

Fragen und Antworten zu Kapitel 22

- (1) Wie ist das Modell der einfachen logistischen Regressionsanalyse für dichotome abhängige Variablen in Form (a) der bedingten Wahrscheinlichkeitsfunktionen, (b) der bedingten Wettquotientenfunktionen und (c) der bedingten Logit-Funktionen definiert?**

(a) Das Modell lautet in der Form der bedingten Wahrscheinlichkeitsfunktionen:

$$P(Y = 1|X) = \frac{e^{\beta_0 + \beta_1 \cdot X}}{1 + e^{\beta_0 + \beta_1 \cdot X}}$$

(b) Das Modell lautet in der Form der bedingten Wettquotientenfunktionen:

$$\frac{P(Y = 1|X)}{1 - P(Y = 1|X)} = e^{\beta_0 + \beta_1 \cdot X} = e^{\beta_0} \cdot e^{\beta_1 \cdot X} = e^{\beta_0} \cdot (e^{\beta_1})^X$$

(c) Das Modell lautet in der Form der bedingten Logit-Funktionen:

$$\ln\left(\frac{P(Y = 1|X)}{1 - P(Y = 1|X)}\right) = \beta_0 + \beta_1 \cdot X$$

- (2) Wie lassen sich die Regressionsparameter β_0 und β_1 in der einfachen logistischen Regressionsanalyse interpretieren?**

Der Parameter β_0 bringt zum Ausdruck, dass generell die Wahrscheinlichkeit, eine bestimmte Kategorie zu wählen, auf einem höheren oder geringeren Niveau liegt. Er entspricht dem Wert der bedingten Logit-Funktion an der Stelle $X = 0$. Der Parameter β_1 zeigt an, wie stark die Wahrscheinlichkeit, die Kategorie zu wählen, mit Zunahme der Werte auf der unabhängigen Variablen ansteigt. Er entspricht der erwarteten Veränderung des Wertes der bedingten Logit-Funktion, wenn sich der Wert auf der Variablen X um eine Einheit erhöht.

- (3) Warum kann die Regressionsanalyse für metrische Variablen nicht auf dichotome und mehrkategoriale Variablen angewendet werden?**

Die Regressionsanalyse für metrische Variablen ist aus drei Gründen für dichotome Variablen nicht geeignet: (1) aufgrund der Annahme der linearen Abhängigkeit, (2) aufgrund der Verletzung der Normalverteilungsannahme der Residualvariablen, (3) aufgrund der Verletzung der Homoskedastizitätsannahme.

- (4) Erläutern Sie den Grundgedanken der Maximum-Likelihood-Schätzung!**

Das ML-Verfahren basiert auf der Likelihood-Funktion. Die Likelihood-Funktion beschreibt die Wahrscheinlichkeit der Daten, die man in einer Untersuchung erhalten hat, als Funktion der Modellparameter unter der Voraussetzung, dass das Modell gilt. Die Modellparameter werden so geschätzt, dass die Likelihood-Funktion ihren maximal möglichen Wert annimmt. Diese geschätzten Parameter sind dann diejenigen aller möglichen Parameter, für die die gefundenen Daten maximal wahrscheinlich sind.

(5) Welche Annahmen müssen getroffen werden, damit die Maximum-Likelihood-Schätzungen und ihre Standardfehler korrekt sind?

Maximum-Likelihood-Schätzungen setzen voraus, dass das Modell korrekt spezifiziert wurde, dass die abhängige Variable bedingt binomialverteilt ist und dass die Beobachtungen voneinander unabhängig sind.

(6) Welche Probleme sind mit dem Devianztest zur Überprüfung der Modellgültigkeit verknüpft?

Bei dem Devianz-Test ist zu beachten, dass der Test große Stichproben voraussetzt. Insbesondere wenn eine Vielzahl von Wertekombinationen der unabhängigen Variablen existiert, die nur wenige Personen aufweisen, ist nicht sichergestellt, dass die Prüfgröße einer χ^2 -Verteilung folgt. Es ist dann fraglich, ob der p -Wert korrekt ist.

(7) Auf welchem Grundgedanken basiert der Hosmer-Lemeshow-Test?

Der Hosmer-Lemeshow-Test basiert auf dem folgenden Grundgedanken: Für jede Kombination von Werten auf den unabhängigen Variablen können anhand der multiplen logistischen Regressionsanalyse die Wahrscheinlichkeiten der beiden Kategorien der abhängigen Variablen bestimmt werden. Multipliziert man diese Wahrscheinlichkeiten mit der Anzahl der Personen, die eine bestimmte Wertekombination der unabhängigen Variablen aufweisen, erhält man die erwarteten Häufigkeiten der beiden Kategorien der abhängigen Variablen. Diese erwarteten Häufigkeiten können mit den beobachteten Häufigkeiten verglichen werden. Bei Modellgültigkeit sollten die beobachteten Häufigkeiten nur zufällig von den erwarteten Häufigkeiten abweichen.

