2.5 Lineare Regressionsmodelle

2.5.1 Wiederholung aus Statistik I

Gegeben Datenpunkte (Y_i, X_i) schätze die beste Gerade $Y_i = \beta_0 + \beta_1 X_i,\ i = 1, \ldots, n.$

Bsp. 2.30. [Kaffeeverkauf auf drei Flohmärkten]

X Anzahl verkaufter Tassen Kaffee

Y zugehöriger Gewinn (Preis Verhandlungssache)

i	x_i	y_i	$y_i - \bar{y}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	10	9	-1	0	0
2	15	21	11	5	25
3	5	0	-10	-5	25
	$\bar{x} = 10$	$\bar{y} = 10$			

Man bestimme die Regressionsgerade und interpretiere die erhaltenen KQ-Schätzungen! Welcher Gewinn ist bei zwölf verkauften Tassen zu erwarten?

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{0 \cdot (-1) + 5 \cdot 11 + (-5) \cdot (-10)}{0 + 25 + 25} = \frac{105}{50} = 2.1\end{aligned}$$

Mit der Erhöhung der Menge X um eine Einheit erhöht sich der Gewinn Y um 2.1 Einheiten, also ist $\hat{b}$ so etwas wie der durchschnittliche Gewinn pro Tasse.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} = 10 - 2.1 \cdot 10 = -11$$

„Grundlevel“, Gewinn bei 0 Tassen (Fixkosten).

Vorhergesagte Werte und Residuen:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i, \qquad \hat{\varepsilon}_i = y_i - \hat{y}_i$$

$$\begin{aligned}
\hat{y}_1 &= -11 + 2.1 \cdot 10 = 10 \quad &\Rightarrow \hat{\varepsilon}_1 = -1 \\
\hat{y}_2 &= -11 + 2.1 \cdot 15 = 20.5 \quad &\Rightarrow \hat{\varepsilon}_2 = 0.5 \\
\hat{y}_3 &= -11 + 2.1 \cdot 5 = -0.5 \quad &\Rightarrow \hat{\varepsilon}_3 = 0.5
\end{aligned}$$

Zur Kontrolle: $\hat{\epsilon_1} + \hat{\epsilon_2} + \hat{\epsilon_3} = 0$

Prognose: $x^* = 12 \quad \Longrightarrow \hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 \cdot x^* = -11 + 2.1 \cdot 12 = 14.2$

Bsp. 2.31. [Arbeitszeit und Einkommen]

Multiples Regressionsmodell:

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

mit

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

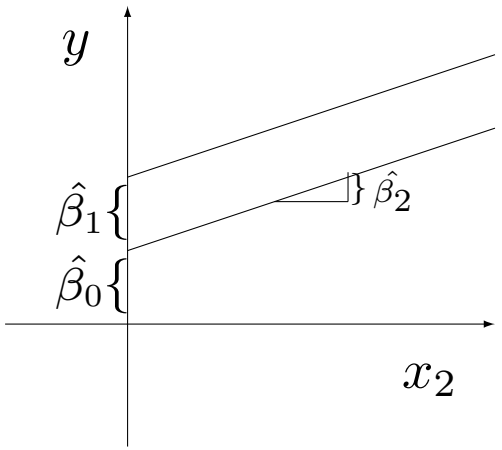
Interpretation:

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

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



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.

Bsp. 2.32. [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.

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

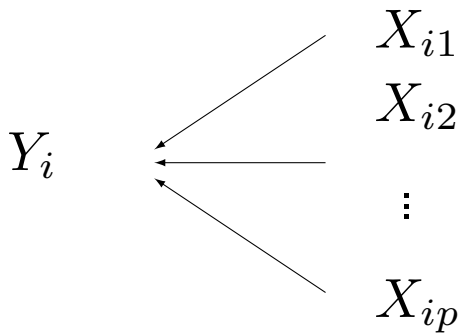
Beispiel zur Interpretation:

- Y: Score auf Autoritarismusskala
- X bzw. X_1, X_2: Parteienpräferenz
- X_3: Einkommen

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

- β_0: Grundniveau
- β_1: ceteris paribus Effekt (Erhöhung des Grundniveaus) von CDU/CSU und FDP
- β_2: ceteris paribus Effekt (Erhöhung des Grundniveaus) von SPD und Grünen
- β_3: ceteris paribus Effekt des Einkommens

Multiples Regressionsmodell:



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

Ansatz:

- linearer Zusammenhang.
- Ermittle aus den Daten „Wirkungsstärke“ der einzelnen Variablen.
- Im Folgenden: Probabilistische Modelle in Analogie zu den deskriptiven Modellen aus Statistik I (damit Verallgemeinerung auf die Grundgesamtheit möglich).

