PROC LOGISTIC: Warum sind die Koeffizienten nicht mit den Odds Ratios konsistent?

Ulrike Braisch, Rainer Muche
Institut für Epidemiologie und Medizinische Biometrie
Universität Ulm
Schwabstraße 13
89075 Ulm
ulrike.braisch@uni-ulm.de

Zusammenfassung

In vielen Analysen wird bei multivariablen Regressionsmodellen häufig mit kategoriellen Prädiktoren gearbeitet. Das Ziel dieses Beitrages ist es, hierbei auf eine potenzielle Falle bei Anwendung einer logistischen Regression mit SAS (PROC LOGISTIC) hinzuweisen.
Ein wichtiges Hilfsmittel zur Interpretation eines logistischen Regressionsmodells ist das Odds Ratio (OR), das das Chancenverhältnis zweier Merkmalsausprägungen im Hinblick auf eine Zielvariable Y angibt. Beschreibt Y beispielsweise das Auftreten (y=1) oder Nicht-Auftreten (y=0) von Lungenkrebs und die Kovariate X, ob die Person ein Raucher (x=1) oder Nichtraucher (x=0) ist, dann deutet ein OR=2 darauf hin, dass Lungenkrebs unter Rauchern zweimal häufiger auftritt als unter Nichtrauchern in der Studienpopulation. Das OR wird typischerweise laut Formel in den einschlägigen Lehrbüchern durch Exponieren des entsprechenden Regressionskoeffizienten β bestimmt (OR=e^{β}). Diese Rechnungsweise ist in SAS allerdings nicht ohne Weiteres gültig, wenn man mit kategoriellen Prädiktoren arbeitet und das CLASS Statement einsetzt. Denn SAS hat bei der logistischen Regression die Effektkodierung als Dummy Kodierung voreingestellt, was dazu führt, dass die von SAS ausgegebenen ORs nicht mit denen übereinstimmen, die man (fälschlicherweise) mit obiger Formel berechnen würde. Die Gültigkeit obiger Formel kann in PROC LOGISTIC in SAS allerdings wieder gewährleistet werden, indem *param=ref* als Option des CLASS Statements hinzugefügt wird. Dadurch wird PROC LOGISTIC gezwungen die Dummy Kodierung gemäß der (üblichen) Referenzkodierung durchzuführen.
Im Rahmen des Beitrages wird an Hand eines Beispiels auf die zwei am häufigsten verwendeten Varianten der Dummy Kodierung eingegangen, die Effektkodierung (*effect coding)* und die Referenzkodierung (*reference coding*), und deren Einfluss auf die Odds Ratio Schätzung.
Bei der Schätzung des Hazard Ratio über die Cox Regression (PROC PHREG) hingegen ist als Dummy Kodierung die Referenzkodierung voreingestellt, so dass hier die oben beschriebene Falle gar nicht vorhanden ist, außer man ändert per Hand die Art der Dummy Kodierung. In dieser Hinsicht sollte also immer Vorsicht geboten sein, denn die Voreinstellung der Dummy Kodierung im CLASS Statement unterscheidet sich in SAS zwischen den unterschiedlichen Regressionsprozeduren.

Schlüsselwörter: Odds Ratio, kategorielle Kovariate, Dummy Kodierung, LOGISTIC-Prozedur, PHREG-, GENMOD-, REG-, GLM-, GLMSELECT-Prozedur

1 Was ist das Problem?

In vielen Analysen wird bei multivariablen Regressionsmodellen häufig mit kategoriellen Prädiktoren gearbeitet. Wird dabei in SAS als Analysemethodik die logistische Regression (PROC LOGISTIC) für eine dichotome Zielgröße gewählt, so könnte man unter Einbezug kategorieller Kovariaten leicht in eine potenzielle Falle treten.
Bindet man eine kategorielle Kovariate X mit dem CLASS Statement ohne zusätzliche Optionen in das logistische Regressionsmodell ein, so sieht die SAS-Syntax folgendermaßen aus [7]:

```
PROC LOGISTIC data=daten;
  CLASS X;
  MODEL Y = X;
RUN;
```

Der Output dieser Prozedur enthält unter anderem die Schätzung der Regressionskoeffizienten und die Odds Ratios. Laut „Lehrbuchwissen“ gilt nun Folgendes (siehe z.B. [4]):

$$OR = e^{\beta}$$

Rechnet man dies allerdings mit den im Output ausgegebenen Regressionskoeffizienten nach, so stellt man fest, dass

$$OR \neq e^{\beta}.$$

Demnach stellt sich die Frage: Was passiert hier im CLASS Statement? Um die Beantwortung dieser Frage geht es in diesem Beitrag. Hierzu wird zunächst im folgenden Abschnitt eine kurze Einführung in die logistische Regression gegeben.

