Predigon-Approximation für die Residuen des j-ten Kovariablenmusters, nämlich*

$$y_j - m_j \hat{\pi}\left(\mathbf{x}_j\right) \approx \left(1 - h_j\right) y_j,$$

dann ist die Varianz des Residuums

$$m_j \hat{\pi}\left(\mathbf{x}_j\right)\left(1 - \hat{\pi}\left(\mathbf{x}_j\right)\right)\left(1 - h_j\right),$$

Dies legt nahe, dass die Pearson-Residuen noch nicht eine Varianz von 1 haben, sondern zusätzlich zu standardisieren sind. Das standardisierte Pearson-Residuum des Kovariablenmusters $\mathbf{x}_j$ ist

$$r_{sj} = \frac{r_j}{\sqrt{1 - h_j}}$$

(r_j von (10) auf S. 98) und das standardisierte Abweichungsresiduum ($d_j := d\left(y_j, \hat{\pi}_j\right)$, s. 8 auf Seite 83) ist

$$d_{sj} = \frac{d_j}{\sqrt{1 - h_j}}$$

Mit dem R-Paket logisticDx erhält man diese Grössen für das R-GLM-Objekt m171 mit

```
dx171=dx(m171)
dx171$sPr
dx171$sdr
```

Delta-Statistiken Weitere nützliche diagnostische Statistiken untersuchen den Effekt des Weglassens jeweils der Daten eines Kovariablenmusters auf die Koeffizienten. Es handelt sich um ein Analogon zur Kennzahl von R. D. Cook (1977) und R. D. Cook (1979) für die lineare Regression. Man erhält sie als standardisierte Differenz zwischen $\hat{\boldsymbol{\beta}}$ und $\hat{\boldsymbol{\beta}}_j$. $\hat{\boldsymbol{\beta}}_j$ ist der Vektor der Koeffizienten ohne das Kovariablenmuster j. Pregibon (1981) zeigte mit linearer Approximation, dass die standardisierte Differenz $\Delta\hat{\boldsymbol{\beta}}_j$ für die Logistische Regression durch

$$\begin{aligned}\Delta\hat{\boldsymbol{\beta}}_j &= \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_j\right)^{\mathsf{T}} \left(\mathbf{X}^{\mathsf{T}}\mathbf{V}\mathbf{X}\right) \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_j\right) \\ &= \frac{r_j^2 h_j}{(1-h_j)^2} \\ &= \frac{r_{sj}^2 h_j}{1-h_j}\end{aligned} \tag{11}$$

dargestellt werden kann.

Das R-Paket LogisticX mit dx(glm-Modell) liefert die Variablenmuster nach dem Wert von $\frac{r_j^2 h_j}{(1-h_j)^2}$ aufsteigend geordnet.

Bemerkung 41. *Im Buch multiplizieren sie die Hebel h_j noch mit der Anzahl Datensätze pro Variablenmuster und ordnen diese Werte den Variablenmustern zu:*

$$\Delta\hat{\boldsymbol{\beta}}_j = \frac{r_j^2 m_j h_j}{(1-m_j h_j)^2}.$$

Die so definierten $\Delta\hat{\boldsymbol{\beta}}_j$ erlaubt es, jene Kovariablenmuster zu identifizieren, die einen grossen Einfluss auf die Parameterschätzungen haben.

Neben den Auswirkungen des Weglassens von Daten oder Datengruppen auf die Koeffizienten, kann man die Auswirkungen auf die summativen Güte-der-Anpassungs-Statistiken begutachten. Indem man eine zu oben analoge lineare Approximationen verwendet, kann man zeigen, dass die Statistik für die Veränderung des Person-Chi-Quadrats dargestellt werden kann durch

$$\Delta X_j^2 = \frac{r_j^2}{1-h_j} = r_{sj}^2. \tag{12}$$

Auch hier rechnen sie mit

$$\frac{r_j^2}{1-m_j h_j}.$$

Damit gilt dann auch

$$\Delta\hat{\boldsymbol{\beta}}_j = \frac{r_{sj}^2 h_j}{1-h_j} = \Delta X_j^2 \frac{h_j}{1-h_j}. \tag{13}$$

wobei $h_j \approx \frac{h_j}{1-h_j}$ für kleine h_j (z.B. 0.001). Für die Veränderung der Abweichung (Devianz) erhält man

$$\Delta D_j = d_j^2 + r_j^2 \frac{h_j}{1-h_j}.$$

Ersetzt man r_j^2 durch d_j^2, erhält man die Näherung

$$\Delta D_j = \frac{d_j^2 (1-h_j) + d_j^2 h_j}{1-h_j} = \frac{d_j^2 - d_j^2 h_j + d_j^2 h_j}{1-h_j} = \frac{d_j^2}{1-h_j}$$

Diese Grössen werden in R für das R-glm-Objekt m171 (s. S. 101) durch z. B.

```
delta=plot(m171)
delta$Bhat
delta$dChisq
delta$dDev
delta$h
```

und vom R-Paket logisticDx durch

```
dx171=dx(m171)
dx171$dBhat
dx171$dChisq
dx171$dDev
```

geliefert. In beiden Ausgaben sind die Werte der Statistiken nach dBhat geordnet.

Um die Grössen des Buches zu erhalten, muss man das individuelle h_j mit m_j (Anzahl Daten pro Kovariablenmuster) multiplizieren und die obigen Grössen mit $m_j h_j$ statt h_j durchrechnen. Mit der folgenden Funktion werden die entsprechenden Spalten zum dx-Dataframe hinzugefügt:

```
deltaHL=function(dxobjekt){ob=dxobjekt
ob$hHL=ob$n*ob$h
ob$dChisqHL=ob$Prˆ2/(1-ob$hHL)
ob$dDevHL=ob$drˆ2/(1-ob$hHL)
ob$dBhatHL=ob$hHL*ob$Prˆ2/(1-ob$hHL)ˆ2
ob}
```

Die diagnostischen Delta-Statistiken sind konzeptuell ziemlich ansprechend, da sie es bei der H&L-Version erlauben, jene Kovariablenmuster zu entdecken, die einen grossen negativen Einfluss auf die Güte der Anpassung aufweisen (grosse ΔX_j^2 oder ΔD_j).

Bemerkung 42. *ΔX_j^2 (Kovariablenmusterversion) ist klein, wenn y_j (die Summe der Einsen des Kovariablenmusters j) und $m_j\hat{\pi}(\mathbf{x}_j)$ nahe beieinander sind. Dies ist i.A. der Fall, wenn $\hat{\pi}(\mathbf{x}_j) < 0.1$ und $y_j = 0$ oder $\hat{\pi}(\mathbf{x}_j) > 0.9$ und $y_j = m_j$. Analog ist ΔX_j^2 gross, wenn $m_j\hat{\pi}(\mathbf{x}_j)$ möglichst weit weg von y_j ist. Dies ist i.A. der Fall, wenn $\hat{\pi}(\mathbf{x}_j) > 0.9$ und $y_j = 0$ oder $\hat{\pi}(\mathbf{x}_j) < 0.1$ und $y_j = m_j$. In diesen zwei Fällen ist $\Delta\hat{\boldsymbol{\beta}}_j$ vermutlich nicht gross, denn dann ist h_j klein (s. obige Graphiken 18 und 19), da gemäss (13) dann*

$$\Delta\hat{\boldsymbol{\beta}}_j \approx \Delta X_j^2 h_j.$$

$\Delta\hat{\boldsymbol{\beta}}_j$ ist gross, wenn sowohl ΔX_j^2 als auch h_j mindestens moderat gross sind. Dies ist vermutlich der Fall, wenn $0.1 < \hat{\pi}(\mathbf{x}_j) < 0.3$ oder $0.7 < \hat{\pi}(\mathbf{x}_j) < 0.9$. Dort ist h_j relativ gross. Im Intervall $0.3 < \hat{\pi}(\mathbf{x}_j) < 0.7$ ist die Chance nicht gross, dass ΔX_j^2 oder h_j gross sind. Diese Ergebnisse kann man in der folgenden Tabelle 9 zusammenfassen — die ausdrückt, was oft der Fall ist, aber nicht sein muss.

$\hat{\pi}$	ΔX_j^2	$\Delta\hat{\boldsymbol{\beta}}_j$	h_j
< 0.1	gross oder klein	klein	klein
$0.1 - 0.3$	mässig	gross	gross
$0.3 - 0.7$	mässig bis klein	mässig	mässig bis klein
$0.7 - 0.9$	mässig	gross	gross
> 0.9	gross oder klein	klein	klein

Tabelle 9: Zusammenstellung über häufige Zusammenhänge zwischen den Delta-Statistiken und h

In der linearen Regression gibt es zwei Vorgehensweisen in der Beurteilung der diagnostischen Statistiken. Die erste ist graphisch. Gibt es in den entsprechenden graphischen Darstellungen extreme Abweichungen? Zweitens kann man auf Grund des Verteilungsmodells unter der Hypothese, dass das Modell passt, die Abweichungen auch quantitativ begutachten: liegen Sie ausserhalb festzulegender Quantile der entsprechenden Verteilung? In der logistischen Regression ist man hingegen vornehmlich auf die graphische Betrachtung verwiesen, da die Verteilung der diagnostischen Statistiken nur in gewissen Spezialfällen bekannt ist. Betrachtet man z.B. das Pearson-Residuum r_j. Von diesem wird oft behauptet, dass bei korrektem Modell dessen Verteilung $\mathrm{N}(0,1)$ ist. Diese Behauptung ist aber nur zutreffend, wenn m_j genügend gross ist. Wenn $m_j = 1$ und es zwei Werte der Zufallsvariable Y_j gibt, kann schwerlich von der Normalverteilung ausgegangen werden (s. Jennings (1986)). Wenn ein Modell mindestens eine stetige Kovariable hat, ist im allgemeinen $m_j = 1$. Entsprechend kann man nicht die m-Asymptotik anwenden und es ist bei der Beurteilung von Abweichungen mit Hilfe von Quantilen der Normalverteilung oder der Chi-Quadrat-Verteilung (mit df $= 1$) äusserste Vorsicht geboten. Hier hilft nur Erfahrung und Menschenverstand.

Es gibt viele Arten der graphischen Darstellung diagnostischer Statistiken. Sie empfehlen die folgenden:

Plot von	versus
ΔX_j^2	$\hat{\pi}_j$
ΔD_j	$\hat{\pi}_j$
$\Delta\hat{\boldsymbol{\beta}}_j$	$\hat{\pi}_j$
ΔX_j^2	h_j
ΔD_j	h_j
$\Delta\hat{\boldsymbol{\beta}}_j$	h_j

wobei die H&L-Varianten der Delta-Statistiken oder die R- oder Dx-Varianten verwendet werden können. Plots von r_j und d_j versus $\hat{\pi}_j$ finden sie aus folgenden Gründen weniger nützlich (S. 178):

1. Wenn $J \approx n$ (d.h. $m_j = 1$), entsprechen die meisten positiven Residuen Mustern, für die $y_j = m_j$, und negative Residuen entsprechen Mustern, für die $y_i = 0$. Das Vorzeichen der Residuen ist also nicht nützlich.

2. Quadrieren hebt Abweichungen besser hervor und löst das Problem mit den Vorzeichen.

3. Die Lage der Punkte erlaubt es zu sehen, welche Muster $y_j = 0$ und welche $y_j = m_j$ entsprechen.

Nützlich sei ferner der Plot von ΔX_j^2 versus $\hat{\pi}_j$ mit Punkten, die proportional zur Grösse von $\Delta\hat{\boldsymbol{\beta}}_j$ sind. Für die UIS-Daten erhält man mit (es werden für die folgenden graphischen Darstellungen die R- oder dx-Deltastatistiken verwendet; das Verfahren für die H&L-Varianten ist völlig analog. Anstatt z.B. dx1$dChisq verwendet man nach der Anwendung des obigen Funktion deltaHL auf das dx-Objekt dx1$dChisqHL).

```
library(LogisticDx)
dx1=dx(m601d,byCov=T)
plot(dx1$P,dx1$dChisq,xlim=c(0,1),
main="DChiquadrat versus Wahrscheinlichkeiten",
xlab="geschaetzte Wahrscheinlichkeiten")
```

die Abbildung 20

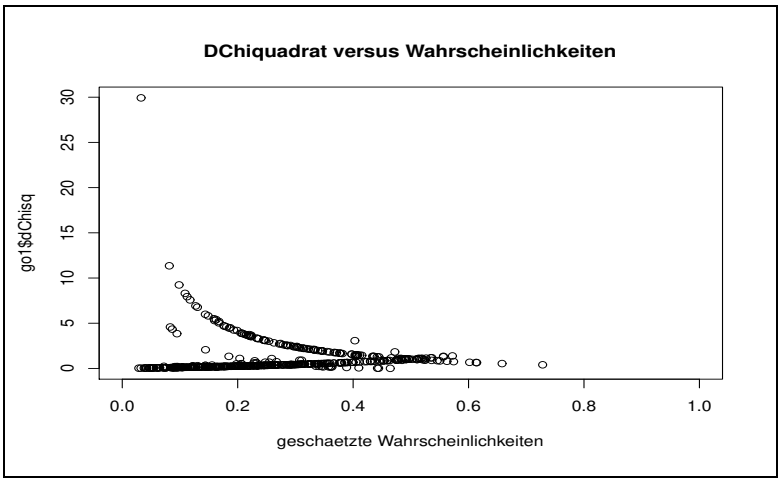

Abbildung 20: Graphische Darstellung ΔX_j^2 versus $\hat{\pi}_j$

Für ΔD_j versus $\hat{\pi}_j$ erhält man mit

```
plot(dx1$P,dx1$dDev,xlim=c(0,1),main="DAbweichung versus
Wahrscheinlichkeiten",xlab="geschaetzte Wahrscheinlichkeiten")
```

die Abbildung 21

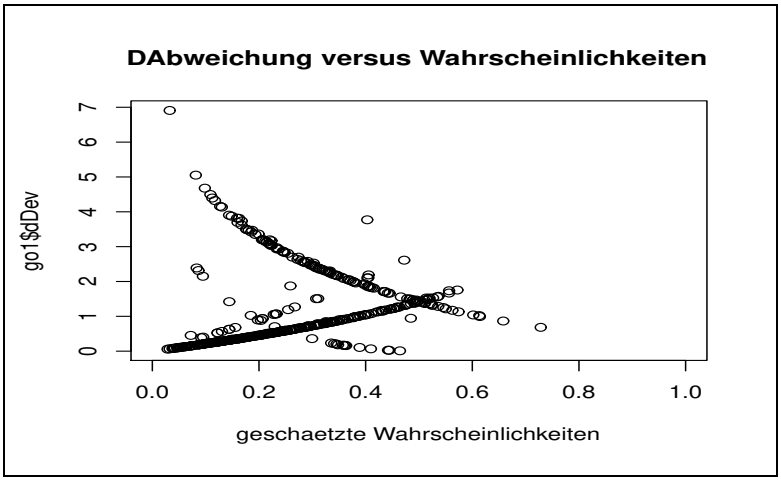

Abbildung 21: Graphische Darstellung ΔD_j versus $\hat{\pi}_j$

Für $\Delta\boldsymbol{\beta}_j$ versus $\hat{\pi}_j$ erhält man mit

```
plot(dx1$P,dx1$dBhat,xlim=c(0,1),main="DKoeffizienten versus
Wahrscheinlichkeiten", xlab="geschaetzte Wahrscheinlichkeiten")
```

die Abbildung 22

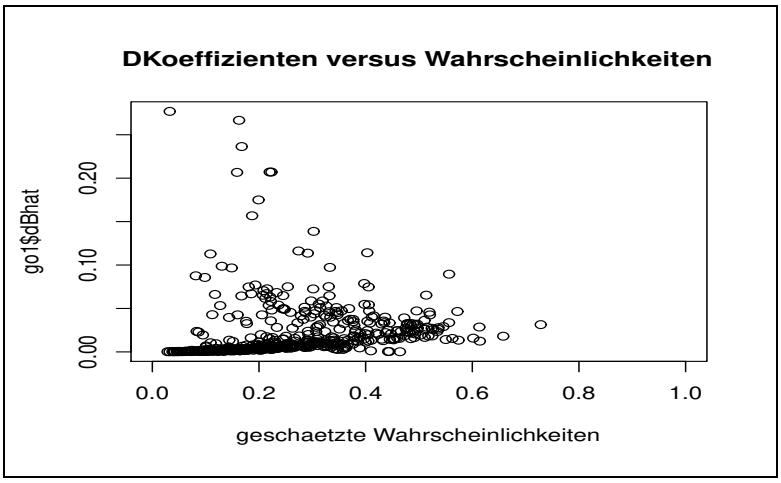

Abbildung 22: Graphische Darstellung $\Delta\boldsymbol{\beta}_j$ versus $\hat{\boldsymbol{\pi}}_j$

Um die Kreise der Graphik 20 in Abhängigkeit von der Grösse von $\Delta\boldsymbol{\beta}_j$ zu zeichnen, gibt man ein:

```
library(LogisticDx)
dx1=dx(m601d,byCov=T)
dfx = data.frame(dx1$P,dx1$dChisq,dx1$dBhat)
with(dfx, symbols(x=dx1$P, y=dx1$dChisq, circles=dx1$dBhat,
inches=1/3,fg="black", ann=T,ylim=c(0,35)))
```

und erhält die Graphik 23

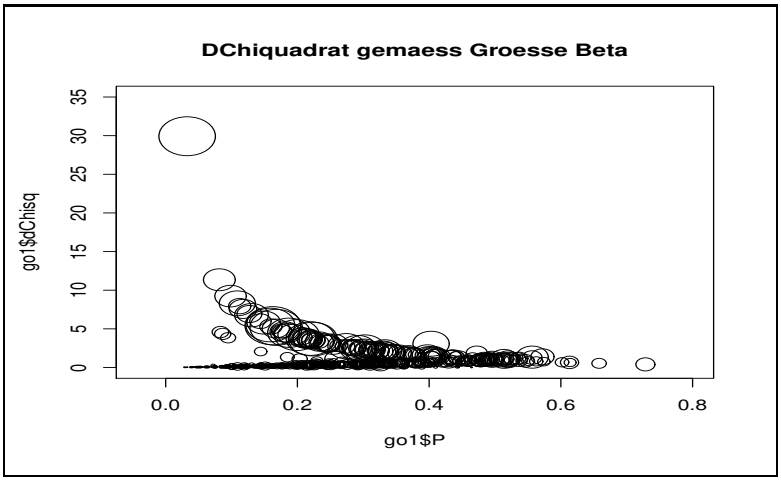

Abbildung 23: Graphische Darstellung ΔX_j^2 versus $\hat{\pi}_j$ mit Kreisen der Grösse $\Delta\boldsymbol{\beta_j}$

In der Graphik des Buches (s. 177 - 179) auf der Seite 180 die Fläche proportional $\Delta\boldsymbol{\beta_j}$ gewählt, hier der Radius. Man könnte aber auch circles=sqrt(dx1$dBhat/pi) im with-

Befehl eingeben. Mit inches=1/4.5 erhält man eine mit der Graphik des Buches auf Seite 180 vergleichbare Graphik.

Da die summativen Güte-der-Anpassung-Überprüfungen keine Abweichung des Modells von den Daten anzeigten, ist nicht zu erwarten, dass es viele Ausreisser gibt. Man kann aber einige wenige Kovariablenmuster entdecken, die nicht passen. Die Kurven der Graphiken 20, 21 und 23 sehen wie quadratische Kurven aus. Die, welche in den zwei Graphiken 20 und 21 von oben rechts nach unten links verlaufen, entsprechen Kovariablenmustern mit $y_j = m_j \geq 1$. Die Ordinate dieser Punkte ist proportional zu $(1 - \pi_j)^2$, da $m_j = 1$ für fast alle Punkte. Die Punkte der anderen Kurven, die von unten links nach oben rechts verlaufen, entsprechen den Kovariablenmustern mit $y_j = 0$. Die Ordinate dieser Punkte ist proportional zu $(0 - \pi_j)^2$. Kovariablenmuster mit schlechter Güte der Anpassung liegen gewöhnlich oben rechts oder oben links in der Graphik. Man hält nach Punkten Ausschau, die nicht zu den allgemeinen Trends passen, wobei man sich auf Inspektion der Graphik und der entsprechenden numerischen Werte stützt. In der Graphik mit ΔX_j^2 versus $\hat{\pi}_j$ fällt der Wert bei ungefähr $\Delta X_i^2 = 30$ stark aus der Tendenz heraus. Der Punkt bei 12 weicht schwach von der Tendenz ab, sollte aber vielleicht auch näher analysiert werden. Die entsprechenden Punkte kann man auch leicht in der Graphik mit ΔD identifizieren. Die ΔX^2 streuen stärker als die ΔD. Sie empfehlen beide Plots zu betrachten.

Abgesehen von den zwei erwähnten Punkten zeigen gemäss H&L die Graphiken, dass das Modell recht gut passt. Die meisten Werte von ΔX^2 und ΔD sind kleiner als 4 (trotz der im Buch vorher geäusserten Kritik verwenden sie dieses Kriterium, um wenigstens eine grobe Abschätzung vornehmen zu können; $\chi^2_{0.95}(1) = 3.84$) .

In der Abbildung 22 ($\Delta\boldsymbol{\beta}_j$ versus π_j) sehen H&L vier Punkte als ausserhalb der allgemeinen Tendenz liegend. Die Werte selber sind allerdings alle nicht besonders gross (kleiner als 0.3). Sie meinen, $\Delta\boldsymbol{\beta}_j$ müsse grösser als 1 sein, damit bezüglich des Musters j ein Problem vorliege. Es gebe aber immer Ausnahmen und es sei eine gute Praxis, alles ausserhalb der Tendenz liegenden $\Delta\boldsymbol{\beta}_j$ zu begutachten.

Ein grosses $\Delta\hat{\boldsymbol{\beta}}$ ist am wahrscheinlichsten, wenn sowohl ΔX^2 als auch der Hebel wenigstens moderat gross sind (s. Formel (13) auf der Seite 104). Ein grosser Wert kann aber auch resultieren, wenn eine der Komponenten gross ist. Dies ist der Fall in der Abbildung 22: das Variablen-Muster mit der grössten Einflussstatistik $\Delta\hat{\boldsymbol{\beta}}$ hat auch den grössten Wert von ΔX^2. Die grössten Werte für $\Delta\hat{\boldsymbol{\beta}}$ und ΔX^2 findet man mit R mittels

```
which(dx1$dChisq==max(dx1$dChisq))
which(dx1$dBhat==max(dx1$dBhat))
```

Man erhält im Beispiel mit den zwei Befehlen jeweils dieselbe Zeile (nämlich die Zeile 521).

