Lineare Regressionsmodelle

Deskriptive Statistik:
Gegeben Datenpunkte (Y_i, X_i) schätze die beste Gerade
$Y_i = \beta_0 + \beta_1 X_i, \ i = 1, \ldots, n.$
(mit der Methode der kleinsten Quadrate)

Statistisches Modell

- linearer Zusammenhang.
- Im Folgenden: Probabilistische Modelle in Analogie zu den deskriptiven Modellen aus Statistik I

Lineare Einfachregression

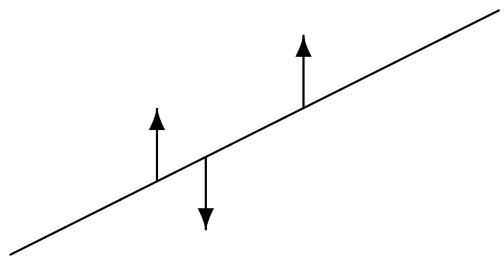
Zunächst Modelle mit nur *einer* unabhängigen Variable.
Statistische Sichtweise:

- Modell

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

  β_1 „Elastizität“: Wirkung der Änderung von X_i um eine Einheit

- gestört durch zufällige Fehler ϵ_i



Modellannahmen

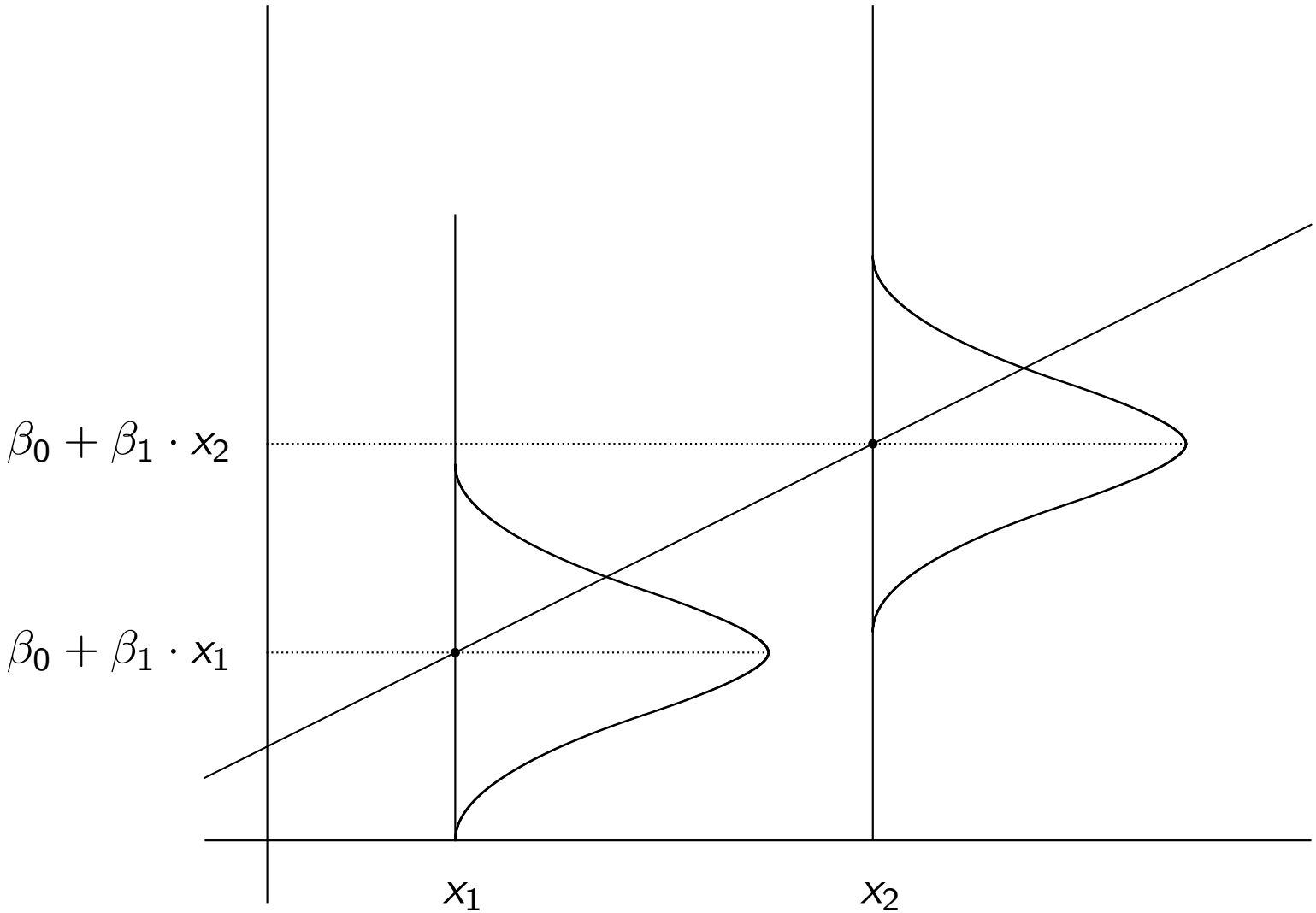
Man beobachtet Datenpaare, (X_i, Y_i), $i = 1, \ldots, n$ mit

