Das logistische Regressionsmodell

3

Die logistische Regression und die lineare Regression haben eine Gemeinsamkeit: Beide versuchen eine Zielgröße durch erklärende Variablen zu modellieren. Wie und auf welcher „Stufe“ dies passiert, schauen wir uns in diesem Kapitel an. Wir wiederholen zuerst die **lineare Regression**. Im Folgenden gehen wir der Einfachheit halber von nur *einer* erklärenden Variable aus. Genau gleich wie ein einfaches lineares Regressionsmodell mit nur einer erklärenden Variablen zu einem multiplen linearen Regressionsmodell erweitert werden kann, ist dies auch bei der logistischen Regression möglich.

3.1 Lineare Regression unter einem neuen Blickwinkel

Das lineare Regressionsmodell für die Daten (x_i, y_i) , $i = 1, \dots, n$ wird typischerweise geschrieben als

$$Y_i = \beta_0 + \beta_1 \cdot x_i + E_i, \quad i = 1, \dots, n, \quad (3.1)$$

wobei E_i unabhängige normalverteilte Fehler sind, d.h. E_i i.i.d. $\sim \mathcal{N}(0, \sigma^2)$. Wir verwenden die englische Abkürzung i.i.d. für „independent and identically distributed“. Die Annahme der Unabhängigkeit bedeutet konkret, dass die Fehler der einzelnen Beobachtungen nichts miteinander zu tun haben (also dass z. B. kein zeitlicher, räumlicher oder sonstiger Zusammenhang zwischen den Fehlern vorhanden ist etc.).

Abb. 3.1 Illustration des datengenerierenden Prozesses bei der einfachen linearen Regression. Für drei verschiedene Werte der erklärenden Variable x ist die entsprechende Verteilung der Zielgröße Y dargestellt. Die Gerade $\beta_0 + \beta_1 x$ ist als durchgezogene Linie eingezeichnet. Die Darstellung basiert auf Meier (2020).

Wenn wir Gl. (3.1) analysieren, können wir folgende Eigenschaft ablesen: An jeder Stelle x_i streut Y_i gemäß einer Normalverteilung um den Wert $\beta_0 + \beta_1 \cdot x_i$ herum, was wir auch als

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 \cdot x_i, \sigma^2)$$

schreiben können (siehe Abb. 3.1). Wir sehen insbesondere, dass die erklärende Variable nur den Erwartungswert dieser Normalverteilung beeinflusst, und zwar durch den Zusammenhang

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 \cdot x_i.$$

Diese Denkweise erlaubt uns, das lineare Regressionsmodell als „zweistufiges“ Modell zu interpretieren. Die zwei Stufen sind:

1. Verteilung der Zielvariable festlegen
2. Geeignete Parameter obiger Verteilung durch erklärende Variable beschreiben

Als Parameter wird dabei in der Regel der Erwartungswert verwendet. Verglichen mit der „direkten“ Schreibweise in Gl. (3.1) erscheint dieses Vorgehen vielleicht auf

den ersten Blick als ein unnötiger Abstraktionsschritt. Es erlaubt aber später eine einfache Erweiterung auf Fälle, bei denen die Normalverteilung nicht mehr angebracht ist. Wir fassen dies nochmals als „neue“ Definition für das lineare Regressionsmodell zusammen.

Definition: Lineare Regression als zweistufiges Modell

Das lineare Regressionsmodell kann folgendermaßen als zweistufiges Modell hingeschrieben werden:

1. Verteilung der Zielvariable festlegen:

$$Y \sim \mathcal{N}(\mu(x), \sigma^2)$$

2. Erwartungswert obiger Verteilung durch erklärende Variable beschreiben:

$$\mu(x) = \beta_0 + \beta_1 \cdot x$$

Für konkret vorliegende Daten (x_i, y_i) , $i = 1, \dots, n$ gehen wir davon aus, dass die Werte y_i jeweils *unabhängige* Realisierungen von obiger Normalverteilung sind.

Obwohl Y von x abhängt, lassen wir diese Abhängigkeit zu Gunsten einer einfachen Notation weg. D.h., wir schreiben bei obigem ersten Punkt jeweils auf der linken Seite nicht $Y(x)$ oder $Y \mid x$, sondern lediglich Y .

Wir fassen nochmals in Worten zusammen:

Der erste Teil legt fest, dass die Beobachtungen Y von einer Normalverteilung stammen. Diese Normalverteilung hat zwei Parameter: Erwartungswert μ und Varianz σ^2 . Durch die Notation $\mu(x)$ drücken wir aus, dass der Erwartungswert μ von der erklärenden Variable x abhängt. Der zweite Parameter, die Varianz σ^2 , wird als konstant angenommen.

