5. ZWEI ODER MEHRERE METRISCHE MERKMALE

wenn an einer Beobachtungseinheit zwei (oder mehr) metrische Variablen erhoben wurden

2 wesentliche Problemstellungen:

Frage nach Zusammenhang:

Bsp.: Duxbury Press (siehe Kap. 1, "Einleitende Beispiele" 3)

Anzahl verschenkte Freiexemplare - Verkaufserlös

Besteht eine direkte Beziehung (ein Zusammenhang) zwischen der Anzahl verschenkter Exemplare und der Anzahl verkaufter Bücher, bzw. der Verkaufserträge?

Frage nach Unterschieden:

Bsp: Einfluß der Helmtragepflicht auf Fahrradfahren

- Diskussion über Einführung einer Helmtragepflicht
- Kritiker behaupteten, diese Pflicht entmutigt Rad zu fahren
- probeweise Einführung der Helmtragepflicht in Testorten
- repräsentative Stichprobe: wieviele km wurde in der Woche vor und der Woche nach Einführung des Gesetzes mit dem Rad zurückgelegt

Hat die Einführung der Helmtragepflicht Einfluß auf die Anzahl gefahrener Kilometer ? (oder anders formuliert)

Besteht ein Unterschied in der Anzahl gefahrener km vor und nach Einführung der Helmtragepflicht ?

WICHTIGE FRAGESTELLUNGEN
BEI ZWEI METRISCHEN MERKMALE

- Wie stark ist der Zusammenhang zwischen zwei metrischen Variablen ?

wenn man wissen möchte, wie eng zwei metrische Variablen mit einander verknüpft sind und ob der Zusammenhang positiv oder negativ ist.

Beispiel: Besteht ein Zusammenhang zwischen den Ausgaben für alkoholische Getränke und Tabakwaren ?

- Welche Form hat der Zusammenhang zwischen zwei Variablen ? Läßt sich der Wert einer Variable anhand des Wertes einer zweiten vorhersagen ?

wenn man wissen möchte, ob eine Variable von einer anderen abhängig ist und wie diese Abhängigkeitsstruktur aussieht.

Beispiel: Ist der Gebrauchtwagenpreis abhängig von der Zahl gefahrener Kilometer ? Kann man den Gebrauchtwagenpreis vorhersagen ?

- Unterscheiden sich die Mittelwerte zweier Variablen, die an einer Beobachtungseinheit erhoben wurden ?

Beispiel: Besteht ein Unterschied in der Anzahl gefahrener km vor und nach Einführung der Helmtragepflicht ?
Sind Dioptrienzahlen an linken und rechten Augen gleich ?

FRAGESTELLUNG 1:

Wie stark ist der Zusammenhang zwischen zwei metrischen Variablen ?

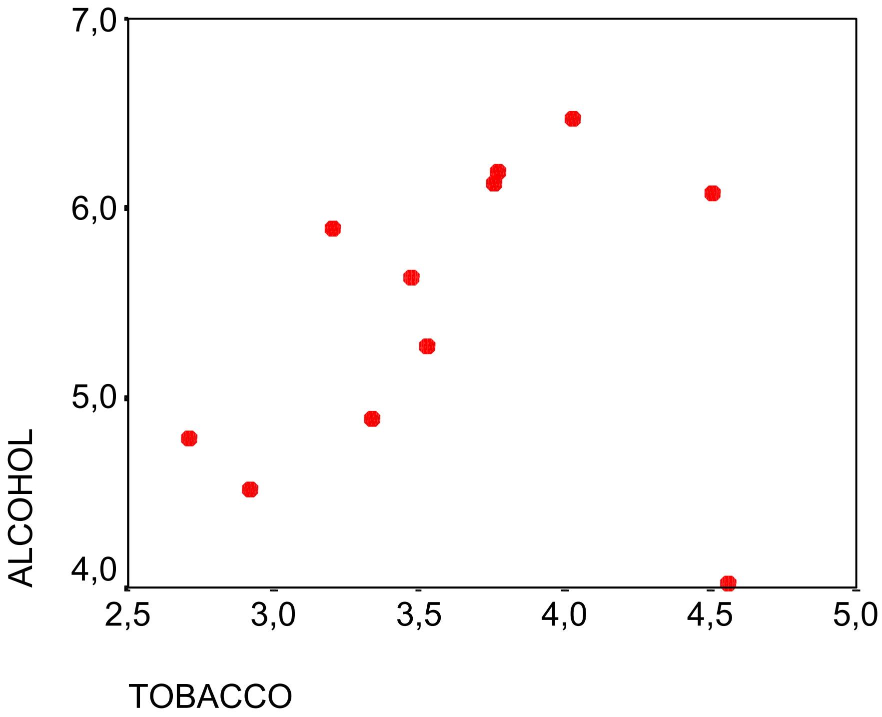
Bsp.: Ausgaben für Alkohol und Tabak

Besteht ein Zusammenhang zwischen den Ausgaben für alkoholische Getränke und Tabakwaren ?

