

Zum Modellieren von binären Zielgrößen bzw. Wahrscheinlichkeiten haben wir die logistische Regression kennengelernt. Alternative Methoden verwenden andere Linkfunktionen (z. B. basierend auf „probit“ oder „complementary log-log“), aber die logistische Regression hat entscheidende Vorteile: Zum einen ist die Interpretation via Odds relativ einfach möglich. Zum anderen ist das Anwendungsspektrum der logistischen Regression sehr breit: Sie kann sowohl auf prospektive und retrospektive Studien als auch auf Querschnittsstudien angewendet werden, während viele alternative Methoden nur auf prospektive Studien angewendet werden können (Wilson and Lorenz 2015).

Die logistische Regression beruht auf mehreren Annahmen. Wenn diese Annahmen nicht erfüllt sind, sind die berechneten Ergebnisse falsch. Leider ist es für die Software in der Regel *nicht* möglich, Verletzungen dieser Annahmen automatisch zu erkennen. Die Überprüfung der Modellannahmen liegt somit in der Verantwortung des Anwenders.

In diesem Kapitel möchten wir diverse Grenzen der logistischen Regression aufzeigen und Hinweise geben, welche Alternativen in solchen Fällen möglich sind.

---

## 6.1 Überprüfung der Modellannahmen

Verglichen mit der linearen Regression ist es bei der logistischen Regression anspruchsvoller, die Modellannahmen zu prüfen. Folgende Punkte sollten überprüft werden:

- Linearität auf Skala Log-Odds: Bei (evtl. von Hand) gruppierten Daten können die empirischen Log-Odds pro Gruppe ermittelt und gegen erklärende Variablen aufgetragen werden. Dabei sollte ein linearer Zusammenhang ersichtlich sein.

- Allgemeine Güte des Modells: Mit dem Hosmer-Lemeshow-Test (Hosmer Jr et al. 2013) kann die Modellgüte einer logistischen Regression überprüft werden (z. B. mit der Funktion `hoslem.test` in Paket `ResourceSelection` (Lele et al. 2019)). Allerdings kann dieser Test nur mit vielen Beobachtungen (mehrere hundert) Modellabweichungen zuverlässig detektieren.
- Auffällige Beobachtungen: Mit der Funktion `residuals` lassen sich verschiedene Arten von Residuen (z. B. sogenannte „Devianz-Residuen“) der einzelnen Beobachtungen berechnen und vergleichen. Vergleichsweise große Absolutbeträge weisen auf Beobachtungen hin, die vom Modell nicht gut erklärt werden. Es gibt noch weitere Varianten von Residuen.

Ausführliche Informationen zu diesem Thema findet man z. B. in Harrell (2015, Abschn. 10–12).

---

## 6.2 Häufige Probleme

### 6.2.1 Korrelierte Beobachtungen

Im Modell der logistischen Regression nehmen wir an, dass die Beobachtungen unabhängig voneinander sind.

In der Praxis trifft dies bei gruppierten Daten häufig nicht mehr zu. Zum Beispiel könnte es mehrere Beobachtungen innerhalb der gleichen Familie oder innerhalb der gleichen Klinik geben. Ein weiteres Beispiel sind sogenannte longitudinale Daten: Pro Patient werden mehrere Beobachtungen in einem Zeitverlauf gemacht.

Dabei sind sich Beobachtungen innerhalb derselben Gruppe möglicherweise ähnlicher als Beobachtungen aus verschiedenen Gruppen. Infolge der nicht mehr gültigen Unabhängigkeit stimmt dann die vom Modell angenommene Varianz nicht mehr (siehe auch die Bemerkung in Abschn. 3.2 mit der Ankoppelung der Varianz an den Erwartungswert). Die Daten zeigen in diesem Falle typischerweise eine größere Streuung als vom Modell erwartet. Dies wird in der Literatur als **Overdispersion** bezeichnet. Entsprechende Erweiterungen, die eine größere Flexibilität bei der Modellierung der Varianz erlauben, sind in der Funktion `glm` schon implementiert, z. B. mit der Familie `quasibinomial`. Man schwächt damit die Ankoppelung der Varianz an den Erwartungswert ab. Weitere Details zu dieser Methode und der Umsetzung in R findet man in Abschn. 4 von Wilson und Lorenz (2015).

