Einfache lineare Regression

statistik-online.ch

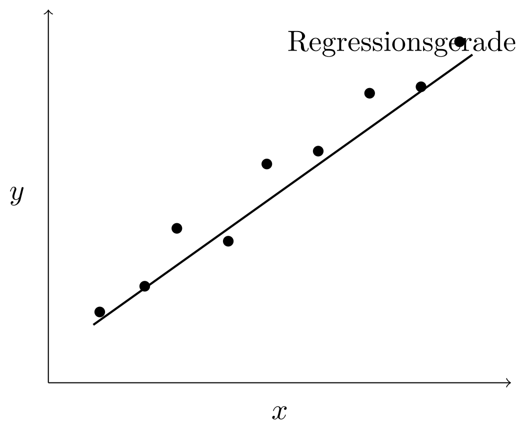
Die **einfache lineare Regression** beschreibt den linearen Zusammenhang zwischen einer erklärenden Variable x und einer Zielvariable y. Mit der Regressionsgerade können Werte beschrieben und vorhergesagt werden.

1 Grundidee

Regression passt eine Gerade so an die Daten an, dass sie die Punkte möglichst gut beschreibt.

Typische Fragen sind:

- Wie verändert sich das Prüfungsergebnis, wenn die Lernzeit steigt?
- Wie kann man y mit Hilfe von x schätzen?
- Wie stark streuen die beobachteten Werte um die Regressionsgerade?
- Wie viel Streuung kann das Modell erklären?



2 Regressionsgleichung

Die einfache lineare Regression verwendet die Gleichung:

$$\hat{y} = b_0 + b_1 x$$

Symbol	Bedeutung
$\hat{y}$	vorhergesagter Wert von y
x	erklärende Variable
b_0	Achsenabschnitt: vorhergesagter Wert bei $x = 0$
b_1	Steigung der Regressionsgerade

Interpretation der Steigung:
Wenn x um 1 Einheit steigt, verändert sich der vorhergesagte Wert $\hat{y}$ um b_1 Einheiten.

3 Steigung und Achsenabschnitt

Die Steigung wird so berechnet:

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Der Achsenabschnitt wird danach berechnet:

$$b_0 = \bar{y} - b_1\bar{x}$$

- $b_1 > 0$: positiver Zusammenhang.
- $b_1 < 0$: negativer Zusammenhang.
- $b_1 = 0$: keine lineare Veränderung von $\hat{y}$, wenn x steigt.

4 Vorhersage

Sobald b_0 und b_1 bekannt sind, kann man für einen neuen x-Wert einen Wert von y vorhersagen.

$$\hat{y} = b_0 + b_1 x$$

Beispiel:

$$\hat{y} = 48 + 4x$$

Wenn eine Person $x = 5$ Stunden lernt:

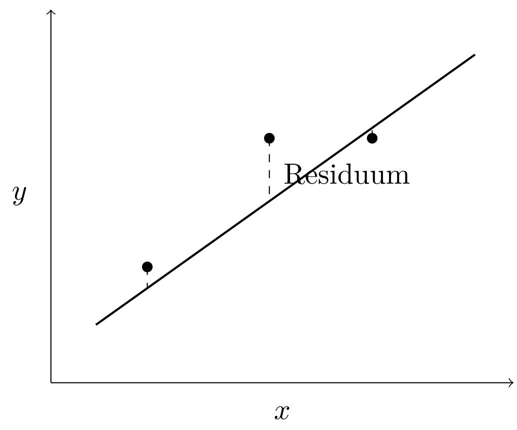
$$\hat{y} = 48 + 4 \cdot 5 = 68$$

Die vorhergesagte Punktzahl beträgt 68 Punkte.

5 Residuen

Ein Residuum ist der Unterschied zwischen dem beobachteten Wert und dem vorhergesagten Wert.

$$e_i = y_i - \hat{y}_i$$



- $e_i > 0$: Der beobachtete Wert liegt über der Gerade.
- $e_i < 0$: Der beobachtete Wert liegt unter der Gerade.
- $e_i = 0$: Der beobachtete Wert liegt genau auf der Gerade.

6 Bestimmtheitsmass R^2

Das Bestimmtheitsmass R^2 beschreibt, welcher Anteil der Streuung von y durch das lineare Regressionsmodell erklärt wird.

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

- $R^2 = 0$: Das Modell erklärt keine Streuung von y.
- $R^2 = 1$: Das Modell erklärt die Streuung von y vollständig.
- $R^2 = 0.64$: Das Modell erklärt 64% der Streuung von y.

Bei der einfachen linearen Regression gilt:

$$R^2 = r^2$$

7 Voraussetzungen und Grenzen

- Der Zusammenhang sollte ungefähr linear sein.
- Ausreisser können die Regressionsgerade stark beeinflussen.
- Vorhersagen ausserhalb des beobachteten Wertebereichs sind unsicher.
- Regression beschreibt einen Zusammenhang, beweist aber keine Kausalität.
- Ein gutes R^2 bedeutet nicht automatisch, dass das Modell sinnvoll ist.