Im zweiten Teil legen wir fest, wie der funktionelle Zusammenhang zwischen dem Erwartungswert $\mu(x)$ und der erklärenden Variable x sein soll. Hier gibt es sehr viele Möglichkeiten, aber wir beschränken uns auf eine Funktion, die in den Parametern β_0 und β_1 linear ist. Diese Funktion nennt man auch **linearer Prädiktor**. Der Begriff „linear“ bezieht sich dabei auf β_0 und β_1 (d.h., wenn man nach β_0 oder β_1 ableitet, verschwindet der entsprechende Parameter) und *nicht* auf die erklärende Variable x . D.h., wir könnten die erklärende Variable x auch durch x^2 oder $\log(x)$

ersetzen und hätten immer noch eine lineare Regression vor uns. Später verwenden wir für den linearen Prädiktor oft den griechischen Buchstaben η , d. h., für die i -te Beobachtung haben wir dann $\eta_i = \beta_0 + \beta_1 x_i$, bzw. etwas allgemeiner eine Funktion von x : $\eta(x) = \beta_0 + \beta_1 x$.

Ein typischer Fehler besteht übrigens darin, im zweiten Teil des zweistufigen Modells einen Fehlerterm hinzuzufügen. Der Fehlerterm ist auf den ersten Blick „verschwunden“. Natürlich ist er immer noch da, denn er ist der Grund für die Normalverteilung im ersten Teil des zweistufigen Modells.

3.2 Logistische Regression als zweistufiges Modell

Die logistische Regression folgt dem gleichen zweistufigen Prinzip. Das Ziel besteht darin, eine binäre Zielgröße $Y \in \{0, 1\}$ zu modellieren. Y kann also nur die beiden Werte 0 oder 1 annehmen, die für die beiden möglichen Zustände der Zielgröße stehen (z. B. 0: „krank“ und 1: „gesund“). Als Verteilung dafür bietet sich die **Bernoulli-Verteilung** an, die nur einen Parameter, die **Erfolgswahrscheinlichkeit** $p \in [0, 1]$ besitzt. Es ist $p = \mathbb{P}(Y = 1)$, d. h.

$$Y = \begin{cases} 1 & \text{Wahrscheinlichkeit } p \\ 0 & \text{Wahrscheinlichkeit } 1 - p. \end{cases}$$

Bemerkung: Direkt verwandt mit der Bernoulli-Verteilung ist die **Binomialverteilung**. Die Binomialverteilung modelliert die Anzahl der Erfolge bei n unabhängigen Bernoulli-Verteilungen („Experimente“) mit Erfolgswahrscheinlichkeit p . Wir schreiben hierzu $\text{Bin}(n, p)$. In diesem Sinne kann die Bernoulli-Verteilung auch als Binomialverteilung mit $n = 1$ interpretiert werden.

Die Idee besteht nun darin, die Erfolgswahrscheinlichkeit p als Funktion der erklärenden Variable x zu modellieren, d. h.

$$Y \sim \text{Bernoulli}(p(x)).$$

Verglichen mit der linearen Regression tritt die Erfolgswahrscheinlichkeit der Bernoulli-Verteilung also an die Stelle des Erwartungswerts der Normalverteilung. Dem mathematisch versierten Leser ist aber vielleicht schon aufgefallen, dass wir eigentlich immer noch den Erwartungswert betrachten, denn es gilt $\mathbb{E}(Y) = p(x)$.

Bemerkungen: (i) Obwohl Y von x abhängt, lassen wir diese Abhängigkeit wie schon bei der linearen Regression zu Gunsten einer einfacheren Notation weg. (ii)

Man kann $p(x)$ auch als bedingte Wahrscheinlichkeit interpretieren: $p(x)$ beschreibt die Wahrscheinlichkeit für das Ereignis $Y = 1$ unter der Annahme, dass die erklärende Variable X den konkreten Wert x annimmt, d.h. $p(x) = \mathbb{P}(Y = 1 \mid X = x)$.

Um die logistische Regression als zweistufiges Modell zu schreiben fehlt nur noch der funktionale Zusammenhang zwischen der Erfolgswahrscheinlichkeit $p(x)$ und der erklärenden Variable x . Naheliegender wäre der gleiche Ansatz wie bei der linearen Regression: Die Gewinnwahrscheinlichkeit wird als lineare Funktion der erklärenden Variable (linearer Prädiktor) modelliert, d.h. $p(x) = \beta_0 + \beta_1 \cdot x$. Allerdings stoßen wir dabei auf ein Problem: Je nach Wert der erklärenden Variable x kann der lineare Prädiktor $\eta = \eta(x) = \beta_0 + \beta_1 \cdot x$ eine beliebige Zahl sein. Die Erfolgswahrscheinlichkeit muss allerdings im Intervall $[0, 1]$ liegen!

