6.2 Regressionsanalyse I: Die lineare Einfachregression

6.2.1 Grundbegriffe und Hintergrund

Bedeutung der Regression:

- Eines der am häufigsten verwendeten statistischen Verfahren. Vielfache Anwendung in den Sozialwissenschaften $\rightarrow$ Analoge Ausdehnung auf viele Variablen möglich!
- Grundidee der Interpretation bleibt in verwandter Weise bei vielen allgemeineren Modellen erhalten, die hier nicht betrachtet werden (können).

Motivation:

- Wir betrachten zunächst zwei metrische Variablen X und Y.
- Der Korrelationskoeffizient nach Bravais-Pearson misst die Stärke des linearen Zusammenhangs zwischen X und Y, beantwortet also die Frage „Wie gut lassen sich Ausprägungen (x_i, y_i), $i = 1, \ldots, n$ durch eine Gerade beschreiben?
- Die Regression geht nun einen Schritt weiter:
  - Wie sieht die am besten passende Gerade aus?
  - $\Rightarrow$ Analyse und Beschreibung des Zusammenhangs.

- Zusätzliche Ziele:
  - „individuelle“ Prognose basierend auf dem x-Wert: gegeben sei ein Punkt x^*. Wo liegt dem Modell nach das dazugehörige $\hat{y}^*$? (z.B. x^* Erwerbsarbeit in Stunden einer neuen Person, wieviel Hausarbeit in Stunden ist zu erwarten?)
  - Elastizität: Wie stark wirkt sich eine Änderung von X um eine Einheit auf Y aus?
    (z.B.: Wird die Erwerbsarbeit um eine Stunde reduziert, wieviel mehr Hausarbeit ist zu erwarten?)
    Entscheidende Grundlage für Maßnahmenplanung

- Die Regression ist ein erster Schritt in die etwas höhere Statistik. Fast alle gängigen Verfahren sind im weiteren Sinne Regressionsmodelle (allerdings oft nicht linear). Viele Grundideen zur Interpretation gelten in verwandter Form auch für andere Regressionsmodelle.
- Bei der Regressionsanalyse wird die Symmetrie des Zusammenhangs i.A. aufgegeben, d.h. nun wird ein gerichteter Zusammenhang der Form $X \longrightarrow Y$ betrachtet.

Bezeichnungen:

X	Y
unabhängige Variable	abhängige Variable
exogene Variable	endogene Variable
erklärende Variable	zu erklärende Variable
Stimulus	Response
Einflußgröße	Zielgröße
	Outcome
Prädiktor	
Kovariable	

6.2.2 Lineare Einfachregression: Grundmodell und Kleinste-Quadrate-Prinzip

Idee: Versuche, Y als einfache Funktion f von X zu beschreiben:

$$Y \approx f(X).$$

Einfachste Möglichkeit: f linear, also

$$Y \approx a + b \cdot X.$$

Für die beobachteten Datenpunkte soll also für jedes $i = 1, \ldots, n$ gelten

$$y_i \approx a + b \cdot x_i$$

Normalerweise besteht kein perfekter linearer Zusammenhang, so dass ein unerklärter Rest ε_i in die Modellgleichung mit aufgenommen wird (In Statistik 2 werden wir ε_i als zufälligen Fehler interpretieren):

$$y_i = a + b \cdot x_i + \varepsilon_i.$$

Dies ist das Modell der linearen Einfachregression.

a und b sind unbekannte Größen, die sogenannten Regressionsparameter oder Regressionskoeffizienten, die anhand der Daten bestimmt werden müssen.
Man beachte hierbei, dass a und b keinen Index tragen; sie werden hier als interindividuell konstant betrachtet und beschreiben den Zusammenhang der für alle Beobachtungen gelten soll.

Methode der kleinsten Quadrate:
Bestimme $\hat{a}, \hat{b}$ so, dass alle Abweichungen der Daten von der Gerade „möglichst klein“ werden, d.h. so, dass die Summe der quadratischen Differenzen zwischen den Punkten y_i und der Gerade $\hat{y}_i = \hat{a} + \hat{b} \cdot x_i$ minimiert wird. D.h. minimiere das $Kleinste\ Quadrate\ Kriterium$ (KQ-Kriterium):

$$\sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2$$

bezüglich $\hat{a}$ und $\hat{b}$.