Die Abbildung 23 kann den Beitrag von Residuum und Hebel zu $\Delta\hat{\boldsymbol{\beta}}$ zeigen. Der grosse Kreis in der oberen linken Ecke entspricht dem grössten ΔX^2 - dort ist der Hebel klein (s. Graphik 18 auf Seite 101 - nahe bei 0 und bei 1 sind die Hebel klein). Ein weiterer grosser Kreis liegt ungefähr bei 4 vor, wo der Hebel allerdings relativ gross ist.

Ein Problem mit der Statistik $\Delta\hat{\boldsymbol{\beta}}$ ist, dass es sich um eine Gesamtkennzahl für die Änderung handelt - bezüglich aller Koeffizienten auf einmal. Aus diesem Grund ist es wich-

tig, die Veränderung in den individuellen Koeffizienten hinsichtlich spezifischer Variablen-Muster zu prüfen. Die vier Kovariablenmuster mit dem grössten $\Delta\hat{\boldsymbol{\beta}}$ erhält man mit

```
dx1[518:521,]
```

Das Resultat stimmt bis auf eine Spalte mit age = 26 und bis auf die Werte für die kategorialen Daten ihvx sowie die transformierte Variable ndrgtx mit den Ergebnissen von S. 182 H&L überein (s. Tabelle 10) — dx liefert die Muster aufsteigend nach dBhat geordnet.

	1	2	3	4
(Intercept)	1.000	1.000	1.000	1.000
age	40.000	40.000	41.000	24.000
ndrgtx1	10.000	10.000	10.000	0.476
ndrgtx2	-23.026	-23.026	-23.026	0.353
ivhx2	1.000	0.000	0.000	1.000
ivhx3	0.000	1.000	1.000	0.000
race	0.000	1.000	1.000	0.000
site	1.000	0.000	0.000	1.000
treat	1.000	0.000	0.000	0.000
y	1.000	1.000	1.000	1.000
P	0.220	0.168	0.163	0.033
n	1.000	1.000	1.000	1.000
yhat	0.220	0.168	0.163	0.033
Pr	1.883	2.229	2.269	5.445
dr	1.740	1.890	1.906	2.616
h	0.052	0.044	0.047	0.009
sPr	1.934	2.279	2.324	5.470
sdr	1.788	1.933	1.952	2.628
dChisq	3.741	5.193	5.403	29.925
dDev	3.195	3.735	3.812	6.909
dBhat	0.207	0.236	0.267	0.277

Tabelle 10: Datenmuster mit den höchsten dBhats

Bemerkung 43.

Da $ndrgfp1 = \frac{10}{ndrgtx+1}$ erhält man mit

```
((10)/(dx1$ndrgtx1[521]))-1
```

für $ndrgtx$ in der letzten Spalte die Zahl 20. Für ihvx steht dort 2 statt 1, da die Kodierung für ihvx2 die Zahl 2 ist. Analog kann man die Zellen in den andere Spalten anpassen, um die Resultate des Buches zu erhalten.

Um den Effekt von extremen Variablenmuster auf die einzelnen Koeffizienten näher zu analysieren, kann man ein Modell ohne die Daten des entsprechenden Musters durchrechnen. So erhält man mit (das Variablenmuster mit dem grössten $\Delta\hat{\boldsymbol{\beta}}$ ist das Muster in der Zeile 521 und wird im folgenden Befehl weggelassen).

```
m601f=glm(cbind(y,n-y)~age+age*ndrgtx1+ndrgtx2+ndrgtx1+ivhx2+ivhx3 +
race*site+race+treat+site, data=dx1[-521,], family="binomial"(link=logit))
```

das Modell ohne das Variablenmuster mit dem grössten $\Delta\hat{\boldsymbol{\beta}}$. Man kann die Änderungen der Koeffizienten in Prozenten ausdrücken (S 183 von H&L). Dabei wird das Modell mit allen Daten als 100% gewählt. Man erhält für den Koeffizienten von age nach dem Weglassen des letzten Musters 0.1269757 und mit allen Daten 0.1166385. Es ergibt sich also $\frac{0.1269757}{0.1166385} = 1.088\,6$, also eine Veränderung von 8.9% mit der restlichen Abweichung (Residual Deviance: 523.6). Die Güte-der-Anpassungstests ergeben:

	gof(m601f)			
H&L (S. 183)		chiSq	df	pVal
D	PrI	489.94	509	0.72044
X^2	drI	523.62	509	0.31752
$\hat{C}$	HL chiSq	3.92661	8	0.8636834

HL chiSq (Testwert des Hosmer-Lemeshow-Tests) stimmt nicht mit dem Resultat im Buch überein (dort $\hat{C} = 5.55$) - bei klassifizierten Daten hängen die Resultate davon ab, wie offene und geschlossene Intervallgrenzen gesetzt werden. H&L meinen, die Änderung von X^2 sei kleiner als auf Grund der Grösse von $\Delta\hat{\boldsymbol{\beta}} = 29.925$ zu erwarten gewesen wäre $(530.74 - 523.62 = 7.12)$. Demgegenüber beträgt der Unterschied bei D $511.87 - 489.94 = 21.\,93$, während ΔD nur 6.909 beträgt. Gemäss ihrer Erfahrung seien ΔX^2 und ΔD mässig mit Änderungen der Koeffizienten korreliert, die durch Weglassen eines Variablenmusters entstehen. Obwohl das Muster mit dem grössten $\Delta\hat{\boldsymbol{\beta}}$ graphisch heraussticht, ist die Grösse von $\Delta\hat{\boldsymbol{\beta}}$ nicht so erheblich, dass grössere Änderungen in den Koeffizienten zu erwarten wären. Die maximale Änderung in allen Koeffizienten beträgt weniger als 10%. Das Muster des Datensatzes mit den grössten $\Delta\hat{\boldsymbol{\beta}}$ und ΔD ist typisch für Muster mit relativ grossen Werten von $\Delta\hat{\boldsymbol{\beta}}$ und ΔD: das angepasste Modell sagt für diese Muster, in denen die Personen den ausgezeichneten Wert annehmen, voraus, dass es unwahrscheinlich ist, dass sie diesen Wert annehmen. Im anderen Fall sagt das Modell für Muster, in denen Personen den nicht-ausgezeichneten Wert annehmen, voraus, dass es unwahrscheinlich ist, dass sie diesen annehmen. Dieser letzte Fall kommt in den uis-Daten nicht vor.

Bei den drei weiteren Mustern, die sie als relevant ansehen, verändern sich die Koeffizienten noch weniger. Die Hebel-Werte sind klein. Wenn man hingegen alle vier Muster weglässt, ergeben sich bedeutender Unterschiede (insgesamt lassen sie 5 Personen weg). Etliche Koeffizienten verändern sich um mehr als 20%. Die Güte-der-Anpassungs-Statistiken verändern sich ebenfalls substantiell. Entsprechend könnte man dafür argumentieren, diese 5 Personen wegzulassen. Nach Diskussionen mit Fachspezialisten verzichteten sie darauf.

Sie weisen darauf hin, dass die H&L-Tests nicht geeignet sind, die verschiedenen Modelle, die nach dem Weglassen von Datenmustern entstehen, miteinander zu vergleichen. Der Effekt des Weglassens der 5 Datenmuster mit extremeren Werten von $\Delta\hat{\boldsymbol{\beta}}$ besteht darin, dass die Koeffizienten für das Alter, die Anzahl der vorausgegangenen Drogenentzugsbehandlungen und die Behandlungsdauer steigen, während der für race sinkt. Der Grund dafür ist, dass 4 der 5 Personen race=other aufweisen und drogenfrei blieben. Sobald man sie weglässt, ergibt sich eine weniger starke Differenz zwischen den Hautfarbengruppen. Alle 5 Personen hatten eine kürzere Behandlungsdauer und waren nach 12 Monaten drogenfrei. Lässt man diese 5 Personen weg, ergibt sich entsprechend ein weniger starker Effekt für die kürzere Behandlungsdauer. Dies führt zu einem stärkeren Unterschied der beiden Ausprägungen bei der Variable race. Der Grund für die Veränderungen in den Variablen age und der Anzahl der vorgängigen Behandlungen sei nicht so klar, da es kein deutliches Muster bei den Variablen Alter und vorgängige Behandlungsanzahlen gibt. Zudem interagieren diese Variablen, und die Zielvariable ist in der Anzahl der vorgängigen Behandlungen im Modell stark nicht-linear.

Das Modell für die uis-Daten ist ein Beispiel für ein gut passendes Modell, und die Verwendung der Diagnostiken identifiziert nur einige wenige Datenmuster, welche nicht sehr gut passen oder die einen gewissen Einfluss auf die Koeffizienten haben. Hat man demgegenüber ein Modell, dessen Überprüfungs-Statistiken substantielle Abweichung etlicher Datenmuster vom Modell anzeigen, kann dies bedeuten:

Bemerkung 44. *1. das logistische Modell liefert nicht eine gute Approximation an die tatsächliche Beziehung zwischen dem Erwartungswert $E(Y|\mathbf{x}_j)$ und $\mathbf{x}_j$,*

2. eine wichtige Kovariable wurde bei der Modellkonstruktion ausser Acht gelassen, oder

3. wenigstens eine Kovariable wurde nicht richtig transformiert.

Das logistische Modell ist bemerkenswert flexibel. Ausser bei einer Menge von Daten, bei denen die meisten Wahrscheinlichkeiten sehr klein oder sehr gross sind, oder bei denen die Modell-Anpassung in einer identifizierbar systematischen Art schlecht ist, ist es unwahrscheinlich, dass irgend ein alternatives Modell besser passen wird. Cox (1970) zeigte, dass logistische und andere ähnliche symmetrische Modelle im Bereich zwischen 0.2 und 0.8 fast identisch sind. Sollte man andere Modell wählen wollen, so sollte man sicherstellen, dass man die Koeffizienten inhaltlich sinnvoll interpretieren kann. In manchen Situationen kann man die Anpassung eines logistischen Modells verbessern, indem man nochmals einen Modell-Bildungs-Durchlauf unternimmt (Variablenselektion und Variablentransformation). Modellanpassung ist eine iterative Prozedur. Es ist selten, dass man ein Endmodell nach nur einem Durchgang erhält.

White (1982) und andere haben gezeigt(s. H&L S. 185), dass das logistische Modell die Kullback-Leibler-Information zwischen dem theoretisch korrekten Modell und dem logistischen Modell minimiert. In diesem Sinn ist das angepasste logistische Regressionsmodell die beste Näherung an das wirkliche Modell. Hat man eine wesentliche Variable nicht erfasst, was bei relativ unbekannten Untersuchungsgebieten vorkommen kann, muss man die Daten nacherheben, was allerdings nicht immer möglich ist.

Bemerkung 45. *Mit R kann man einige diagnostische Graphiken mit plot(glm-Modell) gewinnen.*

Überprüfung mittels externer Validation

In manchen Situationen kann es möglich sein, dass man eine Teilstichprobe der Beobachtungen verwenden kann, um dann für beide Stichproben Modelle zu berechnen und diese zu vergleichen. In anderen Situationen ist es möglich, eine neue Stichprobe zu erheben, um damit die Güte der Anpassung des vorher entwickelten Modells zu testen. Gründe für dieses Vorgehen bestehen darin, dass bei der Anpassung eines Modells immer eine gewissen Überanpassung an die vorliegenden Daten resultiert. Besonders wenn man ein Modell für Voraussagen verwenden will, ist eine Überprüfung mit anderen Daten angesagt. Bei solchen Überprüfungen betrachtet man das bereits geschätzte Modell als das wirkliche und testet, ob die neuen Daten zu diesem Modell passen. Die Methoden, die dabei verwendet werden, entsprechend den bereits dargestellten. Die Validierungsstichprobe bestehe aus n_v Beobachtungen $(y_i, \mathbf{x}_i)$ mit J_v Kovariablen-Datenmustern. y_j ist die Anzahl der ausgezeichneten Antworten der m_j Beobachtungsobjekte des Kovariablenmusters j, $j \in N^*_{J_v}$. Die durch die urprünglichen Daten geschätzte Wahrscheinlichkeit für des j-te Kovariablenmuster sei π_j. Für die Pearson-Statistik erhält man

$$X^2 = \sum_{j=1}^{J_v} \frac{(y_j - m_j\pi_j)^2}{m_j\pi_j\,(1-\pi_j)}$$

Wenn jedes m_j gross genug ist, ist X^2 näherungsweise chi-quadrat-verteilt mit J_v Freiheitsgraden - unter der Null-Hypothese, dass das vorgängig geschätzte Modell korrekt ist. Bei stetigen Kovariablen wird allerdings m_j für die meisten Daten 1 sein, so dass die m-Asymptotik nicht korrekt ist. Entsprechend verwendet man z.B. die Methoden von Osius und Rojek (1992) (s. o.). Man berechnet

$$Z = \frac{X^2 - J_v}{\sigma_v}$$

mit

$$\sigma_v^2 = 2J_v + \sum_{j=1}^{J_v} \frac{1}{m_j\pi_j\,(1-\pi_j)} - 6\sum_{j=1}^{J_v} \frac{1}{m_j}$$

Die Teststatistik ist näherungsweise standardnormalverteilt, und es wird zweiseitig getestet.

Zudem kann man einen H&L-Test verwenden: sei n_k näherungsweise die Anzahl $\frac{n_v}{10}$ der Untersuchungsobjekte im k-ten Dezentil der geschätzten Wahrscheinlichkeiten und o_k die Anzahl Einsen dieses Dezentils, sowie $e_k = \sum m_j\pi_j$ - die Summe läuft über die Kovariablen-Muster des entsprechenden Dezentils. Die Hosmer-Lemeshow Statistik erhält man als Pearson-Chi-Quadrat-Statistik berechnet mit den Differenzen zwischen beobachteten und erwarteten Häufigkeiten:

$$C_v = \sum_{k=1}^{g} \frac{(o_k - e_k)^2}{n_k\bar{\pi}_k\,(1-\bar{\pi}_k)}$$

mit $\bar{\pi}_k := \sum \frac{m_j \pi_j}{n_k}$. Die Statistik ist chi-quadrat-verteilt mit g Freiheitsgraden. Es lohnt sich jeden der g Terme individuell anzuschauen, um den Beitrag zur Statistik zu überprüfen. Spielt Klassifikation eine Rolle, so kann eine Klassifikationstabelle berechnet werden (samt den entsprechenden Kennzahlen wie Sensitivität und 1-Spezifizität).

Interpretation und Präsentation der Resultate eines angepassten logistischen Regressionsmodells

Sobald eine befriedigende Anpassung eines logistischen Regressionsmodells erreicht ist, können wir das Modell verwenden, um Schlüsse daraus zu ziehen. Gewöhnlich verwendet man die Koeffizienten, um Quotenverhältnisse zu berechnen. H&L verwenden die uis-Daten mit dem Modell m601d (s. Seite 76), um die entsprechenden Diskussionen exemplarisch vorzuführen. Sie starten mit der dichotomen erklärenden Variablen treat. Man erhält die Quotenverhältnisse durch Exponenzieren des Koeffizienten. Mit

```
coefftreat=coefficients(m601d)["treat"]
treat_wert=exp(coefficients(m601d)["treat"])
vtreat=vcov(m601d)["treat","treat"]
obg=exp(coefftreat+1.96*sqrt(vtreat))
ung=exp(coefftreat-1.96*sqrt(vtreat))
ausgabetreat=c(treat_wert,ung,obg,sqrt(vtreat))
names(ausgabetreat)=c("OR", "0.95-VI untere Grenze",
"0.95-VI obere Grenze","Standardfehler")
ausgabetreat
```

erhält man die Wald-0.95-VIs für die OR der Variable treat.

OR	0.95-VI untere Grenze	0.95-VI obere Grenze	Standardfehler
1.544848	1.036200	2.303181	0.203758

Mit

```
exp(confint(m601d)["treat",])
```

erhält man die Profile-Likelihood-0.95-VIs:

2.5 %	97.5 %
1.038011	2.309827

Die Interpretation bezüglich treat ist: die Quote, während 12 Monaten drogenfrei zu bleiben (= 1), wird für eine Person bei einer langen Behandlung als 1.54 mal höher geschätzt als für eine Person mit einer kurzen Behandlung - bei konstant gesetzten übrigen Variablen. Oft wird dies nicht ganz korrekt beschrieben als eine 1.54 mal grössere Wahrscheinlichkeit, in diesem Falle drogenfrei zu bleiben. Diese Interpretation ruht auf der Argumentation, dass die Quotenverhältnisse das relative Risiko approximieren. Dies ist aber nur angemessen, wenn das ausgezeichnete Ergebnis selten ist (Faustregel: weniger als 10% der Fälle). Im

vorliegenden Fall bleiben aber 25.6% für 12 Monate drogenfrei, so dass die Interpretation nicht angemessen ist. Wenn es aber nur darum gehe, ein Ansteigen der Wahrscheinlichkeit bei längerer Behandlung auszudrücken und es nicht um exakte Werte gehe, können man auch die Wahrscheinlichkeitsinterpretation wählen. Deshalb verwenden sie in der Folge diese letztere Interpretation.

Als nächstes diskutieren sie die Variablen IVHX_2 und IVHX3 (IV Drug Use Previous und Current) bezüglich der Referenzkategorie Never). Mit

```
exp(confint(m601d)[c("ivhx2","ivhx3"),])
```

erhält man die Profile-Likelihood-0.95-VIs:

	2.5 %	97.5 %
ivhx2	0.2913603	0.9426830
ivhx3	0.2942253	0.8220115

die ORs durch

```
exp(coefficients(m601d)[c("ivhx2","ivhx3")])
```

ivhx2	ivhx3
0.5301313	0.4941345

Für ivhx2 und ivhx3 ergibt sich eine kleinere Wahrscheinlichkeit, drogenfrei zu bleiben (dfree = 1). Die Quote, drogenfrei zu bleiben, ist bei beiden Ausprägen ca. 0.5 mal kleiner als die Quote, nicht drogenfrei zu bleiben. Da die beiden Grössen der beiden Ausprägungen recht nahe beieinander liegen, könnte man unter Umständen die beiden Kategorien zusammenzunehmen. Das Prinzip, möglichst sparsame Modelle zu haben, würde dafür sprechen. Anderseits ist das Ergebnis, dass sich die beiden Kategorien kaum unterscheiden, durchaus interessant.

Als nächstes geht es darum, die Ergebnisse für die Interaktion race und site (bezüglich der Variable dfree) zu untersuchen. Man erhält für die OR für race bei site = 0 (die übrigen Variablen fallen durch die Subtraktion weg, da sie konstant gehalten werden):

$$g(1) - g(0) = 1\beta_{race} + 0\beta_{site} + (1 \cdot 0)\,\beta_{race \times site} - (0\beta_{race} + 0\beta_{site} + (0 \cdot 0)\,\beta_{race \times site}) = \beta_{race}$$

mit

```
exp(coefficients(m601d)["race"])
```

und

```
exp(coefficients(m601d)["race"]+1.96*sqrt(vcov(m601d)["race","race"]))
```

die Ergebnisse

OR	Untere Grenze VI	Obere Grenze VI
1.982001	1.181051	3.326127

.

An diesem Standort weicht die OR signifikant von 1 ab, und zwar nach oben. „Andere“ haben damit an diesem Standort eine höhere Wahrscheinlichkeit, drogenfrei zu sein (0 = weiss, 1 = andere).

Bei site = 1

$$\begin{aligned} g(1) - g(0) &= 1\beta_{race} + 1\beta_{site} + (1 \cdot 1)\beta_{race \times site} - (0\beta_{race} + 1\beta_{site} + (0 \cdot 1)\beta_{race \times site}) \\ &= \beta_{race} + \beta_{race \times site} \end{aligned}$$

und deshalb wegen

$$Var(B_{race} + B_{race \times site}) = Var(B_{race}) + Var(B_{race \times site}) + 2Cov(B_{race}, B_{race \times site})$$

mit

```
rslog=coefficients(m601d)["race"]+coefficients(m601d)["race:site"]
exp(rslog)
var_site_race=vcov(m601d)["race","race"]+vcov(m601d)["race:site","race:site"]+
2*vcov(m601d)["site","race:site"]
exp(rslog-1.96*sqrt(var_site_race))
exp(rslog+1.96*sqrt(var_site_race))
```

das Ergebnis

OR	untere Grenze	obere Grenze
0.474568	0.1871643	1.2033

.

Die obere Grenze stimmt nicht genau mit der im Buch überein. Das VI umfasst 1. Die Abweichung von 1 ist an diesem Standort also nicht signifikant.

