



Allgemeines Lineares Modell

BSc Psychologie SoSe 2023

Prof. Dr. Dirk Ostwald

(1) Regression

Methode der kleinsten Quadrate

Literaturhinweise

Referenzen

Methode der kleinsten Quadrate

Einfache lineare Regression

Selbstkontrollfragen

Anwendungsszenario

Psychotherapie



Mehr Therapiestunden

⇒ Höhere Wirksamkeit?

Unabhängige Variable

- Anzahl Therapiestunden

Abhängige Variable

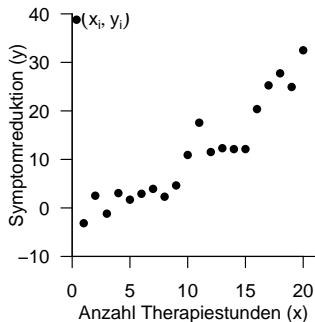
- Symptomreduktion

Beispieldatensatz

$i = 1, \dots, 20$ Patient:innen, y_i Symptomreduktion bei Patient:in i , x_i Anzahl Therapiestunden von Patient:in i

y_i	x_i
-3.15	1
2.52	2
-1.18	3
3.06	4
1.70	5
2.91	6
3.92	7
2.31	8
4.63	9
10.91	10
17.56	11
11.52	12
12.31	13
12.12	14
12.13	15
20.37	16
25.26	17
27.75	18
24.93	19
32.49	20

Beispieldatensatz



Welcher funktionale Zusammenhang zwischen x und y liegt den Daten zugrunde?

Definition (Ausgleichsgerade)

Für $\beta := (\beta_0, \beta_1)^T \in \mathbb{R}^2$ heißt die linear-affine Funktion

$$f_\beta : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f_\beta(x) := \beta_0 + \beta_1 x, \quad (1)$$

für die für einen Datensatz $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}^2$ die Funktion

$$q : \mathbb{R}^2 \rightarrow \mathbb{R}_{\geq 0}, \beta \mapsto q(\beta) := \sum_{i=1}^n (y_i - f_\beta(x_i))^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \quad (2)$$

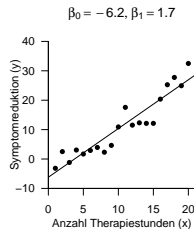
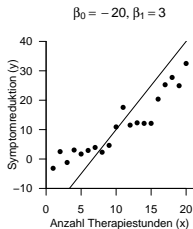
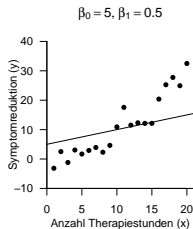
der quadrierten vertikalen Abweichungen der y_i von den Funktionswerten $f_\beta(x_i)$ ihr Minimum annimmt, die *Ausgleichsgerade für den Datensatz* $\{(x_1, y_1), \dots, (x_n, y_n)\}$.

Bemerkungen

- Wir nehmen hier ohne Beweis an, dass das Minimum von q eindeutig ist.

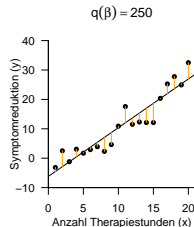
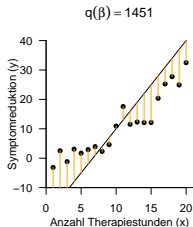
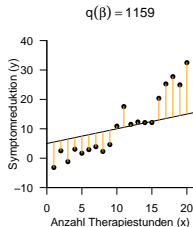
Linear-affine Funktionen $f_{\beta}(x) := \beta_0 + \beta_1 x$

- β_0 : Schnittpunkt von Gerade und y -Achse ("Offset Parameter")
- β_1 : y -Differenz pro x -Einheitsdifferenz ("Steigungsparameter")



Funktion der quadrierten vertikalen Abweichungen

$$q(\beta) := \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \quad (3)$$



— $y_i - (\beta_0 + \beta_1 x_i)$ für $i = 1, \dots, n$

Theorem (Ausgleichsgerade)

Für einen Datensatz $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}^2$ hat die Ausgleichsgerade die Form

$$f_\beta : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f_\beta(x) := \hat{\beta}_0 + \hat{\beta}_1 x, \quad (4)$$

wobei mit der Stichprobenkovarianz c_{xy} der (x_i, y_i) -Werte, der Stichprobenvarianz s_x^2 der x_i -Werte und den Stichprobenmitteln \bar{x} und \bar{y} der x_i - und y_i -Werte, respektive, gilt, dass

$$\hat{\beta}_1 = \frac{c_{xy}}{s_x^2} \text{ und } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (5)$$

Bemerkungen

- Mit den Definitionen von c_{xy} und s_x^2 gilt also

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (6)$$

- Man spricht hier von der Stichprobenkovarianz c_{xy} , auch wenn die Werte x_1, \dots, x_n oft nicht als Realisierungen einer Stichprobe ξ_1, \dots, ξ_n verstanden werden, sondern als gegebene Zahlen.

Beispieldatensatz Analyse

```
# Einlesen des Beispieldatensatzes
fname      = file.path("./Daten/Regression.csv")
D          = read.table(fname, sep = ",", header = TRUE)

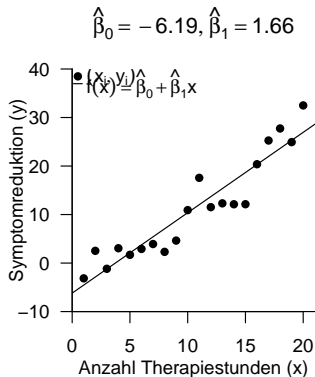
# Stichprobenstatistiken
x_bar      = mean(D$x_i)           # Stichprobenmittel der x_i-Werte
y_bar      = mean(D$y_i)           # Stichprobenmittel der y_i-Werte
s2x        = var(D$x_i)            # Stichprobenvarianz der x_i-Werte
cxy        = cov(D$x_i, D$y_i)     # Stichprobenkovarianz der (x_i,y_i)-Werte

# Ausgleichsgeradenparameter
beta_1_hat = cxy/s2x               #  $\hat{\beta}_1$ , Steigungsparameter
beta_0_hat = y_bar - beta_1_hat*x_bar #  $\hat{\beta}_0$ , Offset Parameter

# Ausgabe
cat("beta_0_hat:", beta_0_hat,
    "\nbeta_1_hat:", beta_1_hat)

> beta_0_hat: -6.19
> beta_1_hat: 1.66
```

Beispieldatensatz Visualisierung



Die Idee der Minimierung einer Summe von quadrierten Abweichungen bei der Anpassung einer Polynomfunktion an beobachtete Werte geht auf die Arbeiten von Legendre (1805) und Gauss (1809) im Kontext der Bestimmung von Planetenbahnen zurück. Eine historische Einordnung dazu gibt Stigler (1981). Der Begriff der Regression geht zurück auf Galton (1886). Stigler (1986) gibt dazu einen ausführlichen historischen Überblick.

- Galton, Francis. 1886. "Regression Towards Mediocrity in Hereditary Stature." *The Journal of the Anthropological Institute of Great Britain and Ireland* 15: 246. <https://doi.org/10.2307/2841583>.
- Gauss, Carl Friedrich. 1809. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. Cambridge: Cambridge University Press.
- Legendre, A. M. 1805. *Nouvelles Methodes Pour La Determination Des Orbites Des Cometes*. Didot Paris.
- Stigler, Stephen M. 1981. "Gauss and the Invention of Least Squares." *The Annals of Statistics* 9 (3). <https://doi.org/10.1214/aos/1176345451>.
- . 1986. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, Mass: Belknap Press of Harvard University Press.