Regression

Dirk Ostwald

2024-04-22

Fundamentales Ziel von Regressionsanalysen ist es, Beziehungen zwischen unabhängigen und abhängigen Variablen zu modellieren. Ein zentrales Thema dabei ist die Anpassung von Funktionen an beobachtete Datensätze. Mit dem Begriff der *Ausgleichsgerade* im Rahmen der *Methode der kleinsten Quadrate* und dem Begriff der *einfachen linearen Regression* wollen wir uns in diesem Abschnitt diesen zentralen Themen der probabilistischen Datenmodellierung schrittweise nähern. Dabei unterscheiden sich die Konzepte von Ausgleichsgerade und einfacher lineare Regression in einem zentralem Aspekt: bei der Ausgleichsgerade werden unabhängige und abhängige Variable nicht als Zufallsvariablen modelliert, im Rahmen der einfachen linearen Regression nimmt die abhängige Variable dann die Form einer Zufallsvariablen an. Im Kontext der Korrelation schließlich werden sowohl abhängige als auch unabhängige Variable als Zufallsvariablen modelliert.

Um die Konzepte dieses Abschnittes zu verdeutlichen, betrachten wir einen Beispieldatensatz in dem die Anzahl an Psychotherapiestunden als unabhängige Variable x der Symptomreduktion einer Gruppe von $n = 20$ Patient:innen als abhängige Variable y gegenüber gestellt wird (Abbildung 1). Die visuelle Inspektion dieses Datensatzes legt nahe, dass ein Mehr an Therapiestunden ein Mehr an Symptomreduktion impliziert. Ziel der Methode der kleinsten Quadrate und der einfachen linearen Regression ist es, diesen intuitiven funktionalen Zusammenhang zwischen unabhängiger und abhängiger Variable auf eine quantitative Basis zu stellen.

Methode der kleinsten Quadrate

Wir definieren zunächst den Begriff der *Ausgleichsgerade*.

Definition 1.1 (Ausgleichsgerade)**.** Für $\beta := (\beta_0, \beta_1)^T \in \mathbb{R}^2$ heißt die linear-affine Funktion

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

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

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

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

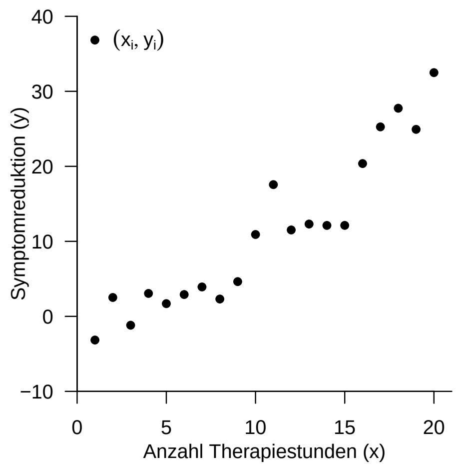

Abbildung 1. Beispieldatensatz

Bei der Ausgleichsgerade handelt es sich also um eine *linear-affine Funktion* der Form

$$f_\beta : \mathbb{R} \to \mathbb{R}, x \mapsto f_\beta(x) := \beta_0 + \beta_1 x. \tag{3}$$

Abbildung 2 zeigt drei durch jeweils andere Werte von β_0 und β_1 parameterisierte linear-affine Funktionen zusammen mit der Wertemenge des Beispieldatensatzes.

Wie bei allen linear-affinen Funktionen entspricht bei f_β der Wert von β_0 dem Wert, den f_β für $x = 0$ annimmt,

$$f_\beta(0) = \beta_0 + \beta_1 \cdot 0 = \beta_0 \tag{4}$$

und damit graphisch dem Schnittpunkt des Funktionsgraphen mit der y-Achse. Da β_0 damit dem Versatz (engl. *offset*) des Funktionsgraphen von $y = 0$ an der Stelle $x = 0$ entspricht, nennt man β_0 auch häufig den *Offsetparameter*. Analog entspricht wie bei allen linear-affinen Funktionen der Wert von β_1 dem Wert der Funktionswertdifferenz pro Argumenteinheitsdifferenz. Beispielsweise gilt etwa für $\beta_0 = 5$ und $\beta_1 = 0.5$, dass

$$\begin{aligned} f_\beta(2) - f_\beta(1) &= (5 + 0.5 \cdot 2) - (5 + 0.5 \cdot 1) = 1 - 0.5 = 0.5 \\ f_\beta(9) - f_\beta(8) &= (5 + 0.5 \cdot 9) - (5 + 0.5 \cdot 8) = 9.5 - 8 = 0.5 \end{aligned} \tag{5}$$