Als nächstes geht es darum, die ORs für das *Alter und die Anzahl der vorgängigen Behandlungen* zu schätzen. Das Modell in diesen beiden Kovariablen ist ziemlich kompliziert, das Verfahren ist aber dasselbe wie bezüglich der Variable race innerhalb der Standorte (sites). Man kann z.B. die Entwicklung der ORs in 5-Jahres-Schritten bezüglich Alter und von Personen mit keiner, einer, etc. vorgängigen Behandlungen darstellen. Man erhält also - die Variable $ndrgtx$ wurde ja in die zwei Variablen $ndrgfp1$ und $ndrgfp2$ differenziert -

$$\begin{aligned} g(age, ndrgtx) &= \beta_0 + (age)\beta_1 + (ndrgfp1)\beta_2 + (ndrgfp2)\beta_3 + (age \times ndrgfp1)\beta_4 \\ g(age+5, ndrgtx) &= \beta_0 + (age+5)\beta_1 + (ndrgfp1)\beta_2 + (ndrgfp2)\beta_3 \\ &\quad + ((age+5) \times ndrgfp1)\beta_4 \end{aligned}$$

Die Differenz ist entsprechend

$$\begin{aligned} & g(age+5, ndrgtx) - g(age, ndrgtx) \\ &= (age+5)\beta_1 - (age)\beta_1 + ((age+5) \times ndrgfp1)\beta_4 - (age \times ndrgfp1)\beta_4 \\ &= 5\beta_1 + (age)(ndrgfp1)\beta_4 + 5\beta_4 ndrgfp1 - (age)(ndrgfp1)\beta_4 \\ &= 5\beta_1 + 5\beta_4 (ndrgfp1) \end{aligned}$$

Bei den Logits handelt sich also um eine linear Funktion in ndrgfp1 - der Faktor 5, der in beide Terme der Funktionsgleichung hineinspielt, rührt vom betrachteten Altersschritt her. Die Varianz beträgt

$$\begin{aligned} V\left(5\beta_1 + 5\beta_4\left(ndrgfp1\right)\right) &= 25V\left(\beta_1 + \beta_4\left(ndrgfp1\right)\right) \\ &= 25\left[V\left(\beta_1\right) + ndrgfp1^2 V\left(\beta_4\right) + 2\left(ndrgfp1\right)Cov\left(\beta_1, \beta_4\right)\right]. \end{aligned}$$

Man kann die ORs in Abhängigkeit von der Anzahl Behandlungen betrachten - und damit eine Tabelle oder eine graphische Darstellung machen. Dabei verwendet man wiederum die Gleichung

$$ndrgfp1 = \frac{10}{ndrgtx + 1} = \left(\frac{ndrgtx + 1}{10}\right)^{-1},$$

um die ORs und deren VIs zu den korrekten Anzahl Behandlungen schreiben zu können. Mit den Befehlen

```
ORage_ndr=rep(NA,11)
for (i in 0:10) ORage_ndr[i+1] = 5*m601d$coefficients["age"]+5*10/(i+1)*
m601d$coefficients["age:ndrgtx1"]
VarORage_ndr=rep(NA,11)
for (i in 0:10) VarORage_ndr[i+1]=25*(vcov(m601d)["age","age"]+(10/(i+1))^2*
vcov(m601d)["age:ndrgtx1","age:ndrgtx1"]+2*((i+1)/10)*vcov(m601d)["age","age:ndrgtx1"])
ORage_ndrall=cbind(c(0:10),ORage_ndr,sqrt(VarORage_ndr),exp(ORage_ndr),
exp(ORage_ndr-1.96*sqrt(VarORage_ndr)), exp(ORage_ndr+1.96*sqrt(VarORage_ndr)))
colnames(ORage_ndrall)=c("Anzahl Behand", "lnOR", "SE(lnOR)", "OR", "VIu","VIo")
plot(1,type="n",xlim=c(0,11),ylim=c(0,2.5))
points(ORage_ndrall[,1], ORage_ndrall[,4])
abline(1,0)
segments(ORage_ndrall[,1],ORage_ndrall[,5],ORage_ndrall[,1],ORage_ndrall[,6])
```

erhält man die Graphik 24

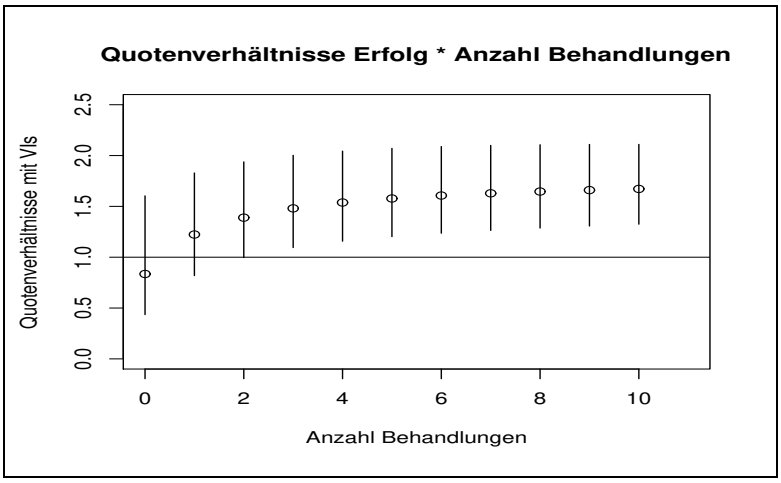

Abbildung 24: Anzahl Behandlungen und Quotenverhältnisse samt VIs

Interpretation: die Quotenverhältnisse geben die Drogenfreiheitschance bezüglich eines Alterszuwachses von 5 Jahren an - kontrolliert bezüglich der Anzahl der Behandlungen: die Chance drogenfrei zu sein, ist für eine fünf Jahre ältere Person mit zwei oder mehr Behandlungen signifikant höher und mit der Anzahl Behandlungen zunehmend höher als für eine 5 Jahre jüngere Person. Bei weniger Behandlungen spielt der 5-Jahres-Unterschied keine Rolle.

Als nächstes möchten sie den Effekt der Anzahl Behandlungen bei Kontrolle des Alters graphisch darstellen. Mit

$$g\left(age, ndrgtx\right) = \beta_0 + \left(age\right)\beta_1 + \left(ndrgfp1\right)\beta_2 + \left(ndrgfp2\right)\beta_3 + \left(age \times ndrgfp1\right)\beta_4$$

ist

$$g\left(age, ndrgtx + 1\right) = \beta_0 + \left(age\right)\beta_1 + \left(ndrgfp11\right)\beta_2 + \left(ndrgfp21\right)\beta_3 + \left(age \times ndrgfp11\right)\beta_4$$

mit

$$ndrgfp11 = \left(\frac{\left(ndrgtx + 1\right) + 1}{10}\right)^{-1}$$

und

$$ndrgpf21 = ndrgfp11 + \ln\left(\frac{ndrgtx + 1 + 1}{10}\right)$$

Die Logit-Differenz ist dann:

$$\begin{aligned} g\left(age, ndrgtx + 1\right) - g\left(age, ndrgtx\right) = {} & \beta_2\left(ndrgfp11 - ndrgfp1\right) + \\ & \beta_3\left(ndrgfp21 - ndrgfp2\right) + \\ & \beta_4\left(age\right) * \left(ndrgfp11 - ndrgfp1\right) \end{aligned}$$

Mit $A := ndrgfp11 - ndrgfp1$ und $B := ndrgfp21 - ndrgfp2$ ist dann die Varianz der Differenz

$$\begin{aligned} V\left(A\beta_2 + B\beta_3 + \beta_4\left(age * A\right)\right) = A^2V\left(\beta_2\right) + B^2V\left(\beta_3\right) + \left(age * A\right)^2 V\left(\beta_4\right) + \\ 2ABCov\left(\beta_2, \beta_3\right) + 2A\left(age * A\right) Cov\left(\beta_2, \beta_4\right) + \\ 2B\left(age * A\right) Cov\left(\beta_3, \beta_4\right) \end{aligned}$$

Daraus wird dann die Standardabweichung berechnet:

$$\sqrt{V\left(A\beta_2 + B\beta_3 + \beta_4\left(age * A\right)\right)}$$

und die VIs von

$$g\left(age, ndrgtx + 1\right) - g\left(age, ndrgtx\right)$$

Die Endpunkte der VIs und

$$g\left(age, ndrgtx + 1\right) - g\left(age, ndrgtx\right)$$

selber werden exponenziert, um die Quotenverhältnisse und die Grenzen der VIs zu erhalten. Sie zeichnen dann die Angelegenheit für verschiedene Alter (20, 25, 30 und 35, S. 197). Schliesslich wählen sei eine Behandlung als Referenz und Berechnen die Quotenverhältnisse für verschiedene Behandlungsanzahlen bezüglich dieser Referenz für verschiedene Alter.

6 Multinomiale logistische Regression

(Im Buch Kapitel 8). Die Zielvariable war bisher dichotom (mit binomialverteilten Zufallsvariablen Y_i), jetzt wird sie als polytom (mit multinomial verteilten Zufallsvariablen Y_i) vorausgesetzt. Die multinomiale logistische Regression, auch polychotome oder polytome Regression genannt, wird in der ökonometrischen Literatur oft „discrete choice model“ genannt: man kann das Wahlverhalten bei Entscheidungssituationen mit mehreren sich ausschliessenden Optionen modellieren. Man kann also untersuchen, wie verschiedene Variablen mit dem Entscheidungsverhalten zusammenhängen. Man wählt eine der Ausprägungen als Referenz (Y=0) und vergleicht dann die anderen Ausprägungen (Y=1), ..., (Y=m-1) mit dieser. In der Folge betrachten sie $m = 3$. Es habe p Kovariablen und eine Konstante. Der Zeilenvektor $\mathbf{x}$ hat entsprechend $p+1$ Komponenten. Es ergeben sich $m-1$ Logit-Funktionen, für $m = 3$

$$g_1(\mathbf{x}) = \ln\left(\frac{P(Y=1|\mathbf{x})}{P(Y=0|\mathbf{x})}\right) = \mathbf{x}^\intercal\boldsymbol{\beta}_1$$
$$g_2(\mathbf{x}) = \ln\left(\frac{P(Y=2|\mathbf{x})}{P(Y=0|\mathbf{x})}\right) = \mathbf{x}^\intercal\boldsymbol{\beta}_2.$$

Durch denselben Nenner in beiden Funktionen und die Bedingung

$$P(Y=0|\mathbf{x}) + P(Y=1|\mathbf{x}) + P(Y=2|\mathbf{x}) = 1$$

wird zwischen den beiden Funktionen ein Zusammenhang hergestellt. Mit $\pi_i := P(Y=i|\mathbf{x})$ und $\exp\left(\ln\frac{\pi_i}{\pi_0}\right) = \frac{\pi_i}{\pi_0} = \exp(g_i(\mathbf{x}))$ ergibt sich daraus

$$\frac{\pi_1}{\pi_0} = \exp(g_1(\mathbf{x})) \Longrightarrow \pi_1 = \pi_0 \exp(g_1(\mathbf{x}))$$
$$\frac{\pi_2}{\pi_0} = \exp(g_2(\mathbf{x})) \Longrightarrow \pi_2 = \pi_0 \exp(g_2(\mathbf{x}))$$

sowie

$$\pi_0 + \pi_0 \exp(g_1(\mathbf{x})) + \pi_0 \exp(g_2(\mathbf{x})) = 1$$

und

$$\pi_0 = \frac{1}{1 + \exp(g_1(\mathbf{x})) + \exp(g_2(\mathbf{x}))}$$
$$\pi_1 = \frac{1}{1 + \exp(g_1(\mathbf{x})) + \exp(g_2(\mathbf{x}))} \exp(g_1(\mathbf{x}))$$
$$\pi_2 = \frac{1}{1 + \exp(g_1(\mathbf{x})) + \exp(g_2(\mathbf{x}))} \exp(g_2(\mathbf{x})).$$

Jede dieser Wahrscheinlichkeiten ist eine Funktion der $2(p+1)$ Parameter $\boldsymbol{\beta}^\intercal = (\boldsymbol{\beta}_1^\intercal, \boldsymbol{\beta}_2^\intercal)$.

Man setzt jeweils genau ein $y_{0i} = 1$, die anderen auf 0. Bei drei Ausprägungen gilt also: wenn $Y_i = 0$, dann ist $y_{0i} = 1$, die anderen sind 0. Wenn $Y_i = 1$, dann ist $y_{1i} =$

1, die anderen 0 und wenn $Y_i = 2$, dann ist $y_{2i} = 1$, die anderen 0. Die Gemeinsame Wahrscheinlichkeitsfunktion ist damit

$$\begin{aligned}\prod_{i=1}^{n} \pi_0^{y_{0i}}(\mathbf{x}_i)\, \pi_1^{y_{1i}}(\mathbf{x}_i)\, \pi_2^{y_{2i}}(\mathbf{x}_i) &= \prod_{i=1}^{n} \pi_0^{1-y_{1i}-y_{2i}}(\mathbf{x}_i)\, \pi_1^{y_{1i}}(\mathbf{x}_i)\, \pi_2^{y_{2i}}(\mathbf{x}_i) \\ &= \prod_{i=1}^{n} \pi_0(\mathbf{x}_i) \left(\frac{\pi_1(\mathbf{x}_i)}{\pi_0(\mathbf{x}_i)}\right)^{y_{1i}} \left(\frac{\pi_2(\mathbf{x}_i)}{\pi_0(\mathbf{x}_i)}\right)^{y_{2i}}\end{aligned}$$

und damit die Log-Likelihoodfunktion

$$\begin{aligned}LL(\boldsymbol{\beta}) &= \ln\left(\prod_{i=1}^{n} \pi_0^{y_{0i}}(\mathbf{x}_i)\, \pi_1^{y_{1i}}(\mathbf{x}_i)\, \pi_2^{y_{2i}}(\mathbf{x}_i)\right) \\ &= \sum_{i=1}^{n} \ln \pi_0(\mathbf{x}_i) + y_{1i} \ln\left(\frac{\pi_1(\mathbf{x}_i)}{\pi_0(\mathbf{x}_i)}\right) + y_{2i} \ln\left(\frac{\pi_2(\mathbf{x}_i)}{\pi_0(\mathbf{x}_i)}\right) \\ &= \sum_{i=1}^{n} -\ln\left[1 + \exp g_1(\mathbf{x}_i) + \exp g_2(\mathbf{x}_i)\right] + y_{1i} g_1(\mathbf{x}_i) + y_{2i}\left(g_2(\mathbf{x}_i)\right).\end{aligned}$$

Die Likelihood-Gleichungen findet man durch Ableiten nach den $2(p+1)$ Parametern. Man erhält mit $\pi_{ij} := \pi_j(\mathbf{x}_i)$ insgesamt $2(p+1)$ Gleichungen

$$\frac{\partial LL(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} x_{ki}\left(y_{ji} - \pi_{ji}\right)$$

für $j = 1, 2$ und $k = 0, 1, 2, ..., p$ mit $x_{0i} = 1$ für alle $1 \leq i \leq n$. Den Maximum-Likelihood-Schätzer für $\boldsymbol{\beta}$ erhält man als Nullstellen dieser Ausdrücke. Die Berechnung erfolgt mittels derselben iterativen Verfahren wie im dichotomen Fall.

Die Matrix der zweiten Ableitungen braucht man, um die Informationsmatrix und den Schätzer der Kovarianzmatrix der Schätzer für $\hat{\boldsymbol{\beta}}$ zu erhalten. Man erhält die $2(p+1) \times 2(p+1)$ - Matrix

$$\begin{aligned}\frac{\partial^2 LL(\boldsymbol{\beta})}{\partial \beta_{jk} \partial \beta_{jk'}} &= -\sum_{i=1}^{n} x_{k'i} x_{ki} \pi_{ji}\left(1 - \pi_{ji}\right) \\ \frac{\partial^2 LL(\boldsymbol{\beta})}{\partial \beta_{jk} \partial \beta_{j'k'}} &= -\sum_{i=1}^{n} x_{k'i} x_{ki} \pi_{ji} \pi_{j'i}\end{aligned}$$

für j und $j' = 1, 2$ sowie k und $k' = 0, 1, 2, ..., p$. Die beobachtete Informationsmatrix $\mathbf{I}\left(\hat{\boldsymbol{\beta}}\right)$ ist die Negation dieser in $\hat{\boldsymbol{\beta}}$ ausgewerteten Matrix. Der Schätzer der Kovarianzmatrix ist die Inverse der Informationsmatrix, d.h.

$$\widehat{V\left(\hat{\boldsymbol{\beta}}\right)} = \mathbf{I}\left(\hat{\boldsymbol{\beta}}\right)^{-1}$$

Wie im dichotomen Fall kann man auch hier eine Darstellung dieser Matrix mittels $\pi_{ji}\left(1 - \pi_{ji}\right)$ angeben: Sei $\mathbf{X}$ die $n \times (p+1)$-Matrix der Kovariablen, $\mathbf{V}_j$ die $n \times n$-Diagonal-Matrix mit den Diagonalkomponenten $\hat{\pi}_{ji}\left(1 - \hat{\pi}_{ji}\right)$ für $j = 1, 2$ und $i = 1, 2, 3, ..., n$, sowie $\mathbf{V}_3$ die

$n \times n$-Diagonal-Matrix mit den Diagonalkomponenten $\hat{\pi}_{1i}\hat{\pi}_{2i}$. Der Schätzer der Informationsmatrix kann dann ausgedrückt werden durch:

$$\hat{\mathbf{I}}\left(\hat{\boldsymbol{\beta}}\right) = \begin{bmatrix} \hat{\mathbf{I}}\left(\hat{\boldsymbol{\beta}}\right)_{11} & \hat{\mathbf{I}}\left(\hat{\boldsymbol{\beta}}\right)_{12} \\ \hat{\mathbf{I}}\left(\hat{\boldsymbol{\beta}}\right)_{21} & \hat{\mathbf{I}}\left(\hat{\boldsymbol{\beta}}\right)_{22} \end{bmatrix}$$

wobei

$$\begin{aligned} \hat{\mathbf{I}}\left(\hat{\boldsymbol{\beta}}\right)_{11} &= \mathbf{X}^{\intercal}\hat{\mathbf{V}}_1\mathbf{X} \\ \hat{\mathbf{I}}\left(\hat{\boldsymbol{\beta}}\right)_{22} &= \mathbf{X}^{\intercal}\hat{\mathbf{V}}_2\mathbf{X} \\ \hat{\mathbf{I}}\left(\hat{\boldsymbol{\beta}}\right)_{21} &= \hat{\mathbf{I}}\left(\hat{\boldsymbol{\beta}}\right)_{12} = -\left(\mathbf{X}^{\intercal}\hat{\mathbf{V}}_3\mathbf{X}\right). \end{aligned}$$

Interpretation und Bewertung der Signifikanz der geschätzten Koeffizienten

Sie besprechen die Interpretation am Beispiel einer Mammographiestudie (s. Seite 265, Datendatei meexp.dat): Zielvariable me: Frauen haben nie eine Mammographie machen lassen, innerhalb eines Jahres oder vor über einem Jahr (Kodierung: 0,1,2), in Abhängigkeit von den folgenden Variablen

1. Meinung, dass man eine Mammographie erst braucht, wenn man Symptome aufweist (Variablenname: sympt; Ausprägungen: stimme stark zu, stimme zu, nicht einverstanden, stark nicht einverstanden; Kodierung: 1,2,3,4)

2. Wahrgenommener Nutzen der Mammographie (Variablenname: pb; Ausprägungen: 4 - 20 - die Summe von 5 Variablen mit je 4 Ausprägungen, je tiefer der Wert, desto höher ist der wahrgenommene Nutzen der Mammographie)

3. Familiengeschichte (Variablenname: hist; Ausprägungen: Mutter oder Schwester hat eine Brustkrebsgeschichte, nein oder ja; Kodierung: 0,1)

4. Aufforderung durch eine Person, eine Mammographie zu machen (Variablennamen: bse; Ausprägungen: nein, ja; Kodierung: 0,1)

5. Einschätzung der Wahrscheinlichkeit, dass eine Mammographie Krebs entdeckt (Variablenname: detc; Ausprägungen: nicht wahrscheinlich, ziemlich wahrscheinlich, sehr wahrscheinlich; Kodierung: 1,2,3)

Um die Diskussion und die Interpretation zu erleichtern, verwenden Sie eine verallgemeinerte Notation für die Quotenverhältnisse. Das Quotenverhältnis der Ergebnisses j der Zielvariable im Vergleich zum Ergebnis 0 der Zielvariable für die Werte der Kovariablen $x = a$ und $x = b$ wird ausgedrückt durch

$$OR_j\left(a, b\right) = \frac{P\left(Y = j|x = a\right)/P\left(Y = 0|x = a\right)}{P\left(Y = j|x = b\right)/P\left(Y = 0|x = b\right)}.$$

Erklärende Variable mit zwei Ausprägungen: H&L starten mit einem Modell mit einer dichotomen Kovariable mit den Ausprägungen 0 und 1. Liegt nur eine erklärende, kategoriale Variable vor, so können die Koeffizienten mit paarweisen logistischen Regressionen berechnet werden. Diese zwei logistischen Regressionen sind allerdings nicht völlig unabhängig voneinander, da die Daten der Referenzkategorie für beide Regressionen identisch sind. Sie exerzieren den Fall an der Zielvariable me und der erklärenden Variable hist durch. Man erhält mit

```
meexp=read.table("pfad.../meexp.dat",header=F)
names(meexp)=c("id","me","sympt","pb","hist","bse","detc")
tab_me_hist=table(meexp$me,meexp$hist)
colnames(tab_me_hist)=c("Nein", "Ja")
rownames(tab_me_hist)=c("Nie","Innerhalb Jahr","Vor mehr als Jahr")
tab_me_hist
```

die Kreuztabelle

	Nein	Ja
Nie	220	14
Innerhalb Jahr	85	19
Vor mehr als Jahr	63	11

und damit die zwei Quotenverhältnisse

$$OR_1 := OR_1(1,0) = \frac{P(Y=1|x=1)/P(Y=0|x=1)}{P(Y=1|x=0)/P(Y=0|x=0)} = \frac{19/85}{14/220} = 3.512\,6$$
$$\ln(OR_1) = \ln(3.512\,6) = 1.256\,4$$
$$OR_2 := OR_2(1,0) = \frac{P(Y=2|x=1)/P(Y=0|x=1)}{P(Y=2|x=0)/P(Y=0|x=0)} = \frac{11/63}{14/220} = 2.743\,8$$
$$\ln(OR_2) = \ln(2.743\,8) = 1.009\,3$$

Die Standardfehler erhält man wie oben paarweise mit

$$SE(B_{11}) = \sqrt{\frac{1}{19} + \frac{1}{220} + \frac{1}{85} + \frac{1}{14}} = 0.374\,66$$
$$SE(B_{21}) = \sqrt{\frac{1}{63} + \frac{1}{11} + \frac{1}{220} + \frac{1}{14}} = 0.427\,50$$

Das 0.95-VIs bei einem Testwert von 3.5126 ist also

$$\exp(\ln(3.\,512\,6) \pm 1.96 \cdot 0.374\,66) = \;]1.\,685\,4, 7.\,320\,6[$$

Mit dichotomer logistischer Regressionen

```
meexp1=meexp[meexp$me==0 | meexp$me ==1,]
summary(glm(me~hist,data=meexp1,family="binomial"))

meexp2=meexp[meexp$me==0 |meexp$me ==2,]
meexp2$me[meexp2$me==2]=1
summary(glm(me~hist,data=meexp2,family="binomial"))
```

erhält man die Ergebnisse des Buches (S. 267).