2 Einführung in die logistische Regression

Die logistische Regression wird herangezogen, wenn die Zielvariable Y dichotom ist, d.h. die Ausprägung 1 für ein Ereignis und die 0 für ein Nicht-Ereignis steht. Modelliert wird in der logistischen Regression die bedingte Wahrscheinlichkeit für das Eintreten des Ereignisses gegeben die Ausprägungen von r Kovariaten:

$$P(y = 1|x_1, x_2, \cdots, x_r) =: \pi(x_1, x_2, \cdots, x_r)$$

Das allgemeine Regressionsmodell sieht dann folgendermaßen aus:

$$\pi(x_1, x_2, \cdots, x_r) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r$$

Das Problem dieses Modells ist allerdings, dass die linke Seite Werte im Bereich [0,1] annimmt, die rechte Seite aber Werte aus dem gesamten Bereich $(-\infty, \infty)$ annehmen kann. Dieses Problem wird meist durch eine Logit-Transformation gelöst, d.h.

$$logit\ p = \ln\left(\frac{p}{1-p}\right).$$

Logistisches Regressionsmodell

Nach der Logit-Transformation erhält man folgendes logistische Regressionsmodell:

$$logit\ \pi(x_1, x_2, \cdots, x_r) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r =: t$$

Daraus kann dann die bedingte Wahrscheinlichkeit für das Auftreten eines Ereignisses folgendermaßen berechnet werden:

$$\pi(x_1, x_2, \cdots, x_r) = logit^{-1}(t) = \frac{1}{1 + e^{-t}}$$

Interpretation eines Regressionskoeffizienten

Ein Regressionskoeffizient beschreibt demnach bei der logistischen Regression die Änderung der Wahrscheinlichkeit $P(y = 1|x_1, x_2, \cdots, x_r)$ **auf der Logit Skala** bei Anstieg des Wertes einer Kovariaten um eine Einheit.

Das Odds Ratio

Ein wichtiges Hilfsmittel zur Interpretation eines logistischen Regressionsmodells stellt das Odds Ratio (OR) dar, das das Chancenverhältnis zweier Merkmalsausprägungen im Hinblick auf eine Zielvariable angibt. Typischerweise findet hierbei ein Vergleich des Erkrankungsrisikos von einer Gruppe von Studienteilnehmern mit einem bestimmten (Risiko)faktor (wie z.B. das Rauchen) zu einer Gruppe ohne diesen Faktor statt. Ein OR von ungefähr 1 würde dabei angeben, dass der Faktor keinen Einfluss auf das Erkrankungsrisiko hat. Ein OR<1 deutet auf eine Schutzwirkung dieses Faktors hin und ein OR>1 auf einen Risikofaktor.
Die allgemeine Definition des Odds Ratio lautet:

$$OR_{a\ vs.b} = \frac{\frac{P(y=1|x_1=a,x_2,\cdots,x_r)}{P(y=0|x_1=a,x_2,\cdots,x_r)}}{\frac{P(y=1|x_1=b,x_2,\cdots,x_r)}{P(y=0|x_1=b,x_2,\cdots,x_r)}} = \frac{\frac{\pi(x_1=a,x_2,\cdots,x_r)}{1-\pi(x_1=a,x_2,\cdots,x_r)}}{\frac{\pi(x_1=b,x_2,\cdots,x_r)}{1-\pi(x_1=b,x_2,\cdots,x_r)}},$$

wobei ein Vergleich nicht nur in x_1 sondern innerhalb jeder Kovariate stattfinden kann.

