2.5 Inferenz bei der linearen Regression

2.5.1 Erinnerung an das Grundmodell

Statistik I

Modellbildung für jedes i

Beispiele

Beispiel 4.11 (Statistik I): Kaffeeverkauf auf drei Flohmärkten
X Anzahl verkaufter Tassen Kaffee
Y zugehöriger Gewinn (Preis Verhandlungssache)
Man bestimme die Regressionsgerade und interpretiere die erhaltenen KQ-Schätzungen! Welcher Gewinn ist bei zwölf verkauften Tassen zu erwarten?

i	x_i	y_i			
1	10	9			
2	15	21			
3	5	0			

Beispiel 4.12 (Statistik I):

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

$$\text{mit} \quad X_1 = \begin{cases} 1 & \text{männlich} \\ 0 & \text{weiblich} \end{cases}$$

$X_2 = \text{(vertragliche) Arbeitszeit}$

$Y = \text{Einkommen}$

Interpretation:

Beispiel zur Dummycodierung (Statistik I)

Nominales Merkmal mit q Kategorien:
z.B X = Parteipräferenz

$$X = \begin{cases} 1 & \text{CDU/CSU oder FDP} \\ 2 & \text{SPD oder Grüne} \\ 3 & \text{Sonstige} \end{cases}$$

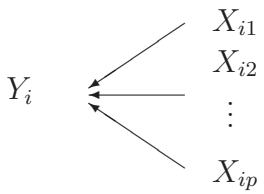
Man darf X nicht einfach mit Werten 1 bis 3 besetzen, da es sich um ein nominales Merkmal handelt

Beispiel zur Interpretation:

Y: Score auf Autoritarismusskala

X bzw. X_1, X_2: Parteienpräferenz

X_3: Einkommen



abhängige Variable	unabhängige Variable
metrisch/quasistetig	metrische/quasistetige oder dichotome (0/1) Variablen (kategoriale Variablen mit mehr Kategorien → Dummy-Kodierung)

Ansatz: linearer Zusammenhang, ermittle aus den Daten „Wirkungsstärke“ der einzelnen Variablen

2.5.2 Lineare Einfachregression

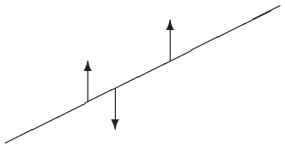
Eine unabhängige Variable:

Statistische Sichtweise:

- wahres Modell

$$y_i = \beta_0 + \beta_1 x_i$$

  β_0 Grundniveau
  β_1 “Elastizität”: Wirkung der Änderung von X_i um eine Einheit



- gestört durch zufällige Fehler ϵ_i

  Man beobachtet Datenpaare, (X_i, Y_i) $i = 1, \ldots, n$

  mit $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ (2.5.1)

  wobei $\epsilon_i \sim N(0, \sigma^2)$ (2.5.2)

  σ^2 für alle i gleich

  $\epsilon_{i1}, \epsilon_{i2}$ stochastisch unabhängig für $i_1 \neq i_2$

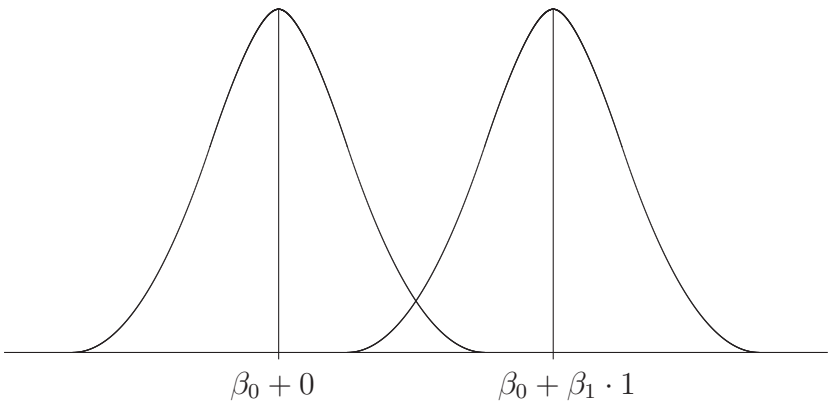
  X_i kann zufällig oder fest sein, hier als fest behandelt (sonst über bedingte Verteilung herleiten); schreibe x_i

  Wegen (2.5.1), (2.5.2) gilt für die bedingte Verteilung von Y_i gegeben $X_i = x_i$:

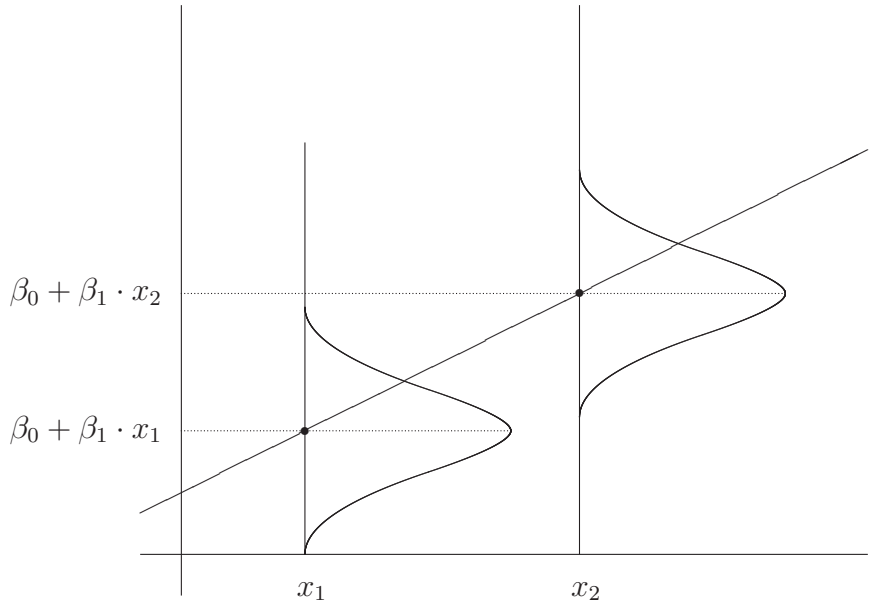
$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \quad i = 1, \ldots, n$$

  Interpretation: verschiedene Normalverteilungen jeweils mit verschobenem Mittelwert $\mu_i = \beta_0 + \beta_1 x_i$, aber gleicher Varianz.

<u>X dichotom:</u>



X metrisch:



Aufgabe: Schätze $\beta_0, \beta_1, \sigma^2$
Schätzwerte und Schätzfunktionen üblicherweise mit $\widehat{\beta_0}, \widehat{\beta_1}, \hat{\sigma}^2$ bezeichnet

Satz 2.22
In der eben beschriebenen Situation gilt:

i) Die Maximum Likelihood Schätzer lauten:

$$\widehat{\beta_1} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}, \tag{2.5.3}$$

$$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}, \tag{2.5.4}$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \tag{2.5.5}$$

$$\text{mit den „Residuen“ } \hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \tag{2.5.6}$$

ii) Mit $\widehat{\sigma_{\hat{\beta}_0}} := \dfrac{\hat{\sigma}\sqrt{\sum_{i=1}^{n} X_i^2}}{\sqrt{n\sum_{i=1}^{n}(X_i - \overline{X})^2}}$ ist

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\sigma_{\hat{\beta}_0}}} \sim t(n-2) \tag{2.5.7}$$

und mit $\widehat{\sigma}_{\hat{\beta}_1} := \dfrac{\widehat{\sigma}}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$ ist

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma_{\hat{\beta}_1}}} \sim t(n-2). \tag{2.5.8}$$

Bem 2.23

i) $\widehat{\beta}_0, \widehat{\beta}_1$ sind die *KQ*-Schätzer aus Statistik I. Unter Normalverteilung fällt hier das *ML*- mit dem *KQ*-Prinzip zusammen.

ii) Man kann unmittelbar Tests und Konfidenzintervalle ermitteln (völlig analog zum Vorgehen in Kapitel 2.3 und 2.4):
Konfidenzintervall zum Sicherheitsgrad γ

$$\begin{aligned} \text{für } \beta_0 &: \quad [\widehat{\beta}_0 \pm \widehat{\sigma}_{\widehat{\beta_0}} \cdot t_{1+\frac{\gamma}{2}}(n-2)] \\ \text{für } \beta_1 &: \quad [\widehat{\beta}_1 \pm \widehat{\sigma}_{\widehat{\beta_1}} \cdot t_{\frac{1+\gamma}{2}}(n-2)] \end{aligned}$$