Logit	Variable	Estimate	Std. Error	z value	Pr(>|z|)
1	(Intercept)	-0.9510	0.1277	-7.446	9.6e-14 ***
	hist	1.2564	0.3747	3.353	0.000798 ***
	residual deviance:	405.93 on 336 df			
2	(Intercept)	-1.2505	0.1429	-8.751	<2e-16 ***
	hist	1.0093	0.4275	2.361	0.0182 *
	residual deviance:	334.39 on 306 df			

Direkt erhält man diese Ergebnisse mit dem multinom-Befehl im R-Paket nnet von Venables und Ripley (2016) (s. auch Venables und Ripley (2002) sowie Thompson (2009); Es gibt weitere Pakete: MULTINOM, MLOGIT (kompliziert, man kann nicht direkt auf Daten anwenden, sondern muss diese in ein spezielles Format einlesen), LCR, VGAM. Das VGAM-Paket ist sehr flexibel. Es erlaubt auch die Schätzung verschiedener Modelle mit ordinal-skalierter Zielvariable)

Mit

```
library(nnet)
multinom(me~hist,data=meexp)
```

erhält man:

	Coefficients:	
	(Intercept)	hist
1	-0.9509794	1.256531
2	-1.2504803	1.009384
	Residual Deviance:	792.3399
	AIC:	800.3399

Mit nnet-Befehl „multinom“

```
m265=multinom(me~hist,data=meexp)
summary(265)
```

erhält man:

	Coefficients:	
	(Intercept)	hist
1	-0.9509794	1.256531
2	-1.2504803	1.009384
	Std. Errors:	
	(Intercept)	hist
1	0.1277115	0.3746633
2	0.1428926	0.4275097

Die 0.95-Vertrauensintervalle erhält man mit:

für die Koeffizienten:	confint(m265)
für die Quotenverhältnisse:	exp(confint(m265))

Mit dem VGAM-Paket erhält man mit

```
m265_vglm=vglm(me~hist,data=meexp,family=multinomial(refLevel=1))
summary(m265_vglm)
```

die Schätzer

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept):1	-0.9510	0.1277	-7.446	9.6e-14 ***
(Intercept):2	-1.2505	0.1429	-8.751	< 2e-16 ***
hist:1	1.2564	0.3747	3.353	0.000798 ***
hist:2	1.0093	0.4275	2.361	0.018225 *

Die Darstellung des Buches von S. 274 erhält man aus dem Resultat des nnet-Befehls „multinom“ mittels

```
summary_MN=function(modell){nk=length(modell$lab)-1
sum=summary(modell); coeff=t(sum$coefficients[1,])
for (i in 2:nk) coeff=c(coeff,t(sum$coefficients[i,]))
mat=matrix(coeff,ncol=1);  colnames(mat)="Coeff"
stabf=t(sum$standard.errors[1,])
for (i in 2:nk) stabf=c(stabf,t(sum$standard.errors[i,]))
mat=cbind(mat,stabf); z=mat[,1]/mat[,2]
mat=cbind(mat,z); n=length(mat[,1]); p_Wert=vector(length=n)
for (i in 1:n) p_Wert[i]=min(1-pnorm(z[i],0,1),pnorm(z[i],0,1))
mat=cbind(mat,p_Wert); rname= paste(1,sum$coefnames)
for (i in 2:nk) rname=c(rname, paste(i,sum$coefnames))
rownames(mat)=rname; mat}
summary_MN(m265)
```

	Coeff	stabf	z	p_Wert
1 (Intercept)	-0.951	0.128	-7.446	0.000
1 hist	1.257	0.375	3.354	0.000
2 (Intercept)	-1.250	0.143	-8.751	0.000
2 hist	1.009	0.428	2.361	0.009

Will man untersuchen, ob die Unterschiede zwischen der Ausprägung 1 und 2 bezüglich der Zielvariable (innerhalb eines Jahre, vor mehr als einem Jahr) signifikant verschieden sind, kann man das entsprechende Quotenverhältnis und dessen Standardfehler berechnen mit:

$$\frac{11/63}{19/85} = 0.781\,12$$

$$\begin{aligned}\ln 0.781\,12 &= -0.247\,03\\ &= \beta_{21} - \beta_{11}\\ &= 1.256531 - 1.009384\\ &= -0.247\end{aligned}$$

Der Standardfehler ist wiederum

$$\sqrt{\frac{1}{85} + \frac{1}{19} + \frac{1}{63} + \frac{1}{11}} = 0.413\,74$$

Dieses Resultat ergibt sich auch mit der dichotomen, logistischen Regression, indem man die dritte Ausprägung mit der zweiten Ausprägung vergleicht und die erste unberücksichtigt lässt, mittels Umkodieren (erste zwei Zeilen)

```
meexp3=meexp[meexp$me==1 |meexp$me ==2,]
meexp3$me[meexp3$me==1]=0; meexp3$me[meexp3$me==2]=1
mod3=glm(me~hist,data=meexp3,family="binomial"); summary(mod3)
```

und man erhält:

Coefficients:	Estimate	Std. Error	z value	$Pr(>\|z\|)$
(Intercept)	-0.2995	0.1662	-1.802	0.0716 .
hist	-0.2470	0.4137	-0.597	0.5505

Mit dem multinomialen Modell erhält man dieselben Resultate mittels der Kovarianz der Differenz der Koeffizientenschätzstatistiken B_{21} und B_{11} der Modelle 1 und 2 (Vergleich der 2. mit der 1. und der 3. mit der 1. Variable). Es gilt

$$\begin{aligned}V(B_{21} - B_{11}) &= V(B_{21}) + V(B_{11}) - 2Cov(B_{21}, B_{11})\\ &= 0.140372582 + 0.182764549 - 2 \cdot 0.075980455\\ &= 0.171\,18\end{aligned}$$

wobei man $Cov\,(B_{21}, B_{11}) = 0.075980455$ mit vcov(m265)) erhält:

	1:(Intercept)	1:hist	2:(Intercept)	2:hist
1:(Intercept)	0.016310221	-0.016310221	0.004545461	-0.004545461
1:hist	-0.016310221	0.140372582	-0.004545461	0.075980455
2:(Intercept)	0.004545461	-0.004545461	0.020418300	-0.020418300
2:hist	-0.004545461	0.075980455	-0.020418300	0.182764549

und daraus mittels

```
vcov(m265)["1:hist","2:hist"]
```

.

Damit erhält man den Standardfehler $\sqrt{0.171\,18} = 0.413\,74$. Dies ergibt das 0.95-VI

$$-0.247\,15 \pm 1.96\sqrt{0.171\,18} = \;]-1.0581, 0.563\,78[.$$

Da das VI 0 enthält, ist die Differenz nicht signifikant von 0 verschieden. Mittels Chi-Quadrat-Test erhält man bestätigend:
Null Deviance: 241.7; Residual Deviance: 241.3; 1-pchisq(241.7-241.3,1) = 0.5270893

Auf Grund dieser Analyse könnte man versucht sein, die beiden von 0 verschiedenen Ausprägungen der Zielvariable zusammenzulegen. Das wäre allerdings vorschnell, da die Unterscheidbarkeit der Ausprägungen bezüglich der anderen Kovariablen oder bezüglich des Gesamtmodells durchaus gegeben sein kann.

Unterschiede zwischen den paarweisen Berechnungen und dem multinomialen Modell ergeben sich bei den Restlichen Abweichungen (s. R-Ergebnisse der glm-Modelle mod1, mod2, mod3 und m265 oben).

	Restliche Abweichung	df	n
Modell 0×1	405.93	336	338
Modell 0×2	334.39	306	308
Modell 1×2	241.32	176	178
Multinomiales Modell	792.3399	408	412

Das multinomiale Modell wird für 412 Daten berechnet. Die Anzahl der geschätzten Parameter beträgt 4, also df = 408. Die Anzahl der Freiheitsgrade für das multinomiale Modell wird von den R-Befehlen nicht angegeben.

Die restliche Abweichung des Nullmodells kann man berechnen lassen mit dem nnet-Befehl „multinom“

```
multinom(me~1,data=meexp)
```

Man erhält 805.198. Damit ergibt sich der Testwert

$$805.198 - 792.3399 = 12.\,858$$

mit zwei Freiheitsgraden (Anzahl Ausprägungen der Zielvariable - 1). p-Wert ist 0.002. Die Variable hist ist also statistisch signifikant mit der Wahrscheinlichkeit, eine Mammographie zu machen, verknüpft. Die Freiheitsgrade werden allgemein wie folgt bestimmt:

$$(J_Y - 1) \cdot (I_X - 1)$$

mit J_Y = Anzahl Ausprägungen Zielvariable und I_X = Anzahl Ausprägungen unabhängige Variable. Bei einer stetigen unabhängigen Variable ergibt der zweite Faktor 1. Bei einer kategorialen Variable mit 4 Ausprägungen ergibt der zweite Faktor 4-1=3. Will man andere Ausprägungen der Zielvariable als Referenz festlegen, muss die Zielvariable anders kodiert werden. Für die erklärenden Variablen können Kontraste wie im Falle der dichotomen logistischen Regression festgelegt werden.

Mit SPSS: es wird die letzte Ausprägung als Referenz gewählt. In den Menus kann dies nicht geändert werden, aber in der ersten Zeile der Syntax kann man in „NOMREG me (BASE=LAST ORDER=ASCENDING) BY hist“ Base=first setzen. Auch bezüglich der erklärenden Variable wird 1 als Referenz gewählt. Entsprechend kehrt sich das Vorzeichen der Koeffizienten um. Dies scheint man ohne Rekodierung nicht ändern zu können. Die Schätzung der Konstanten ist nicht identisch - wegen der anderen Referenz bei hist. Kehrt man die Kodierung um, erhält man dieselben Resultate wie im Buch. Für -2 Log-Likelihood wird 32.515 (Nullmodell) und 19.657 (Modell mit hist) angegeben. Wie diese Zahlen berechnet werden, ist unklar. Die Differenz ist allerdings identisch mit dem obigen Resultat 12. 858.

Erklärende Variable mit mehr als zwei Ausprägungen: H&K verwenden die Variable detc, die in eine Faktorvariable zu verwandeln ist. Man kann die Quotenverhältnisse und deren Standardfehler wie beim ersten Fall berechnen (dichotome erklärende Variable, multinomiale Zielvariable). Man erhält die Tabelle

	Einschätzung Wahrscheinlichkeit, durch Mammographie Krebs zu entdecken		
ME	unwahrscheinlich	eher wahrscheinlich	sehr wahrscheinlich
nie	13	77	144
innerhalb Jahr	1	12	91
mehr als ein Jahr	4	16	54

und damit die Quotenverhältnisse

$$OR_1\,(2,1) = \frac{12/1}{77/13} = 2.026\,0 \quad (\ln 2.026\,0 = 0.706\,06)$$
$$OR_1\,(3,1) = \frac{91/1}{144/13} = 8.215\,3 \; (\ln 8.215\,3 = 2.106\,0)$$
$$OR_2\,(2,1) = \frac{16/4}{77/13} = 0.675\,32 \quad (\ln 0.675\,32 = -0.392\,57)$$
$$OR_2\,(3,1) = \frac{54/4}{144/13} = 1.218\,8 \; (\ln 1.218\,8 = 0.197\,87)$$

Mit R erhält man mittels nnet-Befehl „multinom“

```
m271=multinom(me~factor(detc),data=meexp)
summary(m271)
```

das Ergebnis:

	Coefficients:		
	(Intercept)	factor(detc)2	factor(detc)3
1	-2.565201	0.7062589	2.1062453
2	-1.178617	-0.3926042	0.1977986
	Std. Errors:		
	(Intercept)	factor(detc)2	factor(detc)3
1	1.037867	1.083278	1.0464712
2	0.571762	0.634349	0.5936115

Die ORs erhält man mittels

```
summ271=summary(m271)
exp(summ271$coefficients)
```

Die Vertrauensintervalle für die Quotenverhältnisse erhält man wiederum mit

```
exp(confint(m271))
```

, , 1			
		2.5 %	97.5 %
	(Intercept)	0.01005804	0.5880059
	factor(detc)2	0.24245780	16.9360656
	factor(detc)3	1.05675131	63.8982038
, , 2			
		2.5 %	97.5 %
	(Intercept)	0.1003340	0.9436648
	factor(detc)2	0.1947759	2.3412783
	factor(detc)3	0.3807324	3.9010890

und die Kovarianzmatrix mit vcov(m271). Die restliche Abweichung des Null-Modells ist wiederum 805.198 (Berechnung s. oben), die restliche Abweichung des Modells mit der Variable detc ist 778.4011 - wird durch summary(m271) ausgegeben. Man erhält den Testwert $805.198 - 778.4011 = 26.797$ mit $2 \cdot 2$ Freiheitsgraden und damit den p-Wert 2.184957e-05 (z.B. mit 1-pchisq(805.198-778.4011,4)). Damit besteht ein signifikanter Zusammenhang zwischen der Variable „Einschätzung der Aufdeckungsfähigkeit der Tests“ mit dem Umstand, ob man einen solchen nicht, innerhalb eines Jahres oder vor mehr als einem Jahr gemacht hat. Das grösste Quotenverhältnis besteht bezüglich „sehr wahrscheinlich“ und „innerhalb eines Jahres“. Alle anderen Quotenverhältnisse sind nicht signifikant von 1 verschieden. Die grossen Standardfehler des Quotenverhältnisse, welche „innerhalb eines Jahre“ mit „nie“ vergleichen, sind durch die Häufigkeit 1 in einer der beteiligten Zellen zu erklären. Damit ergeben sich unter den Wurzeln

$$\sqrt{\frac{1}{13} + \frac{1}{77} + \frac{1}{1} + \frac{1}{12}} = 1.083\,2$$

$$\sqrt{\frac{1}{13} + \frac{1}{144} + \frac{1}{1} + \frac{1}{91}} = 1.046\,4$$

relativ grosse Zahlen.

Kann man die zwei Kategorien der Zielvariablen „innerhalb eines Jahres“ und „vor mehr als einem Jahr“ zusammenlegen? Mit (für die Definition von meexp3 s. oben)

```
mod4=glm(me~factor(detc),data=meexp3,family="binomial"); summary(mod4)
```

erhält man

```
Null deviance: 241.68 on 177 degrees of freedom
Residual deviance: 234.71 on 175 degrees of freedom          ,
```

wobei die einzelnen Koeffizienten des Modells nicht signifikant von 0 verschieden sind. Der Chi-Quadrat-Test ergibt aber mit 1-pchisq(241.68-234.71,2) den p-Wert 0.03065376. Die beiden Ausprägungen sollten also nicht zusammengelegt werden. Dies zeigt, dass über die Zusammenlegung von Ausprägungen erst auf dem Hintergund der gesamten Modellbildung diskutiert werden sollte. H&L kommen auf einen p-Wert von 0.045, wobei nicht so klar ist, wie sie auf diese Zahl kommen (S. 272).

Bei Modellen mit mehr als einer kategorialen erklärenden Variable unterscheiden sich die Koeffizienten und Standardfehler des multinomialen Modells von denen, die sich durch die Berechnung zweier dichotomer Modelle ergeben.

Stetige erklärende Variablen: Für stetige Variablen gibt es bei m Ausprägungen der Zielvariable m-1 Koeffizienten - einen für jede Logit-Funktion. Der exponenzierte Koeffizient ergibt das Quotenverhältnis bezüglich der Veränderung um eine Einheit auf der Skala der stetigen Variable. Sind diese Einheiten nicht sinnvoll, so kann man wie Fall der dichotomen Zielvariable andere Distanzen auf der Achse der Variablen wählen und entsprechende Quotenverhältnisse berechnen. Das Verfahren ist völlig analog. Die Koeffizienten und Standardfehler des multinomialen Modells unterscheiden sich auch im bivariaten Fall (eine erklärende und eine erklärte Variable) von denen, die sich durch die Berechnung zweier dichotomer Modelle ergeben. Man kann also festhalten:

- in bivariaten Modellen (eine erklärende und eine erklärte Variable) mit nominalskalierter erklärender Variable liefern die Berechnungen von Modellen, die nur den Referenzwert und einen weiteren Wert der Zielvariable berücksichtigen, dasselbe Ergebnis wie das multinomiale Modell, in bivariaten Modellen mit stetiger erklärender Variable unterscheiden sich diese Resultate.

- in multiplen Modellen (mehrerer erklärende und eine erklärte Variable) liefern die Berechnungen von Modellen, die nur den Referenzwert und einen weiteren Wert der Zielvariable berücksichtigen, nicht dasselbe Ergebnis wie das multinomiale Modell, unabhängig von der Skalierung der erklärenden Variablen.

- Bei der Lieferung unterschiedlicher Resultate liegen diese i.A. nahe beieinander (s. Begg und Gray (1984), die die Konsistenz der dichotomen Schätzer für die Parameter des multinomialen Modells nachweisen).

Modellbildungsstrategien für die multinomiale logistische Regression

Im Prinzip sind die Modellbildungstrategien für die multinomiale logistische Regression identisch zum Modellen mit binärer Zielvariable. Bezüglich des Beispiels stellt sich das Problem, wie die erklärende ordinalskalierte Variable mit vier Ausprägungen sympt einzufügen ist. Traditioneller Weise wurden solche Variablen als stetige oder als kategoriale Variablen ins Modell eingeführt - wobei beides nicht ideal ist. Im ersten Fall werden Abstände in die Daten gelesen, die nicht begründbar sind, im zweiten Fall verzichtet man auf Information. Sie entscheiden sich für Behandlung als kategoriale Variable. Sie begründen dies mit der Möglichkeit, bei diesem Vorgehen die funktionale Form der Logits über die Kategorien zeichnen zu können. Zudem kann man überprüfen, ob man eventuell Kategorien zusammenlegen kann. Die Variable pb behandeln sie als stetige - was auf Grund der Konstruktion der Variable kritisierbar ist (Summe der Punkte von Likert-Skalen-Items). Man erhält mit dem nnet-Befehl „multinom“:

```
m275=multinom(me~factor(sympt)+pb+hist+bse+factor(detc),data=meexp)
summary(m275)
```

das folgende Modell:

Coeff.	(Intercept)	factor(sympt)2	factor(sympt)3	factor(sympt)4	pb
1	-2.9990607	0.1101408	1.9249158	2.457230	-0.2194421
2	-0.9860258	-0.2901346	0.8172916	1.132224	-0.1482079
	hist	bse	factor(detc)2	factor(detc)3	
1	1.366274	1.291772	0.01700706	0.9041717	
2	1.065446	1.052183	-0.92448094	-0.6906153	

Std. Errors:	(Intercept)	factor(sympt)2	factor(sympt)3	factor(sympt)4	pb
1	1.539293	0.9228324	0.7776651	0.775400 7	0.0755147
2	1.111829	0.6440622	0.5397872	0.547666	0.07636886
	hist	bse	factor(detc)2	factor(detc)3	
1	0.4375239	0.5299080 3	1.1619394	1.126864	
2	0.4593986	0.5149934	0.7137332	0.6871025	

Die p-Werte der Wald-Tests werden vom nnet-Paket nicht angeboten. Mit

```
summaryMN(m275)
```

erhält man:

	Coeff	stabf	z	p_Wert
1 (Intercept)	-2.999	1.539	-1.948	0.026
1 factor(sympt)2	0.110	0.923	0.119	0.452
1 factor(sympt)3	1.925	0.778	2.475	0.007
1 factor(sympt)4	2.457	0.775	3.169	0.001
1 pb	-0.219	0.076	-2.906	0.002
1 hist	1.366	0.438	3.123	0.001
1 bse	1.292	0.530	2.438	0.007
1 factor(detc)2	0.017	1.162	0.015	0.494
1 factor(detc)3	0.904	1.127	0.802	0.211
2 (Intercept)	-0.986	1.112	-0.887	0.188
2 factor(sympt)2	-0.290	0.644	-0.450	0.326
2 factor(sympt)3	0.817	0.540	1.514	0.065
2 factor(sympt)4	1.132	0.548	2.067	0.019
2 pb	-0.148	0.076	-1.941	0.026
2 hist	1.065	0.459	2.319	0.010
2 bse	1.052	0.515	2.043	0.021
2 factor(detc)2	-0.924	0.714	-1.295	0.098
2 factor(detc)3	-0.691	0.687	-1.005	0.157

Eine zu den p-Werten gleichwertige Möglichkeit bestünde in der Verwendung von VIs. Mit confint(m275) erhält man diese in informativer Darstellung. Da die erste Ausprägung von sympt nicht signifikant von der Referenz verschieden ist, kann man diese beiden Ausprägungen zusammenlegen. Die Koeffizienten der 3. und 4. Ausprägung weisen das gleiche Vorzeichen auf und unterscheiden sich nicht stark. Dies weist darauf hin, dass man diese zwei Ausprägungen eventuell zusammenlegen könnte: Die Differenz der Koeffizienten beträgt bezüglich „innerhalb eines Jahres“ $2.457 - 1.925 = 0.532$ und die Standardabweichung dieser Differenz erhält man mittels

$$\begin{aligned} V\left(B^1_{s3} - B^1_{s4}\right) &= V\left(B^1_{s3}\right) + V\left(B^1_{s4}\right) - 2Cov\left(B^1_{s3}, B^1_{s4}\right) \\ &= 0.7776^2 + 0.7753^2 - 2 \cdot 0.5598660545 \\ &= 8.602\,0 \times 10^{-2} \end{aligned}$$

und damit den Standardfehler $\sqrt{8.602\,0 \times 10^{-2}} = 0.293\,29$, wobei man $Cov\left(B^1_{s3}, B^1_{s4}\right)$ mit

```
vcov(m275)["1:factor(sympt)3","1:factor(sympt)4"])
```

erhält. Der Testwert ist $\frac{0.532}{0.293\,29} = 1.813\,9$ und $2*(1-pnorm(1.813\,9, 0, 1)$ ergibt den p-Wert 0.06969312. Für „vor mehr als einem Jahr“ erhält man die Differenz $1.132 - 0.817 = 0.315$ und die Standardabweichung dieser Differenz erhält man mittels

$$\begin{aligned} V\left(B^2_{s3} - B^2_{s4}\right) &= V\left(B^2_{s3}\right) + V\left(B^2_{s4}\right) - 2Cov\left(B^2_{s3}, B^2_{s4}\right) \\ &= 0.5398^2 + 0.5477^2 - 2 \cdot 0.2423544205 \\ &= 0.106\,6 \end{aligned}$$

und damit den Standardfehler $\sqrt{0.106\,6} = 0.326\,50$. Der Testwert ist $\frac{0.315}{0.326\,50} = 0.964\,78$ $= 1.5403$ und $2 * (1 - pnorm(0.964\,78, 0, 1)$ ergibt den p-Wert 0.334655. Man kann also auch diese beiden Kategorien zusammenlegen. Man kann entsprechend die Tests dazu verwenden, um sowohl die Unterscheidbarkeit der Kategorien der Zielvariable als auch die der erklärenden Variablen zu analysieren.