Definition: Gegeben seien zwei metrische Merkmale X und Y und das Modell der linearen Einfachregression

$$y_i = a + bx_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

Dann bestimme man $\hat{a}$ und $\hat{b}$ so, dass mit

$$\begin{aligned} \hat{\varepsilon}_i &:= y_i - \hat{y}_i \\ &= y_i - (\hat{a} + \hat{b}x_i) \end{aligned}$$

das Kleinste-Quadrate-Kriterium

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

minimal wird. Die optimalen Werte $\hat{a}$ und $\hat{b}$ heißen KQ-Schätzungen, $\hat{\varepsilon}_i$ bezeichnet das i-te (geschätzte) Residuum.

Bemerkungen:

- Durch das Quadrieren tragen sowohl positive als auch negative Abweichungen von der Regressionsgeraden zum KQ-Kriterium bei.
- Das Quadrieren bewirkt außerdem, dass große Abweichungen überproportional stark berücksichtigt werden. (Die KQ-Schätzer sind in diesem Sinne ausreißeranfällig, da mit aller Macht versucht wird, große Abweichungen zu vermeiden.
  Es gibt robustere Alternativen die z.B. die Summe der absoluten Residuen minimieren ($\mathcal{L}^1$-Regression)

Satz: Für die KQ-Schätzer gilt

i) $$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(X,Y)}{\tilde{s}_X^2} =$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\cdot\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\cdot\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \varrho_{X,Y}\frac{\tilde{s}_Y}{\tilde{s}_X},$$

ii) $$\hat{a} = \bar{y} - \hat{b}\cdot\bar{x},$$

iii) $$\sum_{i=1}^{n}\hat{\varepsilon}_i = 0.$$

Bemerkungen:

- Hat man standardisierte Variablen X und Y (gilt also $\tilde{s}_X = \tilde{s}_Y = 1$), so ist $\hat{b}$ genau $\rho_{X,Y}$.
- Die mittlere Abweichung von der Regressionsgeraden ist Null.
- Diese Eigenschaft kann auch verwendet werden, um die korrekte Berechnung der KQ-Schätzer zu überprüfen.
- Basierend auf den Schätzern $\hat{a}$ und $\hat{b}$ kann der Wert der abhängigen Variablen Y auch für neue, unbeobachtete Werte x^* der Kovariablen X berechnet werden (Prognose):

$$\hat{y}^* = \hat{a} + \hat{b}x^*.$$

- Weiß man, dass $b = 0$ ist, und setzt daher $\hat{b} = 0$, so lautet die KQ-Schätzung $\bar{y}$.
  In der Tat: $\bar{y}$ minimiert $\sum_{i=1}^{n}(y_i - a)^2$, vergleiche Exkurs im Kapitel bei dem Lagemaß.

Interpretation der Regressionsgeraden:

- $\hat{a}$ ist der Achsenabschnitt, also der Wert der Gerade, der zu $x = 0$ gehört. Er lässt sich oft als „Grundniveau“ interpretieren.
- $\hat{b}$ ist die Steigung (Elastizität): Um wieviel erhöht sich y bei einer Steigerung von x um eine Einheit?
- $\hat{y}^*$ (Punkt auf der Gerade) ist der Prognosewert zu x^*.

Fiktives „ökonomisches Beispiel“ zur Klärung: Kaffeeverkauf auf drei Flohmärkten

X Anzahl verkaufter Tassen Kaffee
Y zugehöriger Umsatz (Preis Verhandlungssache)

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

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

6.2.3 Modellanpassung: Bestimmtheitsmaß und Residualplots

- Wie gut lässt sich die abhängige Variable Y durch die Kovariable X erklären?
- Wie gut passt der lineare Zusammenhang zwischen X und Y?

PRE-Ansatz:

Modell 1: Vorhersage von Y ohne X.

Dabei gemachter Gesamtfehler:

$$SQT :=$$

(Gesamtstreuung / Gesamtvariation der y_i: „sum of squares total“).

Modell 2: Vorhersage von Y mit X.

Dabei gemachter Gesamtfehler:

$$SQR := \qquad\qquad = \sum_{i=1}^{n} \varepsilon_i^2$$

(Residualstreuung / Residualvariation: „sum of squared residuals“).

Die Differenz

$$SQE := SQT - SQR$$

nennt man die durch das Regressionsmodel erklärte Streuung („sum of squares explained“).

Man kann zeigen, dass gilt

$$SQE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

Streuungszerlegung:

$$SQT = SQR + SQE$$

(analog zur Streuungszerlegung bei geschichteten Daten).

