Multiple lineare Regression

statistik-online.ch

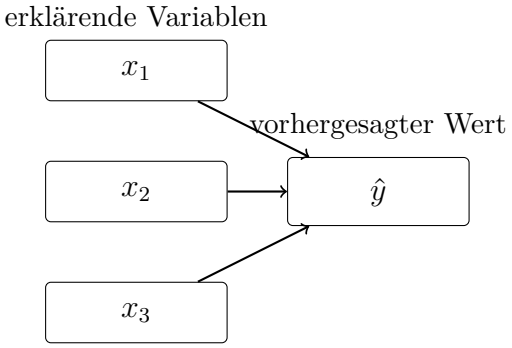
Die **multiple lineare Regression** erweitert die einfache lineare Regression. Statt nur eine erklärende Variable zu verwenden, werden mehrere Variablen gleichzeitig genutzt, um eine Zielvariable vorherzusagen oder zu erklären.

1 Grundidee

Multiple Regression beschreibt, wie mehrere Variablen gemeinsam mit einer Zielvariable zusammenhängen.

Typische Fragen sind:

- Wie hängt die Prüfungsleistung gleichzeitig mit Lernzeit, Vorwissen und Schlaf zusammen?
- Welche Variable erklärt y, wenn die anderen Variablen konstant gehalten werden?
- Wie gut kann das Modell neue Werte vorhersagen?
- Wird das Modell besser, wenn weitere Variablen aufgenommen werden?



2 Modellgleichung

Bei zwei erklärenden Variablen lautet die Gleichung:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

Allgemein gilt:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

Symbol	Bedeutung
$\hat{y}$	vorhergesagter Wert der Zielvariable
$x_1, x_2, \ldots, x_k$	erklärende Variablen
b_0	Achsenabschnitt
$b_1, b_2, \ldots, b_k$	Regressionskoeffizienten
k	Anzahl erklärender Variablen

3 Interpretation der Koeffizienten

Jeder Koeffizient beschreibt den Effekt einer Variable, wenn alle anderen Variablen konstant gehalten werden.

Das ist der wichtigste Unterschied zur einfachen linearen Regression.

- b_1: Veränderung von $\hat{y}$, wenn x_1 um 1 Einheit steigt und alle anderen Variablen gleich bleiben.
- b_2: Veränderung von $\hat{y}$, wenn x_2 um 1 Einheit steigt und alle anderen Variablen gleich bleiben.
- $b_j > 0$: positiver Zusammenhang mit $\hat{y}$.
- $b_j < 0$: negativer Zusammenhang mit $\hat{y}$.

Wichtig: Ein Koeffizient ist kein isolierter Zusammenhang, sondern ein Zusammenhang **unter Kontrolle der anderen Variablen**.

4 Beispiel

Die Prüfungspunktzahl soll durch Lernzeit und Vorwissen vorhergesagt werden.

$$\hat{y} = 35 + 4x_1 + 0.6x_2$$

Variable	Bedeutung	Koeffizient
x_1	Lernzeit in Stunden	$b_1 = 4$
x_2	Vorwissen in Punkten	$b_2 = 0.6$

Interpretation

- Wenn die Lernzeit um 1 Stunde steigt, steigt die vorhergesagte Punktzahl um 4 Punkte, wenn das Vorwissen gleich bleibt.
- Wenn das Vorwissen um 1 Punkt steigt, steigt die vorhergesagte Punktzahl um 0.6 Punkte, wenn die Lernzeit gleich bleibt.
- Der Achsenabschnitt $b_0 = 35$ ist die vorhergesagte Punktzahl, wenn Lernzeit und Vorwissen beide 0 sind.

Vorhersage

Eine Person lernt 6 Stunden und hat 50 Punkte im Vorwissenstest.

$$\hat{y} = 35 + 4 \cdot 6 + 0.6 \cdot 50 = 35 + 24 + 30 = 89$$

Antwort: Die vorhergesagte Prüfungspunktzahl beträgt 89 Punkte.

5 Kategorische Variablen

Multiple Regression kann auch kategoriale Variablen verwenden. Dafür werden sie als **Dummy-Variablen** codiert.

Beispiel: Eine Variable beschreibt, ob eine Person einen Vorbereitungskurs besucht hat.

$$x_3 = \begin{cases} 0, & \text{kein Kurs} \\ 1, & \text{Kurs besucht} \end{cases}$$

Das Modell lautet:

$$\hat{y} = 35 + 4x_1 + 0.6x_2 + 5x_3$$

Interpretation: Personen mit Kursbesuch haben im Modell eine um 5 Punkte höhere vorhergesagte Punktzahl als Personen ohne Kursbesuch, wenn Lernzeit und Vorwissen gleich bleiben.

6 Modellgüte

Bestimmtheitsmass R^2

$$R^2 = \frac{\text{erklärte Streuung}}{\text{gesamte Streuung}}$$

R^2 beschreibt, welcher Anteil der Streuung von y durch das Modell erklärt wird.

$$0 \leq R^2 \leq 1$$

Beispiel:

$$R^2 = 0.64$$

Das bedeutet: 64% der Streuung der Zielvariable werden durch das Modell erklärt.

