6 Korrelationsanalyse: Zusammenhangsanalyse stetiger Merkmale

6.1 Korrelationsanalyse

Jetzt betrachten wir bivariate Merkmale (X, Y), wobei sowohl X als auch Y stetig bzw. quasi-stetig und mindestens ordinalskaliert, typischerweise sogar intervallskaliert, sind. Am Rande wird auch der Fall gestreift, dass nur ein Merkmal quasi-stetig und das andere nominalskaliert ist.

6.1.1 Streudiagramm, Kovarianz- und Korrelationskoeffizienten

Beispiele:

- Nettomiete $\longleftrightarrow$ Wohnfläche
- Autoritarismusscore vor/nach einer Informationsveranstaltung
- Monatseinkommen $\longleftrightarrow$ Alter in Jahren
- Wochenarbeitseinkommen $\longleftrightarrow$ Wochenarbeitsstunden
- Wochenarbeitsstunden $\longleftrightarrow$ Hausarbeit in Stunden pro Woche
- Wochenarbeitsstunden (tatsächlich) $\longleftrightarrow$ Wochenarbeit (vertraglich)

6.1.2 Streudiagramme (Scatterplots)

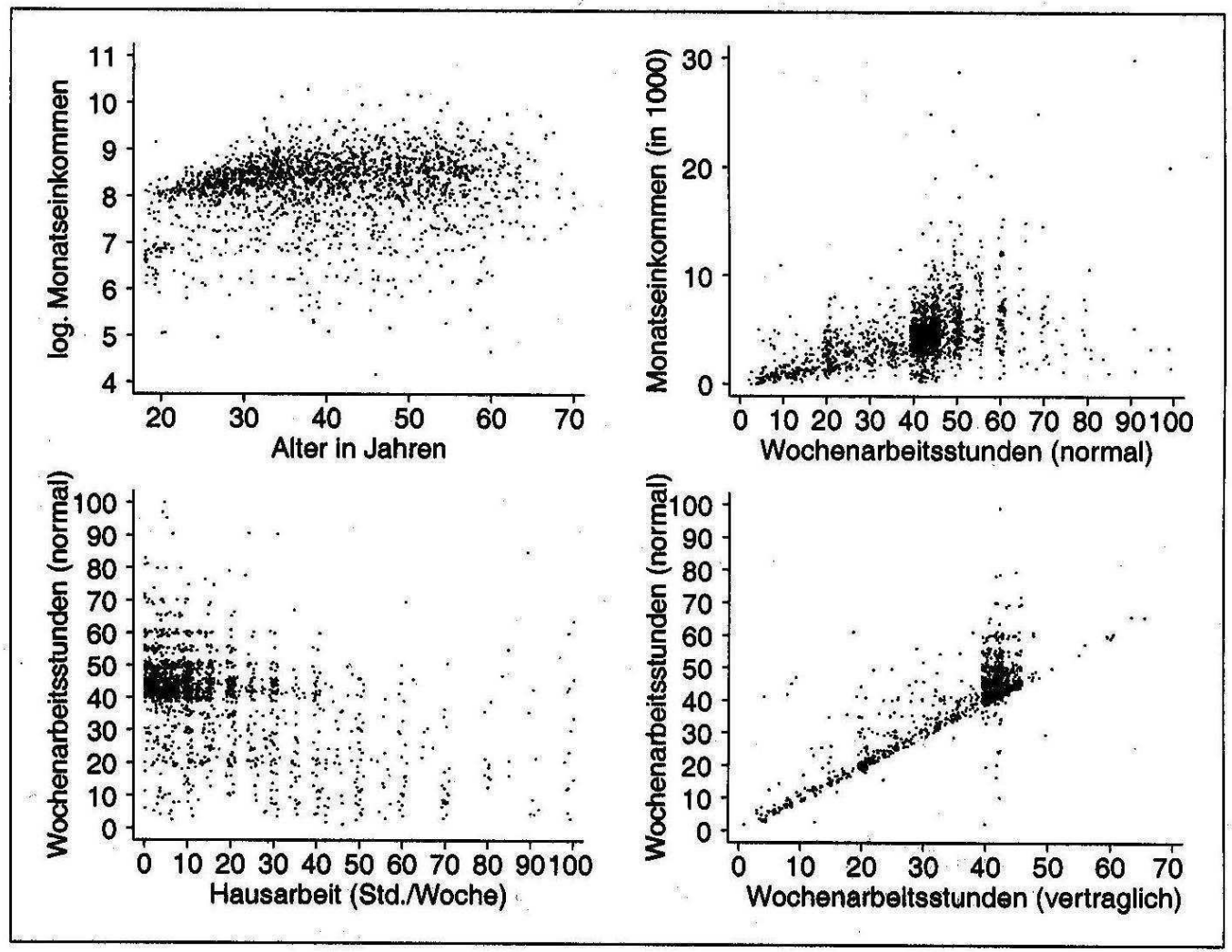
Sind die Merkmale stetig oder zumindestens quasi-stetig (sehr viele verschiedene Ausprägungen), werden Kontingenztabellen sehr unübersichtlich und praktisch aussagelos, da die einzelnen Häufigkeiten in den Zellen der Tabellen natürlicherweise durchwegs sehr klein sind.

