

# **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.

### Beispiele:

- Nettomiete  $\longleftrightarrow$  Wohnfläche
- Monatseinkommen  $\longleftrightarrow$  Alter in Jahren
- Wochenarbeitseinkommen  $\longleftrightarrow$  Wochenarbeitsstunden
- Wochenarbeitsstunden  $\longleftrightarrow$  Hausarbeit in Stunden pro Woche
- Wochenarbeitsstunden (tatsächlich)  $\longleftrightarrow$  Wochenarbeit (vertraglich)

### 6.1.1 Streudiagramme (Scatterplots)

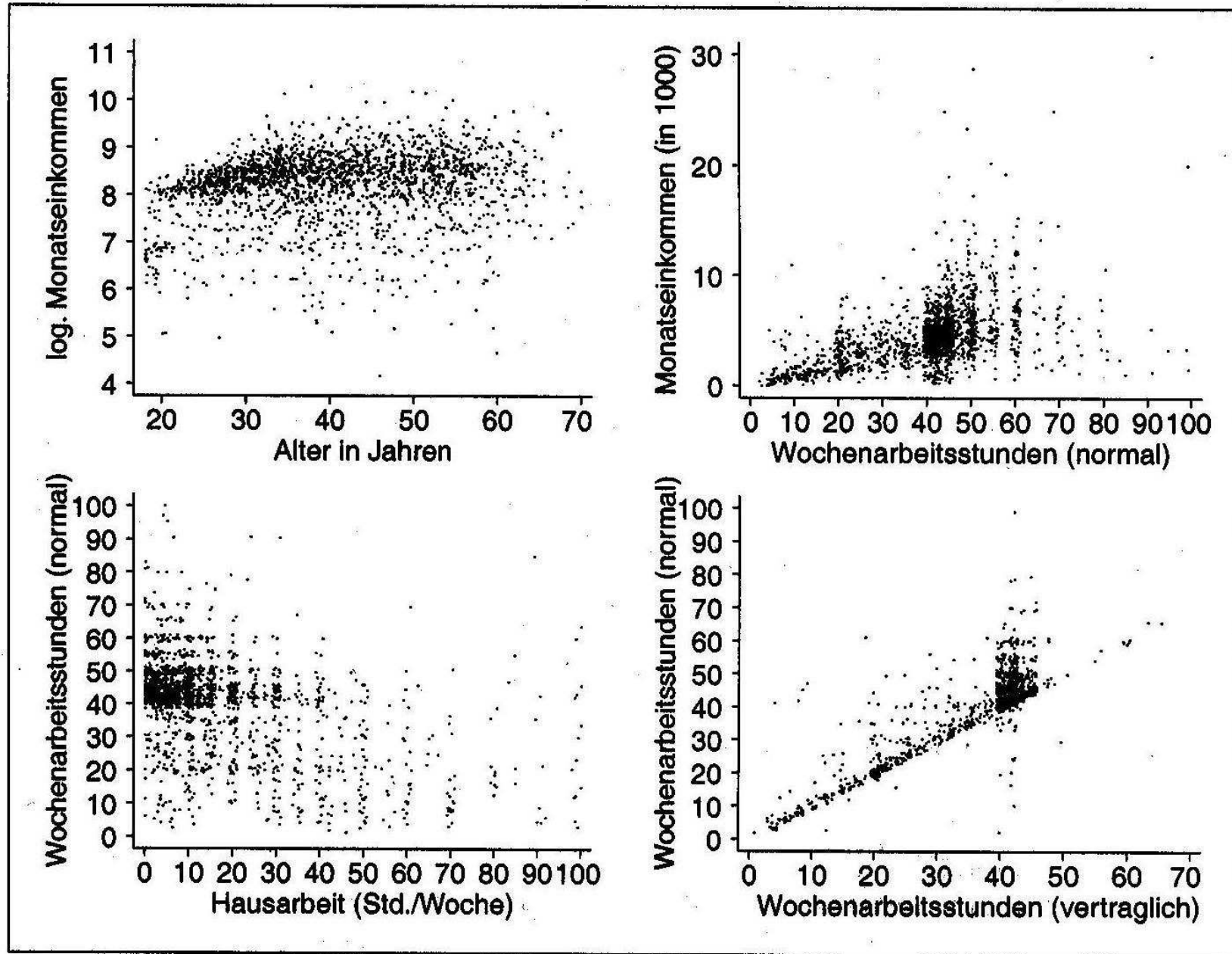
Sind die Merkmale stetig oder zumindestens quasi-stetig (sehr viele verschiedene Ausprägungen), werden Kontingenztabelle sehr unübersichtlich und praktisch aussageelos, da die einzelnen Häufigkeiten in den Zellen der Tabellen natürlicherweise durchwegs sehr klein sind.

Alternative Darstellungsform: *Scatterplot* / *Streudiagramm*:

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

- Guter optischer Eindruck über das Vorliegen, die Richtung und gegebenenfalls die Art eines Zusammenhangs.
- Ausreißer werden leicht erkannt.

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



## 6.1.2 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

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

oder

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

Zusammengefasst: Beobachtung  $i$  ist konkordant zu  $(\bar{x}, \bar{y})$ , wenn

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

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

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

oder

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

Zusammengefasst: Beobachtung  $i$  ist diskordant zu  $(\bar{x}, \bar{y})$ , wenn

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

|            |            |
|------------|------------|
| diskordant | konkordant |
| konkordant | diskordant |

- 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).

⇒ 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 heißt

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

(empirische) Kovarianz von  $X$  und  $Y$ ,

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

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

$$R_{XY}^2 := (\varrho(X, Y))^2 \tag{6.1}$$

Bestimmtheitsmaß von  $X$  und  $Y$ .



## Bemerkungen:

- Die Kovarianz  $Cov(X, Y)$  ist nicht maßstabsunabhängig.
- Das Teilen durch die Standardabweichungen normiert die Kovarianz und macht sie maßstabsunabhängig.

