Logistische Regression

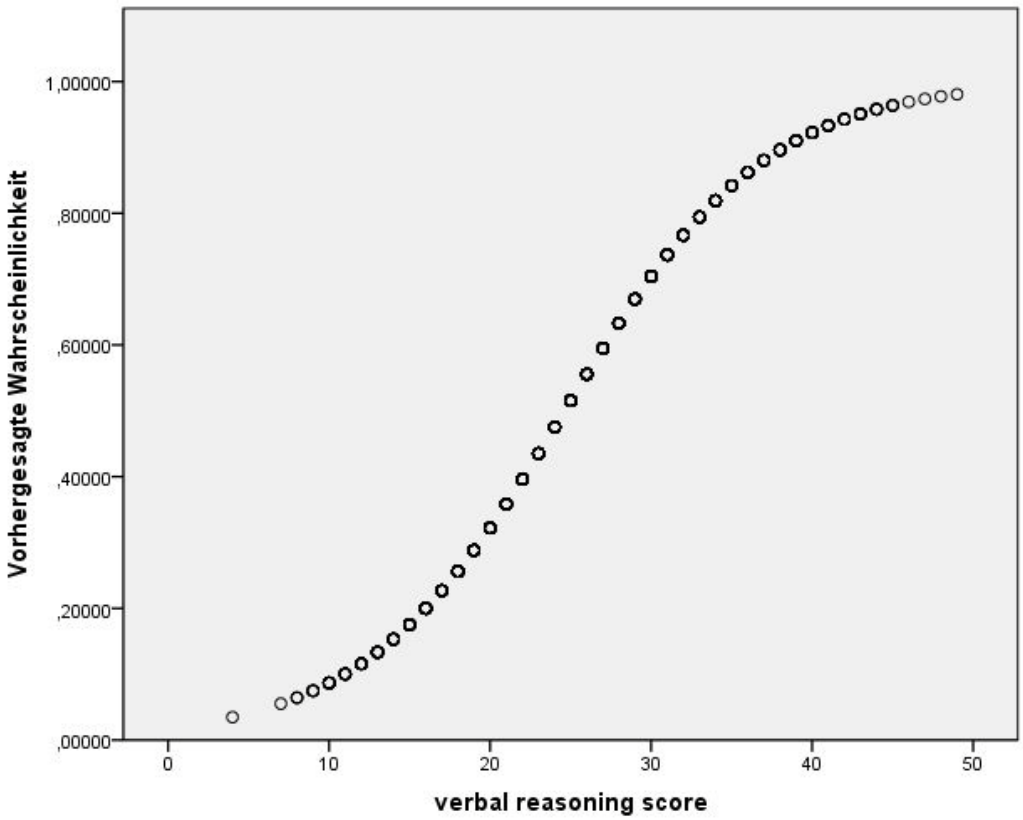
Modell

$$P(Y=1) = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \ldots}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \ldots}}$$

Alternative Schreibweisen der logistischen Funktion:

$$P(Y=1) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \ldots)}}$$

$$\ln\left(\frac{P(Y=1)}{1 - P(Y=1)}\right) = b_0 + b_1 x_1 + b_2 x_2 + \ldots$$



Funktionsverlauf der logistischen Funktion für b>0

Voraussetzungen:

- ein hinreichend großer Stichprobenumfang *n*, mindestens 10 pro Prädiktor bzw. geschätztem Parameter (wobei die Empfehlungen zum Teil stark divergieren). Da nominal skalierte Variable mit *k* Merkmalsausprägungen in *(k-1)* Kontrastvariable transformiert und für die Interaktionen auch deren Produkte als Prädiktoren verwendet werden, bedeutet das *n*: ca. 10*(Anzahl der Zellen).
- ein „vernünftiges“ Modell, d.h. u.a. ohne überflüssige (nicht erklärende) und kollineare Variablen. Diese Forderung erübrigt sich allerdings beim Einsatz als Varianzanalyse.
- wenig leere Zellen.