Nun besteht eine enge Beziehung zwischen Odds Ratios und logistischen Regressionsmodellen, denn wird das Odds Ratio logarithmiert (unter Anwendung der mathematischen Regel $\ln a/b = \ln a - \ln b$), so erhält man:

$$\ln OR_{a\ vs.b} = \ln \frac{\pi(x_1 = a, x_2, \cdots, x_r)}{1 - \pi(x_1 = a, x_2, \cdots, x_r)} - \ln \frac{\pi(x_1 = b, x_2, \cdots, x_r)}{1 - \pi(x_1 = b, x_2, \cdots, x_r)}$$

$$= logit\ \pi(x_1 = a, x_2, \cdots, x_r) - logit\ \pi(x_1 = b, x_2, \cdots, x_r)$$

Wird beispielsweise innerhalb einer dichotomen Kovariate x_1 die Gruppe der Raucher (mit $x_1 = 1$) mit der Gruppe der Nichtraucher (mit $x_1 = 0$) in Bezug auf das Lungenkrebsrisiko verglichen, so würde man das Odds Ratio folgendermaßen bestimmen:

$$OR_{\frac{Raucher}{Nichtraucher}} = \frac{\frac{\pi(1, x_2, \cdots, x_r)}{1 - \pi(1, x_2, \cdots, x_r)}}{\frac{\pi(0, x_2, \cdots, x_r)}{1 - \pi(0, x_2, \cdots, x_r)}}$$

Für nähere Informationen zur logistischen Regression und zu Odds Ratios wird auf [1] und [4] verwiesen.
In den folgenden beiden Abschnitten wird darauf eingegangen, wie das OR im Falle stetiger/dichotomer Kovariaten und im Falle kategorieller (nicht dichotomer) Kovariaten berechnet werden kann.

3 Odds Ratios bei stetigen / dichotomen Kovariaten

In diesem Abschnitt wird folgendes logistisches Modell mit einer zunächst stetigen Kovariate X betrachtet ([4], [6]):

$$logit\ P(y = 1|x) = \beta_0 + \beta_1 x$$

Die allgemeine Formel für das Odds Ratio bei Erhöhung von x um eine Einheit lautet demnach:

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

Hieraus lässt sich folgende Beziehung ableiten:

$$\ln OR = logit\ \pi(x+1) - logit\ \pi(x) = \beta_1\ ,$$

weshalb das OR auch kurz als

$$OR = e^{\beta_1}$$

berechnet werden kann, was im Folgenden als die „Lehrbuch-Formel“ bezeichnet wird.

Dichotome Kovariaten können problemlos wie stetige Kovariaten behandelt werden, sofern zwischen den beiden Merkmalsausprägungen ein Sprung von exakt einer Einheit stattfindet (z.B. 0/1) und die gewünschte Referenzkategorie die kleinere der beiden Ausprägungen ist.

4 Odds Ratios bei kategoriellen Kovariaten

In diesem Abschnitt wird nun die Berechnung des OR bei kategoriellen Kovariaten mit K>2 Ausprägungen vorgestellt (d.h. dichotome Kovariaten bleiben außer Betracht; siehe Abschnitt 3). Beispielhaft wird dazu im Folgenden eine Kovariate X mit den Ausprägungen

1=Nichtraucher, 2=Exraucher, 3=Raucher

und eine dichotome Zielvariable Y mit den Ausprägungen

1=Diagnose Lungenkrebs, 0=keine Diagnose Lungenkrebs

betrachtet. Zunächst wird die Kovariate X als „stetige" Kovariate behandelt, d.h. X wird mit den Ausprägungen 1/2/3 „einfach" in das Modell eingesetzt. Das logistische Regressionsmodell sieht dann wie folgt aus:

$$logit\ P(y = 1|x) = \beta_0 + \beta_1 x$$

Die Odds Ratios würde man dann wie folgt bestimmen:

$$OR_{\frac{Exraucher}{Nichtraucher}} = OR_{\frac{Raucher}{Exraucher}} = e^{\beta_1}$$

Demnach würde man für die Risikoveränderung benachbarter Ausprägungen ein gemeinsames durchschnittliches Odds Ratio bestimmen, falls zwischen benachbarten Ausprägungen jeweils eine Einheit liegt. Das Problem hierbei ist jedoch zum einen, dass die Risikoveränderungen zwischen den verschiedenen Ausprägungen in den seltensten Fällen gleich sind, und zum anderen das Interesse meist gerade in den individuellen Risikoveränderungen liegt. Des Weiteren ist die Kodierung der Merkmalsausprägungen im Falle nominaler Kovariaten nicht eindeutig, was bei verschiedenen Kodierungsvarianten zu unterschiedlichen Ergebnissen führen würde. Aus diesen Gründen ist der Einsatz von kategoriellen Kovariaten als „stetige" Variablen inhaltlich selten sinnvoll. Die Alternative zur Behandlung einer kategoriellen Kovariate als „stetige" Variable ist der Einsatz der sog. Dummy-Kodierung, deren Prinzip im Folgenden näher erläutert wird.