Dieses Problem können wir lösen, indem wir den linearen Prädiktor so transformieren, dass das Ergebnis für beliebige Werte von x immer im gewünschten Intervall $[0, 1]$ liegt. Eine solche Transformation haben wir schon einmal gesehen, nämlich in Abb. 2.2. Diese Funktion hat einen eigenen Namen, es handelt sich um die sogenannte **logistische Funktion**. Formell ist sie gegeben durch

$$h(\eta) = \frac{e^\eta}{1 + e^\eta}, \eta \in \mathbb{R}.$$

Die Funktion h ist (nochmals!) in Abb. 3.2 dargestellt. Wie man in Abb. 3.2 erahnen, bzw. auch formell herleiten kann, gilt

$$\begin{aligned} \lim_{\eta \rightarrow -\infty} h(\eta) &= 0 \\ \lim_{\eta \rightarrow \infty} h(\eta) &= 1 \\ h(0) &= 0.5. \end{aligned}$$

Wir erhalten so

$$\begin{aligned} \mathbb{P}(Y = 1 \mid X = x) &= p(x) = h(\eta(x)) = h(\beta_0 + \beta_1 \cdot x) \\ &= \frac{\exp(\beta_0 + \beta_1 \cdot x)}{1 + \exp(\beta_0 + \beta_1 \cdot x)} \in [0, 1]. \end{aligned}$$

Dies sieht auf den ersten Blick vielleicht etwas kompliziert aus, sorgt aber dafür, dass für beliebige Werte von x immer ein Wert zwischen 0 und 1 für die modellierte Wahrscheinlichkeit resultiert.

Üblicherweise formt man diesen Zusammenhang so um, dass auf der rechten Seite wieder der lineare Prädiktor steht, also so, wie wir es von der linearen Regres-

Abb. 3.2 Darstellung der logistischen Funktion auf dem Bereich $-4 \leq \eta \leq 4$

sion kennen. Das führt zu folgendem Zusammenhang:

$$\log \left(\frac{\mathbb{P}(Y = 1 | X = x)}{1 - \mathbb{P}(Y = 1 | X = x)} \right) = \log \left(\frac{p(x)}{1 - p(x)} \right) = \beta_0 + \beta_1 \cdot x = \eta(x)$$

Erkennen Sie den Ausdruck auf der linken Seite wieder? Es sind die Log-Odds, die wir in Abschn. 2.1 behandelt haben. Die Log-Odds transformieren also die Erfolgswahrscheinlichkeit so, dass man sie ohne technische Probleme direkt mit dem linearen Prädiktor modellieren kann. Allgemein nennt man eine Funktion g , die einen technisch sinnvollen Zusammenhang zwischen Erwartungswert (hier: Erfolgswahrscheinlichkeit p) und linearem Prädiktor η ermöglicht, eine **Linkfunktion**. Oder in anderen Worten: Die Linkfunktion g „verbindet“ den Erwartungswert mit dem linearen Prädiktor. Die logistische Regression verwendet als Linkfunktion die Log-Odds, d. h., es ist

$$g(p) = \log \left(\frac{p}{1 - p} \right),$$

welche auch als Logit-Funktion bezeichnet wird und der Umkehrung der logistischen Funktion entspricht.

Übrigens: Bei der linearen Regression war keine Transformation des Erwartungswerts (dort: μ) nötig. Die lineare Regression verwendet also als Linkfunktion die Identitätsfunktion, d. h. $g(\mu) = \mu$.

Jetzt sind wir in der Lage, die logistische Regression in der üblichen Darstellungsform zu verstehen. Um die Notation kompakt zu halten, bleiben wir im Folgenden allerdings vorwiegend bei der abkürzenden Schreibweise $p(x)$ statt $\mathbb{P}(Y = 1 | X = x)$.