Man erhält mittels Umkodieren (erste drei Zeilen und dem nnet-Befehl „multinom“)

```
meexp$symptd=meexp$sympt
meexp$symptd[meexp$sympt==2]=1
meexp$symptd[meexp$sympt==4]=3
m275sd=multinom(me~factor(symptd)+pb+hist+bse+factor(detc),data=meexp)
summary_MN(m275sd)
```

die Tabelle

	Coeff	stabf	z	p_Wert
1 (Intercept)	-2.703	1.434	-1.885	0.030
1 factor(symptd)3	2.096	0.457	4.581	0.000
1 pb	-0.251	0.073	-3.442	0.000
1 hist	1.293	0.434	2.983	0.001
1 bse	1.244	0.526	2.364	0.009
1 factor(detc)2	0.089	1.161	0.077	0.469
1 factor(detc)3	0.972	1.126	0.863	0.194
2 (Intercept)	-0.998	1.072	-0.931	0.176
2 factor(symptd)3	1.121	0.357	3.139	0.001
2 pb	-0.168	0.074	-2.267	0.012
2 hist	1.014	0.454	2.234	0.013
2 bse	1.029	0.514	2.001	0.023
2 factor(detc)2	-0.902	0.715	-1.263	0.103
2 factor(detc)3	-0.670	0.688	-0.975	0.165

Der Modellvergleich (L-Quotiententest) ergibt

$$697.4959 - 693.9019 = 3.\,594$$

mit 4 Freiheitsgraden. Dies ergibt mit

```
testwert=deviance(m275sd)-deviance(m275)
1-pchisq(testwert,4)
```

den p-Wert 0.4637272 (oder mit anova(m275sd, m275)). Die beiden Modell unterscheiden sich also nicht und man kann das einfachere nehmen. H&L weisen darauf hin, dieser L-Quotiententest sei äquivalent mit dem „kombinierten Wald-Test“ im STATA-Programm.

Als nächstes wird die Rolle von detc im Modell untersucht. Die obige Tabelle zeigt, dass keiner der Wald-Tests signifikant anzeigt. Entsprechend wird ein Modell ohne diese Variable berechnet. Man erhält mit dem nnet-Befehl „multinom“

```
m275sd_o_detc=multinom(me~factor(symptd)+pb+hist+bse,data=meexp)
summary_MN(m275sd_o_detc)
```

die Tabelle

	Coeff	stabf	z	p_Wert
1 (Intercept)	-1.789	0.847	-2.112	0.017
1 factor(symptd)3	2.230	0.452	4.935	0.000
1 pb	-0.283	0.071	-3.960	0.000
1 hist	1.297	0.429	3.020	0.001
1 bse	1.221	0.521	2.343	0.010
2 (Intercept)	-1.742	0.809	-2.154	0.016
2 factor(symptd)3	1.153	0.351	3.282	0.001
2 pb	-0.158	0.071	-2.217	0.013
2 hist	1.061	0.453	2.345	0.010
2 bse	0.960	0.507	1.894	0.029

Die grösste Veränderung liegt bei PB vor (12.7%). Ein Modellvergleich mit

anova(m275sd_o_detc,m275sd)

ergibt den Testwert 8.542124 und den p-Wert Pr(Chi) 0.07362063 (bei 4 Freiheitsgraden). Dies ist ein Hinweis darauf, dass man detc eventuell weglassen könnte, wobei man vorgängig noch überprüfen kann, ob ein Modell mit einer dichotomisierten detc-Variable eventuell besser ist. Man erhält mit (es werden die ersten zwei Kategorien zusammengelegt, unwahrscheinlich und etwas wahrscheinlich, die erste Kategorie hat nur 18 Daten; Umkodieren erste zwei Zeilen; nnet-Befehl „multinom“)

```
meexp$detcd=meexp$detc
meexp$detcd[meexp$detc==2]=1
m275sd_m_detcd=multinom(me~factor(symptd)+pb+hist+bse+factor(detcd),data=meexp)
xtable(summary_MN(m275sd_m_detcd),digits=3)
```

die Tabelle

	Coeff	stabf	z	p_Wert
1 (Intercept)	-2.624	0.926	-2.832	0.002
1 factor(symptd)3	2.095	0.457	4.579	0.000
1 pb	-0.249	0.073	-3.437	0.000
1 hist	1.310	0.434	3.021	0.001
1 bse	1.237	0.525	2.354	0.009
1 factor(detcd)3	0.885	0.356	2.485	0.006
2 (Intercept)	-1.824	0.855	-2.133	0.016
2 factor(symptd)3	1.127	0.356	3.164	0.001
2 pb	-0.154	0.073	-2.125	0.017
2 hist	1.063	0.453	2.348	0.009
2 bse	0.956	0.507	1.884	0.030
2 factor(detcd)3	0.114	0.318	0.359	0.360

Der Modellvergleich

```
anova(m275sd_o_detc, m275sd_m_detcd)
```

ergibt den Testwert 6.905469 und den p-Wert 0.03165895 (2 df). Die Modelle unterscheiden sich also signifikant. Im ersten Logit ist der p-Wert kleiner als 0.05, im zweiten nicht. Die dichotomisierte Variable trägt also signifikant zu Modell bei. Der folgende Modellvergleich

```
anova(m275sd,m275sd_m_detcd)
```

ergibt einen Testwert von 1.636655 und eine p-Wert von 0.441169 (2 df). Das Modell, das alle drei Kategorien von detc verwendet, ist also nicht besser als das Modell, das nur 2 verwendet. Die grösste Veränderung in den Koeffizienten liegt bei pb in logit 2 vor (8%). Dies zeigt, dass die Verwendung der dichotomen Variable eine ebenso gute Anpassung der Effekte der anderen Kovariablen wie die trichotome Variable ergibt.

Als nächste geht es um die Überprüfung von pb. Es geht um die (pseudo)-metrische Variable, die aus der Summe von Likert-Skalen-Items resultiert. Es geht darum, die Linearität der Logit-Kurve zu überprüfen. Es können im Prinzip dieselben Methoden wie ich dichotomen Fall verwendet werden, wobei diese in Software-Paketen noch nicht implementiert seien (Stand 2000). Sie empfehlen deshalb dem Rat von Begg und Gray (1984) zu folgen, für jede Logitfunktion eine dichotome Analyse vorzunehmen. Man vergleicht eine Kategorie mit der Referenzkategorie und lässt die übrigen Kategorien ausser acht. Begg und Gray zeigen, dass die derart berechneten Koeffizientenstatistiken konsistent sind und der Verlust an Effizienz oft klein ist. Gemäss Erfahrung liegen die binomialen in der Nähe der multinomialen Koeffizienten. Entsprechend macht es Sinn, die Linearität der Logits in den dichotomen Modellen mit Hilfe von Quartil-Designvariablen, Glättung und fraktionalen Polynomen zu überprüfen .

Sie wenden zuerst die Methode mit den Designvariablen an (Quartile). Sie teilen die Variable auf in die Werte 5, 6-7, 8-9 und >=10. Mit

```
summary(meexp$pb)
```

erhält man nämlich

Min.	1st Qu.	Median	3rd Qu.	Max.
5.000	6.000	7.000	9.000	17.000

Mit den Befehlen

```
meexp$pbklass=cut(meexp$pb,quantile(meexp$pb),include.lowest=T)
meexp1=meexp[meexp$me==0 |meexp$me ==1,]
m279=glm(me~pbklass+factor(symptd)+hist+bse+factor(detcd),data=meexp1
    ,family = "binomial")
plotKoeff(meexp1$pbklass,m279,plotvi=T,pos="bottomleft")
```

erhält man die Abbildung 25

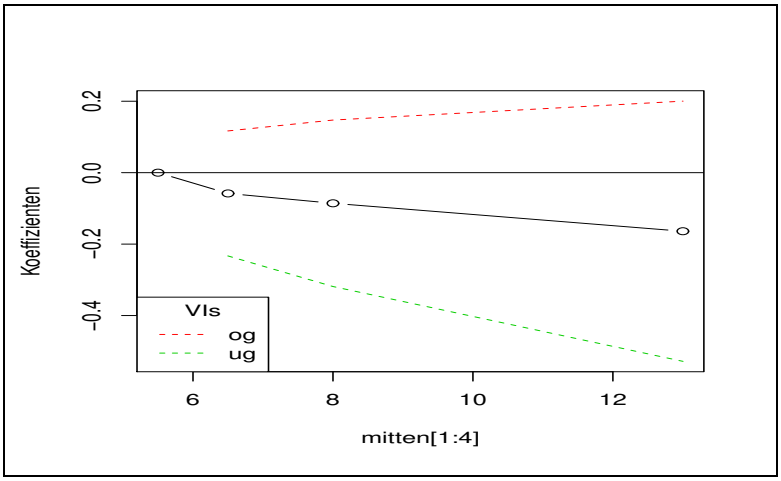

Abbildung 25: Plot der Koeffizienten der klassifizierten Variable pb und dichotomer logistischer Regression, Ausprägung 1 gegen 0

Das Ergebnis entsprechen nicht genau dem Resultat im Buch (S. 278). Mit

```
meexp2=meexp[meexp$me==0 |meexp$me ==2,]; meexp2$me[meexp2$me==2]=1
m279b=glm(me~pbklass+factor(symptd)+hist+bse+factor(detcd),data=meexp2,
    family = "binomial")
plotKoeff(meexp2$pbklass,m279b,plotvi=T,pos="topleft")
```

erhält man

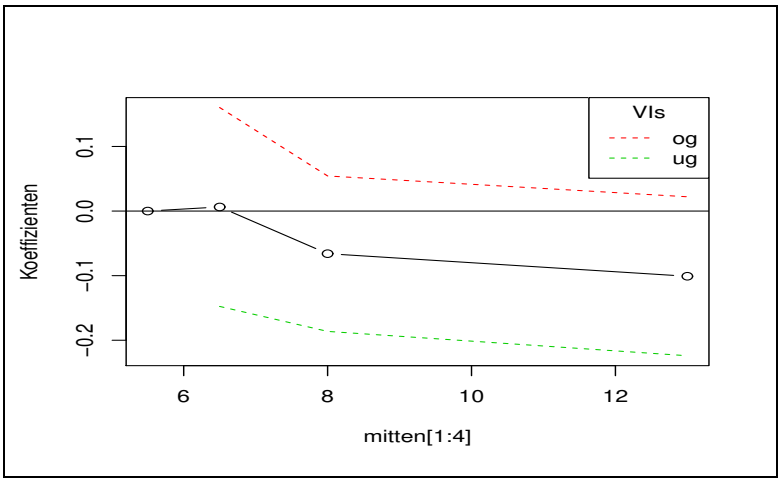

Abbildung 26: Plot der Koeffizienten der klassifizierten Variable pb und dichotomer logistischer Regression, Ausprägung 2 gegen 0

Die Linearität der Logits ist haltbar. Nachher wird mit Fraktionalen Polynomen untersucht, ob man die pb-Variable transformieren sollte. Wir reduzieren das meexp1-Dataframe, um nur mehr die betroffenen Variablen im Modell zu haben und wandeln diese, wenn nötig, in Faktorvariablen um:

```
meexp1red=meexp1[,c("me","pb","symptd","hist","bse","detcd")]
meexp1red$symptd=as.factor(meexp1red$symptd)
meexp1red$detcd=as.factor(meexp1red$detcd)
a=FP(meexp1red,"me","pb")
a
```

Keine Transformation von pb führt zu einem besseren Modell. Analog kann man den zweiten Vergleich (Ausprägung 2 zu 0) analysieren. Man belässt die Variable pb untransformiert im Modell.

Als nächstes ginge es darum, die Variablen auf Interaktion hin zu überprüfen. Gemäss H&L sind alle Interaktionen zwischen den verbleibenden Variablen inhaltlich sinnvoll. Testen ergibt aber, dass keine statistisch signifikant ist. Multinom funktioniert wie glm bezüglich der Eingabe von auf Interaktion zu untersuchenden Variablen.

Um die Methode von Begg und Gray zu illustrieren, werden die Koeffizienten der dichotomen logistischen Regressionen mit denen des multinomialen Modells verglichen (s. Tabelle 11):

		Coeff		stabf	
	MultiN	BinärLog	MultiN	BinärLog	
1	(Intercept)	-2.6237608	-2.76509	0.92639682	0.94213
	factor(symptd)3	2.0947519	2.091	0.45743059	0.46508
	pb	-0.2494748	-0.24261	0.07257904	0.07375
	hist	1.3098659	1.38502	0.43360235	0.46825
	bse	1.2370117	1.36331	0.52542421	0.53388
	factor(detcd)3	0.8851849	0.85269	0.35623804	0.36545
2	(Intercept)	-1.8238869	-1.8381	0.85509328	0.86
	factor(symptd)3	1.1274199	1.153	0.35636231	0.3566
	pb	-0.1543183	-0.1538	0.0726206	0.0726
	hist	1.0631796	1.0977	0.45284125	0.4593
	bse	0.9560139	0.9535	0.50733717	0.5097
	factor(detcd)3	0.114158	0.0987	0.31821222	0.3191

Tabelle 11: Vergleich der Koeffizienten des Multinomialen Modells mit den Koeffizienten zweier logistischer Modelle

Ein Problem, das man bei binärer logistischer Regression nicht antrifft, besteht darin, dass bei der multinomialen Regression eine Variable für manche Logit-Funktionen signifikant ist, für andere nicht. Wenn man das Prinzip anwendet, dass die Anzahl der Parameter möglichst klein sein sollte, möchte man die nicht signifikant von 0 verschiedenen Parameter mit 0 identifizieren und die Werte der übrigen Parameter neu schätzen. Diese Strategie liege im Augenblick (2000) in den zur Verfügung stehenden Software-Paketen nicht vor, kann aber mit der binären logistischen Regression durchgeführt werden. Wie bei aller Modellierung sollten dabei wissenschaftlich-inhaltliche Überlegungen eine Rolle spielen.

Anpassungstests und Diagnostik

Summarische Anpassungstest ergeben mit (g = 12 wird eingestellt, damit keine Fehlermeldung erscheint. Es muss mit verschiedenen g experimentiert werden. Hier scheint ein Fehler im LogisticX-Paket vorzuliegen)

```
m280b=glm(me~factor(symptd)+pb+hist+bse+factor(detcd),data=meexp1,
   family = "binomial")
gof(m280b,g=12)
```

und

```
m280c=glm(me~factor(symptd)+pb+hist+bse+factor(detcd),data=meexp2,
  family = "binomial")
gof(m280c,g=11)
```

unter anderem (für gruppierte Daten, d.h. mittels Kovariatenmuster ausgewertet; im ersten Logit stimmt der Chi-Quadrattest mit dem Wert des Buches überein, der p-Wert aber nicht. Beim HL-Test gibt einen Unterschied wegen der Anzahl Gruppen; beim Stukel-Test stimmt gar nichts überein. Der gof-Befehl liefert etliche Stukel-Tests, kein Ergebnis stimmt mit dem Buch überein. Im zweiten Logit stimmt nur der Stukel-Test mit den Ergebnissen des Buches überein).

Logit		HL chiSq	Chi-Quadratest	Stukel (SllBoth chiSq)
1	Testwert	9.01448	67.836	1.01957
	p-Wert	0.482780	0.53073	0.60063
	df	10	68	2
2	Testwert	63.827	12.81145	1.86440
	p-Wert	0.65349	0.171325	0.393687
	df	69	9	2

Insgesamt muss man die Hypothese des Passens des Modells nicht verwerfen.

Als nächstes geht es darum, den Einfluss einzelner Datenmuster zu überprüfen. Dazu kann man z.B. auf Grund der Graphiken für die verschiedenen Delta-Statistiken vorgängig die Anzahl der zu extrahierenden Zeilen (mit plot(glm-Modell) bestimmen. Man sieht graphisch, dass bezüglich h vermutlich die zwei höchsten Werte von Interesse sind, bezüglich dChisq einer, bezüglich dBhat 4 und bezüglich dDev 2 Werte. Um die zwei Muster mit extremen Werten dD-Werten zu finden rechnet man z.B.

```
m280b=glm(me~factor(symptd)+pb+hist+bse+factor(detcd),data=meexp1,
  family="binomial")
diag1=dx(m280b, byCov=T) #byCov=T ist Voreinstellung!!
which(diag1$dDev==max(diag1$dDev))
```

und erhält die Zeile 74 und mit

```
t(diag1[74,]
```

die Spalte mit den Werten der Statistiken. Die Zeile mit dem zweithöchsten dDev-Wert würde man dann mit

```
which(diag1$dDev==max(diag1$dDev[-74]))
```

finden (ergibt im Beispiel 69) und den Dritthöchsten - den man hier auf Grund der obigen graphischen Überprüfung nicht extrahieren würde - mit

```
which(diag1$dDev==max(diag1$dDev[-c(74,69)])).
```

Dies würde man für diag1$dChisq, diag1$dBhat und diag1$h machen, wobei die dx-Tabelle nach dChisq aufsteigend geordnet ist, so dass die Zeilen mit den grössten dChisq-Werten leicht zu finden sind. Die interessierenden Spalten kann man dann mit cbind verbinden, um die Analysen vorzunehmen.

Für die beiden grössten dDev-Werte, die in denselben Zeilen wie die grössten dChisq- und dChisqHL-Werten liegen, erhält man im Beispiel

	1	2
factor(symptd)3	0.000	1.000
pb	6.000	9.000
hist	0.000	0.000
bse	0.000	1.000
factor(detcd)3	0.000	1.000
y	1.000	11.000
P (π_j)	0.014	0.345
n (m_j)	2.000	18.000
yhat ($m_j\pi_j$)	0.029	6.208
Pr (Pearson-Residuum)	5.749	2.376
dr (Devianz-Residuum)	2.393	2.299
h (Hebel)	0.008	0.011
sPr (stand. Pr)	5.772	2.389
sdr (stand. dr)	2.403	2.312
dChisq	33.316	5.709
dDev	5.773	5.343
dBhat	0.267	0.063
hHL ($m_j h$)	0.016	0.198
dChisqHL	33.585	7.037
dDevHL	5.820	6.586
dBhatHL	0.543	1.733

Tabelle 12: Delta-Statistiken dChisq, dDev und dBhat sowie die HL-Varianten dChisqHL, dDevHL und dBhatHL (samt hHL=n·h).

dChisq ist recht gross in der Zeile 74. Die geschätzte logistische Wahrscheinlichkeit ist ziemlich klein, die relative Häufigkeit ist demgegenüber $\frac{1}{2} = 0.5$. Dies ergibt einen grossen Summanden der Summe der quadrierten Abweichungen. Die dDev ist auch recht gross. Der sehr kleine h-Wert ist dafür verantwortlich, dass dBhat nicht gross wird. Gestützt auf die klinische Plausibilität des beobachteten Kovariablenmusters gibt es wenig Gründe, die beiden Zeilen mit diesem Muster aus den Daten zu entfernen.

Die Zeilen mit den grössten dChisq sind also leicht zu extrahieren. Um die Werte des Buches (S. 283) zu erhalten, kann man mit der oben definierten Funktion deltaHL auf diag1 anwenden.

Als nächstes diskutieren Sie das Zeilenmuster 69 (zweite Spalte in der obigen Tabelle). Sie meinen, das Antwortemuster bezüglich der Variablen symptd, hist, bse und detcd stelle ein Muster dar, das man als modal bezeichnen können. 149 auf 338 weisen dieses Muster auf. Mit (Achtung: statt 0 und 1 sind symptd und detcd mit 1 und 3 kodiert):

```
length(meexp1[meexp1$symptd==3 & meexp1$hist==0 &
meexp1$bse==1 & meexp1$detcd==3][,1])
```

Es ist die Variable pb, die zusätzlich differenziert. dBhat sei ziemlich gross (dBhatHL =

1.733), h sei moderat (hHL= 0.198). Dies entspricht der Erwartung, da auch dChisq moderat ist (s. Tabelle 9 auf S. 105). Sie möchten ein Muster mit 18 Personen mit einem oft vorkommenden Muster nicht weglassen - wobei man noch eine Modellverbesserung versuchen könnte. Andererseits habe man wenig zusätzliche Information in den Kovariablen. Hinzufügen von Interaktionen von pb mit den anderen Kovariablen verbessert das Modell nicht. Es gebe sieben weitere Kovariablen-Muster mit dem „modalen“ Teilmuster. Das Modell passt sich diesen gut an. Deren pb-Werte liegen zwischen 5 und 12, so dass pb=9 im zweiten Muster nicht extrem ist.

Das Zeilenmuster 69 sei ein gutes Beispiel dafür, wieso man die Diagnostik auf Statistiken für Kovariablenmuster stützen sollte - statt auf die für individuelle Muster. Bei 18 Zeilen wäre der Hebel viel kleiner, nämlich $h = \frac{hHL}{18}$. ΔX^2 würde 1.92 und $\Delta\boldsymbol{\beta}$ würde 0.021 betragen (die Zahlen sind von den obigen dChisq und dBhat verschieden, da diese z.B. für dChisq mittels $r_j = \frac{y_j - m_j\pi_j}{\sqrt{m_j\pi_j(1-\pi_j)}}$ und nicht mittels $r_i = \frac{y_i - \pi_i}{\sqrt{\pi_i(1-\pi_i)}}$ berechnet werden. Der Unterschied bei der Berechnung von dChisq und dChisqHL besteht nur in der Verwendung unterschiedlicher Hebel). Man würde gemäss H&L bei der Betrachtung individuelle Muster den doch gewichtigen Einfluss dieses Musters leicht übersehen. Für die zweite dichotome logistische Regression erhält man mit

```
m280c=glm(me~factor(symptd)+pb+hist+bse+factor(detcd),
   data=meexp2,family = "binomial")
diag2=dx(m280c)
diag2=deltaHL(diag2)
tail(diag2,10)[c(8,10),]
```

die ersten zwei Spalten der Tabelle 13 (als letzte und drittletzte Zeile, wobei diag2 nach dBhat geordnet ist). Alle drei Spalten erhält man mit

```
diag3=diag2[order(diag2$dBhatHL),]
tail(diag3)
```

und zwar als die letzten drei Zeilen.