erhoben wurden die durchschnittlichen Haushaltsausgaben pro Woche in Pfund für Alkohol und Tabakwaren in 11 britischen Regionen (1981)

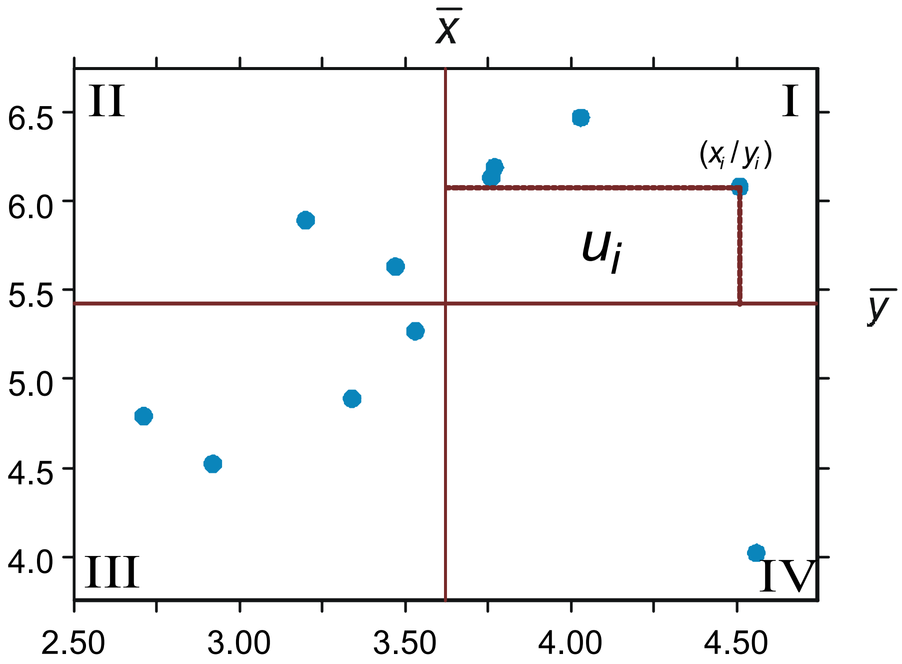
Variablen: - Ausgaben für Tabak (x-Achse)
- Ausgaben für Alkohol (y-Achse)

grafische Darstellung: **Streudiagramm** (*Scattergram*)



KORRELATIONSKOEFFIZIENT

(Korrelation misst Stärke des Zusammenhangs)
im Bsp.: positiver Zusammenhang
kleine y_i – kleine x_i $\longleftrightarrow$ große y_i – große x_i



Bilden der Produkte:

$$u_i = (x_i - \bar{x})(y_i - \bar{y})$$

Eigenschaften d. u_i
(i) liegen die Beobachtungen in I oder III: u_i **positiv**
(ii) in II oder IV: u_i **negativ**

Zusammenhangsmaß: Mittelwert der u_i

KOVARIANZ:

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

Eigenschaften der Kovarianz:

wenn die meisten Beobachtungen
in I, III: $\Rightarrow$ Cov(x,y) > 0
in II, IV $\Rightarrow$ Cov(x,y) < 0
wenn gleichmäßig in I,II,III,IV
$\Rightarrow$ Cov(x,y) $\approx$ 0

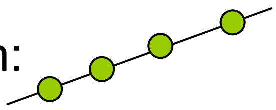
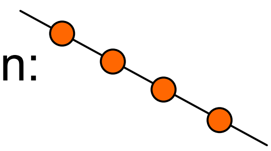
Nachteil: Kovarianz ist abhängig von Größe der Maßeinheit
Lösung: Normieren !

KORRELATIONSKOEFFIZIENT: r_{xy} (von K. Pearson)

$$r_{x,y} = r(x,y) = \frac{Cov(x,y)}{s_x \cdot s_y} = \frac{\frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum (y_i - \bar{y})^2}}$$

Eigenschaften:

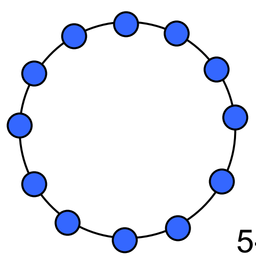
- ist ein normiertes Zusammenhangsmaß, $-1 \leq r_{x,y} \leq 1$
- bildet nur lineare Zusammenhänge ab:

$r_{xy} = 1$ wenn:  $r_{xy} = -1$ wenn: 