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

wobei sich die Annahmne auf den zufälligen Störterm beziehen:

$E(\epsilon_i) = 0$
$Var(\epsilon_i) = \sigma^2$ für alle i gleich
$\epsilon_{i1}, \epsilon_{i2}$ stochastisch unabhängig für $i_1 \neq i_2$
$\epsilon_i \sim ND(0, \sigma^2)$ (zusätzlich, bei großen Stichproben nicht erforderlich)

Einfache lineare Regression



Schätzung der Parameter

Die Schätzwerte werden üblicherweise mit $\hat{\beta}_0, \hat{\beta}_1$ und $\hat{\sigma}^2$ bezeichnet.
In der eben beschriebenen Situation gilt:

- Die (Maximum Likelihood) Schätzer lauten:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \\
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}, \\
\hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2
\end{aligned}$$

mit den geschätzten Residuen

$$\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i.$$

Konstruktion von Testgröße

n

- Mit

$$\hat{\sigma}_{\hat{\beta}_0} := \frac{\hat{\sigma}\sqrt{\sum_{i=1}^n X_i^2}}{\sqrt{n\sum_{i=1}^n (X_i - \bar{X})^2}}$$

gilt

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t^{(n-2)}$$

und analog mit

$$\hat{\sigma}_{\hat{\beta}_1} := \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

gilt

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t^{(n-2)}.$$

Konfidenzintervalle

- $\hat{\beta}_0$ und $\hat{\beta}_1$ sind die *KQ*-Schätzer aus Statistik I. Unter Normalverteilung fällt hier das *ML*- mit dem *KQ*-Prinzip zusammen.
- Man kann unmittelbar Tests und Konfidenzintervalle ermitteln (völlig analog zum Vorgehen in Kapitel 2.3 und 2.4).

  Konfidenzintervalle zum Sicherheitsgrad γ:

$$\text{für } \beta_0 : \quad [\hat{\beta}_0 \pm \hat{\sigma}_{\hat{\beta}_0} \cdot t_{1+\frac{\gamma}{2}}^{(n-2)}]$$

$$\text{für } \beta_1 : \quad [\hat{\beta}_1 \pm \hat{\sigma}_{\hat{\beta}_1} \cdot t_{\frac{1+\gamma}{2}}^{(n-2)}]$$

Tests für die parameter des Modells

Mit der Teststatistik

$$T_{\beta_1^*} = \frac{\hat{\beta}_1 - \beta_1^*}{\hat{\sigma}_{\hat{\beta}_1}}$$

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 $\hat{\beta}_0$).
Von besonderem Interesse ist der Fall $\beta_1^* = 0$: (Steigung gleich 0) Hiermit kann man überprüfen, ob die $X_1, \ldots, X_n$ einen signifikanten Einfluss hat oder nicht.