Bzw: mit der Teststatistik

$$T_{\beta_1^*} = \frac{\widehat{\beta}_1 - \beta_1^*}{\widehat{\sigma_{\hat{\beta}_1}}}$$

ergibt sich

	Hypothesen			kritische Region
I.	$H_0 : \beta_1 \leq \beta_1^*$	gegen	$\beta_1 > \beta_1^*$	$T \geq t_{1-\alpha}(n-2)$
II.	$H_0 : \beta_1 \geq \beta_1^*$	gegen	$\beta_1 < \beta_1^*$	$T \leq t_{1-\alpha}(n-2)$
III.	$H_0 : \beta_1 = \beta_1^*$	gegen	$\beta_1 \neq \beta_1^*$	$\lvert T\rvert \geq t_{1-\frac{\alpha}{2}}(n-2)$

Analog für $\widehat{\beta_0}$.

Von besonderem Interesse ist der Fall $\beta_1^* = 0$:

Bem. 2.24 Typischer SPSS-Output bei Regression

Koeffizienten[a]

			Standardisierte Koeffizienten		
	β	Standardfehler	Beta	T	Signifikanz
Konstante	$\widehat{\beta_0}$	$\widehat{\sigma_0}$	5)	1)	3)
Unabhängige Variable	$\widehat{\beta_1}$	$\widehat{\sigma_1}$	6)	2)	4)

[a] abhängige Variable

1) Wert der Teststatistik

$$T_{\beta_0^*} = \frac{\widehat{\beta_0}}{\widehat{\sigma_0}}$$

Zum Testen von H_0: $\beta_0 = 0$ gegen H_1: $\beta_0 \neq 0$, also Fall $\beta_0^* = 0$

2) Analog: Wert von

$$T_{\beta_1^*} = \frac{\widehat{\beta_1}}{\widehat{\sigma_1}}$$

Zum Testen von H_0: $\beta_1 = 0$ gegen H_1: $\beta_1 \neq 0$, also Fall $\beta_1^* = 0$

3) p-Wert zu 1)

4) p-Wert zu 2)

5), 6) hier nicht von Interesse.

Interpretation
Werte von $\widehat{\beta_0}$, $\widehat{\beta_1}$ signifikant von 0 verschieden? $\longrightarrow$ Einfluss statistisch nachgewiesen ja/nein.

2.5.3 Multiple lineare Regression

- Analoger Ansatz, mit mehreren Variablen

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i$$

- Schätzung von $\beta_0, \beta_1, \ldots, \beta_p, \sigma^2$ sinnvollerweise über Matrixrechnung // Software $\rightarrow$ An dem SPSS Output sind $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ und $\widehat{\sigma}_{\widehat{\beta}_0} =: \widehat{\sigma}_0$, $\widehat{\sigma}_{\widehat{\beta}_1} =: \widehat{\sigma}_1, \ldots$ ablesbar.
  (Outputs lesen können ist absolut klausurrelevant! Matrixrechnung wird nicht verlangt.)

- Es gilt analog für jedes $j = 1, \ldots, p$

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}_j} \sim t(n-p-1)$$

  und man erhält wieder Konfidenzintervalle für β_j:

$$[\widehat{\beta}_j \pm \widehat{\sigma}_j \cdot t_{1+\frac{\gamma}{2}}(n-p-1)]$$

  und entsprechende Tests.

  Von besonderem Interesse ist wieder der Test

$$H_0 : \beta_j = 0, \; H_1 : \beta_j \neq 0.$$

  Der zugehörige p-Wert findet sich im SPSS-Ausdruck (Vorsicht mit Problematik multiplen Testens!).

- Man kann auch simultan testen:

$$\beta_1 = \beta_2 = \ldots = \beta_p = 0.$$

  Dies führt zu einem sog. F-Test ($\longrightarrow$ Software)

- Sind alle $X_{i,j}$ 0/1-wertig, so erhält man eine sog. *Varianzanalyse*, was dem Vergleich von mehreren Mittelwerten entspricht.
  Für Befragte mit $X_{ij} = 0$ für alle j gilt: Mittelwert von $Y = \beta_0$
  Ist $X_{i1} = 1$ und $X_{ij} = 0$ für $j \geq 2$, so ist der Mittelwert $\beta_0 + \beta_1$
  Ist $X_{i1} = 1$ und $X_{i2} = 1$, sowie $X_{ij} = 0$ für $j \geq 3$, so ist der Mittelwert $\beta_0 + \beta_1 + \beta_2$ ...
  (p=1: Zweistichprobenmittelwertsvergleich)

2.5.4 Varianzanalyse

(z.B Fahrmeir et al (2003), Kap 13; Bortz (1999[5]), Kap 7 ff.)