- Unkorreliertheit: ist nicht das Gleiche wie Unabhängigkeit

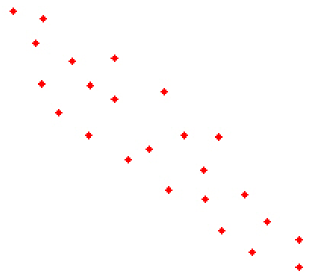
$r_{xy} = 0$

z.B. exakter *nicht-linearer* Zusammenhang

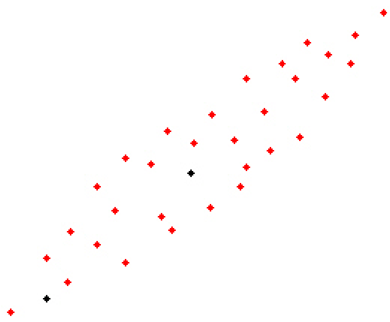


Beispiele für Zusammenhänge

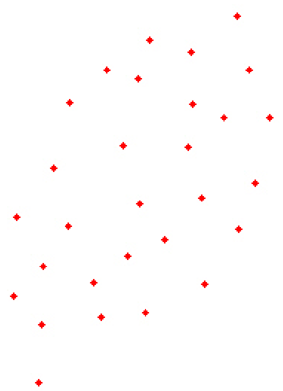
r(x,y) = -0.90



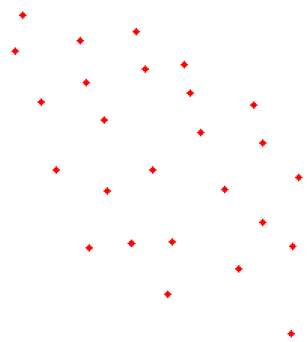
r(x,y) = 0.911



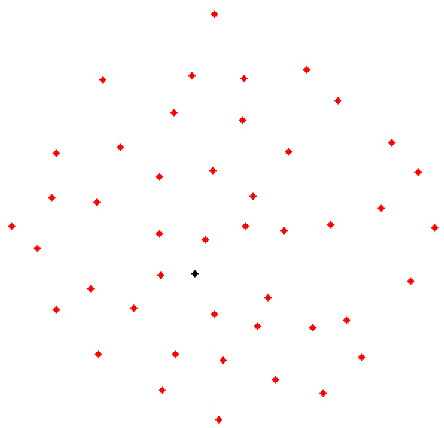
r(x,y) = 0.492



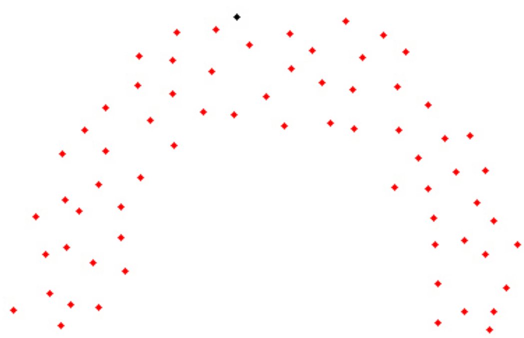
r(x,y) = -0.50



r(x,y) = -0.00



r(x,y) = -0.00



zur Berechnung des Korrelationskoeffizienten:

Beispiel:

gegeben:

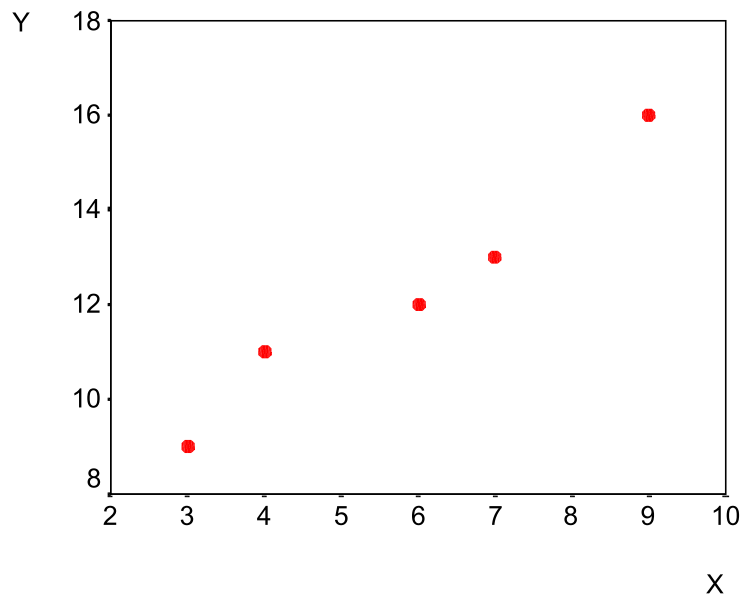
x:	3	6	7	9	4
y:	9	12	13	16	11

Berechnung von r_{XY}

	x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
	3	9	9	81	27
	6	12	36	144	72
	7	13	49	169	91
	9	16	81	256	144
	4	11	16	121	44
Summe	29	61	191	771	378
Mittelwert	5,8	12,2	38,2	154,2	75,6

$$r_{XY} = \frac{\frac{1}{n}\sum x_i y_i - \overline{x}\overline{y}}{\sqrt{\frac{1}{n}\sum x_i^2 - \overline{x}^2}\sqrt{\frac{1}{n}\sum y_i^2 - \overline{y}^2}} = \frac{75{,}5 - 5{,}8 \cdot 12{,}2}{\sqrt{38{,}2 - 5{,}8^2}\sqrt{154{,}2 - 12{,}2^2}} = 0{,}979$$

,



zurück zu Fragestellung 1:

Besteht ein Zusammenhang zwischen den Ausgaben für alkoholische Getränke und Tabakwaren ?

