

Poisson-Regression

statistik-online.ch

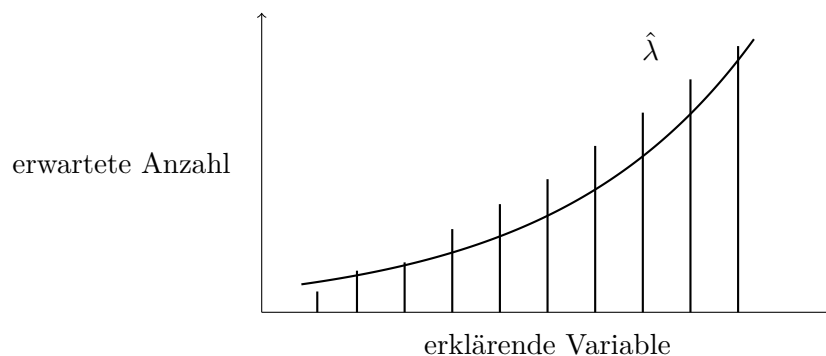
Die **Poisson-Regression** wird verwendet, wenn die Zielvariable eine Anzahl beschreibt. Das Modell sagt also nicht eine beliebige Zahl oder eine Wahrscheinlichkeit voraus, sondern eine **erwartete Anzahl**.

1 Grundidee

Poisson-Regression modelliert erwartete Zählwerte.

Typische Fragen sind:

- Wie viele Arztbesuche hat eine Person in einem Jahr?
- Wie viele Fehler treten auf einer Produktionslinie auf?
- Wie viele Unfälle passieren auf einem Strassenabschnitt?
- Wie verändert sich die erwartete Anzahl, wenn ein Prädiktor steigt?



2 Zählvariable

Die Zielvariable ist eine Anzahl. Deshalb kann sie nur nichtnegative ganze Zahlen annehmen.

$$Y = 0, 1, 2, 3, \dots$$

Beispiele:

- Anzahl Kundenbeschwerden pro Woche

- Anzahl Website-Klicks pro Tag
- Anzahl Krankheitstage pro Jahr
- Anzahl Fehler in einem Text

Das Modell schätzt den Erwartungswert:

$$\lambda = E(Y)$$

3 Warum nicht lineare Regression?

Bei Zähldaten ist eine lineare Regression oft ungeeignet.

- Eine lineare Regression kann negative Werte vorhersagen.
- Zählwerte können aber nicht negativ sein.
- Die Streuung nimmt bei Zählwerten oft mit dem Mittelwert zu.
- Der Zusammenhang zwischen Prädiktoren und erwarteter Anzahl ist oft nicht linear.

Die Poisson-Regression löst dieses Problem mit dem Log-Link. Dadurch bleiben vorhergesagte erwartete Anzahlen immer positiv.

4 Poisson-Modell

Die Poisson-Regression nimmt an, dass die Zielvariable einer Poisson-Verteilung folgt.

$$Y \sim \text{Poisson}(\lambda)$$

λ ist die erwartete Anzahl.

$$E(Y) = \lambda$$

$$\text{Var}(Y) = \lambda$$

Wichtig: In der einfachen Poisson-Verteilung sind Erwartungswert und Varianz gleich gross. Wenn die Varianz deutlich grösser ist als der Mittelwert, spricht man von **Überdispersion**.

5 Log-Link

Die Poisson-Regression modelliert nicht direkt λ , sondern den Logarithmus von λ .

$$\ln(\lambda) = b_0 + b_1x$$

Bei mehreren Prädiktoren:

$$\ln(\lambda) = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k$$

Der lineare Teil wird oft als η geschrieben:

$$\eta = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

Aus η wird die erwartete Anzahl berechnet:

$$\lambda = e^\eta$$

Symbol	Bedeutung
Y	beobachteter Zählwert
λ	erwartete Anzahl
b_0	Achsenabschnitt
b_j	Regressionskoeffizient eines Prädiktors
x_j	Prädiktor

6 Beispiel

Die Anzahl Kundenbeschwerden pro Woche soll mit der Anzahl Lieferungen vorhergesagt werden.

$$\ln(\lambda) = -1 + 0.3x$$

x ist die Anzahl Lieferungen in Hunderten. In einer Woche gibt es $x = 5$.

$$\eta = -1 + 0.3 \cdot 5 = 0.5$$

$$\lambda = e^{0.5} = 1.65$$

Antwort: Das Modell erwartet ungefähr 1.65 Kundenbeschwerden in dieser Woche.