Für eine Argumentdifferenz von 1 ergibt sich also eine Funktionswertdifferenz von 0.5. β_1 enkodiert also die Stärke der Änderung der Funktionswerte pro Argumentseinheitsdifferenz und damit die Steigung (engl. *slope*) des Graphen der linear-affinen Funktion. Entsprechend wird β_1 *Steigungsparameter* oder *Slopeparameter* genannt.

Nach Definition ist die Ausgleichsgerade nun allerdings nicht eine beliebige linear-affine Funktion der Form f_β, sondern eben jene, die für einen gegebenen Datensatze $\{(x_1, y_1), ..., (x_n, y_n)\}$ die Summe der quadrierten vertikalen Abweichnungen

$$q(\beta) := \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{6}$$

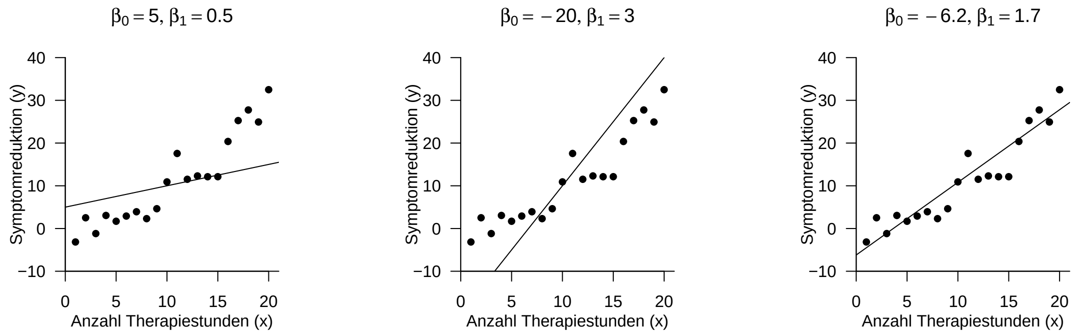

Abbildung 2. Linear-affine Funktionen mit unterschiedlichen Parameterwerten vor dem Hintegrund des Beispieldatensatzes

minimiert. Für eine fest vorgegebenen Datensatz von (x_i, y_i) Paaren ist der Wert dieser Summe abhängig von den Werten von β_0 und β_1 und kann deshalb durch Wahl geeigneter Werte von β_0 und β_1 minimiert werden. Da hierbei eine Summe von quadrierten Abweichungen zwischen Datenpunkten und Werten der Ausgleichsgerade minimiert wird, spricht man auch oft etwas ungenau von der *Methode der kleinsten Quadrate* (engl. *method of least squares*). Abbildung 3 zeigt die vertikalen Abweichungen zwischen y_i und $\beta_0 + \beta_1 x_i$ für $i = 1, ..., n$ des Beispieldatensatzes als orange Linien sowie die Summe ihrer Quadrate $q(\beta)$ im Titel. Für die Parameterwerte $\beta_0 = -6.2$ und $\beta_1 = 1.7$ (vgl. Abbildung 2) nimmt diese Summe ihren kleinsten Wert an.

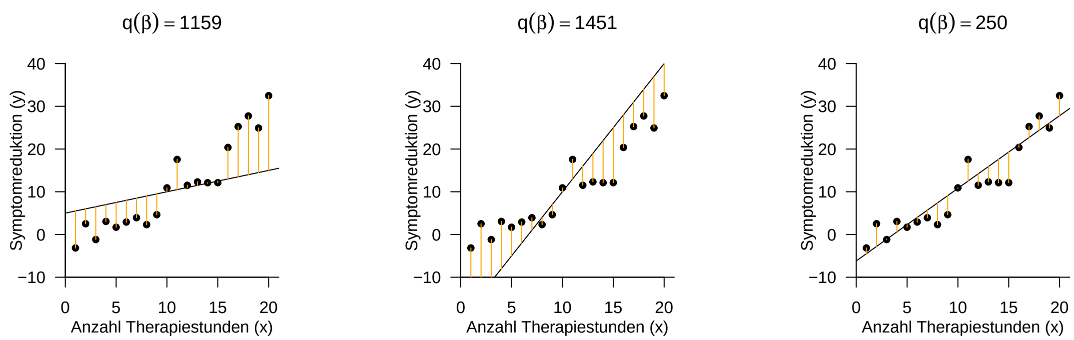

Abbildung 3. Vertikale Abweichungen und Quadratsummen bei unterschiedlichen Parameterwerten