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

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:

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

- Die Größe der Kovarianz für sich genommen ist kaum interpretierbar. 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 ein Verschiebungssatz:

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

und damit

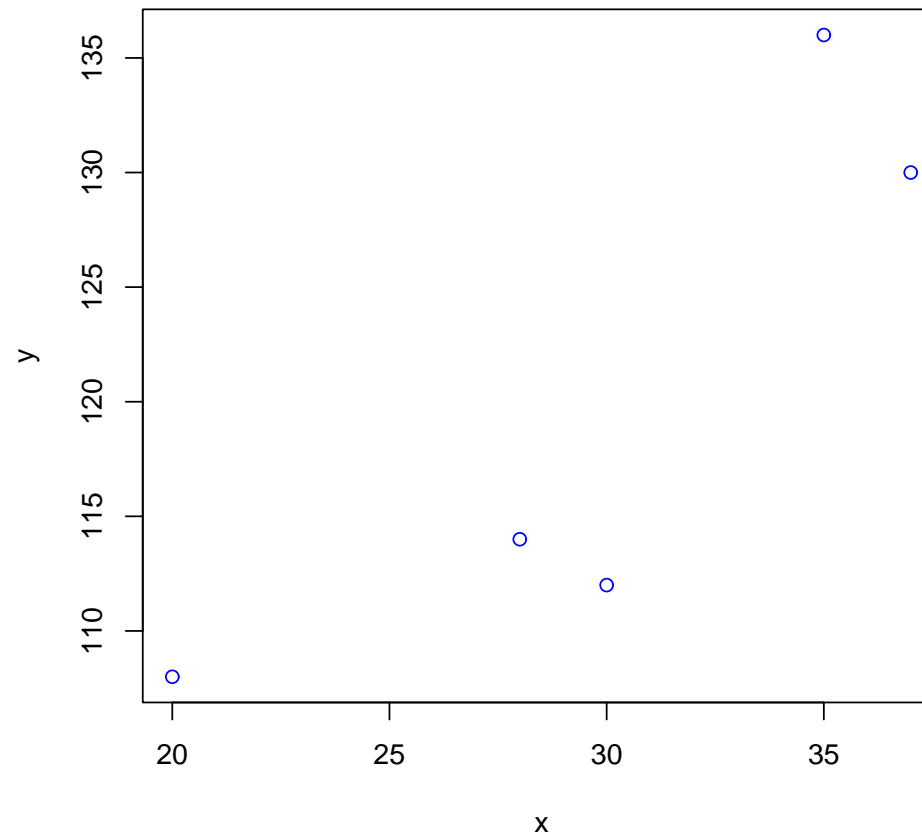
$$\rho(X, Y) = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sqrt{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \cdot \sqrt{\sum_{i=1}^n y_i^2 - n \bar{y}^2}}.$$

Zur Erinnerung:

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

**Beispiel 1:** Gegeben seien die Datenpaare

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



Es gilt:  $n = 5$ ,  $\bar{x} = 30$  und  $\bar{y} = 120$

Tabelle:

|          | $x_i$ | $y_i$ | $x_i \cdot y_i$ | $x_i^2$ | $y_i^2$ |
|----------|-------|-------|-----------------|---------|---------|
|          | 37    | 130   | 4810            | 1369    | 16900   |
|          | 30    | 112   | 3360            | 900     | 12544   |
|          | 20    | 108   | 2160            | 400     | 11664   |
|          | 28    | 114   | 3192            | 784     | 12996   |
|          | 35    | 136   | 4760            | 1225    | 18496   |
| $\Sigma$ | 150   | 600   | 18282           | 4678    | 72600   |

Basierend auf diesen Hilfsgrößen berechnet sich der Korrelationskoeffizient gemäß Verschiebungssatz als

$$\rho(X, Y) = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sqrt{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \cdot \sqrt{\sum_{i=1}^n y_i^2 - n \bar{y}^2}} = 0.863$$

## Beispiel 2:

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

| $x_i$ | $y_i$ | $x_i \cdot y_i$ |
|-------|-------|-----------------|
| 10    | 100   | 1000            |
| 20    | 0     | 0               |
| 30    | 100   | 3000            |

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

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

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

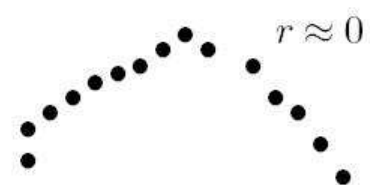
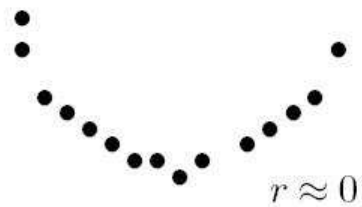
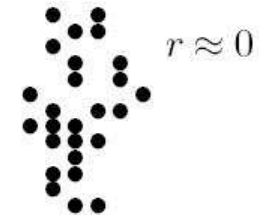
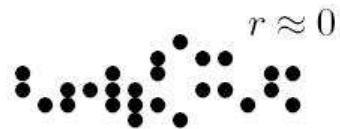
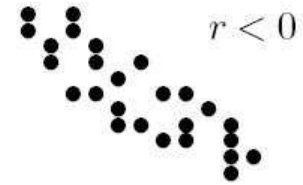
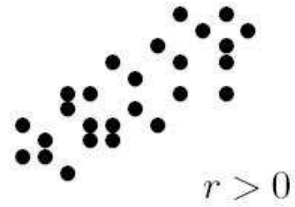
## Bemerkungen:

- Allgemein zeigen  $|\varrho|$  und  $R^2$  die Stärke eines *linearen* Zusammenhangs an, also wie gut sich die Datenpaare  $(x_1, y_1), \dots, (x_n, y_n)$  durch eine *Gerade* beschreiben lassen.
- Die Zusammenhangsmaße sind invariant gegenüber Vertauschen von  $Y$  und  $X$ , unterscheiden also nicht, welche Variable als abhängige, welche als unabhängige gilt:

$$\varrho(X, Y) = \varrho(Y, X), \quad R_{XY}^2 = R_{YX}^2.$$

- 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  $Cov(X, Y) = 0$ ), so nennt man  $X$  und  $Y$  *unkorreliert*. Es besteht dann keinerlei linearer Zusammenhang.