	1	2	3
(Intercept)	1.000	1.000	1.000
factor(symptd)3	1.000	1.000	1.000
pb	9.000	10.000	10.000
hist	1.000	0.000	0.000
bse	1.000	0.000	1.000
factor(detcd)3	0.000	0.000	1.000
y	3.000	1.000	2.000
P	0.496	0.098	0.237
n	3.000	1.000	19.000
yhat	1.487	0.098	4.498
Pr	1.748	3.039	-1.348
dr	2.052	2.157	-1.470
h	0.067	0.027	0.015
sPr	1.809	3.081	-1.358
sdr	2.125	2.186	-1.481
dChisq	3.272	9.490	1.845
dDev	4.514	4.781	2.194
dBhat	0.234	0.264	0.028
hHL	0.200	0.027	0.283
dChisqHL	3.819	9.490	2.534
dDevHL	5.268	4.781	3.014
dBhatHL	0.956	0.264	0.999

Tabelle 13: Delta-Statistiken für die zweite dichotome logistische Regression des Beispiels

Sie meinen, die zweite Spalte zeige, dass das Modell hier schlecht passe. Die geschätzte Wahrscheinlichkeit ist klein, der angenommene Wert ist 1. Dies ergibt eine grosse Abweichung. Da der Hebel aber klein ist, wird der Effekt auf dBhatHL gedämpft. Das Problem bleibe aber bestehen, dass für eine Person 1 eintritt, obwohl die Wahrscheinlichkeit des Eintretens von 1 nur fast 10% beträgt.

Die dritte Spalte hat den grössten dBhatHL-Wert. Sie weist das oben erwähnte „modale“ Muster und pb= 10 auf. Die erste Spalte hat den zweitgrössten dBhatHL-Wert. Drei Zeilen weisen diesen Wert auf, wobei pb = 9. Beide Muster sind einflussreich wegen den moderaten dChisqHL- und hHL-Werten. Während bei den ersten zwei Spalten nur der ausgezeichnete Wert angenommen wird und pb=9 und pb=10, nehmen in der dritten Spalte nur 2 auf 19 Personen den ausgezeichneten Wert an, wobei pb=10. Es scheint, dass die Antwortevariable in diesem Bereich von pb zu variabel ist, als dass sich das Modell daran anpassen könnte. Die Verwendung von Interaktion verbessert das Modell aber nicht.

An diesem Punkt hat man nur wenige Optionen mit den verfügbaren Daten. Kein alternatives Modell verbessert das Schlussmodell. Lässt man die drei einflussreichen Variablenmuster weg, erhält man mit (man lässt das zweite Variablenmuster der ersten Tabelle (1.

Logit) und das erste und dritte Variablenmuster der zweiten Tabelle (2. Logit) weg (symptd und detcd sind statt mit 0 und 1 mit 1 und 3 kodiert. Zu beachten ist, dass Muster von sympdt, pb, hist, bse und detcd sowohl bezüglich me=1 und me=2 vorkommen. Man muss entsprechend darauf achten, die Muster nur in der entsprechenden Untergruppe zu löschen). Man erhält mit

```
meexp7=meexp[!(meexp$symptd==3 & meexp$pb==9 & meexp$hist==0 &
      meexp$bse==1 & meexp$detcd==3 & !meexp$me==2 ),]
meexp8=meexp7[!(meexp7$symptd==3 & meexp7$pb==9 & meexp7$hist==1 &
      meexp7$bse==1 & meexp7$detcd==1 & !meexp7$me==1 ),]
meexp9=meexp8[!(meexp8$symptd==3 & meexp8$pb==10 & meexp8$hist==0 &
      meexp8$bse==1 & meexp8$detcd==3 & !meexp8$me==1 ),]
```

das Resultat:

	Coeff	stabf	z	p_Wert
1 (Intercept)	-2.892	1.042	-2.776	0.003
1 factor(symptd)3	2.125	0.463	4.586	0.000
1 pb	-0.216	0.085	-2.532	0.006
1 hist	1.244	0.442	2.815	0.002
1 bse	1.271	0.531	2.394	0.008
1 factor(detcd)3	0.883	0.369	2.392	0.008
2 (Intercept)	-2.663	0.956	-2.787	0.003
2 factor(symptd)3	1.191	0.361	3.298	0.000
2 pb	-0.080	0.079	-1.023	0.153
2 hist	0.606	0.495	1.223	0.111
2 bse	1.081	0.512	2.110	0.017
2 factor(detcd)3	0.477	0.340	1.401	0.081

Man kann nun die prozentualen Verschiebungen in den Koeffizienten untersuchen. Mit

```
cbind(coef(m280a)[1,],coef(m285)[1,],
(coef(m285)[1,]-coef(m280a)[1,])/abs(coef(m280a)[1,])*100,
coef(m280a)[2,], coef(m285)[2,],
(coef(m285)[2,]-coef(m280a)[2,])/abs(coef(m280a)[2,])*100)
```

erhält man

	M1 L1	M2 L1	Veränderung in %	M1 L2	M2 L2	Veränderung in %
(Intercept)	-2.624	-2.892	-10.230	-1.824	-2.663	-46.032
factor(symptd)3	2.095	2.125	1.429	1.127	1.191	5.607
pb	-0.249	-0.216	13.311	-0.154	-0.080	47.894
hist	1.310	1.244	-5.065	1.063	0.606	-43.013
bse	1.237	1.271	2.780	0.956	1.081	13.089
factor(detcd)3	0.885	0.883	-0.241	0.114	0.477	317.617

Tabelle 14: Veränderung in Prozenten von Modell M1 mit allen Daten zu Modell M2 ohne die 40 Daten der 3 weggelassenen Muster bezüglich Logit 1 (L1) und Logit 2 (L2)

In Logit 1 ist beträgt die maximale Differenz 13% für pb. Die maximale Differenz bezüglich logit 2 beträgt 318% für detcd. Während im ersten Logit die Unterschiede relativ klein sind, sind sie beim zweiten fast immer beträchtlich.

Bezüglich pb weisen Sie darauf hin, dass diese Variable durch 5 andere Variablen gebildet wurde (Summe von Likert-Skalen-Items). Dadurch kann Information bezüglich der Individuen verloren gehen. So kann pb=9 und pb=10 durch sehr unterschiedliche Variablenmuster in den 5 Variablen zustande kommen. Liegen die ursprünglichen Variablen vor, wäre es ratsam, diese einzeln im Modell zu begutachten.

Die Daten der 40 vorläufig ausgeschlossenen Untersuchungsobjekte sind nicht ungewöhnlich, so dass es keine klinische Basis für deren Ausschluss gibt. Entsprechend deklarieren H&L das Modell 280a mit allen Untersuchungsobjekten als das endgültige. Man kann für dieses noch die Quotenverhältnisse und deren VIs berechnen, wobei sie bezüglich pb Zweierschritte nach unten betrachten (darum *-2; ist ja vernünftig, kleinere Zahlen drücken ein grösseres Vertrauen in Mammographie aus). Mit

```
or1=exp(coef(m280a)[1,2:6]);    or1[2]=exp(coef(m280a)[1,3]*-2)
vior1u=exp(coef(m280a)[1,2:6]-1.96*sqrt(diag(vcov(m280a))[2:6]))
vior2u[2]=exp(coef(m280a)[2,3]*(-2)-1.96*2*sqrt(diag(vcov(m280a))[9]))
vior1o=exp(coef(m280a)[1,2:6]+1.96*sqrt(diag(vcov(m280a))[2:6]))
vior1o[2]=exp(coef(m280a)[1,3]*(-2)+1.96*2*sqrt(diag(vcov(m280a))[3]))
```

```
or2=exp(coef(m280a)[2,2:6]);or2[2]=exp(coef(m280a)[2,3]*-2)
vior2u=exp(coef(m280a)[2,2:6]-1.96*sqrt(diag(vcov(m280a))[8:12]))
vior2u[2]=exp(coef(m280a)[2,3]*(-2)-1.96*2*sqrt(diag(vcov(m280a))[9]))
vior2o=exp(coef(m280a)[2,2:6]+1.96*sqrt(diag(vcov(m280a))[8:12]))
vior2o[2]=exp(coef(m280a)[2,3]*(-2)+1.96*2*sqrt(diag(vcov(m280a))[9]))
```

```
cbind(or1,vior1u,vior1o,or2,vior2u,vior2o)
```

erhält man

	OR1	VI OR1 unten	VI OR1 oben	OR2	VI OR2 unten	VI OR2 oben
factor(symptd)3	8.123	3.314	19.912	3.088	1.536	6.208
pb	1.647	1.239	2.189	1.362	1.024	1.810
hist	3.706	1.584	8.669	2.896	1.192	7.034
bse	3.445	1.230	9.649	2.601	0.962	7.031
factor(detcd)3	2.423	1.206	4.872	1.121	0.601	2.091

Die Tabelle zeigt, dass für beide Ausprägungen (innerhalb eines Jahres, vor mehr als einem Jahr) steigende ORs vorliegen. Die Koeffizienten-Schätzer für die erste Ausprägung sind grösser als die für die zweite. Dies ist inhaltlich vernünftig. Alle 5 Kovariablen korrelieren mit einer stärkeren Verwendung der Mammographie.

Sie erwähnen am Schluss einen multivariaten Wald-Test bezüglich Gleichheit der beiden Logits mit dem Testwert 10.87, 5 Freiheitsgraden und einem p-Wert von 0.054. Auf einen Signifikanzniveau von 0.1 spricht diese gegen die Zusammenlegung der Ausprägungen. Sie liefern keine Angaben zur Durchführung des Tests. Rechnen wir wie oben, erhält man mit

```
A=cbind(diag(6),-diag(6))
beta=t(as.vector(cbind(coef(m275sd_m_detcd)[1,], coef(m275sd_m_detcd)[2,])))
Ab = A %*% t(beta)
tw = t(Ab) %*% solve(A %*% vcov(m275sd_m_detcd) %*% t(A)) %*% Ab
```

allerdings den Testwert tw = 11.33504, df= 6, 1-pchisq(tw, 6) = 0.01492061. Rechnet man eine dichotome logistische Regression bezüglich der beiden Ausprägungen durch, so erhält man eher ein Argument für die Zusammenlegung der Ausprägungen:

```
meexp10=meexp[meexp$me==1|meexp$me==2,]
meexp10$me[meexp10$me==1]=0
meexp10$me[meexp10$me==2]=1
m287=glm(me~factor(symptd)+pb+hist+bse+factor(detcd),data=meexp10)
1-pchisq(m287$null.deviance-m287$deviance,5)
```

ergibt den p-Wert 0.7288409. Im Modell ist auf dem 0.05-Niveau ausser der Konstanten kein Koeffizient signifikant von 0 verschieden.

7 Ordinale logistische Regression

(Im Buch 8.2) Bei ordinalskalierten Variablen mit wenig Ausprägungen kann man im Prinzip die multinomiale logistische Regression verwenden, wobei man damit nicht alle Information verwendet und eventuell auch nicht auf die interessierende Fragestellung antworten kann. Es gibt verschiedene Modelle ordinaler logistischer Regression. H&L stellen die drei am häufigsten verwendeten Modelle vor: die Modelle

1. benachbarter Kategorien (adjacent categories)
2. kontinuierlicher Verhältnisse (continuation-ratio)
3. proportionaler Quoten (proportional odds)

Sie erwähnen folgende weiteren Modelle - samt entsprechender Literatur (S. 288): das partiell-proportionale Modell ohne Nebenbedingung (unconstrained partial-proportional), das partiell-proportionale Modell mit Nebenbedingung (constrained) und das stereotype logistische Modell. Im Rahmen der ordinalen logistischen Regression wird das multinomiale Modell oft Basislinien-Logit-Modell (baseline logit model) genannt, da man die Kategorien $Y > 0$ mit einer einzigen Kategorie $Y = 0$, der „Basislinie“, vergleicht. Das multinomiale Modell hat, wie bereits erläutert, $K \cdot (p+1)$ Parameter ($K+1$ = Anzahl der Ausprägungen der erklärten Variable, p Anzahl erklärende Variablen). Die K Logit-Funktionen sind gegeben durch

$$g_k(\mathbf{x}) = \ln \frac{\pi_k(\mathbf{x})}{\pi_0(\mathbf{x})} = \beta_{k0} + \mathbf{x}^\intercal \boldsymbol{\beta}_k$$

für $k \in \mathbb{N}_K^*$ und $\mathbf{x}, \boldsymbol{\beta}_k \in \mathbb{R}^p$. Geht man zu ordinalen Modellen über, muss man entscheiden, welche Kategorien der Zielvariable man vergleichen will.

Das Modell benachbarter Kategorien erhält man, wenn man jede Ausprägung mit der nächsthöheren vergleicht. Geht man davon aus, dass die Log-ORs benachbarter Kategorien konstant und die Log-Odds linear sind, so kann man dieses Modell wie folgt spezifizieren:

$$a_k(\mathbf{x}) = \ln \frac{\pi_k(\mathbf{x})}{\pi_{k-1}(\mathbf{x})} = \alpha_k + \mathbf{x}^\intercal \boldsymbol{\beta}$$

für $k \in \mathbb{N}_K^*$ und $\mathbf{x}, \boldsymbol{\beta}_k \in \mathbb{R}^p$. Es müssen also weniger Koeffizienten als im multinomialen Modell geschätzt werden, nämlich nur $K+p+1$. Dieses Modell ist mit dem multinomialen Modell unter der Nebenbedingung $\boldsymbol{\beta}_k = \boldsymbol{\beta}_{k-1}$ für $k \in \mathbb{N}_K^* \backslash \{1\}$ identisch. Es gilt nämlich durch Erweitern:

$$\ln \frac{\pi_k(\mathbf{x})}{\pi_0(\mathbf{x})} = \ln \left(\frac{\pi_1(\mathbf{x})}{\pi_0(\mathbf{x})} \frac{\pi_2(\mathbf{x})}{\pi_1(\mathbf{x})} \cdot \ldots \cdot \frac{\pi_k(\mathbf{x})}{\pi_{k-1}(\mathbf{x})} \right)$$

und damit

$$\begin{aligned}\ln\frac{\pi_k(\mathbf{x})}{\pi_0(\mathbf{x})} &= \ln\frac{\pi_1(\mathbf{x})}{\pi_0(\mathbf{x})} + \ln\frac{\pi_2(\mathbf{x})}{\pi_1(\mathbf{x})} + ... + \ln\frac{\pi_k(\mathbf{x})}{\pi_{k-1}(\mathbf{x})} \\ &= a_1(\mathbf{x}) + a_2(\mathbf{x}) + ... + a_k(\mathbf{x}) \\ &= (\alpha_1 + \mathbf{x}^\intercal\boldsymbol{\beta}) + (\alpha_2 + \mathbf{x}^\intercal\boldsymbol{\beta}) + ... + (\alpha_k + \mathbf{x}^\intercal\boldsymbol{\beta}) \\ &= (\alpha_1 + \alpha_2 + ... + \alpha_k) + k\mathbf{x}^\intercal\boldsymbol{\beta}\end{aligned}$$

Man erhält das Basislinienmodell mit $\beta_{k0} = (\alpha_1 + \alpha_2 + ... + \alpha_k)$ und $\boldsymbol{\beta}_k = k\boldsymbol{\beta}$. Man kann die Modellparameter also mit Hilfe eines multinomialen Modell unter der erwähnten Nebenbedingung schätzen.

Das Modell kontinuierlicher Verhältnisse erhält man durch den Vergleich jeder Antwort mit allen darunterliegenden Antworten. Man setzt also

$$\begin{aligned}r_k(\mathbf{x}) &= \ln\frac{P(Y=k|\mathbf{x})}{P(Y<k|\mathbf{x})} \\ &= \ln\frac{\pi_k(\mathbf{x})}{\pi_0(\mathbf{x}) + \pi_1(\mathbf{x}) + ... + \pi_{k-1}(\mathbf{x})} \\ &= \theta_k + \mathbf{x}^\intercal\boldsymbol{\beta}_k\end{aligned} \tag{14}$$

mit $\mathbf{x}$, $\boldsymbol{\beta}_k \in \mathbb{R}^p$. Das Modell hat wiederum $K \cdot (p+1)$ Parameter und diese können mittels K binärer logistischer Regressionen geschätzt werden. Man kann auch für dieses Modell die Nebenbedingung $\boldsymbol{\beta}_k = \boldsymbol{\beta}_{k-1}$ für $k \in \mathbb{N}_K^* \backslash \{1\}$ setzen. Man erhält

$$r_k(\mathbf{x}) = \theta_k + \mathbf{x}^\intercal\boldsymbol{\beta} \tag{15}$$

mit $\mathbf{x}$, $\boldsymbol{\beta}_k \in \mathbb{R}^p$. Man könnte auch

$$r'_k(\mathbf{x}) = \ln\frac{P(Y>k|\mathbf{x})}{P(Y=k|\mathbf{x})} \tag{16}$$

für $k \in \mathbb{N}_{K-1}$ setzen, wobei dieses Modell mit jenem kontinuierlicher Verhältnisse nicht äquivalent ist. Sie ziehen die erste Variante vor, da bei $K = 1$ eine normale logistische Regression resultiert. Mit dem R-VGAM-Paket kann man alle drei Varianten (14) - (16) rechnen (s. u.)

Das Modell der proportionalen Quoten schliesslich ist gegeben durch

$$\begin{aligned}c_k(\mathbf{x}) &= \ln\frac{P(Y\leq k|\mathbf{x})}{P(Y>k|\mathbf{x})} \\ &= \ln\frac{\pi_0(\mathbf{x}) + \pi_1(\mathbf{x}) + ... + \pi_k(\mathbf{x})}{\pi_{k+1}(\mathbf{x}) + \pi_{k+2}(\mathbf{x}) + ... + \pi_K(\mathbf{x})} \\ &= \tau_k - \mathbf{x}^\intercal\boldsymbol{\beta}\end{aligned}$$

mit $\mathbf{x}$, $\boldsymbol{\beta}_k \in \mathbb{R}^p$. Bei $K = 1$ resultiert eine gewöhnliche binäre logistische Regression, die eine OR von Y=0 versus Y=1 liefert. Der Term $\mathbf{x}^\intercal\boldsymbol{\beta}$ wird negiert, um in Übereinstimmung mit einem Teil der Software und Literatur zu bleiben: dadurch werden bei der Wahl der tiefsten Kodierung der erklärenden Variablen als Referenz und bei ansteigenden Logits für aufsteigende Kodierung der erklärten Variable positive Koeffzienten berechnet. Das Resultat entspricht der Berechnung des Modells

$$c_k^u(\mathbf{x}) = \ln \frac{P(Y > k|\mathbf{x})}{P(Y \leq k|\mathbf{x})}$$
$$c_k^u(\mathbf{x}) = \tau_k + \mathbf{x}^\intercal\boldsymbol{\beta}.$$

Bevor die drei Modelle „benachbarter Kategorien“, „kontinuierlicher Verhältnisse“ und „proportionaler Quoten“ an einem Beispiel vorgeführt und diskutiert werden, noch ein paar allgemeine Bemerkungen. Die Modell-Anpassungs-Methoden basieren, mit Ausnahme des Modells kontinuierlicher Verhältnisse ohne Nebenbedingung, auf einer Anpassung der Likelihood und der Logitfunktionen des multinomialen Modells. Die grundlegende Prozedur beinhaltet zwei Schritte:

1. die Ausdrücke, welche die modellspezifischen Logits definieren, werden verwendet, um die Gleichungen aufzustellen, welche $\pi_k(\mathbf{x})$ als Funktion der unbekannten Parameter definieren

2. Es werden die Werte K+1 dimensionaler multinomialer Ergebnisse geschaffen - auf der Basis der ordinalen Ergebnisse: man setzt $\mathbf{z}^\intercal = (z_0, z_1, ..., z_K)$, mit $z_k = 1$, wenn $y = k$, sonst $z_k = 0$. Entsprechend ist nur ein Wert von $\mathbf{z}^\intercal$ von 0 verschieden.

Die allgemeine Form der Likelihood für eine Stichprobe von n unabhängigen Beobachtungen $(y_i, \mathbf{x}_i)$ ist dann für $i \in \mathbb{N}_n^*$

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\pi_0(\mathbf{x}_i)^{z_{0i}} \pi_1(\mathbf{x}_i)^{z_{1i}} \cdot ... \cdot \pi_K(\mathbf{x}_i)^{z_{Ki}}\right)$$

und die Log-Likelihoodfunktion

$$LL(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(z_{0i} \ln(\pi_0(\mathbf{x}_i)) + z_{1i} \ln \pi_1(\mathbf{x}_i) + ... + z_{Ki} \ln \pi_K(\mathbf{x}_i)\right)$$

Um die Maximum-Likelihood-Schätzer zu erhalten, werden die Funktionen nach den Parametern abgeleitet. Mit den bekannten Näherungsverfahren werden die Nullstellen gesucht. Mit Hilfe der zweiten Ableitungen werden wiederum die Kovarianzen bestimmt.

Statt in die entsprechenden technischen Details zu gehen, wollen sie die verschiedenen Methoden an einem Beispiel vorführen und Interpretationshilfen gewähren. Dazu wählen sie die „Low Birth Weight“-Studie (die Daten wurden schon oben verwendet). Die metrische

Variable BWT wird klassifiziert und *absteigend* in die Variable BWT4 kodiert:

```
bw=read.table("pfad/lowbwt.txt",header=T)
bw293=bw
bw293$bwt4[bw$bwt>3500]=0
bw293$bwt4[3000 < bw$bwt & bw$bwt <= 3500]=1
bw293$bwt4[2500 < bw$bwt & bw$bwt <= 3000]=2
bw293$bwt4[ bw$bwt <= 2500]=3
```

Sie verwandeln eine metrische Variable in eine ordinale mit wenig Ausprägungen, um zu zeigen, wie man das Modell proportionaler Quoten mit Hilfe der Kategorisierung einer metrischen Variable herleiten kann. Die absteigende Kodierung ist angemessen, wenn man das höchste Gewicht als Referenz wählen will - der tiefste Code ist nämlich gewöhnlich die Referenz.