8 Beispiel

Eine Lehrperson untersucht, ob Lernzeit x Prüfungspunkte y vorhersagen kann.

Person	x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	1	52	-2	-8	16
2	2	56	-1	-4	4
3	3	61	0	1	0
4	4	65	1	5	5
5	5	66	2	6	12
					37

$$\bar{x} = 3 \qquad \bar{y} = 60$$

$$\sum (x_i - \bar{x})^2 = 4 + 1 + 0 + 1 + 4 = 10$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{37}{10} = 3.7$$

$$b_0 = \bar{y} - b_1 \bar{x} = 60 - 3.7 \cdot 3 = 48.9$$

Die Regressionsgleichung lautet:

$$\hat{y} = 48.9 + 3.7x$$

Interpretation: Pro zusätzlicher Lernstunde steigt die vorhergesagte Punktzahl um 3.7 Punkte.
Vorhersage für $x = 6$:

$$\hat{y} = 48.9 + 3.7 \cdot 6 = 71.1$$

Für 6 Stunden Lernzeit werden etwa 71 Punkte vorhergesagt.

9 Zusammenfassung

Begriff	**Bedeutung**	**Formel**
Regressionsgerade	Vorhersage von y	$\hat{y} = b_0 + b_1 x$
Steigung	Veränderung pro Einheit x	b_1
Achsenabschnitt	Vorhersage bei $x = 0$	b_0
Residuum	Beobachtet minus vorhergesagt	$e_i = y_i - \hat{y}_i$
R^2	erklärter Streuungsanteil	$0 \leq R^2 \leq 1$

Aufgaben: Einfache lineare Regression

1 – Regressionsgleichung interpretieren

Gegeben ist:

$$\hat{y} = 35 + 2.5x$$

a) Interpretieren Sie $b_0 = 35$.

b) Interpretieren Sie $b_1 = 2.5$.

c) Berechnen Sie $\hat{y}$ für $x = 8$.

2 – Steigung und Achsenabschnitt berechnen

Berechnen Sie die Regressionsgleichung für die folgenden Daten.

Person	1	2	3
x	1	2	3
y	4	5	7

a) Berechnen Sie $\bar{x}$ und $\bar{y}$.

b) Berechnen Sie b_1.

c) Berechnen Sie b_0.

d) Schreiben Sie die Regressionsgleichung auf.

3 – Residuum

Ein Modell lautet:

$$\hat{y} = 10 + 4x$$

Bei $x = 3$ wurde $y = 25$ beobachtet.

a) Berechnen Sie $\hat{y}$.

b) Berechnen Sie das Residuum.

c) Liegt der beobachtete Wert über oder unter der Regressionsgerade?

4 – Multiple Choice

Welche Aussagen sind korrekt? (Mehrfachantworten möglich)

- ❑ Die Steigung gibt an, wie stark sich $\hat{y}$ verändert, wenn x um 1 steigt.
- ❑ Ein hohes R^2 beweist automatisch Kausalität.
- ❑ Residuen sind Unterschiede zwischen beobachteten und vorhergesagten Werten.
- ❑ Vorhersagen weit ausserhalb des beobachteten Wertebereichs sind besonders sicher.
- ❑ Bei der einfachen linearen Regression gilt $R^2 = r^2$.

Lösungen: Einfache lineare Regression

Lösung zu Aufgabe 1

$$\hat{y} = 35 + 2.5x$$

a) $b_0 = 35$: Wenn $x = 0$, wird $y = 35$ vorhergesagt.

b) $b_1 = 2.5$: Wenn x um 1 Einheit steigt, steigt $\hat{y}$ um 2.5 Einheiten.

c)
$$\hat{y} = 35 + 2.5 \cdot 8 = 55$$

Lösung zu Aufgabe 2

$$\bar{x} = \frac{1+2+3}{3} = 2 \qquad \bar{y} = \frac{4+5+7}{3} = 5.33$$

x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	4	-1	-1.33	1.33
2	5	0	-0.33	0.00
3	7	1	1.67	1.67
				3.00

$$\sum (x_i - \bar{x})^2 = 1^2 + 0^2 + 1^2 = 2$$

$$b_1 = \frac{3}{2} = 1.5$$

$$b_0 = \bar{y} - b_1\bar{x} = 5.33 - 1.5 \cdot 2 = 2.33$$

$$\hat{y} = 2.33 + 1.5x$$

Lösung zu Aufgabe 3

$$\hat{y} = 10 + 4 \cdot 3 = 22$$

$$e = y - \hat{y} = 25 - 22 = 3$$

Der beobachtete Wert liegt 3 Einheiten über der Regressionsgerade.

Lösung zu Aufgabe 4

✓ Richtig – Die Steigung beschreibt die Veränderung von $\hat{y}$ pro Einheit x.

✗ Falsch – Regression beweist keine Kausalität.

✓ Richtig – Residuen sind beobachtete minus vorhergesagte Werte.

✗ Falsch – Vorhersagen ausserhalb des beobachteten Bereichs sind unsicher.

✓ Richtig – Bei einfacher linearer Regression gilt $R^2 = r^2$.