Logistische Regression

statistik-online.ch

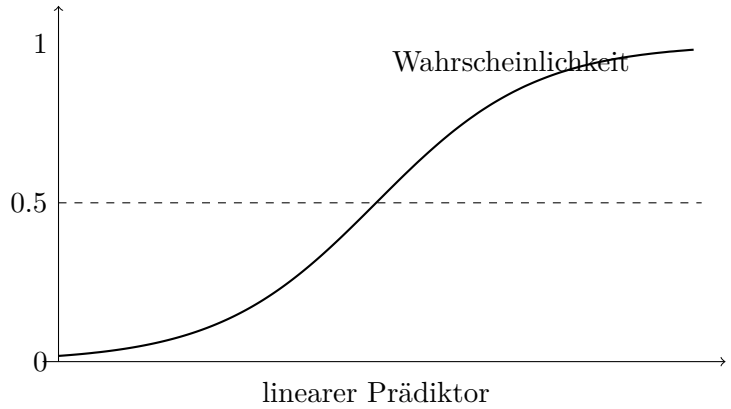
Die **logistische Regression** wird verwendet, wenn die Zielvariable nur zwei mögliche Ausprägungen hat. Das Modell sagt keine beliebigen Zahlen voraus, sondern eine **Wahrscheinlichkeit**.

1 Grundidee

Logistische Regression modelliert die Wahrscheinlichkeit, dass ein Ereignis eintritt.

Typische Fragen sind:

- Besteht eine Person die Prüfung: ja oder nein?
- Kauft eine Person ein Produkt: ja oder nein?
- Tritt ein bestimmtes Ereignis ein: ja oder nein?
- Wie verändern Lernzeit, Vorwissen oder Kursbesuch die Wahrscheinlichkeit für das Ereignis?



2 Binäre Zielvariable

Die Zielvariable wird meistens mit 0 und 1 codiert.

$$Y = \begin{cases} 0, & \text{Ereignis tritt nicht ein} \\ 1, & \text{Ereignis tritt ein} \end{cases}$$

Beispiel:

$$Y = \begin{cases} 0, & \text{Prüfung nicht bestanden} \\ 1, & \text{Prüfung bestanden} \end{cases}$$

Das Modell schätzt:

$$p = P(Y = 1)$$

3 Warum nicht lineare Regression?

Bei einer binären Zielvariable ist eine lineare Regression ungeeignet.

- Eine lineare Regression kann Werte kleiner als 0 oder grösser als 1 vorhersagen.
- Wahrscheinlichkeiten müssen aber zwischen 0 und 1 liegen.
- Der Zusammenhang zwischen Prädiktoren und Wahrscheinlichkeit ist oft nicht linear.

Die logistische Regression löst dieses Problem mit der S-Kurve. Dadurch bleiben die vorhergesagten Wahrscheinlichkeiten immer zwischen 0 und 1.

4 Logit-Modell

Die logistische Regression modelliert nicht direkt p, sondern den **Logit** der Wahrscheinlichkeit.

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

Bei einer erklärenden Variable lautet das Modell:

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

Bei mehreren erklärenden Variablen:

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

Symbol	**Bedeutung**
p	Wahrscheinlichkeit für $Y = 1$
$1 - p$	Wahrscheinlichkeit für $Y = 0$
b_0	Achsenabschnitt
b_j	Regressionskoeffizient einer erklärenden Variable
x_j	erklärende Variable

5 Von Logit zu Wahrscheinlichkeit

Der lineare Teil des Modells wird oft als η geschrieben:

$$\eta = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

Aus η wird die Wahrscheinlichkeit berechnet:

$$p = \frac{1}{1 + \mathrm{e}^{-\eta}}$$

Wichtig: Die vorhergesagte Wahrscheinlichkeit p liegt immer zwischen 0 und 1.