Bestimmtheitsmaß: Der PRE-Ansatz liefert das Gütekriterium

$$\frac{SQT - SQR}{SQT} = \frac{SQE}{SQT}.$$

Diese Größe bezeichnet man als Bestimmtheitsmaß. In der Tat gilt (nach etwas längerer Rechnung):

$$\frac{SQE}{SQT} = R^2_{XY}$$

d.h. dies ist genau das Bestimmtheitsmaß aus Definition (6.1).

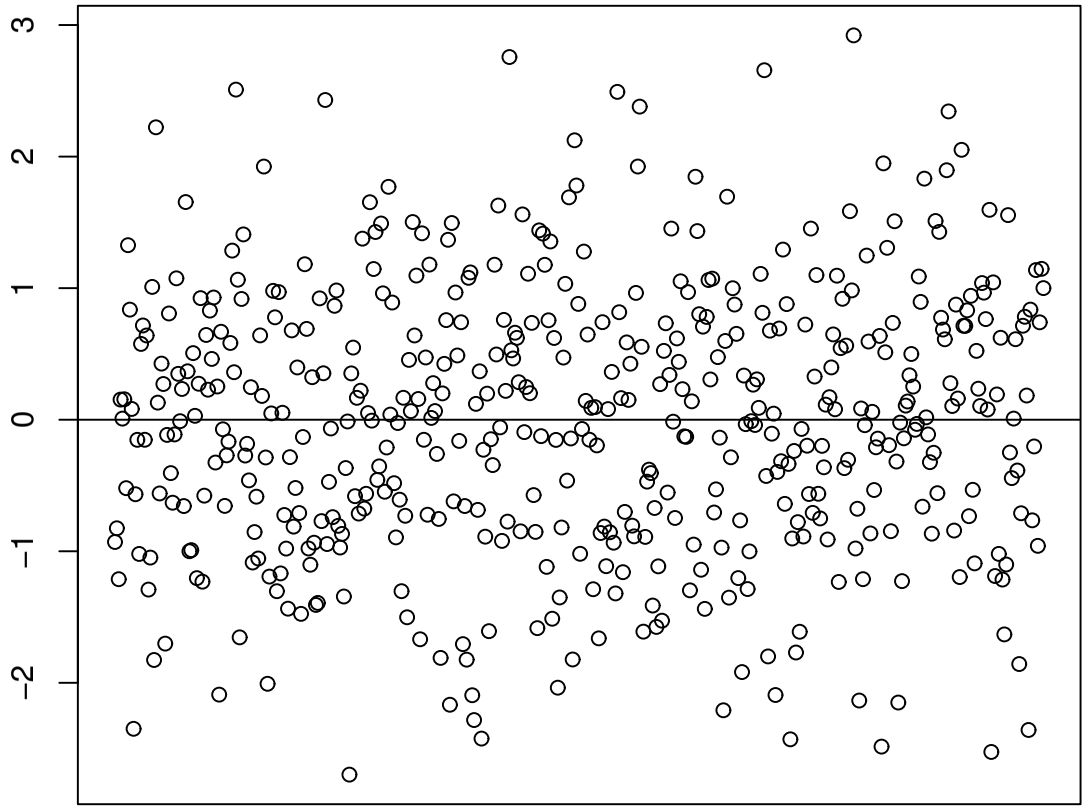
Es gibt also drei Arten, R^2_{XY} zu verstehen:

Eigenschaften:

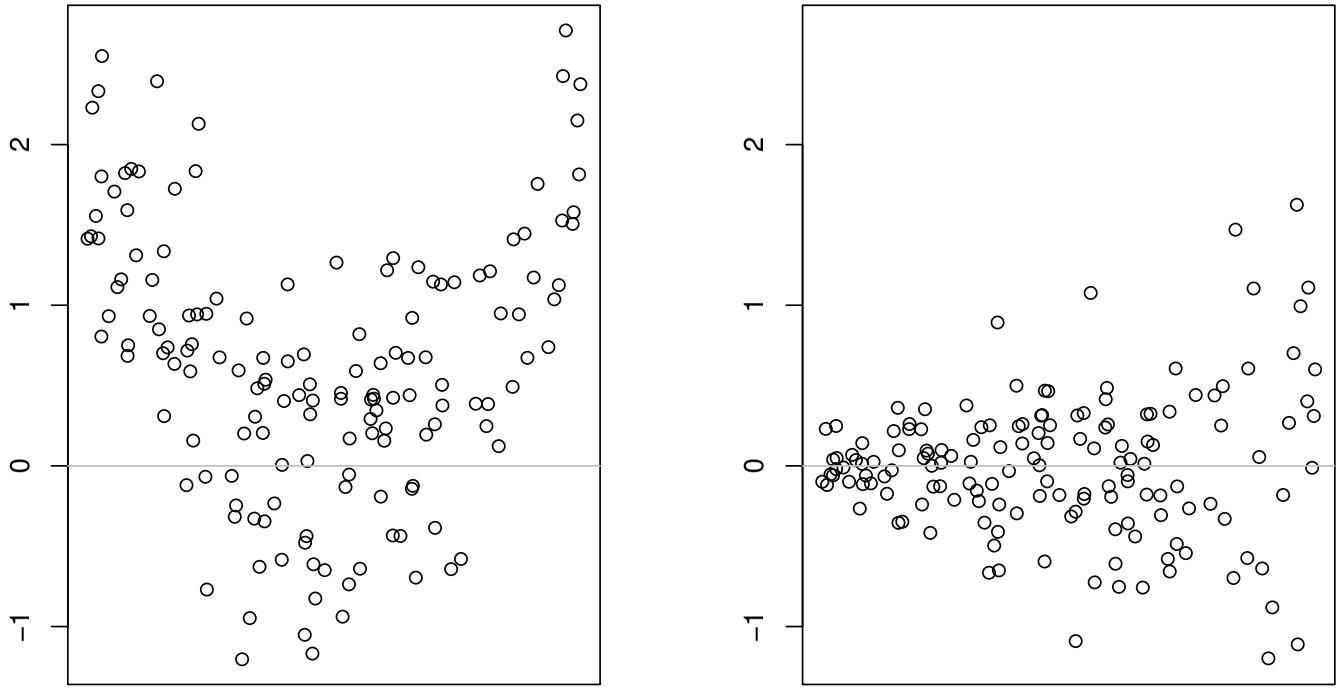
- Es gilt: $0 \leq R^2_{XY} \leq 1$.
- $R^2_{XY} = 0$: Es wird keine Streuung erklärt, d.h. es gibt keinen (linearen) Zusammenhang zwischen X und Y.
- $R^2_{XY} = 1$: Die Streuung wird vollständig erklärt. Alle Beobachtungen liegen tatsächlich auf einer Geraden.

Residualplots

Eine wichtige optische Möglichkeit, die Anpassung zu beurteilen, beruht auf dem Studium der geschätzten Residuen $\hat{\varepsilon}_i$. Sie sollen unsystematisch um 0 streuen.



Zeigt sich eine Systematik, so war der lineare Ansatz unangemessen, und es ist größte Vorsicht bei der Interpretation geboten!



6.2.4 Linearisierende Transformationen:

Sehr häufig wirkt die Variable X nicht „direkt linear“ auf die Variable Y (Streudiagramm anschauen!). Die lineare Regression passt die „optimale Gerade“ an. Was kann man aber tun, wenn selbst diese optimale Gerade nicht passt, da der Zusammenhang eben nicht linear ist.
Bei naiver Anwendung des linearen Ansatzes besteht die Gefahr gravierender Fehlschlüsse.

Viele (nicht alle) der auf den ersten Blick nichtlinearen Modelle lassen sich durch geeignete Variablentransformationen in die lineare Regressionsrechnung einbetten. Entscheidend ist, dass das Wirken der Parameter linear ist!

Der Ansatz

$$g(y_i) = a + b \cdot h(x_i) + \varepsilon_i$$

lässt sich auch völlig analog mit dem KQ-Prinzip behandeln:

Entscheidend ist die Linearität in den Parametern a und b. So ist im Gegensatz zu oben ist der Ansatz

$$y_i = a + b^2 \cdot x_i + \varepsilon_i$$

kein lineares Regressionsmodell.