Definition: Logistische Regression als zweistufiges Modell

Die logistische Regression kann folgendermaßen als zweistufiges Modell hingeschrieben werden:

1. Verteilung der Zielvariable festlegen:

$$Y \sim \text{Bernoulli}(p(x))$$

2. Erwartungswert obiger Verteilung durch linearen Prädiktor beschreiben:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 \cdot x = \eta(x),$$

bzw. äquivalent dazu

$$p(x) = \frac{\exp(\beta_0 + \beta_1 \cdot x)}{1 + \exp(\beta_0 + \beta_1 \cdot x)}.$$

Für konkret vorliegende Daten (x_i, y_i) , $i = 1, \dots, n$ gehen wir davon aus, dass die Werte y_i jeweils *unabhängige* Realisierungen von obiger Bernoulli-Verteilung sind.

Oder nochmals kompakt in Worten: Die Beobachtungen Y stammen von einer Bernoulli-Verteilung mit Erfolgswahrscheinlichkeit $p(x)$. Die Erfolgswahrscheinlichkeit $p(x)$ wird über den „Umweg“ der Log-Odds (Linkfunktion) mit einer linearen Funktion (linearer Prädiktor) modelliert.

Zusammenfassend findet man in Tab. 3.1 die lineare und die logistische Regression im Vergleich.

Die Umkehrung der Linkfunktion wird manchmal auch als **Antwortfunktion** h bezeichnet. Sie berechnet aus dem linearen Prädiktor den entsprechenden Erwartungswert („Umkehrung der Linkfunktion“). Zur Erinnerung: Bei der logistischen Regression haben wir

$$h(\eta) = \frac{e^\eta}{1 + e^\eta}.$$

Dieser Zusammenhang ist auch nochmals in Abb. 3.3 dargestellt.

Tab. 3.1 Vergleich der linearen und der logistischen Regression

Komponente	Lineare Regression	Logistische Regression
Verteilung	Normalverteilung	Bernoulli-Verteilung
Erwartungswert	$\mu(x) \in \mathbb{R}$	$p(x) \in [0, 1]$
Linearer Prädiktor	$\eta(x) = \beta_0 + \beta_1 \cdot x$	$\eta(x) = \beta_0 + \beta_1 \cdot x$
Linkfunktion	Identitätsfunktion: $\mu(x) = \eta(x)$	Logit-Funktion: $\log\left(\frac{p(x)}{1-p(x)}\right) = \eta(x)$

Abb. 3.3 Zusammenhang zwischen Erwartungswert und linearem Prädiktor am Beispiel der logistischen Regression

Bemerkung: Bei der linearen Regression gab es noch einen zweiten Parameter, die Varianz σ^2 . Bei der logistischen Regression ist dies nicht mehr der Fall. Der Grund liegt darin, dass mit der Erfolgswahrscheinlichkeit p bei der Bernoulli-Verteilung sowohl der Erwartungswert als auch die Varianz modelliert werden. In der Tat ist die Varianz gegeben durch $p \cdot (1 - p)$. Diese direkte „Ankoppelung“ der Varianz an den Erwartungswert kann problematisch sein und muss in der Praxis natürlich nicht zwangsläufig erfüllt sein. Mehr dazu in Kap. 6 unter dem Stichwort „quasibinomial“.

Ausblick: Verallgemeinerte lineare Modelle

Dieses zweistufige Schema lässt sich noch auf viele andere Verteilungen anwenden und führt zu den sogenannten **verallgemeinerten linearen Modellen** (auf Englisch: **generalized linear models** oder kurz: **GLM**), die immer aus obigen Komponenten bestehen.

Wenn man Anzahlen modellieren will, bietet sich oft eine Poisson-Verteilung (mit Parameter $\lambda > 0$) an. Man hat dann die in Tab. 3.2 aufgelisteten Komponenten. Man spricht von der sogenannten **Poisson-Regression**.

Tab. 3.2 Komponenten der Poisson-Regression

Komponente	Poisson-Regression
Verteilung	Poisson-Verteilung
Erwartungswert	$\lambda(x) > 0$
Linearer Prädiktor	$\eta(x) = \beta_0 + \beta_1 \cdot x$
Linkfunktion	Logarithmus: $\log(\lambda(x)) = \eta(x)$

3.3 Alternativ: Logistische Regression als latentes Variablenmodell¹

Wir können das logistische Regressionsmodell auch als sogenanntes **latentes Variablenmodell** interpretieren. Als **latente Variable** bezeichnet man eine Variable, deren Wert wir nicht direkt beobachten können. Nur gewisse Eigenschaften der Variable, z. B. ob deren Wert größer oder kleiner gleich Null ist, sind bekannt.

Wir starten mit einem „normalen“ linearen Regressionsmodell für die latente Variable Z_i , d. h.

$$Z_i = \beta_0 + \beta_1 x_i + E_i.$$

Für die Fehler E_i nehmen wir einmal an, dass diese i. i. d. und symmetrisch um Null herum verteilt sind (aber nicht zwangsläufig normalverteilt).

