Tutorium Allgemeines Lineares Modell

BSc Psychologie SoSe 2022

3. Termin: Regression (Teil 2)

Belinda Fleischmann

Wie sieht ein Summe aus Quadratfunktionen aus?

Zur Erinnerung:

Definition (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 eine Wertemenge $\{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathbb{R}^2$ 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, die *Ausgleichsgerade für die Wertemenge* $\{(x_1, y_1), ..., (x_n, y_n)\}$.

Selbstkontrollfragen - Regression

8. Erläutern Sie die Motivation des einfachen linearen Regressionsmodells in Bezug auf die Ausgleichsgerade.
9. Definieren Sie das Generative Modell der einfachen linearen Regression.
10. Geben Sie das Theorem zum Normalverteilungsmodell der einfachen linearen Regression wieder.
11. Skizzieren das Modell der einfachen linearen Regression per Hand.
12. Skizzieren Sie eine Realisierung des Modells der einfachen linearen Regression per Hand.
13. Geben Sie das Theorem zur ML-Schätzung der Parameter der einfachen linearen Regression an.
14. Skizzieren Sie den Beweis des Theorems zur ML-Schätzung der Parameter der einfachen linearen Regression.
15. Bestimmen Sie mithilfe eines R Skripts die ML Parameterschätzer des einfachen linearen Regressionsmodells für den bereitgestellten Beispieldatensatz und visualisieren die entsprechende Regressionsgerade wie für die Ausgleichsgerade gezeigt. Geben Sie weiterhin die Bedeutung der geschätzten Parameterwerte $\hat{\beta}_0$ und $\hat{\beta}_0$ an.

8. Erläutern Sie die Motivation des einfachen linearen Regressionsmodells in Bezug auf die Ausgleichsgerade.

Eine Ausgleichsgerade erlaubt Aussagen über unbeobachtete y Werte für x Werte. Der Wert von $q(\hat{\beta})$ quantifiziert die Güte der Ausgleichsgeradenpassung. Eine Ausgleichsgerade erlaubt allerdings nur implizite Aussagen über die mit der Anpassung verbundene Unsicherheit.

In der einfachen linearen Regression wird die Idee einer Ausgleichsgerade um eine probabilistische Komponente (normalverteilte Fehlervariable) erweitert, um quantitative Aussagen über die mit einer Ausgleichsgeradenanpassung verbundene Unsicherheit machen zu können.

Weiterhin erlaubt die einfache lineare Regression, einen Hypothesentest-basierten Zugang zur Einschätzung der angepassten Parameterwerte $\hat{\beta}_0$ und $\hat{\beta}_1$ sowie das Bestimmen von Konfidenzintervallen, die eine quantitative Aussage über die mit dem Schätzwert assoziierte Unsicherheit ermöglichen.

9. Definieren Sie das Generative Modell der einfachen linearen Regression.

Definition (Generatives Modell der einfachen linearen Regression)

Für $i = 1, ..., n$ sei

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{3}$$

wobei

- $x_i \in \mathbb{R}$ fest vorgegebene sogenannte *Prädiktorwerte* oder *Regressorwerte* sind,
- $\beta_0, \beta_1 \in \mathbb{R}$ wahre, aber unbekannte, Parameterwerte sind und
- $\varepsilon_i \sim N(0, \sigma^2)$ unabhängige und identisch normalverteilte nicht-beobachtbare Zufallsvariablen mit wahrem, aber unbekanntem, Parameter $\sigma^2 > 0$ sind.

Dann heißt (3) *Generatives Modell der einfachen linearen Regression*.

10. Geben Sie das Theorem zum Normalverteilungsmodell der einfachen linearen Regression wieder.

Theorem (Normalverteilungsmodell der einfachen linearen Regression)