Konkrete Formeln zur Bestimmung der Parameterwerte der Ausgleichsgerade stellt Theorem 1.1 bereit.

Theorem 1.1 (Ausgleichsgerade)**.** *Für einen Datensatz* $\{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathbb{R}^2$ *hat die Ausgleichsgerade die Form*

$$f_\beta : \mathbb{R} \to \mathbb{R}, x \mapsto f_\beta(x) := \hat{\beta}_0 + \hat{\beta}_1 x, \tag{7}$$

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*

$$\hat{\beta}_1 = \frac{c_{xy}}{s_x^2} \text{ und } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{8}$$

Beweis. □

Theorem 1.1 besagt, dass die Parameterwerte, die für einen gegebenen Datensatz $\{(x_1, y_1), ..., (x_n, y_n)\}$ die Summe der quadrierten vertikalen Abweichungen für eine linear-affine Funktion minimieren mithilfe der Stichprobenmittel der x_i- und y_i-Werte, der Stichprobenvarianz der x_i-Werte und der Stichprobenkovarianz der x_i- und y_i-Werte berechnet werden können. Die Terminologie orientiert sich hier an den Begrifflichkeiten der deskriptiven Statistik, insbesondere werden die x_i-Werte häufig *nicht* als Realisationen von Zufallsvariablen verstanden, der Begriff der Stichprobe wird jedoch trotzdem verwendet. Aus der Anwendungsperspektive können nach Theorem 1.1 die Parameter der Ausgleichsgerade also mithilfe der bekannten Funktionen für die Auswertung deskriptiver Statistiken bestimmt werden. Folgender **R** Code demonstriert dies.

```r
# 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: -6.194704
beta_1_hat: 1.657055
```

Eine typische Visualisierung der Ausgleichsgerade eines Datensatzes wie in Abbildung 4 implementiert folgender **R** Code.

```r
# Visualisierung der Datenwerte als Punktwolke
plot(
  D$x_i,
  D$y_i,
  pch         = 16,
  xlab        = "Anzahl Therapiestunden (x)",
  ylab        = "Symptomreduktion (y)",
  xlim        = c(0,21),
  ylim        = c(-10, 40),
  main        = TeX("$\\hat{\\beta}_0 =  -6.19, \\hat{\\beta}_1 = 1.66$")
)

# Ausgleichsgerade
abline(
  coef        = c(beta_0_hat, beta_1_hat),
  lty         = 1,
  col         = "black"
)

# Legende
legend(
  "topleft",
  c(TeX("$(x_i,y_i)$"), TeX("$f(x) = \\hat{\\beta}_0 + \\hat{\\beta}_1x$")),
  lty        = c(0,1),
  pch        = c(16, NA),
  bty        = "n"
)
```

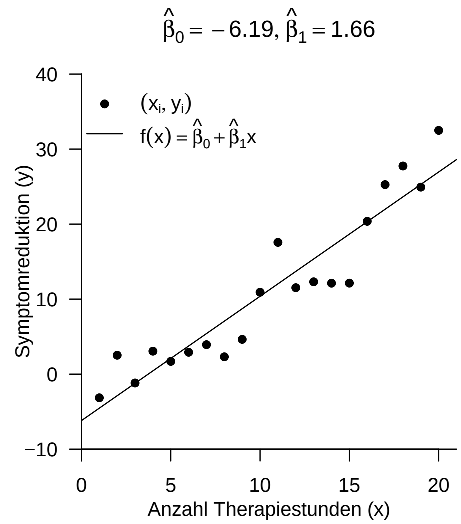

Abbildung 4. Ausgleichsgerade für den Beispieldatensatz

Literaturhinweise

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.

Referenzen

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.