Begriffe und Tests

Wald-Test Test eines oder (simultan) mehrerer Regressionskoeffizienten auf Verschiedenheit von 0

Rao's Score-Test Test des Regressionskoeffizienten auf Verschiedenheit von 0

Odds $\frac{\text{Wahrscheinlichkeit, dass das Ereignis eintritt}}{\text{Wahrscheinlichkeit, dass das Ereignis nicht eintritt}} = \frac{W(1)}{W(0)}$

Regressionskoeffizienten

$$\log\left(\frac{W(1)}{W(0)}\right) = b_0 + b_1 x_1 + b_2 x_2 + \ldots$$

$$\frac{W(1)}{W(0)} = e^{b_0} \cdot e^{b_1 x_1} \cdot e^{b_2 x_2} \cdot \ldots$$

d.h. e^{b_i} ist der Faktor, um den sich das Verhältnis W(1)/W(0) verändert, wenn x_i um eine Einheit zunimmt; er wird *Odds* genannt.

Likelihood Wahrscheinlichkeit, dass vorgegebene Daten mit einem bestimmten vorgegebenen Modell vereinbar sind. Meistens wird angegeben:
LL = -2 log(*Likelihood*), so dass für ein 'gutes' Modell LL~0 sein muss und für ein 'schlechtes' Modell LL sehr groß.

Deviance auch Likelihood ratio: $D = -2\log\frac{Likelihood(model)}{Likehood(saturatedmodel)}$

ist χ^2-verteilt und sollte für ein gutes Modell klein und nicht signifikant sein. Wird vielfach zum Vergleich zweier Modelle verwendet: Signifikante Werte indizieren eine Verbesserung der Modell-Anpassung.

Model χ^2 simultaner Test, dass alle Regressionskoeffizienten gleich 0 sind (sollte in der Regel signifikant sein)

Improvement χ^2 Differenz der LL-Werte (χ^2-Werte) durch Hinzunahme eines oder mehrerer Prädiktoren (sollte signifikant sein, solange Prädiktoren zusätzlich in das Modell aufgenommen werden)

Goodness of Fit $\sum\frac{(y_i - p_i)^2}{(1-p_i)p_i}$ (Güte der Anpassung), wobei
y_i die beobachteten Werte, also 0 oder 1, und
p_i die geschätzte Wahrscheinlichkeit, dass y_i den Wert 1 hat.
(sollte in der Regel nicht signifikant sein)

Pseudo R^2 In Analogie zum R^2 der linearen Regression wurden solche für die logistische Regression nachgebildet und geben in etwa den Anteil der durch das Modell erklärten Streuung wieder. Die bekanntesten sind die von *Cox & Snell*, *Nagelkerke* und das *Likelihood ratio.* Einzelheiten hierzu u.a. unter
http://www.ats.ucla.edu/stat/mult_pkg/faq/general/psuedo_rsquareds.htm
https://en.wikipedia.org/wiki/Logistic_regression

Hosmer-Lemeshow-Test

χ^2-Anpassungstest zur Güte des Modells, sollte daher nicht signifikant sein; hierbei werden intervallweise die beobachteten Werte der abhängigen Variablen mit den auf Grund des Modells vorhergesagten Werten verglichen. (In SPSS über Optionen anfordern)

Interpretation des Regressionskoeffizienten und Odds Ratio bei der logistischen Regression

Odds ist das Verhältnis W(1)/W(0) (s.o.) für einen beliebigen Wert eines Prädiktors *x*. (Dieses Verhältnis ist etwas anderes als die relative Häufigkeit des Zustandes 1.