Das generative Modell der einfachen linearen Regression

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, \ldots, n \tag{4}$$

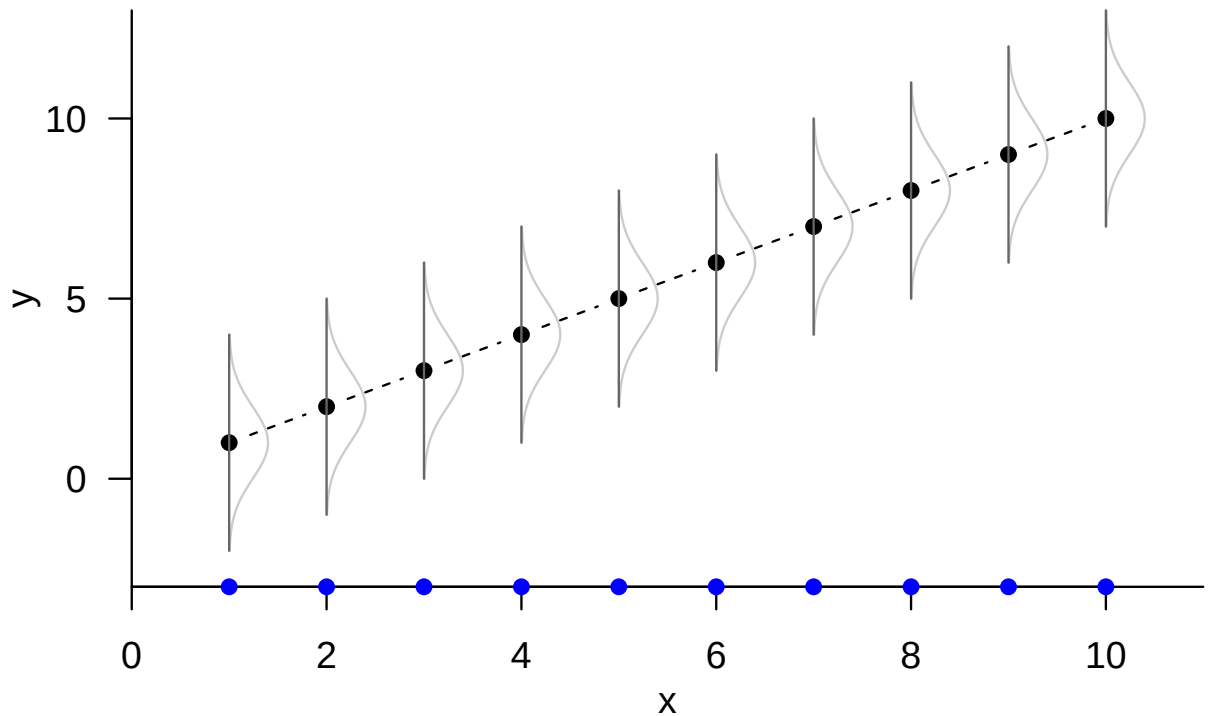
lässt sich äquivalent in der Form

$$Y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right) \text{ u. für } i = 1, \ldots, n \tag{5}$$

schreiben.

11. Skizzieren das Modell der einfachen linearen Regression per Hand.

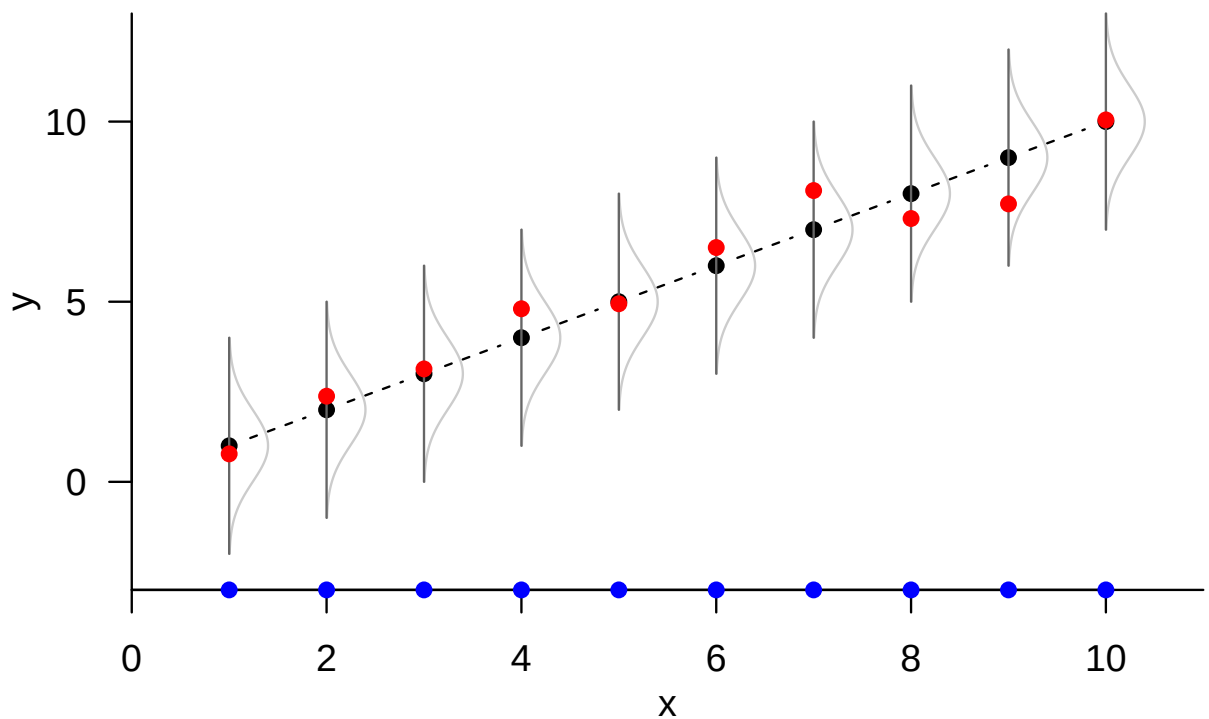
Modell der einfachen linearen Regression