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





- $R^2$  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:
  - $\rho_{XY} \approx 0$ : kein (linearer) Zusammenhang.
  - $\rho_{XY} > 0$ : positive Korrelation, gleichgerichteter (linearer) Zusammenhang.
  - $\rho_{XY} < 0$ : negative Korrelation, gegengerichteter (linearer) Zusammenhang.
  - $|\rho_{XY}| \leq 0.5$ : schwache Korrelation.
  - $0.5 < |\rho_{XY}| \leq 0.8$ : mittlere Korrelation.
  - $|\rho_{XY}| > 0.8$ : starke Korrelation.

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

$$\varrho(\tilde{X}, \tilde{Y}) = \varrho(X, Y) \quad \text{falls } a \cdot c > 0$$

und

$$\varrho(\tilde{X}, \tilde{Y}) = -\varrho(X, Y) \quad \text{falls } a \cdot c < 0.$$

Die Korrelation ist also in der Tat maßstabsunabhängig.

**Beispiel:** MietspiegelKorrelationen ( $n = 150$ )

|            | Nettomiete | Wohnfläche | Baujahr |
|------------|------------|------------|---------|
| Nettomiete |            |            |         |
| Wohnfläche | 0.600      |            |         |
| Baujahr    | 0.223      | -0.174     |         |

## Beispiele aus Jann (2002) S.87ff

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

### 6.1.3 Weitere Korrelationskoeffizienten

#### Anwendung des Korrelationskoeffizienten nach Bravais-Pearson auf dichotome nominale Merkmale

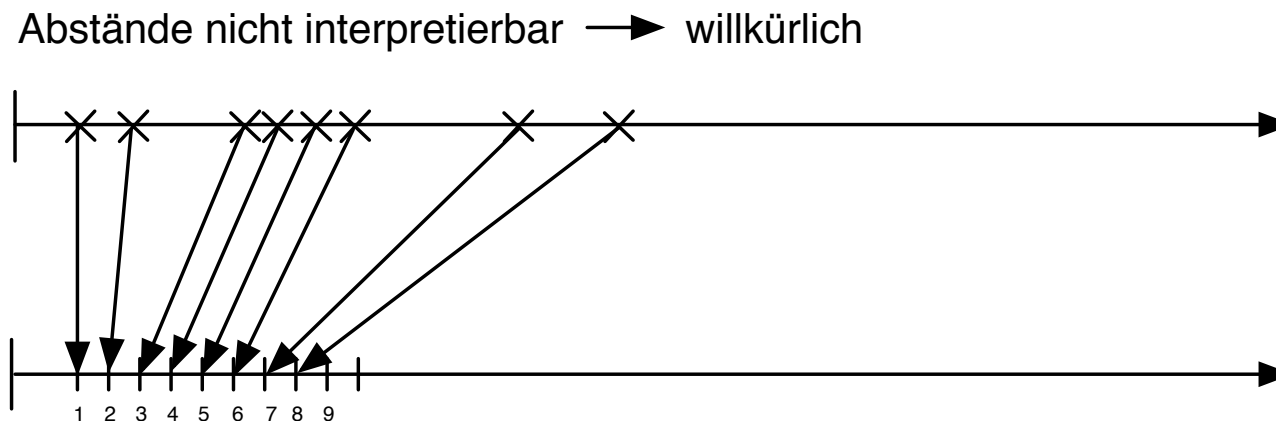
Liegen *dichotome* nominale Merkmale, d.h. Merkmale mit nur zwei ungeordneten Ausprägungen vor (z.B. ja/nein), *und* kodiert man die Ausprägungen mit 0 und 1, so kann man die Formel des Korrelationskoeffizienten nach Bravais-Pearson sinnvoll anwenden. Man erhält den sogenannten *Punkt-Korrelationskoeffizienten*,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

Im Fall einer dichotomen und einer metrischen Variablen ergibt sich bei Anwendung des Korrelationskoeffizienten nach Bravais-Pearson die sogenannte *punkt-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. Die arithmetischen Mittelwerte  $\bar{x}$ ,  $\bar{y}$  wären willkürliche Zahlen, ebenso  $(x_i - \bar{x})$ ,  $(y_i - \bar{y})$ .

### Beispiel



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

$$\text{rg}(x_i) = j : \iff x_i = x_{(j)},$$

d.h. der Rang  $\text{rg}(x_i)$  ist die Nummer, die  $x_i$  in der geordneten Urliste  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  einnimmt (analog für  $\text{rg}(y_i)$ ). Der kleinsten Beobachtung wird also der Wert 1 zugeordnet, der zweitkleinsten der Wert 2, usw., der größten der Wert  $n$ .

Beispiel:

|                  |   |   |   |     |    |
|------------------|---|---|---|-----|----|
| $x_i$            | 1 | 7 | 2 | 5.3 | 16 |
| $\text{rg}(x_i)$ |   |   |   |     |    |

- 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.

Beispiel:

|       |   |   |   |   |    |
|-------|---|---|---|---|----|
| $x_i$ | 1 | 7 | 7 | 3 | 10 |
|       |   |   |   |   |    |
|       |   |   |   |   |    |

- Es gilt

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

Damit ergibt sich für den Korrelationskoeffizienten nach Bravais-Pearson auf die Rangdaten der sogenannte Rangkorrelationskoeffizient nach Spearman.



**Definition:**

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

heißt (empirischer) *Rangkorrelationskoeffizient nach Spearman*.

**Bemerkungen:**

- Liegen keine Bindungen vor, so gilt

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

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

- Wichtig für Interpretation: Da sich  $\rho_S(X, Y)$  aus der Anwendung von  $\rho(X, Y)$  auf Rangdaten ergibt, behalten die entsprechenden Bemerkungen zum Bravais-Pearson-Korrelationskoeffizienten – auf die Ränge bezogen – ihre Gültigkeit. Insbesondere gilt  $-1 \leq \rho_{S,XY} \leq 1$ , und  $\rho_{S,XY}$  ist analog zu interpretieren.
- Die Bildung von Rängen ist unempfindlich gegenüber Ausreißern, so dass auch der Rangkorrelationskoeffizient ausreißerresistent ist.
- 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.