→ Grafik: Nordirland ist ein besonderer Fall (*outlier*), Ausgaben für Tabak hoch, für Alkohol niedrig (Irland billig ?)

Berechnung des Korrelationskoeffizienten einmal mit und einmal ohne Nordirland

$r_{XY} = 0{,}784$ (ohne) bzw. $r_{XY} = 0{,}224$ (mit)

Test eines Korrelationskoeffizienten:

H$_O$: $\rho = 0$ (ρ sprich "rho" ist Korrelation in Population)
H$_A$: $\rho \neq 0$ oder **H$_A$**: $\rho < 0$ oder **H$_A$**: $\rho > 0$

- **Teststatistik**: $T = \frac{r_{XY}}{\sqrt{1 - r_{XY}^2}} \sqrt{n-2}$

für kitischen Wert: *t*-Verteilung mit *df* = *n - 2*
wenn | *T* | größer als kritischer Wert → H$_O$ verwerfen

- zweiseitige **p-values** (Signifikanzwerte) aus SPSS, bzw. R:

  p value = 0,007 (ohne Nordirland)
  p value = 0,509 (mit Nordirland)

Resultat: es besteht ein starker Zusammenhang zwischen durchschnittlichen Ausgaben für Alkohol und Tabakwaren, wenn Nordirland nicht berücksichtigt wird

RANGKORRELATION (*Spearman's Rho*)

Voraussetzung für Pearson's Korrelationskoeffizient r_{XY} :
$\rightarrow$ beide Variablen **intervallskaliert** und **normalverteilt**

wenn diese Voraussetzung nicht erfüllt sind:

Ausweichen auf Methoden für ordinale Daten
diese werden auch *nichtparametrische* oder *parameterfreie Methoden* genannt

wenn Voraussetzungen für bestimmte Methoden für metrische Daten nicht erfüllt sind, verwendet man die entsprechende Methoden für ordinale Daten (wenn möglich)

Idee:

- Rangreihung der Daten (wie bei Median)
- diese Zahlen (Ränge) als Daten verwenden
- Problem: Bindungen (engl. *ties*)
  wenn mehrere Daten gleich groß sind, z.B.

Daten:	1	4	4	6	8	8	8	11	usw.
Ränge:	1	2	3	4	5	6	7	8	...

- Lösung: Vergeben des Mittelwerts der Ränge, d.h.

Ränge:	1	2,5	2,5	4	6	6	6	8	...

SPEARMAN'S RANGKORRELATION:

- jede der beiden Variablen rangreihen
- Berechnung wie r_{XY}, aber mit Rängen (nicht mit Daten)
- Vorgehen beim Testen ebenso wie bei r_{XY}

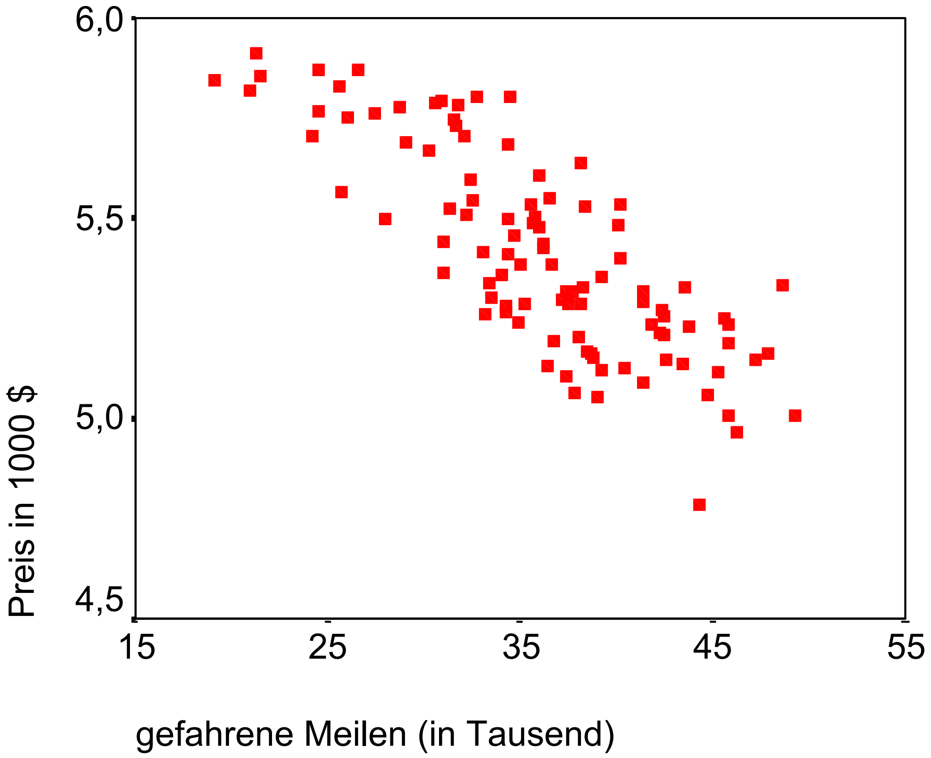
FRAGESTELLUNG 2:

- Welche Form hat der Zusammenhang zwischen zwei Variablen ?
- Läßt sich der Wert einer Variable anhand des Wertes einer zweiten vorhersagen ?

Bsp.: Gebrauchtwagenpreise (USA)

Ist der Gebrauchtwagenpreis abhängig von der Zahl gefahrener Meilen ? Kann man den Gebrauchtwagenpreis aufgrund der gefahrenen Meilen vorhersagen ?

Erstellung einer Richtpreisliste für Gebrauchtwagen
Untersuchung von 100 Ford Taurus, 3 Jahre alt
erhoben wurden: - gefahrene Meilen (x-Achse)
- Preis (y-Achse)



REGRESSION

beschreibt die Form eines Zusammenhangs

Unterscheidung (im Gegensatz zur Korrelation):

Y-Variable: abhängige Variable oder Responsevariable
X-Variable: unabhängige Variable oder erklärende Variable

bei Regression immer folgende Beziehung:

WENN	$\rightarrow$	DANN
X	$\rightarrow$	**Y**
(unabhängig)		(abhängig von X)

Was ist in folgenden Beispielen die abhängige (Y) und die unabhängige (X) Variable ?

- Das Verkehrsministerium möchte das Verhältnis zwischen *Straßenunebenheiten* und *Benzinverbrauch* untersuchen.

- Ein Händler, der seine Waren bei Fußballspielen verkauft, möchte die *Verkaufszahlen* auf die *Anzahl von Siegen* des Heimteams beziehen.

- Ein Soziologe möchte die *Anzahl von Wochenenden*, die ein Student zu Hause verbringt im Verhältnis zur *Entfernung* zwischen Wohn-und Studienort untersuchen.

zur Unterscheidung zwischen Regression und Korrelation:

- wenn man die WENN $\rightarrow$ DANN Beziehung auch umdrehen kann, dann sind beide Variablen gleichwertig, dann **Korrelation**
- wenn man das nicht kann, dann **Regression**

(einfache) lineare Regression:

wir betrachten hier (wie bei Korrelation) nur lineare Zusammenhänge

$$\boxed{Y \;=\; a + b \cdot X}$$

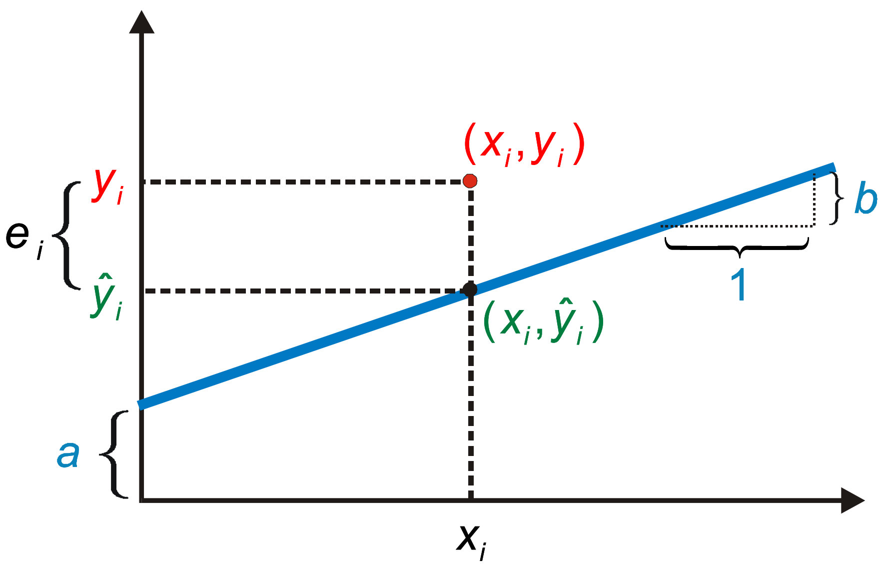
$$\text{Preis} \;=\; a + b \cdot \text{gefahrene Meilen}$$

a und **b** (*Regressionskoeffizienten*) sind die interessierenden Größen

Problemstellung:

aus Werten für **X** und **Y** müssen die unbekannten Größen ***a*** und ***b*** errechnet werden

jedem Punkt (x_i, y_i) wird ein Punkt $(x_i, \hat{y}_i)$ zugeordnet



a ... *intercept*
b ... Anstieg
$\hat{y}_i$ ... geschätztes y_i
e_i ... Residuum

(x_i, y_i):	$y_i = a + bx_i + e_i$	alle beobachteten Punkte
$(x_i, \hat{y}_i)$:	$\hat{y}_i = a + bx_i$	Punkte auf der Geraden