Korrigiertes R^2

Wenn man zusätzliche Variablen in ein Modell aufnimmt, nimmt R^2 immer zu. Deshalb verwendet man bei multipler Regression oft das korrigierte R^2.

$$R^2_{\text{korr}} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

Symbol	Bedeutung
n	Stichprobengrösse
k	Anzahl erklärender Variablen
R^2_{korr}	korrigiertes Bestimmtheitsmass

Interpretation: Das korrigierte R^2 bestraft unnötige Variablen. Es ist deshalb besser geeignet, Modelle mit unterschiedlicher Anzahl erklärender Variablen zu vergleichen.

7 Multikollinearität

Multikollinearität bedeutet, dass erklärende Variablen stark miteinander zusammenhängen.

Das ist problematisch, weil die einzelnen Koeffizienten dann schwer zu interpretieren sind. Das Modell kann zwar gut vorhersagen, aber es ist unklar, welche Variable welchen Anteil erklärt.

Beispiel

Lernzeit und Anzahl gelöster Übungsaufgaben können stark zusammenhängen: Wer viel lernt, löst oft auch viele Aufgaben. Dann ist es schwierig, den separaten Effekt der Lernzeit und den separaten Effekt der Aufgabenanzahl zu trennen.

8 Voraussetzungen

Für multiple lineare Regression sind besonders wichtig:

- **Linearität:** Der Zusammenhang zwischen den Prädiktoren und y sollte ungefähr linear sein.
- **Unabhängigkeit:** Die Beobachtungen sollten voneinander unabhängig sein.
- **Konstante Streuung:** Die Residuen sollten ungefähr gleich stark streuen.
- **Normalverteilung der Residuen:** Die Residuen sollten ungefähr normalverteilt sein.
- **Keine starke Multikollinearität:** Die erklärenden Variablen sollten nicht zu stark miteinander zusammenhängen.

9 Zusammenfassung

Begriff	**Bedeutung**
Multiple Regression	Regression mit mehreren erklärenden Variablen.
$\hat{y}$	vorhergesagter Wert der Zielvariable.
b_j	Effekt einer Variable bei konstanten anderen Variablen.
Dummy-Variable	Codierung einer kategorialen Variable mit 0 und 1.
R^2	Anteil der Streuung von y, der durch das Modell erklärt wird.
R^2_{korr}	korrigiertes R^2, nützlich für Modellvergleiche.
Multikollinearität	starke Zusammenhänge zwischen erklärenden Variablen.

10 Aufgaben

Aufgabe 1: Vorhersage

Ein Modell lautet:

$$\hat{y} = 20 + 3x_1 + 2x_2$$

Berechnen Sie die Vorhersage für $x_1 = 4$ und $x_2 = 7$.

Aufgabe 2: Koeffizient interpretieren

Ein Modell zur Vorhersage der Monatsmiete lautet:

$$\widehat{\text{Miete}} = 500 + 18 \cdot \text{Wohnfläche} + 120 \cdot \text{Zentrum}$$

Zentrum = 1, wenn die Wohnung im Zentrum liegt, sonst Zentrum = 0.
Interpretieren Sie den Koeffizienten 120.

Aufgabe 3: Korrigiertes R^2

Ein Modell hat $R^2 = 0.70$, $n = 80$ und $k = 4$. Berechnen Sie das korrigierte R^2.

Aufgabe 4: Multikollinearität

Warum kann es problematisch sein, wenn zwei erklärende Variablen sehr stark miteinander korrelieren?

Aufgabe 5: Multiple Choice

Kreuzen Sie alle richtigen Aussagen an.

- ❑ Multiple Regression kann mehrere erklärende Variablen gleichzeitig verwenden.
- ❑ Ein Regressionskoeffizient wird interpretiert, während die anderen Variablen konstant gehalten werden.
- ❑ R^2 wird immer kleiner, wenn man weitere Variablen aufnimmt.
- ❑ Dummy-Variablen können kategoriale Variablen in ein Regressionsmodell aufnehmen.

11 Lösungen

Lösung 1

$$\hat{y} = 20 + 3 \cdot 4 + 2 \cdot 7 = 20 + 12 + 14 = 46$$

Antwort: Die vorhergesagte Ausprägung von y beträgt 46.

Lösung 2

Der Koeffizient 120 bedeutet:

Wohnungen im Zentrum haben eine um 120 höhere vorhergesagte Monatsmiete als Wohnungen ausserhalb des Zentrums, wenn die Wohnfläche gleich bleibt.

Lösung 3

$$R^2_{\text{korr}} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

$$= 1 - (1 - 0.70)\frac{80-1}{80-4-1} = 1 - 0.30 \cdot \frac{79}{75}$$

$$= 1 - 0.316 = 0.684$$

Antwort: Das korrigierte R^2 beträgt ungefähr 0.684.

Lösung 4

Wenn zwei erklärende Variablen stark miteinander korrelieren, ist schwer zu erkennen, welche der beiden Variablen welchen Teil der Zielvariable erklärt. Die einzelnen Koeffizienten können dadurch instabil und schwer interpretierbar werden.

Lösung 5

✓ Richtig – Multiple Regression kann mehrere erklärende Variablen gleichzeitig verwenden.

✓ Richtig – Koeffizienten werden bei konstanten anderen Variablen interpretiert.

✗ Falsch – R^2 wird durch zusätzliche Variablen normalerweise nicht kleiner.

✓ Richtig – Dummy-Variablen können kategoriale Variablen codieren.