6 Beispiel

Die Wahrscheinlichkeit, eine Prüfung zu bestehen, soll mit der Lernzeit vorhergesagt werden.

$$\eta = -3 + 0.5x$$

x ist die Lernzeit in Stunden. Eine Person lernt 8 Stunden.

$$\eta = -3 + 0.5 \cdot 8 = 1$$

$$p = \frac{1}{1 + \mathrm{e}^{-1}} = 0.731$$

Antwort: Die vorhergesagte Wahrscheinlichkeit, die Prüfung zu bestehen, beträgt ungefähr 73.1%.

7 Odds und Odds Ratio

Die **Odds** beschreiben das Verhältnis zwischen der Wahrscheinlichkeit für das Ereignis und der Wahrscheinlichkeit gegen das Ereignis.

$$\text{Odds} = \frac{p}{1 - p}$$

Beispiel:

$$p = 0.75 \qquad \text{Odds} = \frac{0.75}{0.25} = 3$$

Das bedeutet: Das Ereignis ist dreimal so wahrscheinlich wie das Nicht-Ereignis.

Odds Ratio

Ein Koeffizient b_j kann in eine Odds Ratio umgerechnet werden:

$$OR = \mathrm{e}^{b_j}$$

Die Odds Ratio ist der **Faktor, mit dem sich die Odds verändern**.

- $OR > 1$: Die Odds steigen.

- $OR < 1$: Die Odds sinken.
- $OR = 1$: Die Odds bleiben gleich.

Im Beispiel ist $b_1 = 0.5$.

$$OR = \mathrm{e}^{0.5} = 1.65$$

Interpretation: Wenn die Lernzeit um 1 Stunde steigt, werden die Odds für das Bestehen mit 1.65 multipliziert. Das gilt, wenn alle anderen Variablen konstant gehalten werden.

8 Mehrere Prädiktoren

Wie bei der multiplen linearen Regression können mehrere Variablen gleichzeitig im Modell stehen.

$$\eta = -4 + 0.45 \cdot \text{Lernzeit} + 0.03 \cdot \text{Vorwissen} + 0.70 \cdot \text{Kurs}$$

$$\text{Kurs} = \begin{cases} 0, & \text{kein Kurs} \\ 1, & \text{Kurs besucht} \end{cases}$$

Interpretation des Kurs-Koeffizienten:

$$OR = \mathrm{e}^{0.70} = 2.01$$

Personen mit Kursbesuch haben etwa doppelt so hohe Odds für das Bestehen wie Personen ohne Kursbesuch, wenn Lernzeit und Vorwissen gleich bleiben.

9 Klassifikation

Aus einer vorhergesagten Wahrscheinlichkeit kann eine Klassifikation gemacht werden. Dafür braucht man einen Schwellenwert.

$$\text{Wenn } p \geq 0.50: \quad \hat{Y} = 1$$

$$\text{Wenn } p < 0.50: \quad \hat{Y} = 0$$

Wichtig: Der Schwellenwert muss nicht immer 0.50 sein. In manchen Anwendungen kann ein anderer Schwellenwert sinnvoll sein, zum Beispiel wenn falsche negative Entscheidungen besonders problematisch sind.

10 Modellgüte

Bei logistischer Regression gibt es kein normales R^2 wie bei linearer Regression. Stattdessen verwendet man zum Beispiel:

- Klassifikationsgenauigkeit
- Sensitivität und Spezifität

- Log-Likelihood
- Pseudo-R^2

Wichtig: Eine hohe Klassifikationsgenauigkeit allein reicht nicht immer aus. Wenn eine Klasse sehr häufig ist, kann ein Modell scheinbar gut wirken, obwohl es die seltene Klasse schlecht erkennt.