1. Fall: dichotomer Prädiktor x (z.B. Raucher/Nichtraucher)

Bei einer Kodierung 0 (z.B. Nichtraucher) bzw. 1 (z.B. Raucher) ist der Odds Ratio

$$\Psi = \frac{\frac{W(1)}{W(0)}(x=1)}{\frac{W(1)}{W(0)}(x=0)}, \quad \Psi = \exp(Koeffizient) \text{ bzw. } e^{b_i}$$

d.h. wie sich der Odds für x=1 zum Odds für x=0 verhält.

Beispiel: Ist die abhängige Variable „Lungenkrebs“ (nein/ja), dann besagt ein $\Psi = 2$, dass Lungenkrebs doppelt so häufig bei Rauchern (x=1) vorkommt wie bei Nichtrauchern (x=0).

2. Fall: polychotomer Prädiktor x mit *k* Kategorien (*Referenz-Kodierung)*
(z.B. Rasse: Weiße (0), Schwarze (1), Latinos (3), Sonstige (4))

Aus dieser werden *k-1* dichtotome Variablen gebildet. Das bei der logistischen Regression übliche Standardverfahren ist die *Referenz-Kodierung* (SPSS: Indikator / R: treatment). Dabei bildet eine Kategorie (standardmäßig die mit dem kleinsten Code, kann aber in SPSS als erste oder letzte gewählt werden) die Referenz-Kategorie (z.B. 0=Weiße). Für jede der übrigen (*k-1*) Kategorien werden wie bei dichotomen Variablen (s.o.) die Statistiken ausgegeben. Für eine Kategorie *i* ist dann der Odds Ratio

$$\Psi_i = \frac{\frac{W(1)}{W(0)}(x=i)}{\frac{W(1)}{W(0)}(x=\text{Referenzkategorie})}$$

Beispiel: Ist die abhängige Variable eine Erkrankung (nein/ja), dann besagt ein $\psi_1 = 3$, daß die Krankheit bei Schwarzen (x=1) 3-mal so häufig vorkommt wie bei Weißen, und ein $\psi_3 = 0,5$, daß die Krankheit bei Sonstigen (x=3) nur halb so häufig auftritt wie bei Weißen.

3. Fall: polychotomer Prädiktor x mit *k* Kategorien (*Marginal-Kodierung)*

Eine gegenüber dem 2. Fall andere Bildung von *k-1* dichotomen Variablen ist die *Marginal-Kodierung* (SPSS: Abweichung / R: sum). Die Referenz ist dann nicht eine ausgewählte Kategorie, sondern das (geometrische) Mittel der Odds Ratio. Für eine Kategorie *i* ist dann der Odds Ratio

$$\Psi_i = \frac{\frac{W(1)}{W(0)}(x=i)}{\frac{\psi_1 + \psi_2 + \ldots + \psi_k}{k}}.$$

Beispiel: Ist die abhängige Variable eine Erkrankung (nein/ja), dann besagt ein $\psi_1 = 3$, dass die Krankheit bei Schwarzen (x=1) 3-mal so häufig vorkommt wie im Durchschnitt aller Rassen.

4. Fall: metrischer/intervallskalierter Prädiktor x (z.B. Alter)

Der Odds Ration gibt an, wie sich das Verhältnis W(1)/W(0) verändert, wenn sich x um eine Einheit vergrößert.
Beispiel: Ist die abhängige Variable eine Erkrankung (nein/ja), dann besagt ein $\psi_1 = 1,17$, dass für jedes Lebensjahr das Krankheitsrisiko um den Faktor 1,17, d.h. um 17 Prozent zu-

nimmt. (Ist der Koeffizient negativ, ψ also kleiner 1, nimmt das Risiko natürlich ab.) Oder in größeren Intervallen ausgedrückt, pro 10 Lebensjahre nimmt das Krankheitsrisiko um den Faktor $\psi^{10} = 1{,}17^{10} = 4{,}8$ zu. Alternative Berechnung, wenn b=0,157 : $e^{10 \cdot b} = e^{10 \cdot 0{,}\dot{1}57} = 4{,}8$.