Berechnung von *a* und *b*: *Kleinstquadrate Prinzip (OLS)*

Lösen der Gleichung:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \text{ min!}$$

die Summe der quadrierten Abstände der Punkte von der Regressionsgeraden soll minimiert werden

Lösungen für *a* und *b*:

$$b = \frac{\text{Kovarianz}}{\text{Varianz von X}} = \frac{\text{cov}(x,y)}{s_x^2} = \frac{\frac{1}{n}\sum x_i y_i - \overline{xy}}{\frac{1}{n}\sum x_i^2 - \overline{x}^2} \qquad \left(= r_{XY}\frac{s_x}{s_y} \right)$$

$a = \overline{y} - b\overline{x}$ (Regressiongerade geht immer durch $\overline{x}$ und $\overline{y}$)

zurück zum Beispiel aus Fragestellung 2:

(*X* ... gefahrene Meilen, *Y* ... Gebrauchtwagenpreis)

b = -0,031	Preis = 6533,38 - 0,031 Meilen
a = 6533,38	*Y* = *a* + *b* *X*

Interpretation: je gefahrener Meile sinkt der Preis um 0,031 Dollar, d.h. ca. 3 Dollar weniger je 100 Meilen

Verwendung der Gleichung zur Vorhersage: welchen Preis erzielt ein 3 Jahre alter Ford Taurus mit 40000 Meilen

x = 40000 $\rightarrow$ $\hat{y} = 6533{,}38 \text{ - } 0{,}031 \cdot 40000 = 5293{,}38$

zur Berechnung der Regressionsparameter

Beispiel: (gleiche Daten wie bei Bsp.Korrelation)

gegeben:

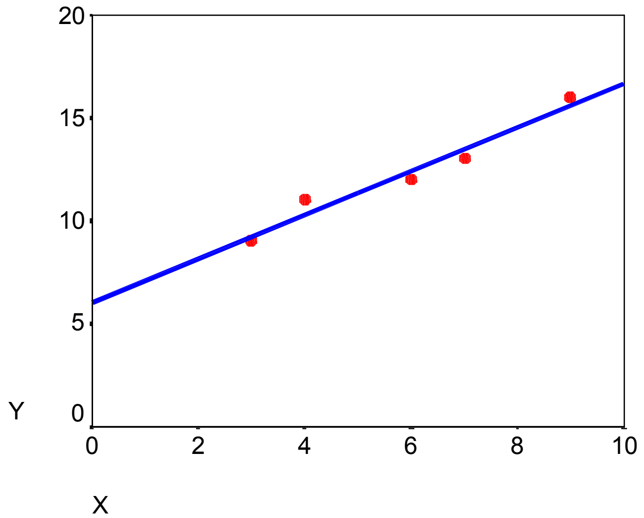
x:	3	6	7	9	4
y:	9	12	13	16	11

Berechnung von b:

	x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
	3	9	9	81	27
	6	12	36	144	72
	7	13	49	169	91
	9	16	81	256	144
	4	11	16	121	44
Summe	29	61	191	771	378
Mittelwert	5,8	12,2	38,2	154,2	75,6

$$b = \frac{\frac{1}{n}\sum x_i y_i - \overline{xy}}{\frac{1}{n}\sum x_i^2 - \overline{x}^2} = \frac{75{,}6 - 5{,}8 \cdot 12{,}2}{38{,}2 - 5{,}8^2} = 1{,}006$$

$$a = \overline{y} - b\overline{x} = 12{,}2 - 1{,}006 \cdot 5{,}8 = 6{,}37$$



Zeichnen der Regressionsgeraden:

2 Punkte notwendig:

1. gemeinsamer Mittelwert
2. entweder: *a* für *x* = 0
   oder
   geeignetes *x* wählen und
   dazugehöriges $\hat{y}$ ausrechnen

Testen im Regressionsmodell:

in Population: $Y = \alpha + \beta X$

Testen von β:
(α meistens nicht so interessant)

wieder 3 mögliche Alternativhypothesen:

H$_O$: $\beta = 0$ (kein linearer Zusammenhang)
H$_A$: $\beta \neq 0$ oder **H$_A$**: $\beta < 0$ oder **H$_A$**: $\beta > 0$

in SPSS:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	6533,383	84,512		77,307	,000
	MEILEN	**-,031**	,002	-,806	-13,495	**,000**

a. Abhängige Variable: PREIS

p-Wert (Signifikanz) < 0,001 (wird zweiseitig ausgegeben)

zum SPSS output:

- "B" sind die Regressionskoeffizienten a, b
- "(Konstante)" ist intercept a
- "Beta" hat nichts mit obigem β zu tun
- "T" ist der *t* - verteilte Wert (deshalb "T") der Teststatistik