Beispiel:  $y = x^3$

|          | $x_i$ | $y_i$  | $x_i \cdot y_i$ | $x_i^2$ | $y_i^2$    |
|----------|-------|--------|-----------------|---------|------------|
|          | -40   | -64000 | 2560000         | 1600    | 4096000000 |
|          | -20   | -8000  | 160000          | 400     | 64000000   |
|          | 0     | 0      | 0               | 0       | 0          |
|          | 5     | 125    | 625             | 25      | 15625      |
|          | 10    | 1000   | 10000           | 100     | 1000000    |
|          | 20    | 8000   | 160000          | 400     | 64000000   |
|          | 25    | 15625  | 390625          | 625     | 244140625  |
| $\Sigma$ | 0     | -47250 | 3281250         | 3150    | 4469156250 |

Es ist  $\bar{x} = 0, \bar{y} = -6750, \tilde{s}_X^2 = 450, \tilde{s}_Y^2 = 592888393, Cov(X, Y) = 468750$  und damit  $\rho_{X,Y} = 0.907504$ .

Es ist  $rg(x_i) = rg(y_i)$  für alle  $i$ , daher ist  $\rho_s(X, Y) = 1$ .

**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 |
| $\text{rg}(x_i)$  |    |    |    |    |    |
| $\text{rg}(y_i)$  |    |    |    |    |    |

$$\begin{aligned}
 \rho_s(X, Y) &= \frac{\sum_{i=1}^n \text{rg}(x_i) \text{rg}(y_i) - n \left(\frac{n+1}{2}\right)^2}{\sqrt{\sum_{i=1}^n (\text{rg}(x_i))^2 - n \left(\frac{n+1}{2}\right)^2} \sqrt{\sum_{i=1}^n (\text{rg}(y_i))^2 - n \left(\frac{n+1}{2}\right)^2}} \\
 &= \frac{(1 \cdot 2 + 2 \cdot 1 + 3.5 \cdot 3 + 3.5 \cdot 4 + 5 \cdot 5) - 5 \cdot \left(\frac{5+1}{2}\right)^2}{\sqrt{1^2 + 2^2 + 3.5^2 + 3.5^2 + 5^2 - 5 \cdot \left(\frac{5+1}{2}\right)^2} \sqrt{2^2 + 1^2 + 3^2 + 4^2 + 5^2 - 5 \cdot \left(\frac{5+1}{2}\right)^2}} \\
 &= 0.872
 \end{aligned}$$

### 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: Die lineare Einfachregression

### 6.2.1 Grundbegriffe und Hintergrund

#### Bedeutung der Regression:

- Eines der am häufigsten verwendeten statistischen Verfahren. Vielfache Anwendung in den Sozialwissenschaften → Analoge Ausdehnung auf viele Variablen möglich!
- Grundidee der Interpretation bleibt bei vielen allgemeineren Modellen erhalten, die hier aber nicht betrachtet werden.

## 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  $(x_i, y_i)$ ,  $i = 1, \dots, n$ , durch eine Gerade beschreiben?“
- Die Regression geht nun einen Schritt weiter:
  - Wie sieht die am besten passende Gerade aus?
  - Ziel ist eine genaue Analyse und Beschreibung des Zusammenhangs.
- Zusätzliche Ziele der Regressionsanalyse:
  - „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? (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).
- 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              |
| Prädiktor            |                        |
| Kovariablen          |                        |



## 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  $i = 1, \dots, n$  gelten

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

Normalerweise besteht kein perfekter linearer Zusammenhang, so dass ein unerklärter Rest  $\varepsilon_i$  in die Modellgleichung mit aufgenommen wird (in Statistik 2 wird  $\varepsilon_i$  als zufälliger Fehler interpretiert werden):

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

Dies ist das Modell der linearen Einfachregression.

$a$  und  $b$  sind unbekannte Größen, die sogenannten **Regressionsparameter**, die anhand der Daten bestimmt werden müssen. (Man beachte hierbei, dass  $a$  und  $b$  keinen Index tragen; sie werden hier als interindividuell konstant betrachtet.)

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 (KQ)-Kriterium**

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

bezüglich  $\hat{a}$  und  $\hat{b}$ .

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

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

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 wie auch negative Abweichungen von der Regressionsgeraden gleichermaßen 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. Eine robuste Alternative ist das Betrachten des Betrags  $\rightarrow$  „ $\mathcal{L}^1$ -Regression“.)

**Satz:** Für die KQ-Schätzer gilt

i)

$$\begin{aligned}\hat{b} &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{Cov}(X, Y)}{\tilde{s}_X^2} \\ &= \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \cdot \bar{y}}{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \varrho(X, Y) \frac{\tilde{s}_Y}{\tilde{s}_X}\end{aligned}$$

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  $\varrho(X, Y)$ .
- 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  $\hat{a} = \bar{y}$ . In der Tat:  $\bar{y}$  minimiert  $\sum_{i=1}^n (y_i - a)^2$ .

## 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 Beispiel:** Kaffeeverkauf an drei Tagen $X$  Anzahl verkaufter Tassen Kaffee $Y$  zugehöriger Gewinn

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

| $i$   | $y_i$ | $y_i - \bar{y}$ | $(y_i - \bar{y})(x_i - \bar{x})$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ | $x_i$ |
|-------|-------|-----------------|----------------------------------|-----------------|---------------------|-------|
| 1     | 9     |                 |                                  |                 |                     | 10    |
| 2     | 21    |                 |                                  |                 |                     | 15    |
| 3     | 0     |                 |                                  |                 |                     | 5     |
| Summe | 30    | 0               | 105                              | 0               | 50                  | 30    |

$$\bar{y} = 10 \text{ und } \bar{x} = 10.$$



$$\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{0 \cdot (-1) + 5 \cdot 11 + 5 \cdot 10}{0 + 5^2 + (-5)^2} = \frac{105}{50} = 2.1$$

$$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x} = 10 - 2.1 \cdot 10 = -11$$