Alternative Darstellungsform: $Scatterplot$ / $Streudiagramm$:

Zeichne die Punkte (x_i, y_i), $i = 1, \ldots, n$, in ein X-Y-Koordinatensystem.

$\Longrightarrow$ Guter optischer Eindruck über das Vorliegen, die Richtung und gegebenenfalls die Art eines Zusammenhangs.

$\Longrightarrow$ Ausreißer werden leicht erkannt.



Quelle für Beispiele: Jann (2002), p. 85 ff.

6.1.3 Kovarianz und Korrelation

Wie misst man den Zusammenhang zwischen metrischen Merkmalen?

- Eine Idee die sogenannte Kovarianz (s.u.) zu konstruieren besteht darin nach Konkordanz/Diskordanz zum Schwerpunkt zu fragen und dabei auch die nun interpretierbaren Abstände zur Messung der „individuellen Konkordanzstärke“ heranzuziehen. Negative Werte sprechen für Diskordanz.
- Betrachte den „Mittelpunkt“ der Daten $(\bar{x}, \bar{y})$ und dazu konkordante/diskordante Paare.

- Eine Beobachtung i mit Ausprägung (x_i, y_i) ist
  - *konkordant* zu $(\bar{x}, \bar{y})$, spricht also für einen gleichgerichteten Zusammenhang, wenn

$$(x_i > \bar{x} \text{ und } y_i > \bar{y}) \text{ oder } (x_i < \bar{x} \text{ und } y_i < \bar{y})$$

    also zusammengefasst wenn

$$(x_i - \bar{x}) \cdot (y_i - \bar{y}) > 0.$$

  - *diskordant* zu $(\bar{x}, \bar{y})$, spricht also für einen gegengerichteten Zusammenhang, wenn

$$(x_i < \bar{x} \text{ und } y_i > \bar{y}) \text{ oder } (x_i > \bar{x} \text{ und } y_i < \bar{y})$$

    also zusammengefasst wenn

$$(x_i - \bar{x}) \cdot (y_i - \bar{y}) < 0.$$

- Wegen des metrischen Skalenniveaus sind auch die Abstände interpretierbar, das Produkt $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ gibt also sozusagen die Stärke der Konkordanz bzw. Diskordanz an.

- $(x_i - \bar{x})(y_i - \bar{y})$ ist positiv, wenn große (kleine) X-Werte mit großen (kleinen) Y-Werten einhergehen (gleichgerichteter Zusammenhang).

- $(x_i - \bar{x})(y_i - \bar{y})$ ist negativ, wenn große (kleine) X-Werte mit kleinen (großen) Y-Werten einhergehen (gegengerichteter Zusammenhang).

$\Longrightarrow$ Definiere als Zusammenhangsmaß die durchschnittliche individuelle Konkordanzstärke.

Definition: Gegeben sei ein bivariates Merkmal (X, Y) mit metrisch skalierten Variablen X und Y mit $\tilde{s}_X^2 > 0$ und $\tilde{s}_Y^2 > 0$. Dann heißen

$$\mathrm{Cov}(X, Y) := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

(empirische) Kovarianz von X und Y,

$$\varrho(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\tilde{s}_Y^2 \tilde{s}_X^2}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

(empirischer) Korrelationskoeffizient nach Bravais und Pearson von X und Y, und

$$R^2_{XY} := (\varrho(X,Y))^2 \tag{6.24}$$

Bestimmtheitsmaß von X und Y.

Bemerkungen:

- Die Kovarianz $\mathrm{Cov}(X,Y)$ ist maßstabsabhängig.

- Das Teilen durch die Standardabweichungen normiert die Kovarianz und macht sie maßstabsunabhängig.

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

  Also ist - im Sinne obiger Interpretation - der Korrelationskoeffizient die *durchschnittliche standardisierte* Konkordanzstärke.

- Die empirische Kovarianz ist eine Verallgemeinerung der empirischen Varianz. Die Kovarianz eines Merkmals mit sich selbst ist genau die empirische Varianz:

$$\begin{aligned} \mathrm{Cov}(X,X) &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\,(x_i-\bar{x}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \tilde{s}_x^2 \end{aligned}$$

- Man sieht hier auch, dass die Größe der Kovarianz für sich genommen unanschaulich zu interpretieren ist. Für den Korrelationskoeffizienten hingegen gilt:

$$-1 \leq \varrho(X,Y) \leq 1.$$

  und insbesondere $\varrho(X,X) = 1$.

- Viele der (un)angenehmen Eigenschaften der Varianz (z.B. Ausreißerempfindlichkeit) gelten in analoger Weise.