Zu Beginn betrachten Sie die Kreuztabelle der Variable BWT4 und der Variable Raucherstatus. Man erhält mit

```
tabord=table(bw293$bwt4,bw293$smoke)
colnames(tabord)=c("Nein", "Ja")
rownames(tabord)=c("0: BWT > 3500", "1: 3000 < BWT <= 3500",
"2: 2500 < BWT <= 3000", "3: <= 2500")
```

die Tabelle:

Geburtsgewicht	Raucher	
BWT	Nein	Ja
0: BWT > 3500	35	11
1: 3000 < BWT <= 3500	29	17
2: 2500 < BWT <= 3000	22	16
3: BWT <= 2500	29	30

Einige interessierende Quotenverhältnisse sind:

$$
\begin{aligned}
OR(1,0) &= \tfrac{17/29}{11/35} = 1.865\,2 \\
OR(2,0) &= \tfrac{16/22}{11/35} = 2.\,314 \quad OR(2,1) = \tfrac{16/22}{17/29} = 1.240\,6 \\
OR(3,0) &= \tfrac{30/29}{11/35} = 3.291\,5 \quad OR(3,2) = \tfrac{30/29}{16/22} = 1.422\,4
\end{aligned}
$$

$$
\begin{aligned}
\ln OR(1,0) &= 0.623\,37 \\
\ln OR(2,0) &= 0.838\,98 \quad \ln OR(2,1) = 0.215\,60 \\
\ln OR(3,0) &= 1.191\,3 \quad \ln OR(3,2) = 0.352\,35
\end{aligned}
$$

$OR(3,0) = 3.291\,5$ gibt dabei z. B. an, dass die Quote Raucherin zu Nicht-Raucherin bei einem Geburtsgewicht von weniger als 2500 3.3 mal so hoch ist als bei einem Geburtsgewicht von mehr als 3500. Da $\frac{30/29}{11/35} = \frac{30/11}{29/35}$ kann man ebenso gut sagen, dass die Quote von

weniger Gewicht als 2500 zu mehr Gewicht als 3500 bei Raucherinnen dreimal höher ist als bei Nicht-Raucherinnen.

Die lnORs steigen bezüglich Basislinie an, vermutlich - bis auf Zufallsstreuung - um einen konstanten Faktor. Die lnORs benachbarter Kategorien scheinen - bis auf Zufallsstreuung - konstant zu sein.

Das Modell benachbarter Kategorien kann man anwenden, wenn das logarithmierte Quotenverhältnis $\ln(OR(k,0))$ das k-fache des ersten logarithmierten Quotenverhältnisse ist, d.h. wenn

$$\ln(OR(k,0)) = k\ln(OR(1,0))$$

für $k \in \mathbb{N}_K^*$. Entsprechend sind die lnORs benachbarter Kategorien konstant. Man erhält die Koeffizientenschätzer mit dem vglm-Befehl des R-Paketes VGAM mittels („acat“ für adjacent categories, parallel=T erzwingt, dass es nur einen Koeffizienten für smoke gibt, der von den Kategorien der Zielvariable unabhängig ist).

```
library(VGAM)
bw203_adj = vglm(bwt4~smoke,data=bw293, family=acat(parallel=TRUE))
summary(bw203_adj)
```

die Schätzer

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept):1	-0.1100	0.2106	-0.522	0.60160
(Intercept):2	-0.3314	0.2249	-1.474	0.14055
(Intercept):3	0.2664	0.2207	1.207	0.22743
smoke	0.3696	0.1332	2.774	0.00553 **

und

```
Residual deviance: 511.3056 on 563 degrees of freedom
Log-likelihood: -255.6528 on 563 degrees of freedom
```

Es handelt sich um die gleichen Schätzer wie im Buch (S. 294), die dort allerdings aus dem Basislinien-Modell mit den erwähnten Nebenbedingungen berechnet werden. Aus den obigen Ergebnissen erhält man drei Logit-Gleichungen:

$$\begin{aligned} a_1(smoke) &= -0.1100 + 0.3692(smoke) \\ a_2(smoke) &= -0.3314 + 0.3692(smoke) \\ a_3(smoke) &= 0.2664 + 0.3692(smoke) \end{aligned}$$

Es gilt

$$OR(k,k-1) = \exp(0.3692) = 1.4466$$

für $k = 1, 2, 3$. Berechnungsbeispiele:

$$\widehat{OR}(1,0) = \exp(-0.1100 + 0.3692 \cdot 1 - (-0.1100 + 0.3692 \cdot 0)) = \exp(0.3692)$$
$$\widehat{OR}(2,1) = \exp(-0.3314 + 0.3692 \cdot 2 - (-0.3314 + 0.3692 \cdot 1)) = \exp(0.3692)$$

Die Quoten, in der nächsttieferen Gewichtsklasse zu sein, sind für Kinder rauchender Frauen jeweils um 1.446 höher als die Quoten von Kindern nicht-rauchender Frauen. Sie vergleichen das Model mit dem multinomialen Modell ohne Nebenbedingungen mittels eines Likelihood-Ratio-Tests. Mit

```
library(nnet)
mmult=multinom(bwt4~smoke,data=bw293, family="binomial")
summary(mmult)
```

erhält man die restliche Abweichung

$$510.9718$$

und damit den Testwert

$$511.3056 - 510.9718 = 0.333\,8$$

(df $= 2$). Die zwei Freiheitsgrade kommen von den Nebenbedingungen für die beieinanderliegenden Kategorien 2 und 3 her. Im Allgemeinen sind die Freiheitsgrade für diesen Test $((K+1)-2)\,p$. 1-pchisq(511.3056-510.9718, 2)= 0.8462842. Die Modelle unterscheiden sich also nicht signifikant. Da das ordinale Modell weniger Parameter enthält, könnte man denken, dieses Modell sei besser. Dies sollte man aber erst auf dem Hintergrund des vollen Modells diskutieren, das auf Anpassung getestet wurde.

Das Modell kontinuierlicher Verhältnisse ergibt fürs Beispiel mit Hilfe des vglm-Befehls („sratio“ für „stop ratio“; es gibt auch eine Familie „cratio“ für „continuation ratio“, ein Modell, das anders definiert ist als bei L&H, s. Hilfe im Paket; reverse = T für Berechnung von unten nach oben, Voreinstellung ist reverse=F; parallel = F für Schätzungen von smoke in Abhängigkeit von Ausprägungen der Zielvariable; ist die Voreinstellung)

```
bw203_sr=vglm(bwt4~smoke, family=sratio(link ="logitlink",
parallel = F, reverse=T), data=bw293)
summary(bw203_sr)
```
(17)

Man erhält Parameterschätzungen des Buches (S. 296):

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept):1	-0.1881	0.2511	-0.749	0.4539
(Intercept):2	-1.0678	0.2471	-4.321	1.56e-05 ***
(Intercept):3	-1.0871	0.2147	-5.062	4.14e-07 ***
smoke:1	0.6234	0.4613	1.351	0.1766
smoke:2	0.5082	0.3991	1.273	0.2029
smoke:3	0.7041	0.3196	2.203	0.0276 *

Die Log-Likelihood ist mit der des multinomialen Modells (ohne Nebenbedingung) identisch (-255.4859 mit df=561).

Man sieht, dass die drei Koeffizienten von smoke recht ähnlich sind. Der Schätzer gibt an, dass die Quote, als Säugling in der nächsttieferen Gewichtsklasse zu sein, bei rauchenden Müttern um ungefähr $\exp(0.6) = 1.8221$ höher ist. Um die smoke-Koeffizienten auf Gleichheit hin zu testen, kann man sich zu Nutze machen, dass die drei Mengen von Parameterschätzern unabhängig sind. Entsprechend kann man einen Chi-Quadrat-Test mit df=2 durchführen:

$$W^2 = \frac{(0.6234 - 0.5082)^2}{0.4613^2 + 0.3991^2} + \frac{(0.6234 - 0.7041)^2}{0.4613^2 + 0.3196^2} = 0.056346$$

und $1 - pchisq(0.056346, 2) = 0.9722202$. Entsprechend ist es sinnvoll, das Modell mit der Nebenbedingung smoke:1=smoke:2=smoke:3 zu berechnen (in (17) setzt man „parallel=T“). Man erhält die Zahlen von S. 297 mit smoke = 0.6266:

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept):1	-0.1890	0.2204	-0.858	0.39107
(Intercept):2	-1.1138	0.2134	-5.220	1.79e-07 ***
(Intercept):3	-1.0523	0.1862	-5.652	1.59e-08 ***
smoke	0.6266	0.2192	2.859	0.00425 **

Damit ist die geschätzte OR $\exp(0.6266) = 1.8712$. Die OR ist grösser als im Modell beieinanderliegender Kategorien (1.45). Die Grund dafür liegt darin, dass im neuen Modell als Referenzgruppe jeweils alle schwereren Kategorien zusammen gewählt werden.

Das Modell der kontinuierlichen Verhältnisse ist geeignet, wenn man es inhaltlich sinnvoll interpretieren kann. Ein gebräuchliches Beispiel dafür ist die Anzahl der Male, die man braucht, um ein Examen zu bestehen. Allgemeiner geht es um die Anzahl der Male, bis man in wiederholten Verfahren mit jeweils binären Resultaten das ausgezeichnete Ergebnis erhält. Das erste Logit modelliert, dass man das erste Mal besteht, das zweite Logit, dass man das zweite Mal besteht, sofern man des erste Mal nicht bestanden hat, etc.

Das Modell der proportionalen Quoten ist das am häufigsten verwendete ordinale Modell . Es ist das in SPSS installierte Modell (Ordinale Regression, PLUM-Befehl). Die Log-Odds hängt im Modell nicht von der Antwortekategorie ab, d.h. $\boldsymbol{\beta}$ ist nicht nach Antwortekategorien differenziert. Das Modell erlaubt entsprechend eine allgemeine Diskussion der Richtung der Antworten und man muss sich nicht mit spezifischen Antwortekategorien beschäftigen. Die Resultate sind einfacher zu beschreiben als jene von Modellen ohne Nebenbedingungen. Die Haltbarkeit der Annahme, dass der Effekt über die Antwortekategorien konstant sind, muss - ebenso wie in anderen Modellen - getestet werden. Modifikationen des Modells, welche für eine oder mehrere Kovariablen kategoriespezifische Effekte erlauben, werden partiell proportionale Quoten-Modelle genannt (s. Ananth und Kleinbaum (1997)).

Eine Herleitungsart für das Modell der proportionalen Quoten besteht in der Kategorisierung einer stetigen Antwortevariable. Um die Diskussion zu führen, kodieren Sie die Variable BWT *aufsteigend*: mehr Geburtsgewicht erhält nun eine höhere Kodierung - mit den Intervallgrenzen 2500, 3000 und 3500. Sie betrachten diese Variable BWT4N in Abhängigkeit von LWT (Gewicht der Mutter). Man kann die Situation wie folgt darstellen (s. Abbildung 27)

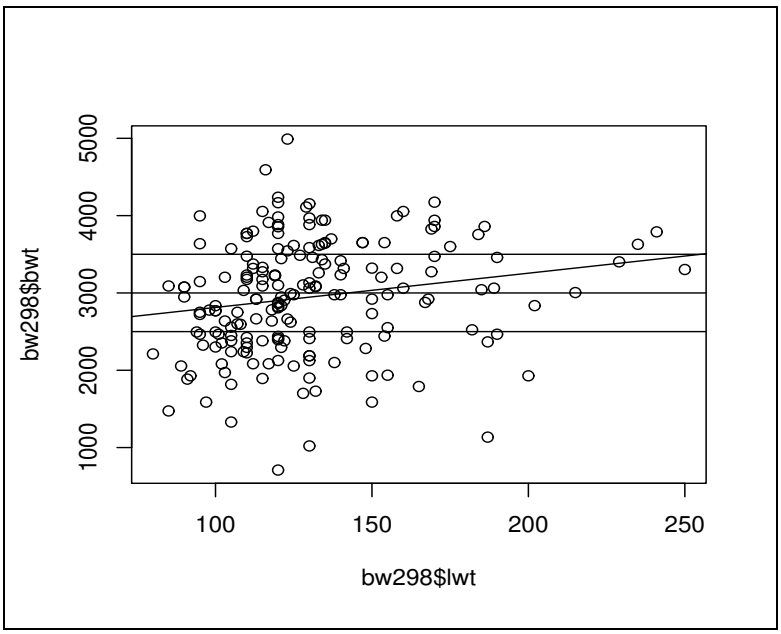

Abbildung 27: Muttergewicht * Säuglingsgewicht, Kleinste-Quadrate-Regression und Unterteilung in 4 Klassen

Man nimmt nun an, die Fehler-Zufallsvariablen wäre logistisch statt normal verteilt. Das Modell wäre also

$$BWT\,(LWT) = \lambda_0 + \lambda_1 LWT + \sigma\varepsilon,$$

wobei σ proportional zur Varianz ist und ε standard-logistisch verteilt ist mit der Verteilungsfunktion

$$P\,(\varepsilon \leq z) = \frac{e^z}{1+e^z}$$

Bei der linearen Regression modellieren wir die Mittel von BWT als Funktion von LWT, bei der ordinalen logistischen Regression modellieren wir die Wahrscheinlichkeit, das BWT in eines der vier durch die Klassengrenzen gegebenen Intervalle fällt. Es gilt (g_i = Klas-

sengrenze $i \in \mathbb{N}_3^*$)

$$\begin{aligned} P\left(BWT4N = 0 | LWT = 125\right) &= P\left(BWT \leq g_1 | LWT = 125\right) \\ &= P\left(\lambda_0 + \lambda_1 125 + \sigma\varepsilon \leq g_1\right) \\ &= P\left(\varepsilon \leq \frac{g_1 - \lambda_0 - \lambda_1 125}{\sigma}\right) \\ &= P\left(\varepsilon \leq \tau_1 - 125\beta\right) \end{aligned}$$

mit

$$\begin{aligned} \tau_1 &:= \frac{g_1 - \lambda_0}{\sigma} \\ \beta &:= \frac{\lambda_1}{\sigma} \end{aligned}$$

(Würde man $\beta := -\frac{\lambda_1}{\sigma}$ setzen, erfolgt das Modell, das von R ausspuckt wird, s.u.). Unter der Voraussetzung, dass die Fehler ε der standardlogistischen Verteilung folgen, folgt

$$P\left(\varepsilon \leq \tau_1 - 125\beta\right) = \frac{e^{\tau_1 - 125\beta}}{1 + e^{\tau_1 - 125\beta}}$$

Damit erhält man

$$\begin{aligned} P\left(\varepsilon > \tau_1 - 125\beta\right) &= 1 - P\left(\varepsilon \leq \tau_1 - 125\beta\right) \\ &= \frac{1}{1 + e^{\tau_1 - 125\beta}} \end{aligned}$$

Damit ist die Log-Odds bezüglich der Grenze 125

$$\begin{aligned} \ln \frac{P\left(bwt4n \leq 0 | LWT = 125\right)}{P\left(bwt4n > 0 | LWT = 125\right)} &= \ln \frac{P\left(\varepsilon \leq \tau_1 - 125\beta\right)}{P\left(\varepsilon > \tau_1 - 125\beta\right)} \\ &= \ln \frac{\frac{e^{\tau_1 - 125\beta}}{1 + e^{\tau_1 - 125\beta}}}{\frac{1}{1 + e^{\tau_1 - 125\beta}}} \\ &= \ln e^{\tau_1 - 125\beta} \\ &= \tau_1 - 125\beta \end{aligned}$$

Mit analogen Rechnungen kann man für BWT4N=1, zeigen, dass

$$\begin{aligned} \ln \frac{P\left(BWT4N \leq 1 | LWT = 125\right)}{P\left(BWT4N > 1 | LWT = 125\right)} &= \ln \frac{P\left(\varepsilon \leq \tau_2 - 125\beta\right)}{P\left(\varepsilon > \tau_2 - 125\beta\right)} \\ &= \tau_2 - 125\beta. \end{aligned}$$

sowie z.B. für BWT4N=1 und LWT=175

$$\begin{aligned} \ln \frac{P\left(BWT4N \leq 1 | LWT = 175\right)}{P\left(BWT4N > 1 | LWT = 175\right)} &= \ln \frac{P\left(\varepsilon \leq \tau_2 - 175\beta\right)}{P\left(\varepsilon > \tau_2 - 175\beta\right)} \\ &= \tau_2 - 175\beta. \end{aligned}$$

Allgemein gilt also

$$\ln \frac{P\left(BWT4N \leq k|LWT = x\right)}{P\left(BWT4N > k|LWT = x\right)} = \ln \frac{P\left(\varepsilon \leq \tau_{k+1} - x\beta\right)}{P\left(\varepsilon > \tau_{k+1} - x\beta\right)}$$
$$= \tau_{k+1} - x\beta.$$

Das Quotenverhältnis für x_1 versus x_0 ist

$$\ln \frac{P\left(BWT4N \leq k|LWT = x_1\right)}{P\left(BWT4N > k|LWT = x_1\right)} - \ln \frac{P\left(BWT4N \leq k|LWT = x_0\right)}{P\left(BWT4N > k|LWT = x_0\right)} \tag{18}$$
$$= \tau_{k+1} - x_1\beta - \left(\tau_{k+1} - x_0\beta\right) = -\beta\left(x_1 - x_0\right).$$

Wendet man das Modell auf die bwt-Daten an, erhält man mit dem VGAM-Paket (Familie cumulative, parallel=T, reverse=F) für die Zielvariable bwt4n und die erklärende Variable LWT mit („parallel = F“ und „reverse = F“ sind wiederum die Voreinstellungen)

```
bwt303=vglm(bwt4N~lwt, family=cumulative(link = "logitlink",
  parallel = T, reverse = F), data=bw298)
summary(bwt303)
```

das Ergebnis

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept):1	0.831577	0.585664	1.420	0.15564
(Intercept):2	1.706925	0.593670	2.875	0.00404 **
(Intercept):3	2.831087	0.617570	4.584	4.56e-06 ***
lwt	-0.012737	0.004447	-2.864	0.00418 **

(im Buch wird derselben Wert mit positivem Vorzeichen angegeben, in SPSS erhält man ebenfalls -0.013.). Wie man in der Abbildung 27 sieht, steigt die Kodierung von bwt in Anhängigkeit von der steigenden Kodierung von LWT. Das Vorzeichen des Koeffizienten im Buch (und z.B. in STATA) scheint dies zu reflektieren. Dies ist allerdings das Resultat des Vorzeichens in $c_k\left(\mathbf{x}\right) = \tau_k - \mathbf{x}^\intercal\boldsymbol{\beta}$. $c_k\left(\mathbf{x}\right)$ ist ja definiert durch $\ln \frac{P(Y \leq k|\mathbf{x})}{P(Y > k|\mathbf{x})}$. Man misst die tieferen Kategorien an den höheren. Ein Ansteigen (gemäss ansteigender Kodierung) den ordinalen Stufen entlang sollte entsprechend ein negatives Vorzeichen aufweisen. Dadurch dass man

$$c_k\left(\mathbf{x}\right) = \tau_k - \mathbf{x}^\intercal\boldsymbol{\beta}$$

definiert, wird dies umgekehrt. Ein Ansteigen der Logits mit der Kategorien-Kodierung wird bei $c_k\left(\mathbf{x}\right)$ also durch ein positives Vorzeichen angegeben.

Bemerkung 46. *In SPSS wird auch mittels* $\ln \frac{P(Y \leq k|\mathbf{x})}{P(Y > k|\mathbf{x})}$ *und* $c_k\left(\mathbf{x}\right) = \tau_k - \mathbf{x}^\intercal\boldsymbol{\beta}$ *gerechnet, das Vorzeichen wird aber erneut umgekehrt, da die höchste Kategorie als Referenz gewählt wird.*

Im VGAM-Paket erhält man positive Vorzeichen für ein Ansteigen in den Kategorien der Zielvariable, indem man das Modell

$$c_k^u(\mathbf{x}) = \ln \frac{P(Y > k|\mathbf{x})}{P(Y \leq k|\mathbf{x})} \tag{19}$$

berechnet (durch „reverse=T“). Eine Alternative bei der VGAM-Voreinstellung „reverse=F“ besteht darin, einfach die Vorzeichen der Koeffizienten umzukehren.

Um die OR zu berechnen, dass die Kinder mit leichteren Müttern ein leichteres Gewicht haben, verwendet man -0.012737. Man erhält für Zehnereinheiten in LWT:

$$\exp(-0.012737 \cdot 10) = 0.880\,41$$

Sinkt LWT um 10 Gewichtseinheiten, beträgt das Quotenverhältnis für ein leichteres Kind gegenüber einem schwereren Kind 0.88041 (d.h. leichtere Mütter haben leichtere Kinder). Umgekehrt ist bei einer Steigung von LWT um 10 das Quotenverhältnis für ein schwereres Kind gegenüber einem leichteren Kind

$$\exp(0.012737 \cdot 10) = 1.1358$$

Die Quote, dass eine schwerere Mutter ein schwereres Kind hat, ist um 1.1358 grösser als bei leichteren Müttern.

Um die Ergebnisse der verschiedenen Modelle zu vergleichen, wird nun die Methode der proportionalen Quoten auf bwt4N und smoke angewendet. Mit

```
bw203_po2=vglm(bwt4N~smoke, family=cumulative(link = "logitlink",
  parallel = T, reverse = F), data=bw298)
summary(bw203_po2)
```

erhält man das Ergebnis:

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept):1	-1.1163	0.1981	-5.636	1.74e-08 ***
(Intercept):2	-0.2477	0.1808	-1.370	0.17074
(Intercept):3	0.8667	0.1928	4.495	6.96e-06 ***
smoke	0.7608	0.2728	2.789	0.00528 **

samt

```
Residual deviance: 511.345 on 563 degrees of freedom
Log-likelihood: -255.6725 on 563 degrees of freedom
```
(20)

Die Ergebnisse stimmen mit denen im Buch (S. 304) bis auf das Vorzeichen von smoke überein. Das Quotenverhältnis einer rauchenden Frau, ein leichteres Kind zu haben, gegenüber einer nicht-rauchenden Frau beträgt $\exp(0.7608) = 2.\,14$.