4.1 Einführung in die Dummy-Kodierung

Bei der Dummy-Kodierung wird zunächst eine der K Kategorien einer kategoriellen Kovariate X als Referenzkategorie gewählt. Im Anschluss wird X typischerweise in K-1 Dummy-Variablen zerlegt, wobei es verschiedene Arten der Dummy-Kodierung gibt ([3], [4], [6], [7]).

Um das Prinzip der Dummy-Kodierung anschaulich zu erläutern, wird obiges Beispiel zum Rauchverhalten erneut aufgegriffen, wobei im Folgenden die „klassische" Art der Dummy-Kodierung dargestellt wird.

Zunächst wird beispielsweise die Gruppe der Nichtraucher als Referenzkategorie gewählt, so dass man folgende Dummy-Variablen erhält:

$$x_{(1)} = \begin{cases} 1, Exraucher \\ 0, sonst \end{cases} \qquad x_{(2)} = \begin{cases} 1, Raucher \\ 0, sonst \end{cases}$$

Für Nichtraucher gilt: $x_{(1)} = 0$ und $x_{(2)} = 0$

Das logistische Regressionsmodell sieht dann folgendermaßen aus:

$$logit\, P\left(y = 1 \middle| x_{(1)}, x_{(2)}\right) = \beta_0 + \beta_{(1)} x_{(1)} + \beta_{(2)} x_{(2)} \,,$$

womit sich die jeweiligen Odds Ratios wie folgt berechnen lassen:

$$OR_{\frac{Exraucher}{Nichtraucher}} = e^{\beta_{(1)}}, OR_{\frac{Raucher}{Nichtraucher}} = e^{\beta_{(2)}}, OR_{\frac{Raucher}{Exraucher}} = e^{\beta_{(2)} - \beta_{(1)}}$$

Demnach können mit Hilfe der Dummy-Variablen die individuellen Odds Ratios für alle möglichen Kombinationen von Ausprägungen berechnet werden und nicht nur ein gemeinsames mittleres Odds Ratio für alle Risikoveränderungen benachbarter Ausprägungen so wie bei stetigen Kovariaten.
Die Konsequenz des Einsatzes von Dummy-Variablen liegt allerdings darin, dass sich die Anzahl der Regressionskoeffizienten pro Kovariate von einem auf K-1 erhöht, wodurch die Modellstabilität abnimmt.

Übliche Arten der Dummy-Kodierung

Es gibt viele Arten der Dummy-Kodierung, die üblichen Arten sind jedoch die Referenzkodierung und die Effektkodierung ([3], [6]), die im Folgenden in der allgemeinen Form kurz vorgestellt werden.

Die Referenzkodierung

Mit Referenzkategorie K ergeben sich wie oben beschrieben folgende binäre Dummy-Variablen [3]:

$$x_{(k)} = \begin{cases} 1\ falls\ Kategorie\ k\ mit\ k = 1, \ldots, K-1\ vorliegt \\ 0\ sonst \end{cases}$$

Für die Referenzkategorie K erhält man: $x_{(1)} = \cdots = x_{(K-1)} = 0$

Die Effektkodierung

Mit Referenzkategorie K ergeben sich folgende Dummy-Variablen [3]:

$$x_{(k)} = \begin{cases} 1\ falls\ Kategorie\ k\ mit\ k = 1, \ldots, K-1\ vorliegt \\ -1\ falls\ Referenzkategorie\ K\ vorliegt \\ 0\ sonst \end{cases}$$

Für die Referenzkategorie K erhält man: $x_{(1)} = \cdots = x_{(K-1)} = -1$

Logistische Regression mit Dummy-Variablen

Das logistische Regressionsmodell mit einer kategoriellen Kovariate X mit K Ausprägungen sieht folgendermaßen aus [1]:

$$logit\,\pi(x) = \beta_0 + \beta_{(1)}x_{(1)} + \beta_{(2)}x_{(2)} + \cdots + \beta_{(K-1)}x_{(K-1)}$$

Interpretation der Regressionskoeffizienten

Die Interpretation der einzelnen Regressionskoeffizienten hängt von der jeweiligen Dummy-Kodierung ab und wird im Folgenden für die beiden üblichen Arten, der Referenzkodierung und der Effektkodierung, angegeben.