Typischer Output

Koeffizienten[a]

			Standardisierte Koeffizienten		
	β	Standardfehler	Beta	T	Signifikanz
Konstante	$\hat{\beta}_0$	$\hat{\sigma}_{\hat{\beta}_0}$	5)	1)	3)
Unabhängige Variable	$\hat{\beta}_1$	$\hat{\sigma}_{\hat{\beta}_1}$	6)	2)	4)

Output Erklärung

1) Wert der Teststatistik

$$T_{\beta_0^*} = \frac{\hat{\beta}_0}{\hat{\sigma}_{\hat{\beta}_0}}.$$

zum Testen von H_0: $\beta_0 = 0$ gegen H_1: $\beta_0 \neq 0$.

2) Analog: Wert von

$$T_{\beta_1^*} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$$

zum Testen von H_0: $\beta_1 = 0$ gegen H_1: $\beta_1 \neq 0$.

3) p-Wert zu 1)

4) p-Wert zu 2)

5), 6) hier nicht von Interesse.

- Die Testentscheidung „$\hat{\beta}_1$ signifikant von 0 verschieden“ entspricht dem statistischen Nachweis eines Einflusses von X.

Das multiple Regressionsmodell

Beispiel: Arbeitszeit und Einkommen

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

mit

$$\begin{aligned} X_1 &= \begin{cases} 1 & \text{männlich} \\ 0 & \text{weiblich} \end{cases} \\ X_2 &= \text{(vertragliche) Arbeitszeit} \\ Y &= \text{Einkommen} \end{aligned}$$

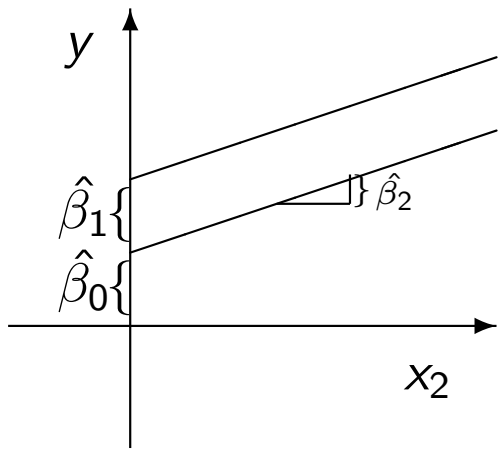
Interpretation:
Die geschätzte Gerade für die Männer lautet

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot x_{2i}$$

für die Frauen hingegen erhält man

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot x_{2i} \\ &= \hat{\beta}_0 + \hat{\beta}_2 \cdot x_{2i} \end{aligned}$$

Grundidee (ANCOVA)



Modellbildung

β_0 Grundlevel
β_2 durchschnittlicher Stundenlohn
β_1 Zusatzeffekt des Geschlechts zum Grundlevel.

Die 0-1 Variable dient als Schalter, mit dem man den Männereffekt an/abschaltet.

Dummykodierung

Nominales Merkmal mit q Kategorien, z.B. $X =$ Parteipräferenz mit

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

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

Dummycodierung (2)

Idee: Mache aus der einen Variable mit q (hier 3) Ausprägungen $q-1$ (hier 2) Variablen mit den Ausprägungen ja/nein ($\hat{=}0/1$). Diese *Dummyvariablen* dürfen dann in der Regression verwendet werden.

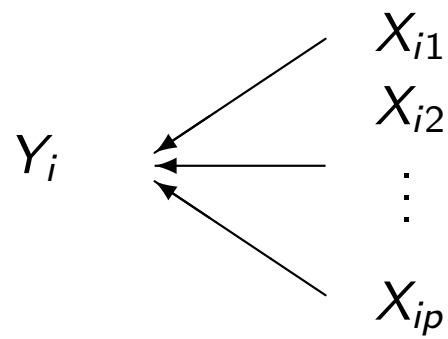
$$X_1 = \begin{cases} 1 & \text{CDU/CSU oder FDP} \\ 0 & \text{andere} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{SPD, Grüne} \\ 0 & \text{andere} \end{cases}$$

Durch die Ausprägungen von X_1 und X_2 sind alle möglichen Ausprägungen von X vollständig beschrieben:

X	Text	X_1	X_2
1	CDU/CSU, FDP	1	0
2	SPD, Grüne	0	1
3	Sonstige	0	0

Multiples Regressionsmodell



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

Multiple lineare Regression

- Analoger Modellierungsansatz, aber mit mehreren erklärenden Variablen:

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

- Schätzung von $\beta_0, \beta_1, \ldots, \beta_p$ und σ^2 sinnvollerweise über Matrixrechnung bzw. Software.