Wenn wir nicht den effektiven Wert von Z_i beobachten können, sondern nur, ob Z_i größer als Null ist oder nicht, erhalten wir als „beobachtbare“ Zielgröße Y_i , wobei

$$Y_i = \begin{cases} 1 & Z_i > 0 \\ 0 & Z_i \leq 0. \end{cases}$$

Y_i folgt als binäre Variable also einer Bernoulli-Verteilung mit Erfolgswahrscheinlichkeit

$$\mathbb{P}(Y_i = 1) = \mathbb{P}(\beta_0 + \beta_1 x_i + E_i > 0) = \mathbb{P}(E_i > -(\beta_0 + \beta_1 x_i)) = \mathbb{P}(E_i < \beta_0 + \beta_1 x_i),$$

wobei die letzte Gleichung aus der Symmetrie der Verteilung der Fehler E_i folgt.

Wenn wir annehmen, dass die Fehler einer sogenannten logistischen Verteilung mit Dichte f und kumulativer Verteilungsfunktion F folgen, wobei

¹Dieser Abschnitt kann beim ersten Lesen auch übersprungen werden.

$$f(x) = \frac{e^x}{(1 + e^x)^2}, \quad F(x) = \frac{e^x}{1 + e^x}, \quad x \in \mathbb{R},$$

so erhalten wir

$$\mathbb{P}(Y_i = 1) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}.$$

Dies hat die gleiche Form wie das logistische Regressionsmodell! Das logistische Regressionsmodell entspricht also einem linearen Regressionsmodell für eine latente Variable mit logistischer Fehlerverteilung!

Bemerkung: Die logistische Verteilung (siehe Abb. 3.4) ist symmetrisch um Null und hat qualitativ eine ähnliche Form wie eine Normalverteilung.

Intuition: Latentes Variablenmodell

Das latente Variablenmodell kann nützlich sein, um die logistische Regression zu verstehen oder zu motivieren. Oft hat die latente Variable die Art eines „Potentials“, das wir nicht direkt beobachten können, sondern nur, ob es realisiert oder umgesetzt wurde. Einige Beispiele:

- Baby lernt gehen: Wir können die neuromotorischen Fähigkeiten nicht direkt messen, sehen aber, ob es mit dem Gehen schon klappt oder nicht.
- Fahrprüfung bestehen: Durch Lernen und Üben werden Fähigkeiten verbessert, die wir nicht direkt messen können. Allerdings helfen größere Fähigkeiten, die Fahrprüfung zu bestehen.

Abb. 3.4 Dichte der logistischen Verteilung auf dem Bereich $-4 \leq x \leq 4$

3.4 Interpretation der Parameter

In der Praxis besteht das Ziel darin, anhand von beobachteten Daten Werte für β_0 und β_1 zu schätzen. Wie dies gemacht wird, schauen wir in Abschn. 3.5 kurz vom mathematischen Standpunkt aus an. Die Umsetzung in der Praxis mit der Statistiksoftware R ist dann das Thema in Kap. 4. Weil das Schätzen der Parameter dank des Computers sehr einfach ist, liegt die Hauptaufgabe darin, die (geschätzten) Modellparameter richtig zu interpretieren.

Um den Einfluss des Achsenabschnitts β_0 und der Steigung β_1 besser zu verstehen, sind in Abb. 3.5 sowohl die Log-Odds als auch die Wahrscheinlichkeiten für verschiedene Parametersettings eingezeichnet. Auf der Skala der Log-Odds haben wir das von der linearen Regression bekannte Bild mit Geraden. Dies führt schon jetzt zur Faustregel: „Die Interpretation der Modellparameter der logistischen Regression auf der Skala der Log-Odds ist genau gleich wie bei der linearen Regression“.

Auf der Skala der Wahrscheinlichkeiten können wir folgendes ablesen: Mit dem Achsenabschnitt β_0 findet nur eine *Verschiebung* der Kurven nach links oder rechts statt. Die Steigung β_1 steuert die Trennschärfe: Für betragsmäßig große Werte von β_1 wechselt die Kurve schnell von sehr kleinen Wahrscheinlichkeiten zu sehr großen (d. h., das Modell ist sehr trennscharf). Genau umgekehrt sieht es aus für betragsmäßig kleine Werte von β_1 . Die Kurve ist dann eher flach. Natürlich hängt die Größe des Koeffizienten β_1 auch direkt davon ab, in welchen Einheiten die erklärende Variable gemessen wird (z. B. cm vs. mm).