Referenzkodierung [3]:

β_0: Mittelwert der Referenzgruppe
$\beta_{(j)}(j \geq 1)$: Differenz zwischen Mittelwert der jeweiligen Gruppe und dem Mittelwert der Referenzgruppe

Effektkodierung ([2], [3], [6]):

β_0: Gesamtmittelwert aller Kategorien
$\beta_{(j)}(j \geq 1)$: Differenz zwischen Mittelwert der jeweiligen Gruppe und dem Gesamtmittelwert

Die geeignete Dummy-Kodierung

Die Referenzkodierung ist die geeignete Dummy-Kodierung, wenn es um den Vergleich einer oder mehrerer Gruppen zu einer Referenzgruppe geht, z.B. in der Medizin zum Vergleich verschiedener neuer Therapien mit einer etablierten Standardtherapie [3].
Die Effektkodierung hingegen ist die geeignete Dummy-Kodierung, wenn die Differenz der einzelnen Mittelwerte zum Gesamtmittelwert von Interesse ist, z.B. zur Analyse des Unterschieds verschiedener OP-Methoden im Hinblick auf die Kosten (analog zur ANOVA [2]).

4.2 Anwendungsbeispiel

Um die Problematik kategorieller Kovariaten im Zusammenhang mit dem CLASS Statement der Prozedur LOGISTIC an einem Beispiel zu veranschaulichen, wird im Folgenden mit Daten des Surveillance, Epidemiology and End Results (SEER) Program vom National Cancer Institute der USA gearbeitet [8]. Hierbei handelt es sich um Daten von Krebspatienten, bei denen Lungenkrebs als Ersttumor innerhalb des Beobachtungszeitraumes 1973 und 2008 diagnostiziert wurde. Die Zielgröße stellt das Ereignis Fol-

getumor (ja/nein) dar und die kategorielle Kovariate soll der Familienstand bei Diagnose (nominal) mit folgenden Ausprägungen sein:

1=single, 2=verheiratet, 3=getrennt, 4=geschieden, 5=verwitwet .

Im Folgenden wird mit Hilfe dieser Daten der Output von PROC LOGISTIC unter Anwendung des CLASS Statements analysiert, und zwar sowohl unter Einsatz der Referenzkodierung als auch der Effektkodierung, wobei mit der Voreinstellung Effektkodierung begonnen wird.

4.2.1 Effektkodierung – Default bei PROC LOGISTIC

Wird nun mit der Variable Folgetumor (FT) als Zielvariable und Familienstand bei Diagnose (FamStand) als Kovariate in SAS eine logistische Regression unter Anwendung des CLASS Statements aber ohne Angabe von zusätzlichen Optionen durchgeführt, d.h. mit den Voreinstellungen der Prozedur LOGISTIC, so ergibt sich folgende SAS-Syntax:

```
PROC LOGISTIC data=lungenkrebs;
  CLASS FamStand;
  MODEL FT=FamStand;
RUN;
```

Die im zugehörigen Output geschätzten Regressionskoeffizienten sind in Tabelle 1 aufgeführt.

Tabelle 1: Analysis of Maximum Likelihood Estimates
(unter Default-Dummy-Kodierung von PROC LOGISTIC)

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	2.8686	0.0151	36239.4115	<.0001
FamStand 1	1	***0.0792***	0.0265	8.9363	0.0028
FamStand 2	1	-0.1810	0.0168	116.4010	<.0001
FamStand 3	1	-0.0193	0.0498	0.1498	0.6988
FamStand 4	1	0.0377	0.0251	2.2580	0.1329

Die Referenzkategorie ist im CLASS Statement in PROC LOGISTIC per Voreinstellung die höchstkodierte Ausprägung [7], was sich im Output (Tabelle 1) widerspiegelt. Da im Output der Parameter „FamStand 5“ fehlt, werden hier die verwitweten Patienten (Kategorie 5) als die Referenzkategorie angenommen.
Möchte man nun das Odds Ratio z.B. für den Vergleich von Singles (Katgorie 1) zu verwitweten Lungenkrebspatienten (Kategorie 5) laut „Lehrbuch-Formel“ bestimmen, so würde sich folgender Rechenweg ergeben:

$$OR_{1\ vs\ 5} = OR_{\frac{single}{verwitwet}} = \exp(0.0792) = \mathbf{1.0824}$$

Das bedeutet, Singles hätten im Vergleich zu verwitweten Patienten ein etwas höheres Risiko an einem Folgetumor nach einem bereits diagnostizierten Lungenkrebs zu erkranken.
Schaut man sich nun aber die Ergebnisse des Odds Ratio im Output der Prozedur LOGISTIC an, so erhält man Tabelle 2.