2.5.2 Lineare Einfachregression

Zunächst Modelle mit nur $einer$ unabhängigen 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

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

wobei

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$
σ^2 für alle i gleich
$\epsilon_{i1}, \epsilon_{i2}$ stochastisch unabhängig für $i_1 \neq i_2$

Nach den Modellannahmen gilt für die bedingte Verteilung von Y_i gegeben $X_i = x_i$:

$$Y_i | X_i = x_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), \quad i = 1, \ldots, n.$$

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

Aufgabe: Schätze die Parameter β_0, β_1 und σ^2. Die Schätzwerte und Schätzfunktionen werden üblicherweise mit $\hat{\beta}_0, \hat{\beta}_1$ und $\hat{\sigma}^2$ bezeichnet.

In der eben beschriebenen Situation gilt:

1. 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

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

2. 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

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

Bem. 2.33.

- $\hat{\beta}_0$ und $\hat{\beta}_1$ sind, wenn man die Realisationen (x_i, y_i) von (X_i, Y_i) einsetzt, 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:

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

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

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

(analog für $\hat{\beta}_0 \to T_{\beta_0^*}$).

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

- Typischer SPSS-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)

[a] abhängige Variable

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.
- Man kann analog zu Kap. 2.4.7 auch einseitige Hypothesen testen

2.5.3 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.

  (Outputs lesen können ist absolut klausurrelevant! Matrixrechnung wird nicht verlangt.)

- 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 SPSS-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).

2.5.4 Varianzanalyse (Analysis of Variance, ANOVA)

- Sind alle X_{ij} $0/1$-wertig, so erhält man die sogenannte $Varianzanalyse$, was dem Vergleich von mehreren Mittelwerten entspricht.

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

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

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

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

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

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

* etc.

- Vor allem in der angewandten Literatur, etwa in der Psychologie, wird die Varianzanalyse unabhängig vom Regressionsmodell entwickelt. Diese Sichtweise soll auch hier jetzt kurz besprochen werden.
- Ziel: Mittelwertvergleiche in mehreren Gruppen, häufig in (quasi-) experimentellen Situationen.
- Verallgemeinerung des t-Tests. Dort nur zwei Gruppen.
- Hier nur *einfaktorielle Varianzanalyse* (*Eine* Gruppierungsvariable).

Bsp. 2.34.

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}

$j = 1, \ldots, J$ Faktorstufen

$i = 1, \ldots, n_j$ Personenindex in der j-ten Faktorstufe

Zwei äquivalente Modellformulierungen:

a) Modell in Mittelwertsdarstellung:

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

mit

μ_j faktorspezifischer Mittelwert

ϵ_{ij} zufällige Störgröße

ϵ_{ij} $\sim \mathcal{N}(0, \sigma^2)$, $\epsilon_{11}, \epsilon_{12}, \ldots, \epsilon_{Jn_J}$ unabhängig.

Testproblem:

$$H_0 \quad : \quad \mu_1 = \mu_2 = \ldots \mu_J$$

gegen

$$H_1 \quad : \quad \mu_\ell \neq \mu_q \quad \text{für mindestens ein Paar } (\ell, q)$$

b) Modell in Effektdarstellung:

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

wobei α_j so, dass

$$\sum_{j=1}^{J} n_j \alpha_j = 0.$$

μ	globaler Erwartungswert
α_j	Effekt in der j-ten Faktorstufe, faktorspezifische systematische Abweichung vom gemeinsamen Mittelwert μ

Testproblem:

$$H_0 \quad : \quad \alpha_1 = \alpha_2 = \ldots \alpha_J = 0$$

gegen

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

Die beiden Modelle sind äquivalent: setze $\mu_j := \mu + \alpha_j$.

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:

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

Die Testgröße

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

ist geeignet zum Testen der Hypothesen

$$H_0 \quad : \quad \mu_1 = \mu_2 = \ldots \mu_J$$

gegen

$$H_1 \quad : \quad \mu_\ell \neq \mu_q \quad \text{für mindestens ein Paar } (\ell, q)$$

beziehungsweise

$$H_0 \quad : \quad \alpha_1 = \alpha_2 = \ldots \alpha_J = 0$$

gegen

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

Sie besitzt eine sog. $F\text{-}Verteilung$ mit $(J-1)$ und $(n-J)$ Freiheitsgraden.

Die kritische Region besteht aus den $gro\ss en$ Werten von F (Vorsicht: obwohl H_0 von „Gleichheitsform“).

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 α_j bzw. die Mittelwerte μ_j. Dabei tritt allerdings wieder Problematik des multiplen Testens auf.