- Es gilt auch einen analogen Verschiebungssatz für die Kovarianz:

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \; - \; \bar{x}\bar{y}$$

und damit

$$\varrho(X,Y) = \frac{\displaystyle\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sqrt{\displaystyle\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}\;\cdot\;\sqrt{\displaystyle\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2}}.$$

Zur Erinnerung:

$$\tilde{s}_X^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.$$

Beispiel: Zunächst inhaltsleere Zahlenbeispiele, zur Interpretation später.

- Gegeben seien die Datenpaare

x_i	37	30	20	28	35
y_i	130	112	108	114	136

Es gilt: $\bar{x} = 30$ und $\bar{y} = 120$, sowie

$$\sum_{i=1}^{n} x_i^2 = 4678 \qquad \sum_{i=1}^{n} y_i^2 = 72600$$

$$\sum_{i=1}^{n} x_i y_i = 18282$$

$$n = 5$$

Tabelle:

	x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
	37	130			
	30	112			
	20	108			
	28	114			
	35	136			
$\sum$					

Basierend auf diesen Hilfsgrößen berechnet sich der Korrelationskoeffizient gemäß Verschiebungssatz als

$$\varrho(X,Y) = \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2}} =$$

- Gegeben sei ein Merkmal X und das Merkmal $Y = (X - 20)^2$ mit den Datenpaaren.

x_i	10	20	30
y_i	100	0	100
$x_i y_i$	1000	0	3000

Es gilt: $\bar{x} = 20$ und $\bar{y} = \frac{200}{3}$ und damit

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y} \\ &= \frac{1}{3}(1000 + 0 + 3000) - 20 \cdot \frac{200}{3} \\ &= \frac{4000}{3} - \frac{4000}{3} = 0 \end{aligned}$$

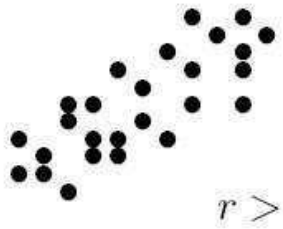
Für den Korrelationskoeffizienten ergibt sich damit ebenfalls $\varrho(X,Y) = 0$!

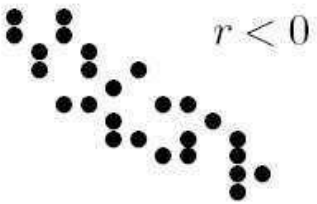
Bemerkungen:

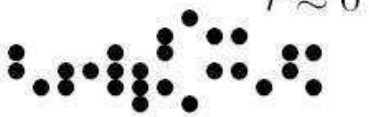
- Es gilt $|\varrho| = 1$ genau dann wenn $Y = aX + b$ mit $a \neq 0$, d.h. X und Y stehen in einem perfekten linearen Zusammenhang.

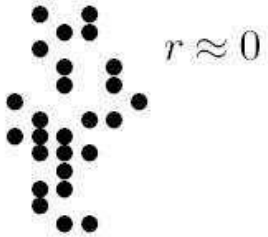
- Ist $\varrho = 0$ (und äquivalent dazu $\mathrm{Cov}(X, Y) = 0$), so nennt man X und Y *unkorreliert*. Es besteht dann keinerlei linearer Zusammenhang.

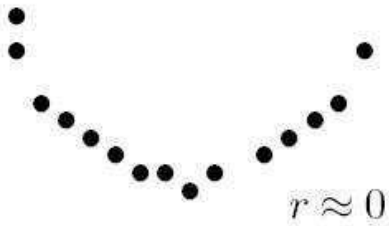
- Die Betonung der *Linearität* des Zusammenhangs ist wesentlich.