Wie bei anderen Modellen mit Nebenbedingungen sollte man die Voraussetzungen des Modells überprüfen. Dies wird typischerweise mittels eines Vergleichs des Modells mit Nebenbedingungen mit dem multinomialen Modell ohne Nebenbedingungen mit Hilfe eines

LR-Tests gemacht. Das Problem mit diesem Test besteht darin, dass das Modell proportionaler Quoten nicht als multinomiales Modell unter Nebenbedingungen dargestellt werden kann. Entsprechend ist ein solcher Test statistisch nicht völlig korrekt. Er kann aber trotzdem einen Hinweis auf die Adäquatheit des Modell liefern. $df = (K+1)-2)p$. Fürs Beispiel erhält man fürs multinomiale Modell mit dem VGAM-Paket

```
smokelwtMN= vglm(bwt4N~smoke, data=bw298,
    family= multinomial(refLevel = 1));
summary(smokelwtMN)
```

das Ergebnis

```
Residual deviance: 510.9718 on 561 degrees of freedom
Log-likelihood: -255.4859 on 561 degrees of freedom
```

Mit (20) erhält man

$$511.345 - 510.9718 = 0.373\,2$$

mit zwei Freiheitsgraden. $1 - pchisq(0.3732, 2) = 0.8297756$. Man kann also nicht sagen, dass die beiden Modell verschieden sind. Vergleicht man die drei behandelten Hauptmodelle bezüglich des LWT-smoke-Beispiels ergibt sich

Modell	Variable	Koeffizient	Std. Fehler	z Wert	Pr(>|z|)
benachbarte Kategorien	smoke	0.3696	0.1332	2.774	0.00553 **
kontinuierlich mit NB	smoke	0.6266	0.2192	2.859	0.00425 **
proportionale Quoten	smoke	0.7608	0.2728	2.789	0.00528 **

Die Wahl des endgültigen Modells hängt davon ab, welche ORs für die untersuchten Fragen am informativsten sind und ob das Modell zu den Daten passt.

Beispiel 47. *Um bezüglich Interpretation die Sicherheit zu stärken, ein Beispiel mit klaren Tendenzen: Mit folgenden Befehlen werden künstliche Daten geschaffen mit klaren Effekten auf die Zielvariable („alter“ für „Alter in Jahren“. „Geschlecht“ (0 = männlich, 1 = weiblich), „zivilstand“ (1 = ledig, 2 = verheiratet und 3 = geschieden oder verwitwet), „zufri“ für eine metrische Konstruktionsvariable „Zufriedenheit“, die dazu dient, ordinal geordnete Zufriedenheitsdaten mit vier Ausprägungen zu kontruieren).*

```
alter=rnorm(600,35,10) #Altersvariable
alter=alter+10 #Altersvariable + 10
zivilstand=rbinom(600,2,0.5) #Zivilstandsvariable
Geschlecht=rbinom(600,1,0.5) #Geschlechtsvariable
zufrie=alter+2*zivilstand+3*Geschlecht+rnorm(600,0,3)
   #lineares Modell ansteigend in allen Variablen
daten=data.frame(Alter=alter,Zivilstand=zivilstand,
   Geschlecht=Geschlecht,Zufriedenheit=zufrie)
```

Dies ergibt mit

```
lm(Zufriedenheit~Alter+as.factor(Zivilstand)+
as.factor(Geschlecht),data=daten)
```

Coefficients:	
(Intercept)	*-0.323*
Alter	*1.012*
as.factor(Zivilstand)1	*1.989*
as.factor(Zivilstand)2	*3.803*
as.factor(Geschlecht)1	*2.684*

und mittels Klassifikation der Kontruktions-Zielvariable „zufrie“ in die ordinale Variable „zuklss:

```
daten$zuklss=as.numeric(cut(zufrie,quantile(zufrie),include.lowest=T))
```

(Daten abgespeichert unter „Hosmer_Lemeshow/beispiel_interpretation.txt“) erhält man mit

```
library(VGAM)
bei2=vglm(zuklss~as.factor(Geschlecht)+as.factor(Zivilstand)+
Alter,family=cumulative(link = "logitlink",parallel = T,
reverse = T), data=daten)
summary(bei2)
```

die Ergebnisse:

Coefficients:	*Estimate*	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept):1	*-23.30972*	*1.43544*	*-16.239*	*< 2e-16 ****
(Intercept):2	*-27.43383*	*1.64365*	*-16.691*	*< 2e-16 ****
(Intercept):3	*-31.48709*	*1.84360*	*-17.079*	*< 2e-16 ****
as.factor(Geschlecht)1	*1.12410*	*0.21814*	*5.153*	*2.56e-07 ****
as.factor(Zivilstand)1	*1.13441*	*0.26943*	*4.210*	*2.55e-05 ****
as.factor(Zivilstand)2	*1.90195*	*0.32217*	*5.904*	*3.56e-09 ****
Alter	*0.57476*	*0.03403*	*16.892*	*< 2e-16 ****

Die Quote von Zufriedenheitsstufe $> x$ zu Zufriedenheitstufe $\leq x$ für beliebige x (x für Ausprägungen der ordinalskalierten Zielvariable) ist bei Frauen $\exp(1.12410) = 3.077\,4$ *mal grösser als bei Männern. Die Quote von Zufriedenheitsstufe $> x$ zu Zufriedenheitstufe $\leq x$ ist bei Verheirateten, ist* $\exp(1.13441) = 3.109\,3$ *mal grösser als bei Ledigen, etc. Die Quote, von Zufriedenheitsstufe $> x$ zu Zufriedenheitstufe $\leq x$, ist bein einer ein Jahr älteren Person* $\exp(0.57476) = 1.776\,7$ *mal grösser als bei einer ein Jahr jüngeren Person. Positive Koffizienten zeigen auf, dass die Zufriedenheit bei den Kategorien mit höherer*

Kodierung ansteigt.
Mit SPSS erhält man mittels

```
PLUM zuklss BY Zivilstand Geschlecht WITH Alter
/CRITERIA=CIN(95) DELTA(0) LCONVERGE(0) MXITER(100)
      MXSTEP(5) PCONVERGE(1.0E-6) SINGULAR(1.0E-8)
/LOGIT
/PRINT=FIT PARAMETER SUMMARY.
```

die Ergebnisse:

							VI 95%	
		Schätzer	*Stdfehler*	*Wald*	*df*	*Sig.*	*Unten*	*Oben*
Schwelle	*[zuklss = 1]*	*20.284*	*1.278*	*251.822*	*1*	*.000*	*17.778*	*22.789*
	[zuklss = 2]	*24.408*	*1.481*	*271.480*	*1*	*.000*	*21.504*	*27.311*
	[zuklss = 3]	*28.461*	*1.682*	*286.446*	*1*	*.000*	*25.165*	*31.757*
Lage	*Alter*	*.575*	*.034*	*285.336*	*1*	*.000*	*.508*	*.641*
	[Zivilstand=0]	*-1.902*	*.322*	*34.853*	*1*	*.000*	*-2.533*	*-1.271*
	[Zivilstand=1]	*-.768*	*.257*	*8.888*	*1*	*.003*	*-1.272*	*-.263*
	[Zivilstand=2]	*0a*	*.*	*.*	*0*	*.*	*.*	*.*
	[Geschlecht=0]	*-1.124*	*.218*	*26.555*	*1*	*.000*	*-1.552*	*-.697*
	[Geschlecht=1]	*0a*	*.*	*.*	*0*	*.*	*.*	*.*

Es ergeben sich ausser bei den Ordinatenabschnitten (Schwelle) und bis auf die Wahl der Referenzen bei den erklärenden Variablen bis aufs Vorzeichen identische Werte für die Koeffizienten. Als Referenz wird in SPSS ja die letzte Kategorie gewählt. Dies sieht man auch daran, dass bei der entsprechenden Kategorie kein Koeffizient erscheint. Es scheint keine Möglichkeit zu geben, die Vorzeichen umzukehren ausser durch eine Rekodierung der Variablen - bei den p-Werten sieht man, dass es eine Rolle spielen kann, was als Referenz gewählt wird. Die Interpretationen der Variablen und den gegebenen Interpretationen der Ausprägungen sind also in beiden Modellen identisch (im SPSS-Modell ist z.B. fürs Geschlecht = 0: -1.124, d.h. die Quote, als Mann auf einer höheren Zufriedenheitsstufe zu sein, ist $\exp(-1.124) = 0.324\,98$ *mal die der Frauen, oder anders ausgedrückt:* $\frac{1}{\exp(-1.124)} = 3.077\,1$ *mal tiefer als bei Frauen (also bei den Frauen 3.0771 mal grösser als bei Männern).*

Modell-Bildungs-Strategien für Modelle ordinaler Regression Die Etappen für die Modellbildung für die ordinale Regression sind dieselben wie die bezüglich der binären logistischen Regression. Da entsprechende Methoden nicht in den gängigen Software-Paketen implementiert sind (Stand 2000), verwendet man am besten die Methoden für die binäre logistische Regression (M-1 mal, M = Anzahl Ausprägungen der ordinalen Zielvariable). Man startet mit einer stufenweisen Auswahl der Haupteffekte. Man untersucht die Linearität der Logits bezüglich der stetigen erklärenden Variablen und wendet, wenn nötig, die Methode der Fraktionalen Polynome an. Nichtlineare Transformationen sollten klinisch Sinn machen, genügend ähnlich über die Kategorien sein und eine wesentliche Verbesserung

der Anpassung des Modells an die Daten bewirken. Dann untersucht man, ob weggelassene Variablen im Modell mit den Haupteffekten signifikant oder störend (confounder) sind. Zuletzt untersucht man, ob Interaktionen vorliegen. Dann untersucht man die Konstanz der Koeffizienten (sind die Nebenbedingungen des ordinalen Modells haltbar oder nicht), indem man die Modelle mit und ohne Nebenbedingungen vergleicht. Oft wird dabei das Modell mit Nebenbedingungen mit dem Basislinienmodell verglichen. Weiterhin begutachtet man, ob es einflussreiche Datenmuster hat. Gibt es solche, kann man diese weglassen, das Modell neu anpassen und die Veränderungen abschätzen. Schliesslich wird man die ORs samt VIs angeben und klar festhalten, welches ordinale Modell man verwendet hat.

Sie starten bezüglich der BWT-Daten unmittelbar mit einem Modell, das alle signifikanten Effekte und inhaltlich sinnvollen Variablen enthält (da das Verfahren der Auswahl bereits mehrere Male vorgeführt wurde). Die im Buch gelieferten Koeffizienten sind mit dem VGAM-Paket nicht reproduzierbar. VGAM ergibt allerdings dieselben Resultate wie SPSS. Mit

```
m306=vglm(bwt4N~age+lwt+as.factor(race)+smoke+ht+ui+ptd, data=bw298,family=
      cumulative(link="logitlink",reverse=T,parallel=T))
summary(m306)
```

erhält man

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept):1	-0.547791	0.898351	-0.610	0.542012	
(Intercept):2	0.449534	0.897122	0.501	0.616311	
(Intercept):3	1.731939	0.906956	1.910	0.056183	.
age	-0.004881	0.027162	0.180	0.857381	
lwt	0.013309	0.005003	-2.661	0.007802	**
as.factor(race)2	-1.515224	0.446471	3.394	0.000689	***
as.factor(race)3	-0.904849	0.333168	2.716	0.006610	**
smoke	-1.049432	0.313143	3.351	0.000804	***
ht	-1.225975	0.589102	2.081	0.037426	*
ui	-0.976758	0.409860	2.383	0.017165	*
ptd	-0.293842	0.303719	0.967	0.333305	

Mit SPSS erhält man - wenn man race umgekehrt kodiert und auch das von unten nach oben kodierte bwt4N verwendet - mit

```
PLUM bwt4n BY race_recod WITH age lwt smoke ht ui ptd
/CRITERIA=CIN(95) DELTA(0) LCONVERGE(0)
MXITER(100) MXSTEP(5) PCONVERGE(1.0E-6) SINGULAR(1.0E-8)
/LINK=LOGIT
/PRINT=FIT PARAMETER SUMMARY.
```

das VGAM-Ergebnis

Parameterschätzer

		Schätzer	Standardfehler	Wald	Freiheitsgrade	Sig.	Konfidenzintervall 95% Untergrenze	Konfidenzintervall 95% Obergrenze
Schwelle	[bwt4n = .00]	-.548	.898	.372	1	.542	-2.309	1.213
	[bwt4n = 1.00]	.449	.897	.251	1	.616	-1.309	2.208
	[bwt4n = 2.00]	1.732	.907	3.646	1	.056	-.046	3.509
Lage	age	-.005	.027	.032	1	.857	-.058	.048
	lwt	.013	.005	7.077	1	.008	.004	.023
	smoke	-1.049	.313	11.231	1	.001	-1.663	-.436
	ht	-1.226	.589	4.331	1	.037	-2.381	-.071
	ui	-.977	.410	5.679	1	.017	-1.780	-.173
	ptd	-.294	.304	.936	1	.333	-.889	.301
	[race_recod=1.00]	-.905	.333	7.376	1	.007	-1.558	-.252
	[race_recod=2.00]	-1.515	.446	11.517	1	.001	-2.390	-.640
	[race_recod=3.00]	0[a]	.	.	0	.	.	.

Verknüpfungsfunktion: Logit.

a. Dieser Parameter wird auf Null gesetzt, weil er redundant ist.

Abbildung 28:

Das VGAM-Paket liefert unter „Exponentiated coefficients“ auch die ORs und mit exp(confint(m306)) deren VIs. Die Variable Alter berücksichtigen sie aus inhaltlichen Gründen. HT belassen sie ebenfalls aus klinischen Gründen im Modell, wobei ein LQ-Test einen p-Wert von 0.046 ergibt (bei ihnen beim Wald-Test auf dem 0.05-Niveau knapp nicht signifikant, in VGAM jedoch signifikant). PDT ist bei ihnen knapp signifikant, bei VGAM bei weitem nicht.

Sie verwendeten drei separate binäre Regressionen, um die Skalierung von Alter und LWT zu überprüfen. Die Ergebnisse zeigten, dass es vernünftig sei, die Linearität in den Logits zu postulieren. Die Quartilmethode auf LWT angewandt zeigte, dass die Koeffizienten der neuen Variable recht ähnlich sind (d.h. die ORs bezüglich der Referenzkategorie unterscheiden sich kaum). Dies legte es nahe, BWT in eine dichotome Variable zu rekodieren (Referenzkategorie gegen den Rest). Die Anpassung des neuen Modells ergab aber wesentliche Änderungen in UI (18%). Deshalb behalten Sie die ursprüngliche Variable LWT bei.

Schliesslich überprüfen Sie die Interaktionen. Es gibt nur eine signifikante Interaktion und zwar zwischen LWT und HT (p = 0.044 beim LQ-Test und 0.062 beim Wald-Test). Der geschätzte Koeffizient ergibt für den Haupteffekt von HT bei diesem Modell -5.648 mit einem geschätzten Standardfehler von 2.5353. Dies legt numerische Instabilität nahe, wohl dem Umstand zuzuschreiben, dass es nur 12 Datensätze mit HT=1 hat. Deshalb nehmen

sie diese Interaktion nicht ins Modell auf.

Als nächstes untersuchen Sie die Voraussetzung des Modells proportionaler Quoten m306. Dazu vergleichen Sie das obige Modell mit dem Basislinien-Modell. Der Vergleich ergibt mit

Proportionale Quoten:
Residual deviance: 474.4531 on 556 degrees of freedom Log-likelihood: -237.2265
Basislinien-Modell:
Residual deviance: 510.9718 on 561 degrees of freedom Log-likelihood: -255.4859

$$510.9718 - 474.4531 = 36.519$$

und mit 16 Freiheitsgraden den p-Wert 1-pchisq(36.519,16)=0.002449619 ($\mathrm{df} = ((K+1)-2)p = (4-2)\,8 = 16$). Sie erhalten den Testwert 21.7432 und den p-Wert 0.152.

Um die Anpassung und den Einfluss individueller Datenmuster zu überprüfen führten sie drei verschiedene volle Untersuchungen auf der Grundlager der binären logistischen Regression bezüglich BWT4N=k versus BWT4N=0. Es wurden die Datensätze 98, 132, 133, 138 und 188 mit grossem $\Delta\boldsymbol{\beta} > \mathbf{1}$ in einer der drei Regressionen entdeckt. Überprüfung der Werte dieser Datensätze ergaben keine ungewöhnlichen Werte. Zwei dieser Datensätze entsprechen aber Individuen, die zu den 12 mit HT=1 gehören. Wenn man die 5 Datensätze weglässt, ergeben sich Änderungen von mehr als 20% in dreien der acht Koeffizienten. Allerdings führt ihr Wegfallen zu zwei leeren Zellen (BWT4N versus HT). Deshalb kann man bezüglich dieses Datensatzes das Basislinienmodell nicht anpassen und man kann die Konstanz der Koeffizienten des Modells proportionaler Quoten nicht überprüfen. Deshalb entschliessen sie sich, die Daten drin zu lassen.

Die Resultate der Studie zeigen, dass nach der Kontrolle des Alters und des Gewichtes der Mutter, die übrigen Variablen die Quotenverhältnisse von leichteren gegenüber schwereren Kindern erhöhen. Sie finden für die vorliegenden Daten das Modell proportionaler Quoten aus inhaltlichen Gründen am angemessensten.

Literatur

Agresti, A. (2013). *Categorical Data Analysis* (3. Aufl.). New York: John Wiley.

Albert, A. & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, *71* (1), 1-10.

Ananth, C. V. & Kleinbaum, D. G. (1997). Regression models for ordinal responses: a review of methods and applications. *International Journal of Epidemiology*, *26* (6), 1323-1333.

Begg, C. B. & Gray, R. (1984). Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika*, *71* (1), 11-18.

Breslow, N. (1996). Statistics in Epidemiology: The Case-Controle Study. *Journal of the American Statistical Association*, *91* (433), 14-28.

Breslow, N. E. & Day, N. E. (2000). *Statistical Methods in Cancer Research - The analysis of case-control studies* (9. Aufl.; W. DAVIS, Hrsg.). Lyon: International Agency for Research on Cancer.

Cleveland, W. (1993). *Visualising Data*. Summit: Hobart Press.

Cook, E. D. (2001). *Solutions Manual to Accompany "Applied Logistic Regression", Second Edition, by David W. Hosmer and Stanley Lemeshow*. New York: Wiley.

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics*, *19* (1), 15-18.

Cook, R. D. (1979). Influential Observations in Linear Regression. *Journal of the American Statistical Association*, *74* (365), 169-174.

Cox, D. (1970). *The Analysis of Binary Data*. London: Methuen.

Cox, D. & Snell, E. (1989). *Analysis of Binary Data, Second Edition*. Abingdon: Taylor & Francis.

Dardis, C. (2015). Logisticdx: Diagnostic tests for models with a binomial response [Software-Handbuch]. Zugriff auf https://CRAN.R-project.org/package=LogisticDx (R package version 0.2)

Dümbgen, L. (2009). *Lineare Modelle und Regression I-II*. Bern: Vorlesungsskript Universität Bern.

Harrell, F. E., Lee, K. L. & Mark, D. B. (1996). Tutorial in Biostatistics Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medecine*, *15*, 361-387.

Hoaglin, D. C. & Welsch, R. E. (1978). The Hat Matrix in Regression and ANOVA. *The American Statistician*, *32* (1), 17–22.

Hosmer, D. W., Hosmer, T., Le Cessie, S. & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, *16* (9), 965–980.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression*. New York: Wiley.

Huber, P. J. (1981). *Robust Statistics*. New York: John Wiley.

Igo, R. P. (2010). Influential Data Points. In Salkind (Hrsg.), *Encyclopedia of Research Design* (Bd. 2, S. 600–602). Los Angeles: Sage.

Jennings, D. E. (1986). Outliers and Residual Distributions in Logistic Regression. *Journal of the American Statistical Association*, *81* (396), 987-990.

Lawless, J. & Singhal, K. (1987a). ISMOD: An all-subsets regression program for generalized linear models II. Statistical and computational background. *Computer Methods and Programs in Biomedicine*, *24* (2), 125-134.

Lawless, J. & Singhal, K. (1987b). ISMOD: An all-subsets regression program for generalized linear models I. Statistical and computational background. *Computer Methods and Programs in Biomedicine*, *24* (2), 117-124.

Mantel, N. & Haenszel, W. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *Journal of the National Cancer Institute*, *22*, 719-748.

McLeod, A. & Xu, C. (2014). bestglm: Best Subset GLM [Software-Handbuch]. Zugriff auf https://CRAN.R-project.org/package=bestglm (R package version 0.34)

Mittlböck, M. & Schemper, M. (1996). Explained Variation for Logistic Regression. *Statistics in Medicine*, *15* (19), 1987–1997.

Nagelkerke, N. J. D. (1991). A Note on a General Definition of the Coefficient of Determination. *Biometrika*, *78* (3), 691-692.

Osius, G. & Rojek, D. (1992). Normal Goodness-of-Fit Tests for Multinomial Models with Large Degrees of Freedom. *Journal of the American Statistical Association*, *87* (420), 1145-1152.

Pregibon, D. (1981). Logistic Regression Diagnostics. *Annals of Statistics*, *9* (4), 705-724.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications* (2. Aufl.). New York: Wiley.

Sauerbrei, W., Meier-Hirmer, C., Benner, A. & Roystond, P. (2006). Multivariable regression model building by using fractional polynomials: Description of SAS, STATA and R programs. *Computational Statistics and Data Analysis*, *50*, 3464 - 3485.

Thompson, L. A. (2009). R (and S-PLUS) Manual to Accompany Agresti's Categorical Data Analysis - 2002 [Software-Handbuch]. Zugriff auf http://www.stat.ufl.edu/~aa/cda/Thompson_manual.pdf

Tsiatis, A. A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika*, *67* (1), 250-251.

Velleman, P. F. & Welsch, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician*, *35* (4), 234–242.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4. Aufl.). New York: Springer.

Venables, W. N. & Ripley, B. D. (2016). nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models [Software-Handbuch]. Zugriff auf https://CRAN.R-project.org/package=nnet (R package version 7.3-12)

Wald, A. (1943). Tests of Statistical Hypotheses concerning several Parameters when the number of Observations is large. *Transactions of the American Mathematical Society*, *38*, 426-482.

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, *50* (1), 1-25.

Wilks, S. S. (1938). The Large-sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Annals of the Institute of Statistical Mathematics*, *9*, 60-62.