Wie der Effekt der Modellparameter auf den verschiedenen Skalen genau quantifiziert und interpretiert wird, schauen wir uns nun anhand eines Beispiels genauer an.

3.4.1 Bedeutung der Modellparameter: Skala Log-Odds

Besonders einfach ist wie oben schon erwähnt die Interpretation der Parameterwerte auf der Skala der Log-Odds. Die Log-Odds werden in unserem Modell durch eine Gerade modelliert. Der Achsenabschnitt ist β_0 und die Steigung β_1 . Das heißt, für $x = 0$ sind die Log-Odds gleich β_0 . Wenn man x um eine Einheit erhöht, erhöhen sich die Log-Odds um den Wert β_1 . Mit den Faustregeln für die Umrechnung von Log-Odds in Wahrscheinlichkeiten (siehe Abschn. 2.1) gelingt eine rasche Interpretation der geschätzten Parameter.

Abb. 3.5 Abhängigkeit der Log-Odds (links) und der Wahrscheinlichkeit (rechts) für verschiedene Parametersettings: $\beta_0 = 0, \beta_1 = 1$ (durchgezogen), $\beta_0 = -1, \beta_1 = 1$ (gestrichelt), $\beta_0 = 0, \beta_1 = 2$ (gepunktet) und $\beta_0 = 0, \beta_1 = 0.5$ (strich-punktiert)

Beispiel: Diagnostischer Test – Skala Log-Odds

Wir betrachten einen (medizinischen) diagnostischen Test, der mittels eines Blutwertes den Gesundheitszustand angibt: Die erklärende Variable x sei der gemessene Blutwert. Die Zielgröße Y ist 1, falls die Person krank ist und 0, falls die Person gesund ist. Wir nehmen an, dass folgendes logistisches Regressionsmodell gilt: $Y \sim \text{Bernoulli}(p(x))$, wobei

$$\log \left(\frac{\mathbb{P}(Y = 1 | X = x)}{1 - \mathbb{P}(Y = 1 | X = x)} \right) = \log \left(\frac{p(x)}{1 - p(x)} \right) = -2 + 0.5 \cdot x.$$

Welche Schlüsse können wir daraus ziehen?

- Der **Achsenabschnitt** ist $\beta_0 = -2$: Wenn der Blutwert $x = 0$ gemessen wurde, sind die Log-Odds für Krankheit gleich -2 . Das entspricht (z.B. gemäß Tabelle in Abschn. 2.1) einer Wahrscheinlichkeit von etwa 10 %, krank zu sein.
- **Effektstärke**: Die Steigung ist $\beta_1 = 0.5$. Wenn der Blutwert um eine Einheit größer wird, dann werden gemäß Modell die Log-Odds für Krankheit um 0.5 größer. Je höher der Blutwert, desto größer also die Wahrscheinlichkeit, krank zu sein.
- **Vorhersage**: Wir können mit diesem Modell Vorhersagen für *beliebige* Werte von x machen. Wenn der Blutwert z.B. den Wert $x = 6$ annimmt, sind die

Log-Odds für Krankheit $-2 + 0.5 \cdot 6 = 1$, was einer Wahrscheinlichkeit von (gerundet) 75 % entspricht, krank zu sein.

3.4.2 Bedeutung der Modellparameter: Skala Odds

Auch auf der Skala der Odds ist eine einfache Interpretation der Parameter möglich. Dazu lösen wir die Modellgleichung mit der Exponentialfunktion nach den Odds auf:

$$\begin{aligned} \text{odds}(Y = 1 | X = x) &= \frac{\mathbb{P}(Y = 1 | X = x)}{1 - \mathbb{P}(Y = 1 | X = x)} = \frac{p(x)}{1 - p(x)} \\ &= \exp(\beta_0 + \beta_1 \cdot x) \\ &= \exp(\beta_0) \cdot \exp(\beta_1 \cdot x) \end{aligned}$$

Für $x = 0$ sind die Odds für Krankheit gleich $\exp(\beta_0)$. Wenn man x um eine Einheit erhöht, dann wird der Faktor $\exp(\beta_1 \cdot x)$ zu

$$\exp(\beta_1 \cdot x) \rightarrow \exp(\beta_1 \cdot (x + 1)) = \exp(\beta_1 \cdot x) \cdot \exp(\beta_1).$$

Das bedeutet, dass sich die Odds um den *Faktor* $\exp(\beta_1)$ ändern, d. h.

$$\text{odds}(Y = 1 | X = x + 1) = \text{odds}(Y = 1 | X = x) \cdot \exp(\beta_1).$$