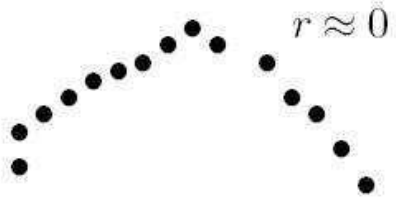

$r > 0$


$r < 0$

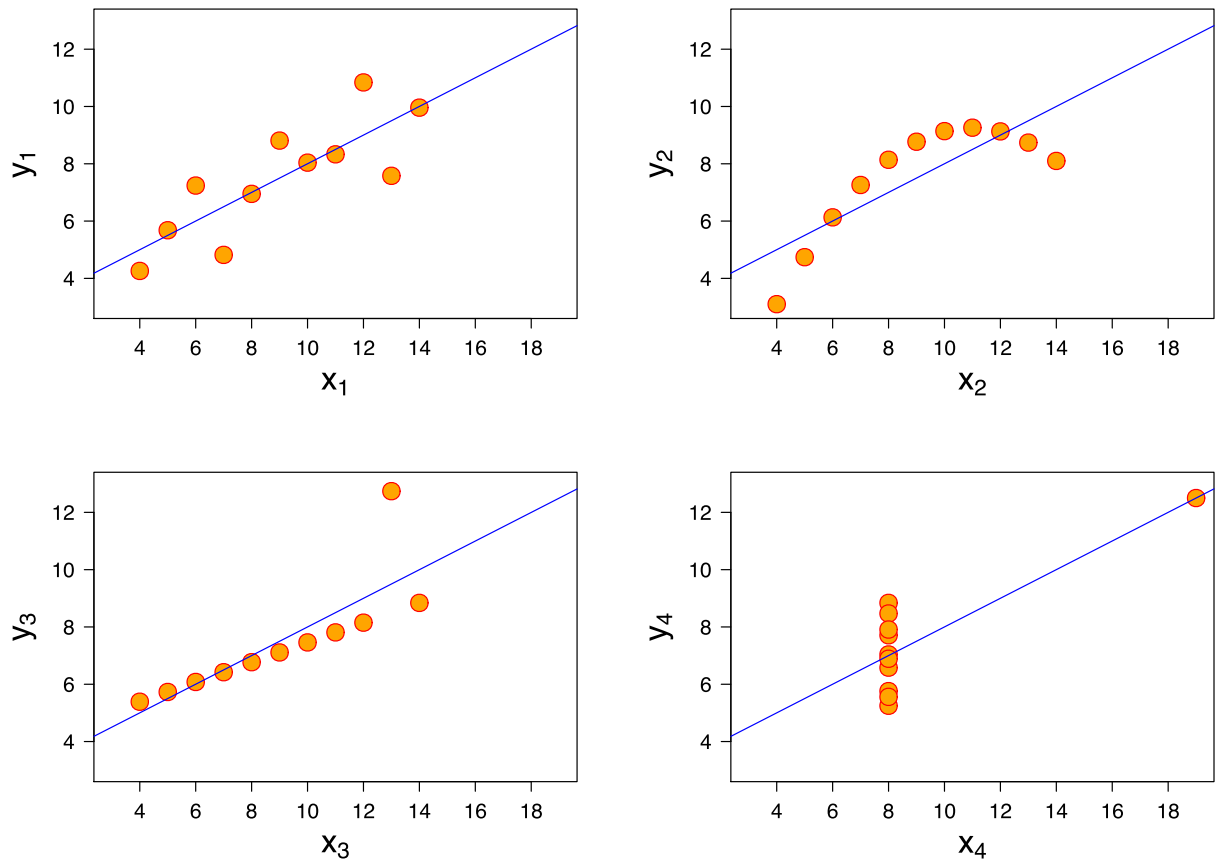

$r \approx 0$


$r \approx 0$


$r \approx 0$


$r \approx 0$

Beispiel: Anscombe's Quartet:



(Quelle: Wikipedia; Anscombe's quartet)

- Allgemein zeigt $|\varrho|$ und R^2 die Stärke eines *linearen* Zusammenhangs an, also wie gut sich die Datenpaare $(x_1, y_1), \ldots, (x_n, y_n)$ durch eine *Gerade* beschreiben lassen.
- R^2 ist ein PRE-Maß, das misst, welchen Anteil der gesamten Variation sich durch einen linearen Zusammenhang beschreiben lässt. (Näheres dazu im Abschnitt über die Regression.)
- Gelegentlich wird der Wert des Korrelationskoeffizienten folgendermaßen schematisch interpretiert:
  - $\varrho_{XY} \approx 0$: kein (linearer) Zusammenhang.
  - $\varrho_{XY} > 0$: positive Korrelation, gleichgerichteter (linearer) Zusammenhang.
  - $\varrho_{XY} < 0$: negative Korrelation, gegengerichteter (linearer) Zusammenhang.
  - $|\varrho_{XY}| \leq 0.5$: schwache Korrelation.
  - $0.5 < |\varrho_{XY}| \leq 0.8$: mittlere Korrelation.
  - $|\varrho_{XY}| > 0.8$: starke Korrelation.

- Die Zusammenhangsmaße sind invariant gegenüber Vertauschen von Y und X, unterscheiden also nicht welche Variable als abhängige, wele als unabhängige gilt:

$$\varrho(X,Y) = \varrho(Y,X) \qquad R_{XY} = R_{YX}.$$

- Im Gegensatz zur Kovarianz sind $\varrho(X,Y)$ und R^2_{XY} invariant gegenüber streng monoton steigenden linearen Transformationen. Genauer gilt mit $\tilde{X} := a \cdot X + b$ und $\tilde{Y} := c \cdot Y + d$

  $$\varrho(\tilde{X},\tilde{Y}) = \varrho(X,Y)$$

  falls $a \cdot c > 0$ und

  $$\varrho(\tilde{X},\tilde{Y}) = -\varrho(X,Y)$$

  falls $a \cdot c < 0$. Die Korrelation ist also in der Tat maßstabsunabhängig.