Tabelle 2: Odds Ratio Estimates
(unter Default-Dummy-Kodierung von PROC LOGISTIC)

Effect	**Point Estimate**	**95% Wald Confidence Limits**	
FamStand 1 vs 5	***0.996***	0.931	1.065
FamStand 2 vs 5	0.768	0.736	0.801
FamStand 3 vs 5	0.902	0.796	1.023
FamStand 4 vs 5	0.955	0.897	1.018

Das laut „Lehrbuch-Formel“ berechnete Odds Ratio und das Odds Ratio im Output der Prozedur LOGISTIC stimmen demnach nicht überein und kommen sogar auf unterschiedliche Tendenzen. Denn auf Grund von Tabelle 2 weisen Singles im Vergleich zu verwitweten Patienten die Tendenz eines etwas niedrigeren Risikos auf, nach einer Lungenkrebsdiagnose an einem Folgetumor zu erkranken.
Nun stellt sich die Frage, was innerhalb des CLASS Statements passiert. Um diese Frage zu beantworten, sollte man sich die Information über die Dummy-Kodierung, die ebenfalls standardmäßig im Output der Prozedur LOGISTIC ausgegeben wird, genauer anschauen (siehe Tabelle 3).

Tabelle 3: Class Level Information
(unter Default-Dummy-Kodierung von PROC LOGISTIC)

Class	**Value**	**Design Variables**			
FamStand	1	1	0	0	0
	2	0	1	0	0
	3	0	0	1	0
	4	0	0	0	1
	5	**-1**	**-1**	**-1**	**-1**

Aus Tabelle 3 ist ersichtlich, dass PROC LOGISTIC als Voreinstellung die Effektkodierung mit der letzten Kategorie als Referenzkategorie verwendet. Würde man nun das Odds Ratio per Hand bestimmen, so würde man folgendermaßen vorgehen.
Das logistische Regressionsmodell für das Anwendungsbeispiel sieht wie folgt aus:

$$logit\ \pi(x) = \beta_0 + \beta_{(1)}x_{(1)} + \beta_{(2)}x_{(2)} + \beta_{(3)}x_{(3)} + \beta_{(4)}x_{(4)}$$

Möchte man nun das Odds Ratio für den Vergleich Kategorie 1 zu Kategorie 5 berechnen, so würde man zunächst den Logarithmus des Odds Ratios bestimmen:

$$\ln OR_{1\ vs\ 5} = logit\ \pi(1) - logit\ \pi(5)$$

$$= [\beta_{(0)} + \beta_{(1)}] - [\beta_{(0)} - \beta_{(1)} - \beta_{(2)} - \beta_{(3)} - \beta_{(4)}]$$

$$= 2 * \beta_{(1)} + \beta_{(2)} + \beta_{(3)} + \beta_{(4)}$$

$$= 2 * 0.0792 + (-0.1810) + (-0.0193) + 0.0377$$

$$= \mathbf{-0.0042}$$

Damit erhält man für das Odds Ratio:

$$OR_{1\ vs\ 5} = OR_{\frac{single}{verwitwet}} = exp(\mathbf{-0.0042}) = \mathbf{0.9958}\ ,$$

was dem Odds Ratio in Tabelle 2 entspricht.
Wird also die Effektkodierung als Dummy-Kodierung verwendet, kann das Odds Ratio nicht nach der „Lehrbuch-Formel“ berechnet werden ([4], [6]), sondern muss entweder per Hand berechnet oder das Odds Ratio aus dem Output von PROC LOGISTIC angegeben werden.

4.2.2 Referenzkodierung

Möchte man nun aber im CLASS Statement die Referenzkodierung erzwingen und eine andere als die von SAS festgelegte Referenzkategorie wählen, so muss bei der Prozedur LOGISTIC im CLASS Statement zusätzlich eine Option wie folgt gesetzt werden:

```
PROC LOGISTIC data=lungenkrebs;
  CLASS FamStand (param=ref ref='2');
  MODEL FT=FamStand;
RUN;
```

Die Option `param=ref` fordert die Referenzkodierung und mit der Option `ref='2'` kann die gewünschte Referenzkategorie (hier: Kategorie 2, da hier die meisten Beobachtungen enthalten sind; zudem ist ein Vergleich von single zu verheiratet viel interessanter) gewählt werden ([5], [6]).
Die verwendeten Dummy-Variablen sind in Tabelle 4 laut SAS-Output angegeben.