Daraus lässt sich das entsprechende **Odds-Ratio** berechnen:

$$\text{OR}(Y = 1 | X = x + 1 \text{ vs. } X = x) = \frac{\text{odds}(Y = 1 | X = x + 1)}{\text{odds}(Y = 1 | X = x)} = \exp(\beta_1).$$

Wir sehen insbesondere: Unabhängig vom Wert von x hat eine Erhöhung von x um eine Einheit immer den gleichen *multiplikativen* Effekt auf die Odds. Oder: Das entsprechende Odds-Ratio ist immer $\exp(\beta_1)$. Das Odds-Ratio lässt sich also sehr einfach aus dem Parameter β_1 der logistischen Regression ermitteln.

Beispiel: Diagnostischer Test (Fortsetzung) – Skala Odds

Auf der Skala der Odds können wir folgende Schlüsse ziehen:

- **Effektstärke:** Es ist $\exp(\beta_1) = \exp(0.5) \approx 1.65$. Wenn sich der Blutwert x um eine Einheit erhöht, dann erhöhen sich die Odds für Krankheit um den *Faktor* 1.65. Das entsprechende Odds-Ratio ist also 1.65.
- **Vorhersage:** Auch auf der Skala der Odds können wir Vorhersagen für beliebige Werte von x machen. Für $x = 6$ sind die Log-Odds gleich 1 (siehe das ursprüngliche Beispiel) und somit sind die Odds $\exp(1) \approx 2.718$.

3.4.3 Bedeutung der Modellparameter: Skala Wahrscheinlichkeiten

Auf der Skala der Wahrscheinlichkeiten ist die Interpretation der Parameter schwieriger. Die Effektstärke lässt sich nicht mehr „universell“ quantifizieren. Wir haben gesehen: Wenn wir die erklärende Variable x um eine Einheit erhöhen, hat dies auf der Skala der Log-Odds eine Verschiebung um eine Konstante (β_1) zur Folge. Unabhängig vom Startwert der Log-Odds wird eine Erhöhung von x um eine Einheit die Log-Odds also immer um den Wert β_1 erhöhen. Der Zusammenhang zwischen Log-Odds und Wahrscheinlichkeit ist allerdings *nicht* linear. Eine fixe Erhöhung der Log-Odds um den Wert β_1 führt daher zu *unterschiedlichen* Erhöhungen der Wahrscheinlichkeit, je nachdem, bei welchem Wert der Wahrscheinlichkeit man startet. Dies können wir auch an folgender Tabelle einsehen:

Log-Odds	-2	-1	0
Wahrscheinlichkeit	10 %	25 %	50 %

Die Log-Odds von -2 entsprechen einer Wahrscheinlichkeit von etwa 10 %. Wenn wir die Log-Odds um 1 auf den Wert -1 erhöhen, verändert sich die dazugehörige Wahrscheinlichkeit auf den Wert 25 %. Die Wahrscheinlichkeit wurde also um 15 % größer. Wenn wir die Log-Odds nochmals um 1 auf den Wert 0 erhöhen, verändert sich die dazugehörige Wahrscheinlichkeit auf den Wert 50 %. Die Wahrscheinlichkeit hat sich diesmal also um 25 % und nicht wie vorher um 15 % verändert. Wir sehen: Je nach Startwert hat die Erhöhung der Log-Odds um eine additive Konstante also eine *unterschiedliche* Auswirkung auf die dazugehörige Wahrscheinlichkeit!

Was kann man trotzdem aussagen? Die „Richtung“ des Effekts ist universell gültig und durch das Vorzeichen von β_1 gegeben. Wenn β_1 positiv ist, bedeutet dies: Eine Erhöhung von x hat zur Folge, dass sich die Wahrscheinlichkeit für $Y = 1$ erhöht (hierzu ist es auch nützlich, sich das Ganze mit dem latenten Variablenmodell aus Abschn. 3.3 vorzustellen). Genau umgekehrt geht es mit negativem Vorzeichen.

Weiterhin problemlos möglich sind Vorhersagen, weil wir zu jedem vorhergesagten Wert der Log-Odds durch Umformen die dazugehörige Wahrscheinlichkeit berechnen können.