• x_i • $\beta_0 + \beta_1 x_i$ für $\beta_0 := 0, \beta_1 := 1$ — $N(y_i; \beta_0 + \beta_1 x_i, \sigma^2)$ für $\sigma^2 := 1$.

12. Skizzieren Sie eine Realisierung des Modells der einfachen linearen Regression per Hand.

Realisierung des Modells der einfachen linearen Regression



• x_i • $\beta_0 + \beta_1 x_i$ für $\beta_0 := 0, \beta_1 := 1$ — $N(y_i; \beta_0 + \beta_1 x_i, \sigma^2)$ für $\sigma^2 := 1$ • (x_i, y_i)

13. Geben Sie das Theorem zur ML-Schätzung der Parameter der einfachen linearen Regression an.

Theorem (Maximum Likelihood Schätzung)

Es sei

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, \ldots, n \tag{6}$$

das Modell der einfachen linearen Regression. Dann sind Maximum Likelihood Schätzer der Modellparameter β_0, β_1 und σ^2 gegeben durch

$$\hat{\beta}_1 := \frac{c_{xy}}{s_x^2}, \quad \hat{\beta}_0 := \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{und } \hat{\sigma}^2 := \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right)^2. \tag{7}$$

14. Skizzieren Sie den Beweis des Theorems zur ML-Schätzung der Parameter der einfachen linearen Regression.

Teil 1/2: $\hat{\beta}_0$ und $\hat{\beta}_1$

Wir wollen zunächst zeigen, dass die Ausgleichsgerardenparameter $\hat{\beta}_0$ und $\hat{\beta}_1$ den entsprechenden ML Schätzern gleichen.

Um die ML Schätzer zu bestimmen, formulieren wir zunächst die Likelihood-Funktion des Modells der einfachen linearen Regression in Abhängigkeit von β_0 und β_1.

Die Likelihood-Funktion ist definiert als der Wert der gemeinsamen Verteilung der $Y_1, \ldots, Y_n$ in Abhängigkeit von den Parametern β_0 und β_1.

Aufgrund der Unabhängigkeit der $Y_1, \ldots, Y_n$ können wir die gemeinsame Verteilung als Produkt der einzelnen Wahrscheinlichkeitsdichtefunktionen, also als Produkt von Dichtefunktionen der univariaten Normalverteilung aufschreiben.

Die funktionale Form der Dichtefunktionen der univariaten Normalverteilung enthält eine Exponentialfunktion. Mit den Eigenschaften einer Exponentialfunktion können wir dieses Produkt umschreiben zu einer Exponentialfunktion von einem Term, der im Wesentlichen aus der negativen Summe der quadrierten Abweichungen (i.e. der Funktion q) besteht.

Weil für eine Exponentialfunktion gilt, dass für $a < b \leq 0$ gilt, dass $\exp(a) < \exp(b)$, wird der Exponentialterm der Likelihood-Funktion maximal, wenn q minimal und entsprechend $-q$ maximal wird.

Wie im Beweis der Ausgleichsgeradenform gezeigt, wissen wir, dass q für $\hat{\beta}_0$ und $\hat{\beta}_1$, wie sie auch im Theorem zur ML-Schätzung der Parameter der einfachen linearen Rregression angegeben sind, minimal wird, und damit $\hat{\beta}_1$ und $\hat{\beta}_0$ die Likelihood-Funktion maximieren.