Tabelle 4: Class Level Information
(unter Referenzkodierung)

Class	Value	Design Variables			
FamStand	1	1	0	0	0
	2	**0**	**0**	**0**	**0**
	3	0	1	0	0
	4	0	0	1	0
	5	0	0	0	1

Die geschätzten Regressionskoeffizienten zu obiger SAS-Syntax sind in Tabelle 5 zu finden.

Tabelle 5: Analysis of Maximum Likelihood Estimates (unter Referenzkodierung)

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	2.6876	0.00953	79593.7568	<.0001
FamStand 1	1	***0.2602***	0.0297	76.7708	<.0001
FamStand 3	1	0.1618	0.0620	6.8104	0.0091
FamStand 4	1	0.2187	0.0275	63.0290	<.0001
FamStand 5	1	0.2645	0.0216	150.0076	<.0001

Bestimmt man nun laut „Lehrbuch-Formel“ das Odds Ratio für den Vergleich Kategorie 1 zu Kategorie 2, so ergibt sich folgende Berechnung:

$$OR_{1\ vs\ 2} = OR_{\frac{single}{verheiratet}} = \exp(0.2602) = 1.297$$

Dieses Odds Ratio stimmt diesmal mit dem Odds Ratio im Output (siehe Tabelle 6) überein. Denn rechnet man dies nach, so ergibt sich Folgendes:

$$\ln OR_{1\ vs\ 2} = logit\ \pi(1) - logit\ \pi(2) = \left[\beta_{(0)} + \beta_{(1)}\right] - \left[\beta_{(0)}\right] = \beta_{(1)}\ ,$$

was exakt der „Lehrbuch-Formel“ entspricht.
Auf Grund von Tabelle 6 scheinen also Singles ein signifikant höheres Risiko im Vergleich zu Verheirateten zu haben an einem Folgetumor nach einem bereits diagnostizierten Lungenkrebs zu erkranken.

Tabelle 6: Odds Ratio Estimates (unter Referenzkodierung)

Effect	Point Estimate	95% Wald Confidence Limits	
FamStand 1 vs 2	***1.297***	1.224	1.375
FamStand 3 vs 2	1.176	1.041	1.327
FamStand 4 vs 2	1.244	1.179	1.313
FamStand 5 vs 2	1.303	1.249	1.359

Wird also die Referenzkodierung angewendet, so kann die bekannte „Lehrbuch-Formel“ für die Berechnung des Odds Ratios herangezogen werden [4], falls ein Vergleich zur gewählten Referenzkategorie gewünscht ist.

4.3 Default der Dummy-Kodierung bei verschiedenen Regressionsmodellen

In diesem Beitrag wurde speziell auf die Prozedur LOGISTIC eingegangen, da hier die erwähnte potenzielle Falle möglich ist. Wie ist aber die Voreinstellung der Dummy-Kodierung im CLASS Statement bei anderen Regressionsprozeduren?

Hierzu findet man in Tabelle 7 eine Übersicht über die Defaults der Dummy-Kodierung bei den klassischen Regressionsmodellen zusammen mit der zugehörigen Prozedur in SAS [7] und der zugehörigen Verhältnismaßzahl, für dessen Berechnung die Dummy-Kodierung eine Rolle spielt.

Tabelle 7: Defaults der Dummy-Kodierung im CLASS Statement bei verschiedenen Regressionsmodellen [7]

Regressions-typ	**SAS Prozedur**	**Verhältnis-maßzahl**	**CLASS Statement vorhanden?**	**Default Dummy-Kodierung im CLASS Statement [SAS-Syntax]**
Logistische Regression	LOGISTIC	Odds Ratio	ja	**Effektkodierung** [`param=effect`]
Cox Regression	PHREG	Hazard Ratio	ja	Referenzkodierung [`param=ref`]
Poisson Regression	GENMOD	Relatives Risiko	ja	Referenzkodierung [`param=glm`]
Lineare Regression	REG/ GLM/ GLMSELECT		nein/ ja/ ja	- / Referenzkodierung??/ Referenzkodierung [`param=glm`]