Beispiel: Diagnostischer Test (Fortsetzung) – Skala Wahrscheinlichkeit

- **Effektstärke:** Nicht einfach quantifizierbar. Aber, weil hier β_1 positiv ist, gilt: Je größer x , desto größer die Wahrscheinlichkeit, krank zu sein ($Y = 1$).
- **Vorhersage:** Auch auf der Skala der Wahrscheinlichkeiten können wir Vorhersagen für beliebige Werte von x machen. Für $x = 6$ sind die Log-Odds gleich 1 (siehe vorangehendes Beispiel) und somit ist die Wahrscheinlichkeit

$$p(x) = \frac{\exp(1)}{1 + \exp(1)} \approx 0.73.$$

3.4.4 Überblick

Zusammenfassend erhalten wir also bei einer Veränderung von x nach $x + 1$ auf den verschiedenen Skalen die in Tab.3.3 aufgelisteten Auswirkungen.

Eine entsprechende Visualisierung für das Modell des Beispiels findet man in Abb.3.6 auf allen drei Skalen.

Tab. 3.3 Übersicht über die Bedeutung der Modellparameter auf den verschiedenen Skalen

Skala	Veränderung, wenn x zu $x + 1$ wird
Log-Odds	Additive Veränderung um den Wert β_1
Odds	Multiplikative Veränderung um den Faktor $\exp(\beta_1)$
Wahrscheinlichkeit	Nicht universell quantifizierbar, Richtung gegeben durch das Vorzeichen von β_1

Abb. 3.6 Zusammenhang zwischen der erklärenden Variable x und Log-Odds, Odds bzw. Wahrscheinlichkeit

3.4.5 Mehrere erklärende Variablen

Die multiple logistische Regression ist eine Erweiterung auf mehrere erklärende Variablen. Wie bei der multiplen linearen Regression werden nun mehrere erklärende Variablen verwendet. In der Regel werden numerische und kategorielle Variablen, also Faktoren, verwendet. Auch Wechselwirkungen (Interaktionen) zwischen erklärenden Variablen sind möglich.

Vorsicht ist bei der Interpretation der Parameter geboten. Bei einer multiplen Regression (egal ob linear oder logistisch) werden *bereinigte* Zusammenhänge berechnet. Also der Zusammenhang zwischen einer erklärenden Variable und der Zielgröße, wenn die *übrigen erklärenden Variablen konstant bleiben*.

Beispiel: Diagnostischer Test: Einfache vs. multiple logistische Regression

Bisher haben wir in diesem Beispiel eine einfache logistische Regression mit einer einzigen erklärenden Variable (Blutwert x) verwendet. Die Steigung ist $\beta_1 = 0.5$. D.h., wenn der Blutwert um eine Einheit größer wird, dann werden gemäß Modell die Log-Odds für Krankheit um 0.5 größer.

Nun erweitern wir das Modell zu einer multiplen logistischen Regression und nehmen zusätzlich die erklärende Variable z auf, die das Alter beschreibt. Wir nehmen an, dass folgendes logistisches Regressionsmodell gilt:

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot z = -2 + 0.3 \cdot x + 0.1 \cdot z.$$

Die Steigung bzgl. dem Blutwert x ist nun $\beta_1 = 0.3$. Dieser Zusammenhang ist für das Alter *bereinigt*, weil Alter eine weitere erklärende Variable im Modell ist.

Es ist wichtig, diese Zusatzinformation in der Interpretation klar auszuweisen: Wenn der Blutwert um eine Einheit größer wird *und das Alter gleich bleibt*, dann werden gemäß Modell die Log-Odds für Krankheit um 0.3 größer. Oder auf der Skala der Odds: Wenn sich der Blutwert x um eine Einheit erhöht *und das Alter gleich bleibt*, dann erhöhen sich die Odds für Krankheit um den Faktor $\exp(0.3) \approx 1.35$. Das entsprechende Odds-Ratio ist also 1.35. ◀

3.5 Ausblick: Parameterschätzung und statistische Inferenz

Die Parameter werden bei der logistischen Regression mit der Maximum-Likelihood-Methode geschätzt. Im Gegensatz zur linearen Regression gibt es keine geschlossene Lösung mehr („Lösungsformel“), sondern es muss ein numerisches Maximierungsverfahren verwendet werden. Wir verzichten hier auf Details.

Auch die statistische Inferenz ist etwas komplizierter als bei der linearen Regression. Während bei der linearen Regression die Verteilung der geschätzten Parameter (exakt) hergeleitet werden kann, ist dies bei der logistischen Regression nicht mehr der Fall, sondern es sind nur asymptotische Resultate vorhanden. Dies bedeutet, dass die berechneten Standardfehler, Vertrauensintervalle und p-Werte nur genähert gelten und die Näherung mit steigender Anzahl Beobachtungen besser wird. Details werden ausführlich z. B. in McCullagh und Nelder (1989) besprochen.

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.