Vohergesagte Werte  $\hat{y}_i = \hat{a} + \hat{b}x_i$  und Residuen  $\hat{\varepsilon}_i = y_i - \hat{y}_i$ :

$$\begin{array}{ll} \hat{y}_1 = -11 + 2.1 \cdot 10 = 10 & \hat{\varepsilon}_1 = 9 - 10 = -1 \\ \hat{y}_2 = & \hat{\varepsilon}_2 = \\ \hat{y}_3 = & \hat{\varepsilon}_3 = \end{array}$$

Prognose:

$$x^* = 12 \quad \Longrightarrow \quad \hat{y}^* = \hat{a} + \hat{b} \cdot x^* =$$

### 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:

$$SQT := \sum_{i=1}^n (y_i - \bar{y})^2$$

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

Modell 2: Vorhersage von  $Y$  mit  $X$  **basierend auf den KQ-Schätzern:**

$$\hat{y}_i = \hat{a} + \hat{b} \cdot x_i$$

Dabei gemachter Gesamtfehler:

$$SQR := \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \varepsilon_i^2$$

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

Die Differenz

$$SQE := SQT - SQR$$

nennt man die durch das Regressionsmodell 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):

$$\frac{SQE}{SQT} = R_{XY}^2$$

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

Man kann also  $R_{XY}^2$  auf 3 Arten verstehen:

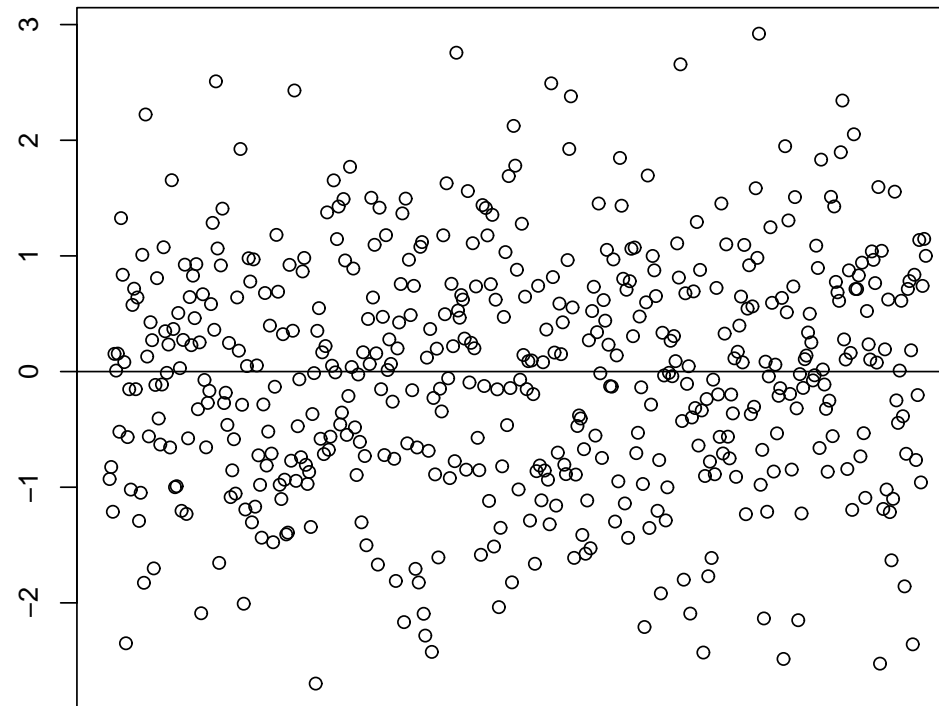
1. über den Korrelationskoeffizienten  $R_{XY}^2 = (\rho(X, Y))^2$  (vgl. (6.1))
2. als PRE-Maß
3. als Verhältnis der durch die Regression erklärten Variation und der Gesamtvariation.