14. Skizzieren Sie den Beweis des Theorems zur ML-Schätzung der Parameter der einfachen linearen Regression.

Teil 2/2: $\hat{\sigma^2}$

Als nächstes wollen wir zeigen, dass $\hat{\sigma^2}$ dem ML-Schätzer entspricht.

Dazu betrachten wir analog zu oben die Likelihood-Funktion des Modells der einfachen linearen Regression, jedoch als Funktion von σ^2 und formulieren die entsprechende log-Likelihood-Funktion.

Wir wollen das $\hat{\sigma^2}$ bestimmen, für das die (log-)Likelihood-Funktion maximial wird.

Um die log-Likelihood-Funktion zu maximieren, bilden wir die 1. Ableitung, setzen diese gleich 0 und lösen nach σ^2 auf. Durch umstellen erhalten wir dann die Formel zur Schätzung von σ^2, also $\hat{\sigma^2}$, wie sie im Theorem angeben ist.

15. Bestimmen Sie mithilfe eines R Skripts die ML Parameterschätzer des einfachen linearen Regressionsmodells für den bereitgestellten Beispieldatensatz und visualisieren die entsprechende Regressionsgerade wie für die Ausgleichsgerade gezeigt. Geben Sie weiterhin die Bedeutung der geschätzten Parameterwerte $\hat{\beta}_0$ und $\hat{\beta}_0$ an.

15. Bestimmen Sie mithilfe eines R Skripts die ML Parameterschätzer . . .

Lösungsmöglichkeit A) mithilfe der Formeln für Stichprobenstatistik und ML Schätzer

(vollständiges Skript in der Datei Lösung_Aufg_15.R)

```
# Einlesen des Beispieldatensatzes
fname       = file.path(getwd(), "1_Daten", "1_Regression.csv")
D           = read.table(fname, sep = ",", header = TRUE)

# Stichprobenstatistiken
n           = length(D$y_i)                                    # Anzahl Datenpunkte
x_bar       = mean(D$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

# Parameteterschätzer (nach dem Theorem der ML Schätzung)
beta_1_hat  = cxy/s2x                                          # \hat{\beta}_1, Steigungsparameter
beta_0_hat  = y_bar - beta_1_hat*x_bar                         # \hat{\beta}_0, Offset Parameter
sigsqr_hat  = (1/n)*sum((D$y_i-(beta_0_hat+beta_1_hat*D$x_i))^2) # Varianzparameter

# Ausgabe
cat("beta_0_hat:"  , beta_0_hat,
    "\nbeta_1_hat:", beta_1_hat,
    "\nsigsqr_hat:", sqrt(sigsqr_hat))
```

```
> beta_0_hat: -6.19
> beta_1_hat: 1.66
> sigsqr_hat: 3.54
```

15. Bestimmen Sie mithilfe eines R Skripts die ML Parameterschätzer . . .

Lösungsmöglichkeit B) mithilfe der Funktion `lm()` der R package (`car`)

(vollständiges Skript in der Datei Lösung_Aufg_15.R)

```r
# Laden des 'car' package
library(car)

# Einlesen des Beispieldatensatzes
fname       = file.path(getwd(), "1_Daten", "1_Regression.csv")
D           = read.table(fname, sep = ",", header = TRUE)

# Analyse mit lm()
model       = lm(formula = D$y_i ~ D$x_i, data = D)
print(model)
```

```
>
> Call:
> lm(formula = D$y_i ~ D$x_i, data = D)
>
> Coefficients:
> (Intercept)        D$x_i
>       -6.19         1.66
```

15. Bestimmen Sie mithilfe eines R Skripts die ML Parameterschätzer . . .

R Code zur Visualisierung

(vollständiges Skript in der Datei Lösung_Aufg_15.R)

```r
# Datenwerte
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,  # linetype (0: Punkt, 1:Linie, 2:gestrichelte Linie, usw.)
  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),  # plot character (0: Vierreck, 1:Kringel, ..., 16: ausgefüllter Kreis, usw.)
  bty         = "n")  # boxtype ("o": komplette box, "n": keine box,... usw.)
```

15. Bestimmen Sie mithilfe eines R Skripts die ML Parameterschätzer ...

Visualisierung