Vor allem in der angewandten Literatur wird die Varianzanalyse unabhängig vom Regressionsmodell entwickelt.
Ziel: Mittelwertvergleiche in mehreren Gruppen, häufig in (quasi-) experimentellen Situationen. (Verallgemeinerung des t-Tests; dort nur zwei Gruppen)

Hier nur *einfaktorielle Varianzanalyse* („Einfachklassifikation“)

typische Beispiele (Fahrmeir et al (2003), S.515)
Einstellung zu Atomkraft anhand eines Scores, nachdem ein Film gezeigt wurde
3 Gruppen („Faktorstufen“)

- Pro-Atomkraft-Film
- Contra-Atomkraft-Film
- ausgewogener Film

Name Varianzanalyse: Vergleich der Variabilität in und zwischen den Gruppen

Beobachtungen: Y_{ij}
$i = 1, \ldots, I$ Faktorstufen
$j = 1, \ldots, n_i$ Personenindex in der i-ten Faktorstufe

Zwei äquivalente Modellformulierungen

a) $Y_{ij} = \mu_i + \epsilon_{ij} \qquad i = 1, \ldots, I,\ j = 1, \ldots, n_i,$
mit

μ_i faktorspezifischer Mittelwert
ϵ_{ij} zufällige Störgröße
$\epsilon_{ij} \sim N(0, \sigma^2), \quad \epsilon_{11}, \epsilon_{12}, \ldots, \epsilon_{In_I}$ unabhängig

Typische Testsituation:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \ldots \mu_I \\ \text{gegen} & \\ H_1 &: \mu_l \neq \mu_q \quad \text{für mindestens ein Paar } (l, q) \end{aligned} \tag{2.5.9}$$

b) Modell in Effektdarstellung

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

wobei α_i so, dass $\sum_{i=1}^{I} n_i \alpha_i = 0$.

μ globaler Erwartungswert

α_i Effekt in der i-ten Faktorstufe, systematische Abweichung von μ

Testproblem

$$\begin{aligned} H_0 &: \alpha_1 = \alpha_2 = \ldots \alpha_I = 0 \\ \text{gegen} & \\ H_1 &: \alpha_i \neq 0 \text{ für mindestens ein } i \end{aligned} \tag{2.5.10}$$

Modelle sind äquivalent: setze $\mu_i := \mu + \alpha_i$

Streuungszerlegung (vgl. Statistik I)

Mittelwerte:

$\bar{Y}_{\bullet\bullet}$ Gesamtmittelwert in der Stichprobe

$\bar{Y}_{i\bullet}$ Mittelwert in der i-ten Faktorstufe

Es gilt: (vgl. Statistik I: $SQT = SQE + SQR$, mit dortigen $\widehat{y_i} = \bar{y}_{i\bullet}$)

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{I} \underbrace{n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2}_{= SQE} + \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{n_j}(Y_{ij} - \bar{Y}_{i\bullet})^2}_{= SQR}$$

Variabilität **der** Gruppen

Variabilität **in den** Gruppen

Die Testgröße

$$F = \frac{SQE/(I-1)}{SQR/(n-I)}$$

ist geeignet zum Testen der Hypothesen (2.5.9) und (2.5.10). Sie besitzt eine sog. F-Verteilung mit $(I-1)$ und $(n-I)$ Freiheitsgraden.

Die kritische Region besteht aus den *großen* Werten von F (Vorsicht: obwohl H_0 von „Gleichheitsform“); also H_0 ablehnen falls

$$T > F_{1-\alpha}(I-1, n-I),$$

dem entsprechenden $(1-\alpha)$-Quantil der F-Verteilung mit $(I-1)$ und $(n-I)$ Freiheitsgraden.

(„Klar“: Je größer die Variabilität zwischen den Gruppen im Vergleich zu der Variabilität in den Gruppen, desto unplausibler ist die Nullhypothese, dass alle Gruppenmittelwerte gleich sind.)

Bei Ablehnung oft dann noch von Interesse, welche Gruppen sich unterscheiden. (Allerdings wieder Problematik des multiplen Testens.)