### Eigenschaften:

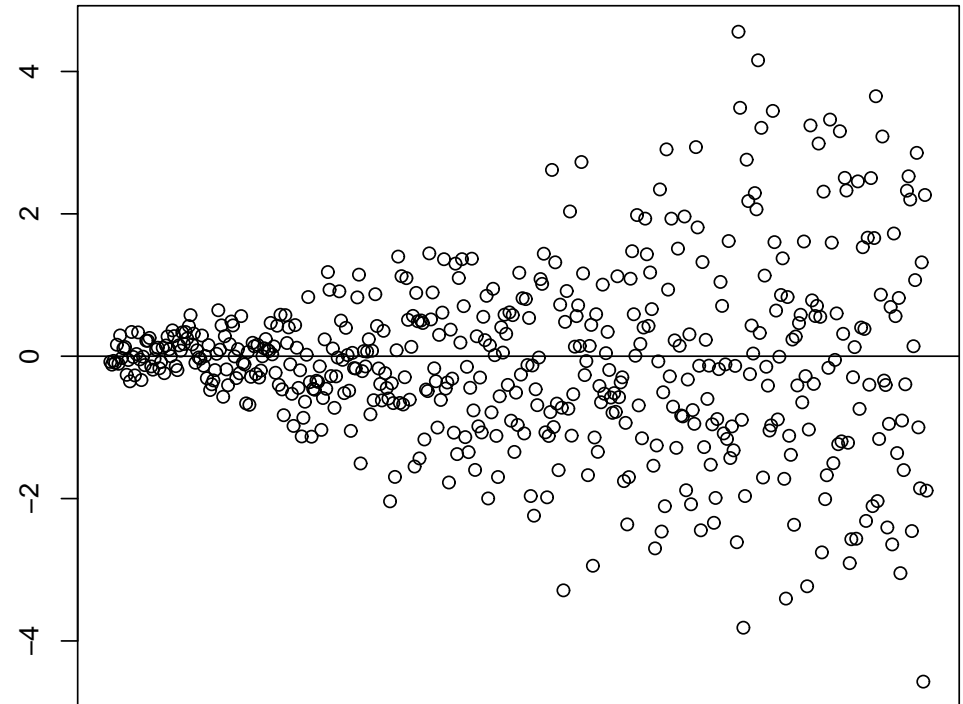
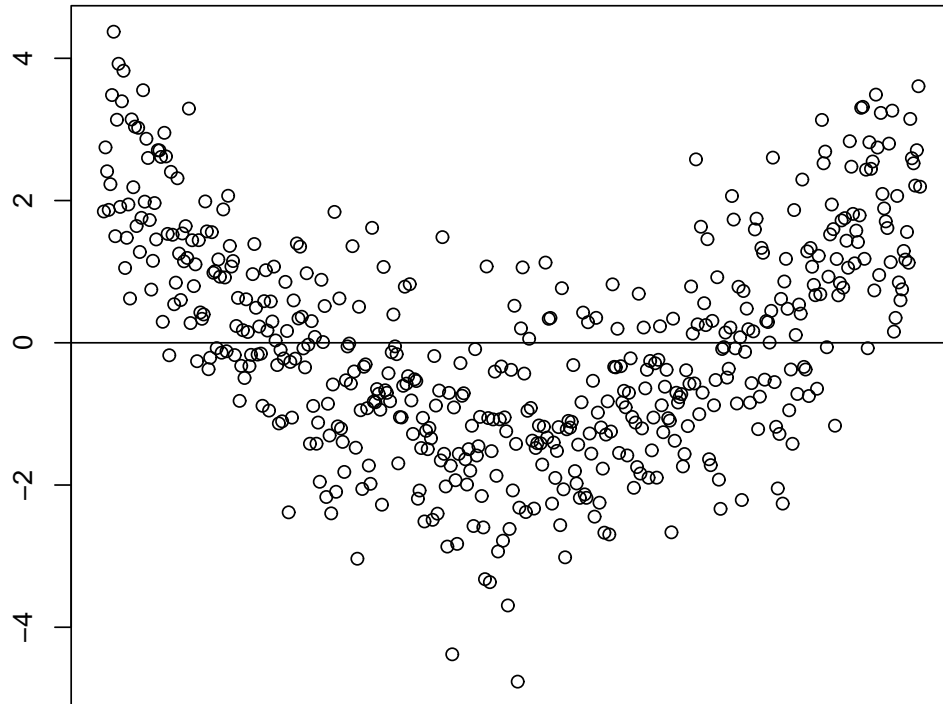
- Es gilt:  $0 \leq R_{XY}^2 \leq 1$ .
- $R_{XY}^2 = 0$ : Es wird keine Streuung erklärt, d.h. es gibt keinen (linearen) Zusammenhang zwischen  $X$  und  $Y$ .
- $R_{XY}^2 = 1$ : Die Streuung wird vollständig erklärt. Alle Beobachtungen liegen tatsächlich auf einer Geraden.

## 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.



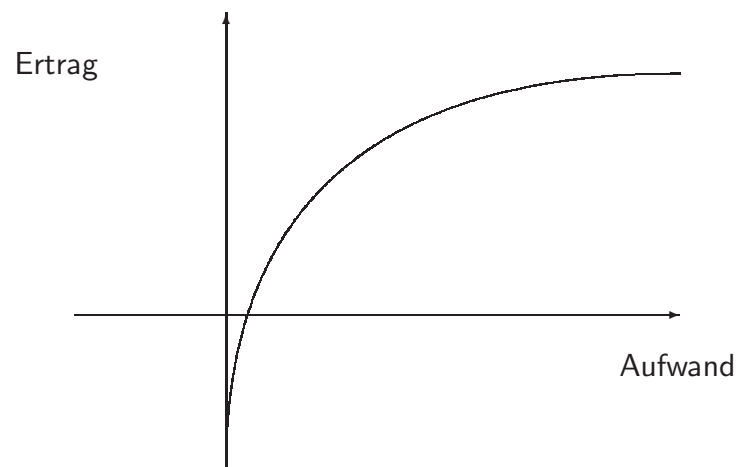
Zeigt sich eine Systematik, so war der lineare Ansatz unangemessen, und es ist größte Vorsicht bei der Interpretation geboten!



## 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.



Erstes Gossensches Gesetz:  
Abnehmender Grenznutzen

oder: Engelsches Gesetz: Aufwendungen für  
Lebensmittel in Abhängigkeit vom Einkommen



Viele (nicht alle) der auf den ersten Blick nichtlinearen Modelle lassen sich durch geeignete Variablentransformationen in die lineare Regressionsrechnung einbetten.

Der Ansatz

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

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

Definiere dazu die Merkmale  $Y^* = g(Y)$  und  $X^* = h(X)$  und betrachte

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

$b$  bzw.  $\hat{b}$  gibt dann allerdings nicht direkt die Stärke der Elastizität von  $Y$  bezüglich  $X$  an, sondern die von  $Y^*$  bezüglich  $X^*$ . Eine geeignete Interpretation erhält man über den Ansatz:

$$b = \frac{\Delta Y^*}{\Delta X^*} = \frac{\text{Änderung in } Y^*}{\text{Änderung in } X^*}$$

Entscheidend ist die Linearität. Der Ansatz

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

ist zwar nicht linear bzgl.  $x_i$ , wohl aber linear bzgl.  $x_i^2$  (vgl. vorherige Folie).

Ein „echtes“ nichtlineares Modell ist beispielsweise

$$y_i = a + x_i^b + \varepsilon_i,$$

viele weitere nichtlineare Modelle ergeben sich aus der Theorie der generalisierten linearen Modelle, die insbesondere auch für kategoriales oder ordinales  $Y$  geeignet sind.

Sehr häufiger Ansatz:

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

Interpretation von  $b$ :

Erhöht man einen Wert von  $X$  um  $p$  Prozent, so erhöht sich der entsprechende  $Y$ -Wert etwa um  $b \cdot p$  Prozent, denn

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

## 6.2.5 Multiple lineare Regression

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

### Modellgleichung:

$$y = a + b_1x_{1i} + b_2x_{2i} + \dots + b_px_{pi} + \varepsilon_i$$

Dabei bezeichnet  $x_{1i}$  den für die  $i$ -te Beobachtung beobachteten Wert der Variablen  $X_1$ ,  $x_{2i}$  den Wert der Variablen  $X_2$ , usw.