Beispiel: Mietspiegel (SPSS-Ausdruck)

Korrelationen

		Nettomiete	Wohnfläche	Baujahr
Nettomiete	Korrelation nach Pearson	1	.600	.223
	Signifikanz (2-seitig)		.000	.006
	N	150	150	150
Wohnfläche	Korrelation nach Pearson	.600	1	-.174
	Signifikanz (2-seitig)	.000		.033
	N	150	150	150
Baujahr	Korrelation nach Pearson	.223	-.174	1
	Signifikanz (2-seitig)	.006	.033	
	N	150	150	150

Zur Interpretation der einzelnen Zellen: In der entsprechenden Zelle stehen Informationen zur Korrelation der Variablen, in der entsprechenden Zeile mit der Variable der jeweiligen entsprechenden Spalte. In der ersten Zeile stehen jeweils die Korrelationskoeffizienten, N ist der Stichprobenumfang, der bei uns mit n bezeichnet wird.
Die zweite Zeile „Signifikanz“ wird erst in Statistik II verständlich. Grob gesprochen gilt: Je kleiner diese Zahl ist, desto sicherer ist man sich, dass der errechnete Korrelationskoeffizient nicht nur zufällig von 0 abweicht.

Beispiele aus Jann (2002) S.87ff

- Arbeitsstunden und Erwerbseinkommen: 0.495
  moderater positiver Zusammenhang.
- Arbeitsstunden und Haushalt: -0.434
  moderater negativer Zusammenhang.
- Vertragliche und geleistete Wochenarbeitsstunden: 0.868
  hoch positiv korreliert (Punkte liegen sehr nahe an „bester Gerade“).

6.1.4 Weitere Korrelationskoeffizienten

Anwendung des Korrelationskoeffizienten nach Bravais-Pearson auf dichotome nominale Merkmale

Liegen $dichotome$ nominale Merkmale, d.h. Merkmale mit nur zwei ungeordneten Ausprägungen vor (z.B. ja/nein), und kodiert man die Ausprägung mit 0 und 1, so kann man die Formel des Korrelationskoeffizienten nach Bravais-Pearson sinnvoll anwenden. Man erhält den sogenannten $Punkt\text{-}Korrelationskoeffizienten$, der identisch zu Φ_s aus ($\rightarrow$ Kapitel 5.3) ist.

Im Fall einer dichotomen und einer metrischen Variablen ergibt sich bei Anwendung des Korrelationskoeffizienten nach Bravais-Pearson die sogenannte $Punkt\text{-}biseriale\ Korrelation$. (vgl. etwa Jann (2002, S.90f) oder Wagschal (1999, Kap 10.8).)

Rangkorrelationskoeffizient nach Spearman

- Wir betrachten ein bivariates Merkmal (X, Y), wobei X und Y nur ordinalskaliert sind, aber viele unterschiedliche Ausprägungen besitzen.
- Der Korrelationskoeffizient von Bravais-Pearson darf nicht verwendet werden, da hier die Abstände nicht interpretierbar sind. $(\bar{x}, \bar{y})$ wären willkürliche Zahlen, ebenso $(x_i - \bar{x}), (y_i - \bar{y})$.

  Beispiel

- Liegen keine Bindungen vor, dann rechnet man statt mit $(x_i, y_i)_{i=1,\ldots,n}$ mit $(\mathrm{rg}(x_i), \mathrm{rg}(y_i))$ $i = 1, \ldots, n$. Dabei ist

$$\mathrm{rg}(x_i) = j :\Longleftrightarrow x_i = x_{(j)},$$

d.h. der Rang $\mathrm{rg}(x_i)$ ist die Nummer, die x_i in der geordneten Urliste $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ einnimmt (analog für $\mathrm{rg}(y_i)$). Der kleinsten Beobachtung wird also der Wert 1 zugeordnet, der zweitkleinsten der Wert 2, usw., der grö§ten der Wert n. Beispiel:

x_i	1	7	2	5.3	16
$\mathrm{rg}(x_i)$					

- Liegen sogenannte Bindungen vor, d.h. haben mehrere Einheiten dieselbe Ausprägung der Variablen X oder der Variablen Y, so nimmt man den Durchschnittswert der in Frage kommenden Ränge (Achtung: etwas anderer Begriff der Bindung als in Kapitel 5).

  Beispiel:

x_i	1	7	7	3	10
Rang					
$\mathrm{rg}(x_i)$					

- Wende nun den Korrelationskoeffizienten nach Bravais-Pearson auf die Rangdaten an. Nach Umformung ergibt sich unter Benutzung von

$$\sum_{i=1}^{n} rg(x_i) = \sum_{i=1}^{n} i = \frac{n(n+1)}{2} = \sum_{i=1}^{n} rg(y_i)$$

folgende Formel:

Definition:

$$\varrho_S(X,Y) := \frac{\sum_{i=1}^{n} \mathrm{rg}(x_i) \cdot \mathrm{rg}(y_i) - n\left(\frac{n+1}{2}\right)^2}{\sqrt{\sum_{i=1}^{n} (\mathrm{rg}(x_i))^2 - n\left(\frac{n+1}{2}\right)^2}\ \sqrt{\sum_{i=1}^{n} (\mathrm{rg}(y_i))^2 - n\left(\frac{n+1}{2}\right)^2}}$$

heißt (empirischer) *Rangkorrelationskoeffizient nach Spearman.*

Bemerkungen:

- Liegen keine Bindungen vor, so gilt

$$\varrho_{S,XY} = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2-1)}.$$

wobei $d_i := \mathrm{rg}(x_i) - \mathrm{rg}(y_i)$.