  Aus dem SPSS-Output sind $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ sowie $\hat{\sigma}_{\hat{\beta}_0}, \hat{\sigma}_{\hat{\beta}_1}, \ldots, \hat{\sigma}_{\hat{\beta}_p}$ ablesbar.

- Es gilt für jedes $j = 0, \ldots, p$

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

und man erhält wieder Konfidenzintervalle für β_j:

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

sowie 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 Ausdruck (Vorsicht mit Problematik des multiplen Testens!).
Man kann auch simultan testen, z.B.

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

Dies führt zu einem sogenannten F-Test ($\longrightarrow$ Software).
Sind alle X_{ij} 0/1-wertig, so erhält man eine sogenannte *Varianzanalyse*, was dem Vergleich von mehreren Mittelwerten entspricht.

- Für Befragte mit $X_{ij} = 0$ für alle j gilt:

$$\mathbb{E}(Y) = \beta_0$$

- Ist $X_{i1} = 1$ und $X_{ij} = 0$ für $j \geq 2$, so gilt

$$\mathbb{E}(Y) = \beta_0 + \beta_1$$

- Ist $X_{i1} = 1$ und $X_{i2} = 1$, sowie $X_{ij} = 0$ für $j \geq 3$, so gilt

$$\mathbb{E}(Y) = \beta_0 + \beta_1 + \beta_2$$

- etc.

Varianzanalyse (Analysis of Variance, ANOVA)

- Vor allem in der angewandten Literatur, etwa in der Psychologie, 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* (Eine Gruppierungsvariable).

Einstellung zu Atomkraft anhand eines Scores, nachdem ein Film gezeigt wurde.
3 Gruppen („Faktorstufen“):

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

Varianzanalyse: Vergleich der Variabilität in und zwischen den Gruppen
Beobachtungen: Y_{ij}

$$\begin{array}{ll} j = 1, \ldots, J & \text{Faktorstufen} \\ i = 1, \ldots, n_j & \text{Personenindex in der } j\text{-ten Faktorstufe} \end{array}$$

Modellformulierung

- Modell (Referenzcodierung):

$$Y_{ij} = \mu_J + \beta_j + \epsilon_{ij} \qquad j = 1, \ldots, J,\ i = 1, \ldots, n_j,$$

mit

μ_J	Mittelwert der Referenz
β_j	Effekt der Kategorie j im Vergleich zur Referenz J
ϵ_{ij}	zufällige Störgröße
ϵ_{ij}	$\sim N(0, \sigma^2)$, $\epsilon_{11}, \epsilon_{12}, \ldots, \epsilon_{Jn_J}$ unabhängig.

Testproblem:

$$H_0 : \beta_1 = \beta_2 = \ldots \beta_{j-1} = 0$$

gegen

$$H_1 : \beta_j \neq 0 \quad \text{für mindestens ein j}$$

Streuungszerlegung

Mittelwerte:

$\bar{Y}_{\bullet\bullet}$ Gesamtmittelwert in der Stichprobe
$\bar{Y}_{\bullet j}$ Mittelwert in der j-ten Faktorstufe

Es gilt (vgl. Statistik I) die Streuungszerlegung:

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

Variabilität **der** Gruppen

Variabilität **in den** Gruppen

F-Test

Die Testgröße

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

ist geeignet zum Testen der Hypothesen

$$H_0 \quad : \quad \beta_1 = \beta_2 = \ldots \beta_{j-1} = 0$$

gegen

$$H_1 \quad : \quad \beta_j \neq 0 \quad \text{für mindestens ein j}$$

Testprozedur

Die kritische Region besteht aus den *großen* Werten von F
Also H_0 ablehnen falls

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

mit dem entsprechenden $(1-\alpha)$-Quantil der F-Verteilung mit $(J-1)$ und $(n-J)$ Freiheitsgraden.
(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 des globalen Tests ist dann oft von Interesse, welche Gruppen sich unterscheiden.
$\Rightarrow$ Testen spezifischer Hypothesen über die Effekte β_j. Dabei tritt allerdings die Problematik des multiplen Testens auf.