7 Interpretation der Koeffizienten

Ein Koeffizient b_j wird in der Poisson-Regression meist mit e^{b_j} interpretiert.

$$\text{Rate Ratio} = e^{b_j}$$

Die Rate Ratio ist der **Faktor, mit dem sich die erwartete Anzahl verändert**.

- $e^{b_j} > 1$: Die erwartete Anzahl steigt.
- $e^{b_j} < 1$: Die erwartete Anzahl sinkt.
- $e^{b_j} = 1$: Die erwartete Anzahl bleibt gleich.

Im Beispiel ist $b_1 = 0.3$.

$$e^{0.3} = 1.35$$

Interpretation: Wenn x um 1 Einheit steigt, wird die erwartete Anzahl mit 1.35 multipliziert. Die erwartete Anzahl steigt also um etwa 35%.

8 Mehrere Prädiktoren

Wie bei anderen Regressionsmodellen können mehrere Prädiktoren gleichzeitig verwendet werden.

$$\ln(\lambda) = -2 + 0.25 \cdot \text{Lieferungen} + 0.60 \cdot \text{Wochenende}$$

$$\text{Wochenende} = \begin{cases} 0, & \text{Werktag} \\ 1, & \text{Wochenende} \end{cases}$$

Interpretation des Wochenende-Koeffizienten:

$$e^{0.60} = 1.82$$

Am Wochenende ist die erwartete Anzahl etwa 1.82-mal so hoch wie an einem Werktag, wenn die Anzahl Lieferungen gleich bleibt.

9 Offset und Bezugsgrösse

Manchmal werden Zählwerte für unterschiedlich lange Beobachtungszeiten oder unterschiedlich grosse Gruppen erhoben.

Beispiele:

- Unfälle pro Strassenkilometer
- Beschwerden pro 1000 Lieferungen
- Krankheitstage pro Personenjahr

Dann modelliert man nicht nur die Anzahl der Ereignisse, sondern eine **Rate**: Wie viele Ereignisse treten pro Bezugsgrösse auf?

Dafür wird ein Offset verwendet. Die Bezugsgrösse beschreibt, worauf sich die Anzahl der Ereignisse bezieht, zum Beispiel auf eine Beobachtungsdauer, eine Strecke oder eine Gruppengrösse.

$$\ln(\lambda) = b_0 + b_1 x + \ln(\text{Bezugsgrösse})$$

$\ln(\text{Bezugsgrösse})$ ist der **Offset**. Er sorgt dafür, dass Beobachtungen mit unterschiedlich grossen Bezugsgrössen fair miteinander verglichen werden können.

Interpretation: Je grösser die Bezugsgrösse ist, desto mehr Gelegenheit gibt es, dass ein Ereignis auftritt. Der Offset korrigiert für diesen Unterschied.

10 Modellgüte und Probleme

Bei Poisson-Regression sind besonders wichtig:

- **Residualanalyse:** Passen die vorhergesagten Werte zu den beobachteten Zählwerten?

- **Überdispersion:** Ist die Varianz deutlich grösser als der Mittelwert?
- **Viele Nullen:** Gibt es mehr Nullen, als das Poisson-Modell erwartet?
- **Einflussreiche Beobachtungen:** Gibt es einzelne Beobachtungen mit starkem Einfluss?

Wichtig: Bei starker Überdispersion kann eine negative Binomialregression passender sein als eine einfache Poisson-Regression.