**Interpretation:** Die Interpretation von  $a$  und  $b_1, \dots, b_p$  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:

$$\begin{array}{rcl} a & \rightarrow & \beta_0 \\ b_1 & \rightarrow & \beta_1 \\ & & \vdots \\ b_p & \rightarrow & \beta_p \end{array}$$

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

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

der Ausdruck

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

minimal wird.

Die Schätzungen  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$  sind nur mit Matrizenrechnung einfach darzustellen und insbesondere schwierig „von Hand“ zu berechnen.

Im Rahmen dieser Veranstaltung müssen Sie bei der multiplen Regression daher 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 ein Preis zu bezahlen ist.)

## Multiple lineare Regression in R:

```
> fuel = read.table("cars.txt", header=T)
> attach(fuel)
> names(fuel)
[1] "PRICE"      "WEIGHT"     "CITYMPG"    "HWYMPG"    "DISP"       "COMP"       "HP"
[8] "TORQUE"     "AUTO"       "CYLIN"     "COUNTRY"
```

# Umrechnen von Meilen pro Gallone in Liter pro 100 km  
# 1 Meile = 1.609344 km  
# 1 Gallone = 3.785412 Liter

```
> HWYLPK = 100/(HWYMPG*1.609344)*3.785412
```

# Umrechnen von Pfund in kg  
# 1 Pfund = 0.453592 kg

```
> WEIGHT2 = WEIGHT * 0.453592
```



```
# Lineares Modell für Verbrauch auf 100 km gegen Gewicht und PS
```

```
> fit1 = lm(HWYLPK~WEIGHT2+HP)
```

```
> summary(fit1)
```

Call:

```
lm(formula = HWYLPK ~ WEIGHT2 + HP)
```

Residuals:

| Min      | 1Q       | Median  | 3Q      | Max     |
|----------|----------|---------|---------|---------|
| -1.70257 | -0.43936 | 0.03178 | 0.49258 | 2.15776 |

Coefficients:

|             | Estimate  | Std. Error | t value | Pr(> t ) |     |
|-------------|-----------|------------|---------|----------|-----|
| (Intercept) | 4.2277560 | 0.3802245  | 11.12   | <2e-16   | *** |
| WEIGHT2     | 0.0008859 | 0.0003543  | 2.50    | 0.014    | *   |
| HP          | 0.0181969 | 0.0012873  | 14.14   | <2e-16   | *** |

## SPSS-Output einer multiplen Regression:

**Coefficients<sup>a</sup>**

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

<sup>a</sup> Dependent Variable: Y

## 6.3 Nominale Einflussgrößen in Regressionsmodellen, Varianzanalyse

### 6.3.1 Dichotome Kovariablen

Bisher wurden  $Y, X_1, X_2, \dots, X_p$  als metrisch vorausgesetzt.

Ähnlich wie für Korrelationskoeffizienten können dichotome Variablen, sofern sie mit 0 und 1 (wichtig!) kodiert sind, ebenfalls als Einflussgrößen zugelassen werden.

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.

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

mit

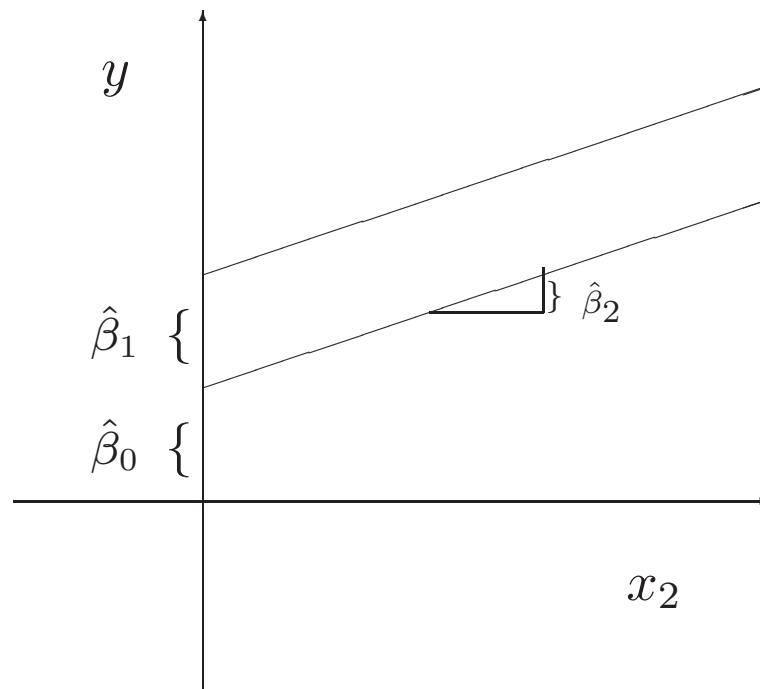
$$X_1 = \begin{cases} 1 & \text{männlich} \\ 0 & \text{weiblich} \end{cases}$$
$$X_2 = (\text{vertragliche}) \text{ Arbeitszeit}$$
$$Y = \text{Einkommen}$$

Interpretation: Die geschätzte Gerade für die Männer lautet:

$$\hat{y}_i =$$

Für die Frauen hingegen gilt

$$\hat{y}_i =$$



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

so ergäbe sich dieselbe Schätzung für  $\beta_2$ , die Schätzung für  $\beta_1$  wäre betragsmäßig gleich, aber mit umgekehrtem Vorzeichen (z.B. positiver Männereffekt  $\iff$  negativer Fraueneffekt).