- Wichtig für Interpretation: Da $\varrho_S(X, Y)$ sich aus der Anwendung von $\varrho(X, Y)$ auf Rangdaten ergibt, behalten die entsprechenden Bemerkungen zum Bravais-Pearson-Korrelationskoeffizienten – auf die Ränge bezogen – ihre Gültigkeit. Insbesondere gilt $-1 \leq \varrho_{S,XY} \leq 1$, und $\varrho_{S,XY}$ ist analog zu interpretieren.

- Im Gegensatz zum Korrelationskoeffizienten von Bravais-Pearson misst der Rangkorrelationskoeffizient nicht nur lineare, sondern allgemeiner monotone Zusammenhänge. Die Anwendung der Rangtransformation bewirkt in gewisser Weise eine Linearisierung monotoner Zusammenhänge.
  Tabelle für $y = x^3$:

	x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
	10	1000	10000	100	1000000
	10	1000	10000	100	1000000
	0	0	0	0	0
	-20	-8000	160000	400	64000000
$\sum$	0	-6000	180000	600	66000000

  Also ist $\varrho(X, Y) = 0.973$, und, da hier in der Tat $rg(x_i) = rg(y_i)$ für alle i, $\varrho_s(X, Y) = 1$.

- Die Bildung von Rängen ist unempfindlich gegenüber Ausreißern, so dass auch der Rangkorrelationskoeffizient ausreißerresistent ist.

Beispiel: (fiktiv, Zahlen aus Jann, 2002/2005)

Zwei Gutachter sollen das autoritäre Verhalten von 5 Gruppenmitgliedern vergleichen, indem sie Scores auf einer Skala zwischen 0 und 100 vergeben. (Dies ist ein typischer Fall einer Ordinalskala; die Abstände sind nicht direkt interpretierbar, sondern nur die Reihenfolge!)

Man berechne den Rangkorrelationskoeffizienten nach Spearman für die Merkmale X und Y mit

X Einstufung durch Gutachter 1

Y Einstufung durch Gutachter 2

Person i	1	2	3	4	5
X: Gutachter 1	10	15	20	20	30
Y: Gutachter 2	20	10	30	40	60
$\mathrm{rg}(x_i)$					
$\mathrm{rg}(y_i)$					

Bemerkung:

- Analog zur punkt-biserialen Korrelation gibt es auch eine $biseriale\ Rangkorrelation$ zur Beschreibung des Zusammenhangs zwischen einer $0-1$-kodierten dichotomen nominalen und einer quasi-stetigen ordinalen Variable (vgl. Wagschal, 1999, Kap 10.7).

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

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

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

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

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

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 (6.24).

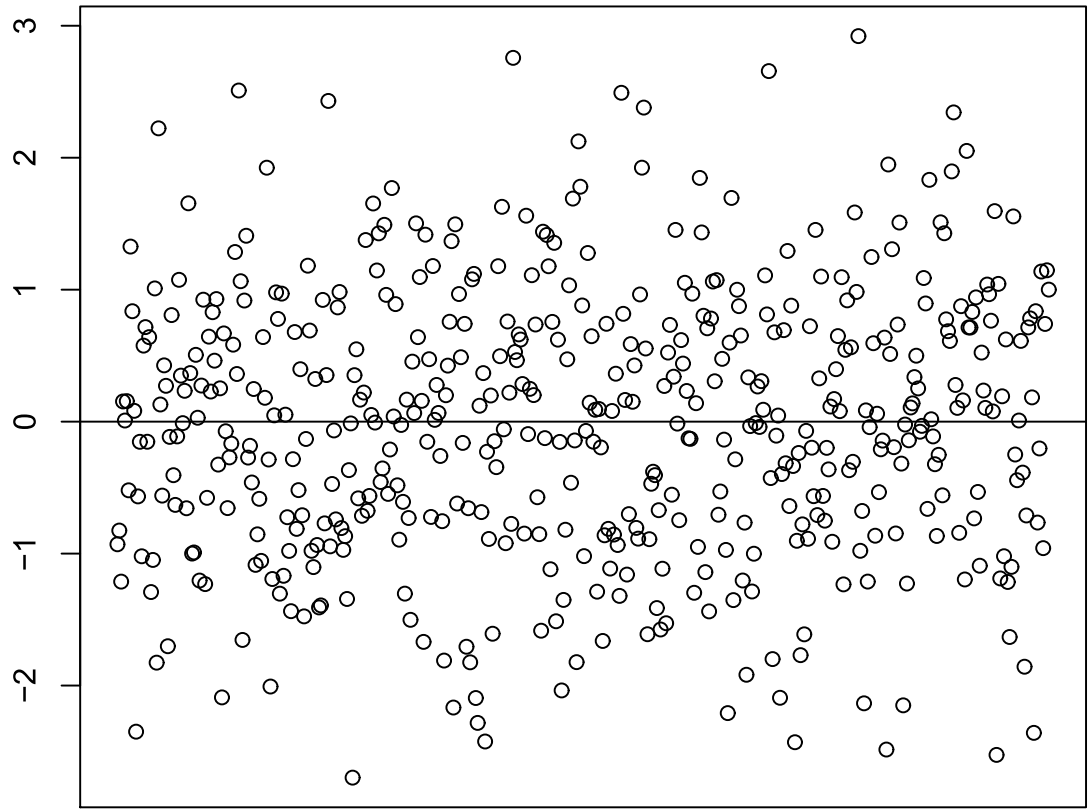
Es gibt also drei Arten, R^2_{XY} zu verstehen:

Eigenschaften:

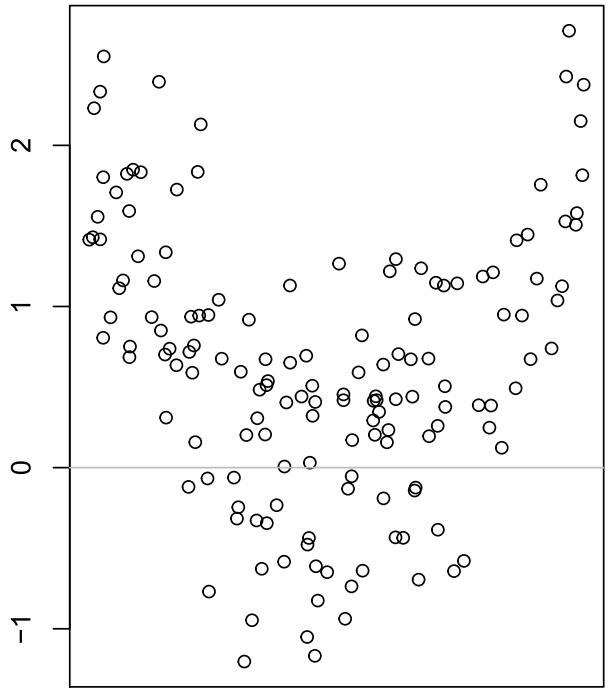
- Es gilt: $0 \leq R^2_{XY} \leq 1$.
- $R^2_{XY} = 0$: Es wird keine Streuung erklärt, d.h. es gibt keinen (linearen) Zusammenhang zwischen X und Y.
- $R^2_{XY} = 1$: Die Streuung wird vollständig erklärt. Alle Beobachtungen liegen tatsächlich auf einer Geraden.

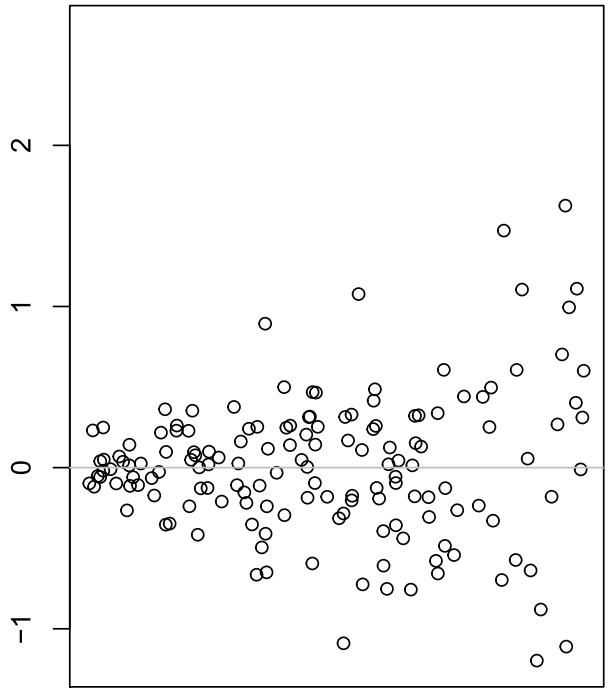
Residualplots

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.

6.3 Nominale Einflussgrößen in Regressionsmodellen, Varianzanalyse

6.3.1 Dichotome Kovariablen

Bisher wurden $Y, X_1, X_2, \ldots, X_p$ als metrisch vorausgesetzt. Ähnlich wie für Korrelationskoeffizienten können dichotome Variablen, sofern sie mit 0 und 1 (wichtig!) kodiert sind, ebenfalls als Einflussgrößen zugelassen werden können.

Die zugehörigen Koeffizienten geben dann an, um wieviel sich Y – ceteris paribus – erhöht, wenn die entsprechende Kovariable den Wert 1 statt 0 hat.

Beispiel: Einfluss von Arbeitszeit und Geschlecht auf das Einkommen.