11 Voraussetzungen

Für Poisson-Regression sind besonders wichtig:

- **Zählvariable:** Die Zielvariable besteht aus nichtnegativen ganzen Zahlen.
- **Unabhängigkeit:** Die Beobachtungen sollten voneinander unabhängig sein.
- **Log-linearer Zusammenhang:** Die Prädiktoren wirken linear auf $\ln(\lambda)$.
- **Keine starke Überdispersion:** Mittelwert und Varianz sollten ungefähr zusammenpassen.
- **Keine starke Multikollinearität:** Die Prädiktoren sollten nicht zu stark zusammenhängen.

12 Zusammenfassung

Begriff	Bedeutung
Poisson-Regression	Regression für Zählvariablen.
Y	beobachtete Anzahl.
λ	erwartete Anzahl.
Log-Link	Modelliert $\ln(\lambda)$ statt λ direkt.
Rate Ratio	Faktor, mit dem sich die erwartete Anzahl verändert.
Offset	Korrektur für unterschiedliche Beobachtungszeiten oder Gruppengrößen.
Überdispersion	Varianz ist deutlich grösser als der Mittelwert.

13 Aufgaben

Aufgabe 1: Erwartete Anzahl berechnen

Ein Modell lautet:

$$\ln(\lambda) = -0.5 + 0.2x$$

Berechnen Sie die erwartete Anzahl für $x = 4$.

Aufgabe 2: Rate Ratio interpretieren

Ein Koeffizient beträgt $b = 0.4$. Berechnen und interpretieren Sie die Rate Ratio.

Aufgabe 3: Dummy-Variable

Ein Modell enthält die Variable Wochenende, wobei Wochenende = 1 für Wochenende und Wochenende = 0 für Werktag steht. Der Koeffizient ist $b = 0.8$.

Interpretieren Sie diesen Koeffizienten mit Hilfe der Rate Ratio.

Aufgabe 4: Offset

Warum braucht man in manchen Poisson-Modellen einen Offset?

Aufgabe 5: Multiple Choice

Kreuzen Sie alle richtigen Aussagen an.

- ☐ Poisson-Regression wird für Zählvariablen verwendet.
- ☐ Eine Poisson-Regression kann negative erwartete Anzahlen vorhersagen.
- ☐ Eine Rate Ratio von 1.5 bedeutet, dass die erwartete Anzahl mit 1.5 multipliziert wird.
- ☐ Überdispersion bedeutet, dass die Varianz deutlich grösser ist als der Mittelwert.

14 Lösungen

Lösung 1

$$\eta = -0.5 + 0.2 \cdot 4 = 0.3$$

$$\lambda = e^{0.3} = 1.35$$

Antwort: Die erwartete Anzahl beträgt ungefähr 1.35.

Lösung 2

$$\text{Rate Ratio} = e^{0.4} = 1.49$$

Interpretation: Wenn die Variable um 1 Einheit steigt, wird die erwartete Anzahl ungefähr mit 1.49 multipliziert. Sie steigt also um etwa 49%.

Lösung 3

$$\text{Rate Ratio} = e^{0.8} = 2.23$$

Interpretation: Am Wochenende ist die erwartete Anzahl etwa 2.23-mal so hoch wie an einem Werktag, wenn alle anderen Variablen gleich bleiben.

Lösung 4

Ein Offset wird verwendet, wenn Beobachtungen unterschiedlich viel Gelegenheit haben, Ereignisse zu erzeugen. Zum Beispiel ist bei mehr Beobachtungszeit oder einer grösseren Gruppe auch eine höhere Anzahl von Ereignissen zu erwarten. Der Offset macht die Werte vergleichbarer.

Lösung 5

- ✓ Richtig – Poisson-Regression wird für Zählvariablen verwendet.
- ✗ Falsch – Erwartete Anzahlen sind durch den Log-Link immer positiv.
- ✓ Richtig – Eine Rate Ratio von 1.5 multipliziert die erwartete Anzahl mit 1.5.
- ✓ Richtig – Überdispersion bedeutet, dass die Varianz deutlich grösser ist als der Mittelwert.