11 Voraussetzungen

Für logistische Regression sind besonders wichtig:

- **Binäre Zielvariable:** Die Zielvariable hat zwei Ausprägungen.
- **Unabhängigkeit:** Die Beobachtungen sollten voneinander unabhängig sein.
- **Linearität im Logit:** Metrische Prädiktoren sollten linear mit dem Logit zusammenhängen.
- **Keine starke Multikollinearität:** Die Prädiktoren sollten nicht zu stark miteinander zusammenhängen.
- **Ausreichend grosse Stichprobe:** Für stabile Schätzungen braucht es genügend Beobachtungen in beiden Gruppen.

12 Zusammenfassung

Begriff	**Bedeutung**
Logistische Regression	Regression für eine binäre Zielvariable.
p	Wahrscheinlichkeit, dass $Y = 1$ eintritt.
Logit	$\ln\left(\frac{p}{1-p}\right)$.
Odds	Verhältnis $p/(1-p)$.
Odds Ratio	Faktor, mit dem sich die Odds verändern.
Schwellenwert	Grenze, ab der eine Beobachtung als $Y = 1$ klassifiziert wird.
Dummy-Variable	Codierung einer kategorialen Variable mit 0 und 1.

13 Aufgaben

Aufgabe 1: Wahrscheinlichkeit berechnen

Ein Modell lautet:

$$\eta = -2 + 0.4x$$

Berechnen Sie die vorhergesagte Wahrscheinlichkeit für $x = 5$.

Aufgabe 2: Odds Ratio interpretieren

Ein Koeffizient beträgt $b = 0.7$. Berechnen und interpretieren Sie die Odds Ratio.

Aufgabe 3: Dummy-Variable

Ein Modell enthält die Variable Kurs, wobei Kurs = 1 für Kursbesuch und Kurs = 0 für keinen Kurs steht. Der Koeffizient ist $b = 0.6$.
Interpretieren Sie diesen Koeffizienten mit Hilfe der Odds Ratio.

Aufgabe 4: Klassifikation

Ein Modell verwendet den Schwellenwert 0.50. Klassifizieren Sie:

Person	Vorhergesagte Wahrscheinlichkeit	Klasse
A	0.82	
B	0.47	
C	0.51	

Aufgabe 5: Multiple Choice

Kreuzen Sie alle richtigen Aussagen an.

- ❑ Logistische Regression wird für binäre Zielvariablen verwendet.
- ❑ Vorhergesagte Wahrscheinlichkeiten können kleiner als 0 werden.
- ❑ Eine Odds Ratio von 2 bedeutet, dass sich die Odds verdoppeln.
- ❑ Koeffizienten werden bei konstanten anderen Variablen interpretiert.

14 Lösungen

Lösung 1

$$\eta = -2 + 0.4 \cdot 5 = 0$$

$$p = \frac{1}{1 + \mathrm{e}^{-0}} = \frac{1}{2} = 0.50$$

Antwort: Die vorhergesagte Wahrscheinlichkeit beträgt 50%.

Lösung 2

$$OR = \mathrm{e}^{0.7} = 2.01$$

Interpretation: Wenn die Variable um 1 Einheit steigt, werden die Odds ungefähr mit 2.01 multipliziert. Die Odds verdoppeln sich also ungefähr.

Lösung 3

$$OR = \mathrm{e}^{0.6} = 1.82$$

Interpretation: Personen mit Kursbesuch haben 1.82-mal so hohe Odds für das Ereignis wie Personen ohne Kursbesuch, wenn alle anderen Variablen gleich bleiben.

Lösung 4

Bei einem Schwellenwert von 0.50 gilt:

Person	Vorhergesagte Wahrscheinlichkeit	Klasse
A	0.82	1
B	0.47	0
C	0.51	1

Lösung 5

✓ Richtig – Logistische Regression wird für binäre Zielvariablen verwendet.

✗ Falsch – Vorhergesagte Wahrscheinlichkeiten liegen immer zwischen 0 und 1.

✓ Richtig – Eine Odds Ratio von 2 bedeutet eine Verdopplung der Odds.

✓ Richtig – Koeffizienten werden bei konstanten anderen Variablen interpretiert.