Aus Tabelle 7 ist ersichtlich, dass die Effektkodierung als Default lediglich bei der Prozedur LOGISTIC verwendet wird, bei den anderen Prozeduren ist die Referenzkodierung die Voreinstellung der Dummy-Kodierung. Es gibt zwar kleine Unterschiede zwischen den Optionen `param=ref` und `param=glm`, aber im Prinzip wird bei beiden Varianten eine Referenzkodierung durchgeführt, was sich auch im Output durch exakt dieselben Ergebnisse bei beiden Varianten widerspiegelt.
Da die Prozedur REG kein CLASS Statement anbietet, wird im Falle kategorieller Kovariaten üblicherweise die Prozedur GLM verwendet. Bei der Prozedur GLM ist jedoch nicht ganz sicher, was im CLASS Statement tatsächlich passiert. Laut SAS-Dokumentation sollte eine Referenzkodierung im CLASS Statement durchgeführt werden [7], allerdings konnte dies auf Grund des Outputs nicht nachgeprüft werden. Zudem erweckt der Output den Eindruck, PROC GLM würde trotz CLASS Statement überhaupt keine Dummy-Kodierung anwenden, sondern eine kategorielle Kovariate als stetige Variable behandeln. Hier besteht Bedarf an einer transparenteren Dokumentation und einem standardmäßig detaillierteren Output der Prozedur GLM! Daher erscheint die Prozedur GLMSELECT für eine lineare Regression mit kategoriellen Kovariaten geeigneter.
Insbesondere bei der logistischen Regression ist also Vorsicht bei der Berechnung des Odds Ratios geboten, die anderen beiden Verhältnismaßzahlen können nach der bekannten „Lehrbuch-Formel“ bestimmt werden, falls jeweils ein Vergleich zur Referenzkategorie von Interesse ist.
Die Referenzkodierung kann bei allen Prozeduren aus Tabelle 7 (außer PROC REG und PROC GLM) mit Wahl der gewünschten Referenzkategorie beispielsweise für die kate-

goriellen Kovariaten „Familienstand bei Diagnose“ und „Geschlecht“ wie folgt erzwungen werden:

```
CLASS FamStand (param=ref ref='2') sex (param=ref ref='1');
```

5 Fazit

Im Rahmen dieses Beitrages wurde festgestellt, dass sich in SAS die Voreinstellung der Dummy-Kodierung im CLASS Statement zwischen den verschiedenen Regressionstypen unterscheidet. Nur unter der Referenzkodierung jedoch erhält man durch exponieren des jeweiligen Regressionskoeffizienten das zugehörige Odds Ratio/Hazard Ratio/Relative Risiko.
Insbesondere bei der Prozedur LOGISTIC ist Vorsicht geboten, denn hier ist die Effektkodierung die Voreinstellung der Dummy-Kodierung, wodurch die bekannte „Lehrbuch-Formel“ für die Berechnung des Odds Ratio nicht ohne Weiteres angewendet werden kann.
Im Rahmen von Publikationen sollte einem daher bewusst sein, welche Dummy-Kodierung dahintersteckt, denn werden beispielsweise die Odds Ratios aus Platzmangel nicht publiziert, käme der an Odds Ratios interessierte Leser ohne Kenntnis der verwendeten Dummy-Kodierung auf falsche Ergebnisse.

Literatur

[1] Agresti A (2007). An Introduction to Categorical Data Analysis. Hoboken: John Wiley and Sons.

[1] FAQ: What is effect coding? UCLA: Academic Technology Services, Statistical Consulting Group. *http://www.ats.ucla.edu/stat/mult_pkg/faq/general/effect.htm* (letzter Zugriff: Januar 2013).

[2] Fenske N (2010). Kodierung von kategorialen Kovariablen im Regressionsmodell. *http://www.statistik.lmu.de/institut/ag/leisch/teaching/stat3nf1011/uebung/blatt4/KodierungKategorial.pdf* (letzter Zugriff: Januar 2013).

[3] Hosmer D W, Lemeshow S (2000). Applied Logistic Regression, 2nd edition. New York: Wiley.

[4] In PROC LOGISTIC why aren`t the coefficients consistent with the odds ratios? UCLA: Academic Technology Services, Statistical Consulting Group. *http://www.ats.ucla.edu/stat/sas/faq/proc_logistic_coding.htm* (letzter Zugriff: Januar 2013).

[5] Lewis T (2007). PROC LOGISTIC: The Logistics Behind Interpreting Categorical Variable Effects. *http://www.nesug.org/proceedings/nesug07/sa/sa11.pdf* (letzter Zugriff: Januar 2013).

[6] SAS/STAT 9.3 User`s Guide. *http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#titlepage.htm* (letzter Zugriff: Januar 2013).

[7] SEER Research Data 1973-2008: Surveillance, Epidemiology and End Results (SEER) Program. www.seer.cancer.gov (Datenzugriff: März 2012).