## 6.3.2 Interaktionseffekte

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

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

$\beta_3$  gibt den Interaktions- oder Wechselwirkungseffekt an. Dieser lässt sich insbesondere bei dichotomen Kovariablen einfach interpretieren:

## Fortsetzung des Beispiels:

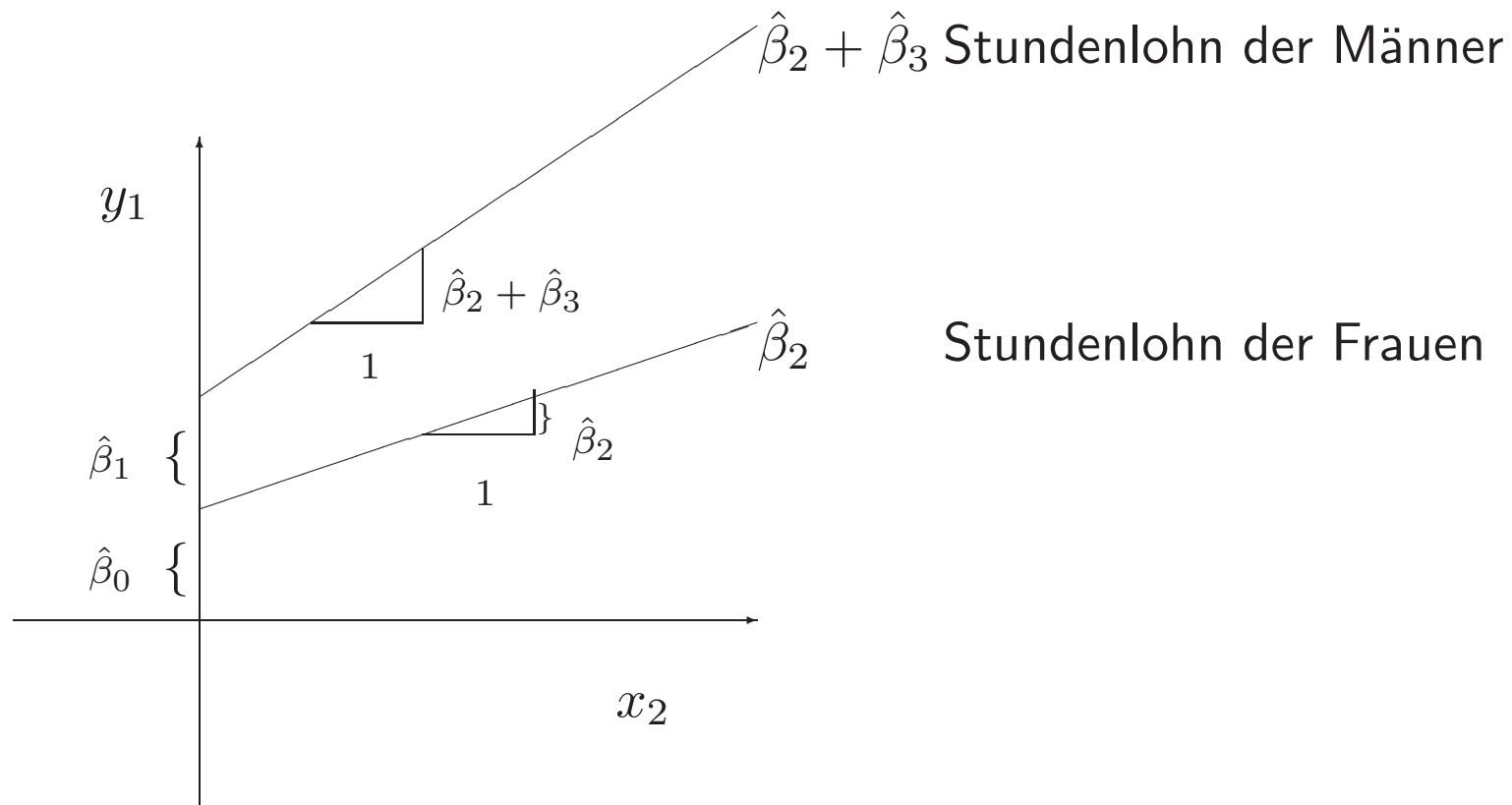
Die geschätzte Regressionsgerade hat bei den Männern die Form

$$\hat{y}_i =$$
$$=$$

und bei den Frauen

$$\hat{y}_i =$$
$$=$$

## Regressionsgerade:





### 6.3.3 Dummykodierung

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

Man beachte, dass man unbedingt  $q - 1$  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:

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

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

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

| $X$ | Parteien     | $X_1$ | $X_2$ |
|-----|--------------|-------|-------|
| 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

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

### 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 den Einheiten gehören, die dieselbe Ausprägung bei dem Merkmal  $X$ , also dieselben Ausprägungen bei den zugehörigen Dummyvariablen  $X_1, \dots, X_{k-1}$  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 Streuungszersetzung

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

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  $y_{1j}, y_{2j}, \dots, y_{n_j j}$  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 Streuungszersetzung aus Kapitel 3.2!

Das zugehörige Bestimmtheitsmaß wird üblicherweise mit  $\eta^2$  bezeichnet:

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

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

## 6.4 Korrelation und „Kausalität“

### Tücken bei der Interpretation von Zusammenhängen und Regressionsmodellen

Alle Zusammenhangsmaße messen rein 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.

⇒ Störche bringen die Kinder...?

Weiteres Beispiel aus Gemeindestudie:

$X$  Alter

$Y$  Anzahl ausgeliehener Bücher in Bibliothek

$X$  und  $Y$  stark negativ korreliert

⇒ Angebot für alte Gemeindemitglieder schlecht?



## Rechnerischer Ausweg

- Bei metrischem  $Y$ : multiple Regression, also mehrere Variablen einbeziehen; geschätzte Effekte sind „ceteris-paribus-Effekte“, also unter Konstanthalten aller anderen Variablen; Wirkung aller Variablen im Modell kontrolliert
- Auch bei kategorialem/ordinalem  $Y$  möglich: „Drittvariablenkontrolle“

Betrachte nicht nur Zusammenhang zwischen  $X$  und  $Y$ , sondern auch die „bedingten Zusammenhänge“ bei festen Werten von  $Z$ . Insbesondere beim Korrelationskoeffizient nach Bravais-Pearson spricht man von partieller Korrelation.

Beispiel: Zusammenhang zwischen Störche/Geburten **getrennt** auf dem **Land** und in der **Stadt**. (Die Korrelation verschwindet dann in dem Beispiel.)