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \\
\text{mit} \qquad X_1 &= \begin{cases} 1 & \text{männlich} \\ 0 & \text{weiblich} \end{cases} \\
X_2 &= \text{(vertragliche) Arbeitszeit} \\
Y &= \text{Einkommen}
\end{aligned}
$$

Interpretation:

Würde man ansetzen $X_1 = \begin{cases} 1 & \text{weiblich} \\ 0 & \text{männlich} \end{cases}$,

so ergäben sich dieselben Schätzungen für $\hat{\beta}_0$ und $\hat{\beta}_2$, die Schätzung für $\hat{\beta}_1$ wäre betragsmäßig gleich, aber mit umgekehrten Vorzeichen. (also: positiver Männereffekt $\Longleftrightarrow$ negativer Fraueneffekt)

6.3.2 Interaktionseffekte

Wechselwirkung zwischen Kovariablen lassen sich durch den Einbezug des Produkts als zusätzliche Kovariable modellieren

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

β_3 gibt den Interaktions- oder Wechselwirkungseffekt an. Dieser lässt sich insbesondere bei dichotomen Kovariablen einfach interpretieren:

Fortsetzung des Beispiels:

Die geschätzte Regressionsgerade hat bei den Männern die Form

$$\begin{aligned} \hat{y}_i &= \\ &= \end{aligned}$$

und bei den Frauen die Form

$$\begin{aligned} \hat{y}_i &= \\ &= \end{aligned}$$

6.3.3 Dummykodierung

Betrachten wir nun ein nominales Merkmal X mit q Kategorien, z.B. Parteipräferenz

Man beachte, dass man umbedingt $q-1$ und nicht q Dummyvariablen verwendet, da sonst die Schätzwerte völlig willkürlich und unsinnig werden.

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

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

Idee:

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

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

Beispiel zur Interpretation:

Y: Score auf Autoritarismusskala

X bzw. X_1, X_2: Parteienpräferenz

X_3: Einkommen

6.3.4 Varianzanalyse

Ist ein nominales Merkmal X mit insgesamt k verschiedenen Ausprägungen die einzige unabhängige Variable, so führt die Regressionsanalyse mit den entsprechenden $k-1$ Dummyvariablen auf die sogenannte (einfaktorielle) Varianzanalyse, die insbesondere in der Psychologie als Auswertungsmethode sehr verbreitet ist.

Als Schätzwert $\hat{y}_i$ ergibt sich für jede Einheit i genau der Mittelwert aller Werte y_i, die zu Einheiten l gehören, die dieselben Ausprägungen bei dem Merkmal X, also den zugehörigen Dummyvariablen $X_1, \ldots, X_{k-1}$, haben. Man bildet also k Gruppen bezüglich X, und $\hat{y}_i$ ist der Mittelwert der Gruppe, zu der i gehört.

Beispiel:

Y Autoritarismusscore

X Parteienpräferenz

X_1 CDU/CSU oder FDP, X_2 SPD oder Grüne, X_3 Sonstiges

Die Streuungszerlegung

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 \quad = \quad \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

der linearen Regression vereinfacht sich in diesem Fall und hat eine ganz charakteristische Form:

Indiziert man die Beobachtungen um und betrachtet die k Gruppen, so hat man in der j-ten Gruppe n_j Beobachtungen $y_{1j}, y_{2j}, \ldots, y_{n_j j}$ und den Gruppenmittelwert $\bar{y}_j$. Damit erhält man:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-\bar{y})^2=\sum_{j=1}^{k} n_j\cdot(\bar{y}_j-\bar{y})^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2.$$

Dies ist genau die Streuungszerlegung aus Kapitel 3!

Das zugehörige Bestimmtheitsmaß wird üblicherweise mit η^2 bezeichnet:

$$\eta^2 = \frac{SQE}{SQT} = \frac{\sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2}.$$

η^2 und $\eta = \sqrt{\eta^2}$ werden auch als Maße für den Zusammenhang zwischen einer metrischen Variable und einer nominalen Variable verwendet.

Mehr dazu am Ende von Statistik II...

6.4 Korrelation und „Kausalität“

Tücken bei der Interpretation von Zusammenhängen und Regressionsmodellen

Alle Zusammenhangsmaße messen ausschließlich die statistische Koinzidenz von Variablenwerten. Ob tatsächlich eine echte wirkende Beziehung vorliegt, kann – bestenfalls – aufgrund substanzwissenschaftlicher Überlegungen entschieden werden. In einem strengen Sinn bedürfen Kausalaussagen ohnehin eines experimentellen Designs.

- erstes (kleineres) Problem:
  Viele Zusammenhangsmaße sind symmetrisch, Kausalität ist eine gerichtete Beziehung.

- zweites, sehr schwerwiegendes Problem:
  Die falsche Beurteilung von Zusammenhängen entsteht insbesondere dadurch, dass entscheidende Variablen nicht in die Analyse miteinbezogen werden.

klassisches (fiktives) Beispiel:
Erhebung aus den 60er Jahren von Gemeinden:

X	Anzahl der Störche
Y	Anzahl der neugeborenen Kinder

X und Y sind hochkorreliert.
$\Rightarrow$ Störche bringen die Kinder...?

Weiteres Beispiel aus Gemeindestudie:

X	Alter
Y	Anzahl ausgeliehener Bücher in Bibliothek

X und Y stark negativ korreliert
$\Rightarrow$ Angebot für alte Gemeindemitglieder schlecht?

Rechnerischer Ausweg