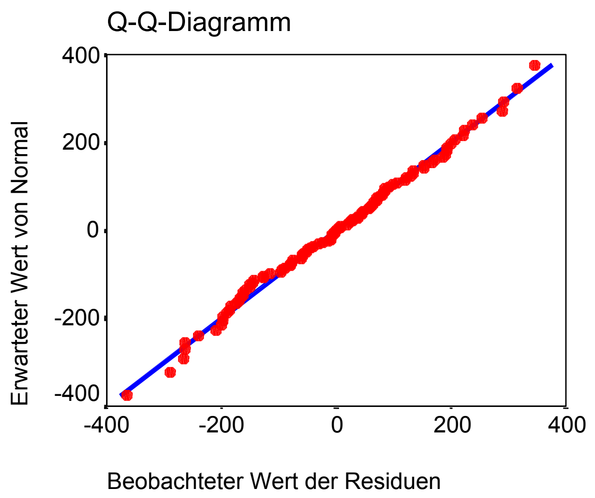
Voraussetzungen eines linearen Regressionsmodells

Voraussetzungen ähnlich wie bei Korrelation:

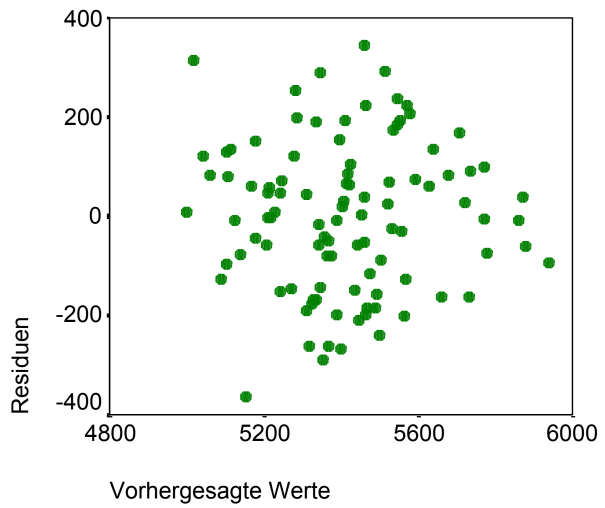
- Linearität der Beziehung
- Intervallskala für die abhängige Variable y
- y normalverteilt $\Leftrightarrow$ Residuen e normalverteilt

- Achten auf outlier !

Voraussetzungen lassen sich grafisch prüfen:
(für Gebrauchtwagenbeispiel)



QQ-Plot der Residuen:
(zur Überprüfung der Normal-verteilungsannahme)
Punkte sollen entlang einer 45° Geraden liegen



Residuen Plot:

y-Achse: Residuen
x-Achse: vorhergesagte Werte $\hat{y}$
soll kein wie immer geartetes Muster zeigen

wenn Voraussetzungen nicht erfüllt: eventuell Daten transformieren, sonst keine einfachen Alternativen !

Wie gut ist ein Regressionsmodell ?

Regressionsmodell dient dazu eine abhängige Variable zu erklären bzw. vorherzusagen

1. Voraussetzungen sollten erfüllt sein

2. Residuen sollten klein sein:
   je kleiner die Residuen umso exakterer Vorhersagen möglich

$$\underbrace{y}_{\substack{\text{beobachteter}\\ \text{Wert}}} = \underbrace{a+bx}_{\substack{\text{vorhergesagter}\\ \text{Wert}}} + \underbrace{e}_{\substack{\text{unerklärter}\\ \text{Rest}}}$$

allgemeines Maß zur Beurteilung des Erklärungswertes eines Regressionsmodells:

$$R^2 = \frac{Var(y) - Var(e)}{Var(y)} \qquad \text{"Bestimmtheitsmaß"}$$

- ist Anteil der erklärten Varianz von y
- ist quadrierter Korrelationskoeffizient

in SPSS: (Gebrauchtwagenbeispiel)

Modellzusammenfassung[b]

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,806[a]	,650	,647	151,57

a. Einflußvariablen : (Konstante), MEILEN

b. Abhängige Variable: PREIS

"Korrigiertes R^2" ist R^2, um die Freiheitsgrade korrigiert, damit die Stichprobengröße berücksichtigt wird

$R^2_{korr} = R^2 - [(k-1)/(n-k)] \cdot R^2$, k...Anzahl erklärender Variablen $\boldsymbol{x}$ inkl. $\boldsymbol{a}$