(8) Warum lassen sich Residuen bei der logistischen Regressionsanalyse nur schwer zur Beurteilung von Ausreißern heranziehen?

Dies liegt daran, dass die abhängige Variable im dichotomen Fall nur zwei Werte annehmen kann. Das Ausreißerkonzept lässt sich daher nur schwer auf die abhängige Variable übertragen, da die möglichen Abweichungen von den vorhergesagten Werten eine obere und untere Grenze haben.

(9) Was versteht man unter dem Nullzellenproblem?

Ein Nullzellenproblem besteht, wenn Kategorien der abhängigen Variablen für eine Wertekombination der unabhängigen Variablen nicht besetzt sind. Führt diese Datensituation zur vollständigen Separierbarkeit, dann können im Falle der einfachen logistischen Regression die Regressionsparameter nicht geschätzt werden bzw. im Fall der multiplen logistischen Regression nicht zuverlässig geschätzt werden.

(10) Was versteht man unter vollständiger Separierbarkeit, und was sind ihre Konsequenzen?

Vollständige Separierbarkeit liegt im Fall der einfachen logistischen Regression dann vor, wenn alle Personen mit Werten unterhalb eines spezifischen Wertes die eine Kategorie der abhängigen Variablen, alle Personen mit Werten oberhalb dieses spezifischen Wertes hingegen die andere Kategorie der abhängigen Variablen wählen. In diesem Fall können die Regressionsparameter nicht geschätzt werden. Wie bei einer unabhängigen Variablen liegt auch im Falle mehrerer unabhängiger Variablen vollständige Separierbarkeit dann vor, wenn die Werte der abhängigen Variablen aufgrund der unabhängigen Variablen perfekt vorhergesagt werden können. Für jede Wertekombination der unabhängigen Variablen ist dann eine der beiden Kategorien der abhängigen Variablen

nicht besetzt. Zwar werden bei der multiplen logistischen Regressionsanalyse im Falle vollständiger Separierbarkeit die Regressionsparameter geschätzt, sie nehmen aber extrem große Werte an. Dies gilt auch für die Standardfehler.

(11) Wie ist das logistische Regressionsmodell für eine nominalskalierte abhängige Variable definiert?

Zur Definition des Modells muss eine Kategorie der abhängigen Variablen als Referenzkategorie ausgewählt werden. Alle anderen Kategorien werden mit der Referenzkategorie kontrastiert. Für das Beispiel einer einfachen logistischen Regressionsmodell und eine nominalskalierte abhängige Variable mit vier Kategorien lauten die Modellgleichungen:

$$\ln\left(\frac{P(Y=1|X)}{P(Y=0|X)}\right) = \beta_{0(1,0)} + \beta_{1(1,0)}X$$

$$\ln\left(\frac{P(Y=2|X)}{P(Y=0|X)}\right) = \beta_{0(2,0)} + \beta_{1(2,0)}X$$

$$\ln\left(\frac{P(Y=3|X)}{P(Y=0|X)}\right) = \beta_{0(3,0)} + \beta_{1(3,0)}X$$

(12) Unter welchen Bedingungen ist die Überprüfung der Modellangepassungsgüte des Regressionsmodells für nominalskalierte abhängige Variablen mit dem Pearson- χ^2 -Test und dem Likelihood-Ratio-Test problematisch?

Beide Tests machen strenge Voraussetzungen in Bezug auf die benötigte Stichprobengröße, die bei der logistischen Regression mehrkategorialer Variablen häufig nicht erfüllt sind. Es ist dann fraglich, ob die Prüfgrößen des Pearson- χ^2 -Tests und des Likelihood-Ratio-Tests einer χ^2 -Verteilung folgen und ob daher der p -Wert und darauf aufbauend der statistische Schluss korrekt sind.

(13) Wie ist das logistische Regressionsmodell für ordinalskalierte abhängige Variablen definiert?

Das Modell ist wie folgt definiert:

$$\ln\left(\frac{P(Y \leq i|X)}{P(Y > i|X)}\right) = \beta_{0i} + \beta_1 X; \quad 0 \leq i \leq c-1$$

(14) Was kann getan werden, wenn die Annahme der Parallelität der bedingten Wahrscheinlichkeitsfunktionen im proportional odds model verworfen werden muss?

Eine Möglichkeit besteht darin, für jede der bedingten Logit-Funktionen eine eigene logistische Regression für dichotome Variablen anzunehmen. Dieser Ansatz wird als *nested dichotomies approach* bezeichnet, da die Dichotomisierung der abhängigen Variablen sukzessive anhand der Ordnung der Kategorien erfolgt und die dichotomen Variablen somit ineinander verschachtelt sind. Eine andere Möglichkeit besteht darin, auf die Ordnung der Kategorien ganz zu verzichten und das Modell für nominalskalierte Variablen anzuwenden.