Alternativ gibt es noch zwei weit verbreitete Methoden, mit denen die logistische Regression auf solche Datenstrukturen erweitert werden kann: Die Generalized Linear Mixed Models (Jiang 2007), kurz GLMMs, sind im Paket `lme4` (Bates et al.

2015) implementiert. Die Generalized Estimation Equations (Ziegler 2011), kurz GEE, sind im Paket `gee` (Carey 2019) implementiert.

Weitere Methoden zum Umgang mit korrelierten binären Beobachtungen findet man in Wilson und Lorenz (2015).

### 6.2.2 Wenige Beobachtungen

Die in R produzierten Schätzwerte basieren auf der Annahme, dass sehr viele Beobachtungen zur Verfügung stehen („asymptotische Resultate“). Falls die Anzahl der Beobachtungen „zu klein“ ist, liegen die Schätzwerte der logistischen Regression systematisch daneben, siehe z. B. Nemes et al. (2009).

Es stellt sich natürlich die Frage, ab wann die Anzahl der Beobachtungen „groß genug“ ist. Für diese Frage gibt es leider noch keine einfache und praxistaugliche Antwort. Eine ausführliche Diskussion des Themas findet man in van Smeden et al. (2016).

Falls die Anzahl der Beobachtungen „zu klein“ ist, könnte die sogenannte „exakte logistische Regression“ verwendet werden. Während die Theorie zu dieser Methode existiert (siehe z. B. Abschn. 8.4 in Hosmer Jr et al. (2013) oder Abschn. 8 in Wilson und Lorenz (2015)), ist eine zuverlässige Implementierung in der Software R zur Zeit nicht verfügbar.

### 6.2.3 Perfekte Separierung

Sogenannte „perfekte Separierung“ tritt dann auf, wenn die beiden Gruppen der Zielgröße perfekt durch eine erklärende Variable (oder einer Linearkombination von mehreren erklärenden Variablen) getrennt werden können. Intuitiv scheint diese Situation sehr erstrebenswert, allerdings führt sie zu technischen Problemen bei der Parameterschätzung. Das Problem äußert sich häufig dadurch, dass manche geschätzte Parameterwerte (betragsmäßig) unendlich groß werden.

Dieses Problem tritt besonders häufig auf, wenn es wenige Beobachtungen gibt oder wenn eine der beiden Gruppen sehr selten ist.

Eine mögliche Lösung ist die logistische Regression nach Firth und wird in Heinze und Schemper (2002) diskutiert. Weitere Verbesserungen dieser Methode (FLIC und FLAC) werden in Puhr et al. (2017) vorgestellt. Alle genannten Methoden sind im Paket `logistf` (Heinze et al. 2020) implementiert.

### 6.3 Erweiterungen auf mehr als zwei Klassen

Bei der logistischen Regression besteht die Zielgröße aus einer Faktorvariable mit genau zwei Levels (z. B. Spendebereitschaft „nein“ oder „ja“).

Es gibt Erweiterungen der logistischen Regression für mehr als zwei Levels. Dabei unterscheidet man, ob die Levels *ungeordnet* (z. B. bei Wahlen „Partei A“, „Partei B“, „Partei C“) oder *geordnet* (z. B. bei Krankheitssymptomen „leicht“, „mittel“, „schwer“) sind. Diese Unterscheidung spielt übrigens bei nur zwei Levels keine Rolle.

Bei mehr als zwei *ungeordneten* Levels kann die multinomiale logistische Regression verwendet werden. Der theoretische Hintergrund wird in Abschn. 5.2 von Fahrmeir et al. (2009) illustriert. In R kann die Funktion `multinom` aus dem Paket `nnet` (Venables und Ripley 2002) verwendet werden. Erweiterungen findet man im Paket `mlogit` (Croissant 2020).

Bei mehr als zwei *geordneten* Levels kann die „proportional odds logistic regression (POLR)“ verwendet werden. Mehr Informationen dazu findet man in Abschn. 5.3 von Fahrmeir et al. (2009). In R kann die Funktion `polr` aus dem Paket `MASS` (Venables und Ripley 2002) verwendet werden. Erweiterungen gibt es im Paket `ordinal` (Christensen 2019).

**Open Access** Dieses Kapitel wird unter der Creative Commons Namensnennung 4.0 International Lizenz (<http://creativecommons.org/licenses/by/4.0/deed.de>) veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Die in diesem Kapitel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