Sehr häufiger Ansatz:

$$Y = a + b \cdot \ln X + \varepsilon\,,$$

hier kann b wie folgt interpretiert werden:
Erhöht man einen Wert von X um p Prozent, so erhöht sich der entsprechende Y-Wert etwa um $b \cdot p$ Prozent, denn

$$\begin{aligned}
&\triangle Y^* = b \cdot \triangle X^* = \\
&= b \cdot (\ln((1+p) \cdot x) - \ln(x)) \\
&= b \cdot (\ln(1+p) + \ln(x) - \ln(x)) \\
&= b \cdot \ln(1+p) \approx b \cdot p, \text{ falls } p \text{ klein.}
\end{aligned}$$

„Echte“ nichtlineare Modelle ergeben sich aus der Theorie der generalisierten linearen Modelle und generalisierten additiven Modelle. Erstere sind insbesondere auch für kategoriales oder ordinales Y geeignet, letztere erlauben es, Modelle zu schätzen die sogar die geeignetste Transformation der Kovariablen in sehr allgemeiner Form aus den Daten mitschätzen.
($\rightarrow$ Nebenfach Statistik)

Beide Ansätze sind direkte Verallgemeinerungen und Erweiterungen der linearen Regressionsmodells.

6.2.5 Multiple lineare Regression

Verallgemeinerung der linearen Einfachregression: Betrachte mehrere unabhängige metrische Variablen $X_1, X_2, \ldots, X_p$ gemeinsam, da typischerweise ja kein monokausaler Zusammenhang vorliegt.

Modellgleichung:

$$y = a + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_p x_{pi} + \varepsilon_i.$$

Dabei bezeichnet x_{i1} den für die i-te Beobachtung beobachteten Wert der Variablen X_1, x_{i2} den Wert der Variablen X_2, usw.

Interpretation: Die Interpretation von a und $b_1, \ldots, b_p$ erfolgt analog zu oben, insbesondere ist b_j die Änderung in Y, wenn X_j um eine Einheit vergrößert wird — und alle anderen Größen gleich bleiben („*ceteris paribus Effekt*“).

Üblich ist allerdings eine andere Notation für die Regressionskoeffizienten:

$$\begin{aligned} a &\rightarrow \beta_0, \\ b_1 &\rightarrow \beta_1, \\ &\vdots \\ b_p &\rightarrow \beta_p, \end{aligned}$$

KQ-Prinzip: Die Schätzung von $\beta_0, \beta_1, \ldots, \beta_p$ erfolgt wieder über das KQ-Prinzip: Bestimme $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ so, dass mit

$$\hat{\varepsilon}_i = y_i - \hat{y}_i := y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_p x_{pi})$$

der Ausdruck

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

minimal wird.

Die Schätzungen $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ sind nur mit Matrizenrechnung einfach darzustellen und insbesondere nur noch schwierig „von Hand“ zu berechnen. Im Rahmen dieser Veranstaltung brauchen Sie bei der multiplen Regression nicht mehr rechnen, sondern „nur“ typische Outputs korrekt interpretieren können.

Bestimmtheitsmaß: Analog zur linearen Einfachregression lässt sich ein Bestimmtheitsmaß

$$R^2 = \frac{SQE}{SQT}$$

über die Streuungszerlegung definieren. In der multiplen Regression verwendet man allerdings meistens das korrigierte Bestimmtheitsmaß

$$\tilde{R}^2 := 1 - \frac{n-1}{n-p-1}(1-R^2)$$

das die Anzahl der in das Modell mit einbezogenen Variablen mit berücksichtigt. (Das übliche R^2 würde ja auch durch das Einführen irrelevanter Variablen ansteigen, während bei $\tilde{R}^2$ sozusagen für jede aufgenommene Variable einen Preis zu bezahlen ist.)

SPSS-Output einer multiplen Regression:

Coefficients[a]

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	$\hat{\beta}_0$	$\hat{\sigma}_0$	T_0	p-Wert
	X_1	$\hat{\beta}_1$	$\hat{\sigma}_1$	T_1	"
	X_2	$\hat{\beta}_2$	$\hat{\sigma}_2$	T_2	"
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	"
	X_p	$\hat{\beta}_p$	$\hat{\sigma}_p$	T_p	"

[a] Dependent Variable: Y

Im Rahmen von Statistik 1 ist nur die Spalte „B“ mit den unstandardisierten Koeffizienten $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ relevant.

Anmerkung: SPSS gibt auch noch die „standardisierten Koeffizienten“ β aus, das sind nicht etwa die $\hat{\beta}$'s im Sinne der Vorlesung, sondern die Schätzer, wenn man die Variablen vorher standardisiert. Bei der linearen Einfachregression findet man hier den Korrelationskoeffizienten von Bravais Pearson wieder.