Multiple Regression

manchmal läßt sich Modell verbessern, wenn man **zusätzliche erklärende Variablen** berücksichtigt

multiples Regressionsmodell:

$$y_i = a + b_1 x_{i1} + b_2 x_{i2} + \cdots + e_i$$

Berechnung von *a* und den *b* 's ohne Computer nicht mehr so einfach

Beispiel: Gebrauchtwagen

zusätzliche Variable: Anzahl der Serviceüberprüfungen

mittels SPSS:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	6206,128	24,966		248,581	,000
	MEILEN	-,031	,001	-,814	-49,788	,000
	SERVICE	135,837	3,903	,569	34,807	,000

a. Abhängige Variable: PREIS

- korrigiertes R^2 jetzt 0,974
- beide Einflußgrößen MEILEN und SERVICE signifikant
- zusätzliche Interpretation für SERVICE:
  pro durchgeführtem Service erhöht sich durchschnittlich der Preis um ca. 136 $

- Überprüfen der Voraussetzungen wie bei einfachem linearen Modell

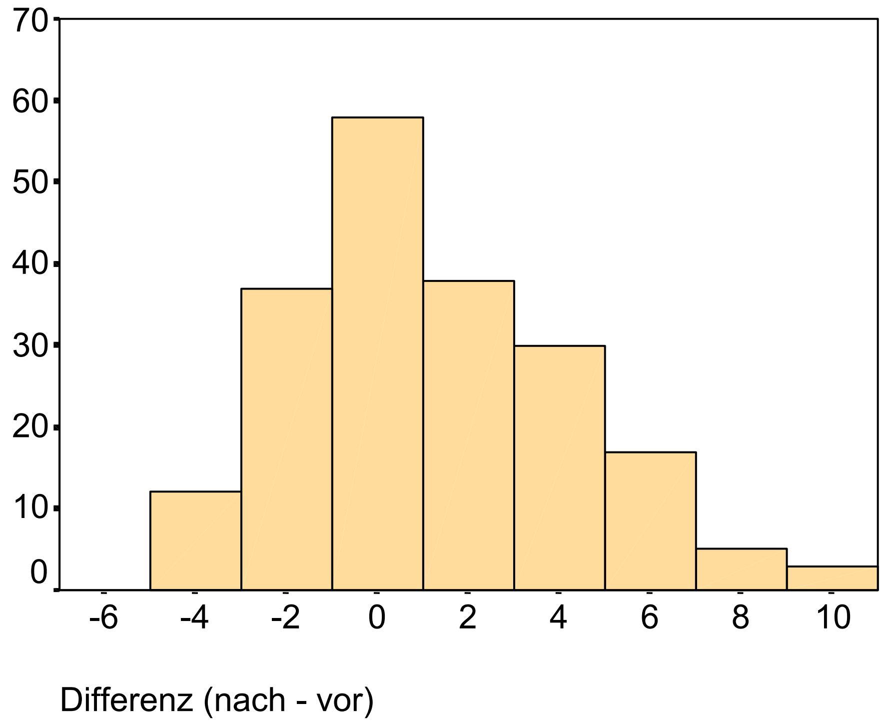
FRAGESTELLUNG 3A:

- Unterscheiden sich die Mittelwerte zweier Variablen, die an einer Beobachtungseinheit erhoben wurden ?

Bsp.: Helmpflicht (USA)

- Diskussion über Einführung einer Helmtragepflicht
- Kritiker behaupteten, diese Pflicht entmutigt, Rad zu fahren
- probeweise Einführung der Helmtragepflicht in Testorten
- repräsentative Stichprobe: wie viele km wurde in der Woche vor und der Woche nach Einführung des Gesetzes mit dem Rad zurückgelegt

- Besteht ein Unterschied in der Anzahl gefahrener km vor und nach Einführung der Helmtragepflicht ?



Mittelwert von 200 Personen: 0,77 (s = 3,07)

t - Test für abhängige Stichproben

"abhängig" heißen Stichproben dann, wenn zwei oder mehrere Variablen an einer Beobachtungseinheit erhoben worden sind (in SPSS: "gepaarte" Stichproben)

funktioniert wie der *t* -Test für eine Stichprobe, allerdings prüft man die Differenz der Mittelwerte

H$_O$: $\mu_2 - \mu_1 = 0$ (oder: **H$_O$:** $\mu_1 = \mu_2$)
H$_A$: $\mu_1 \neq \mu_2$ oder **H$_A$**: $\mu_1 > \mu_2$ oder **H$_A$**: $\mu_1 < \mu_2$

Voraussetzung: Intervallskala
Normalverteilung der Differenzen

Beispiel: Helmtragepflicht

SPSS Output: (hier Zeilen und Spalten vertauscht)

Test bei gepaarten Stichproben

			Paaren KMNACH - KMVOR
Gepaarte Differenzen	Mittelwert		,7700
	Standardabweichung		3,0650
	Standardfehler des Mittelwertes		,2167
	95% Konfidenzintervall der Differenz	Untere	,3426
		Obere	1,1974
T			3,553
df			199
Sig. (2-seitig)			,000

Ergebnis: die Einführung der Helmtragepflicht hat keine negativen Auswirkungen, es werden nach Einführung durchschnittlich um 0,77 km mehr (pro Woche) mit dem Fahrrad zurückgelegt

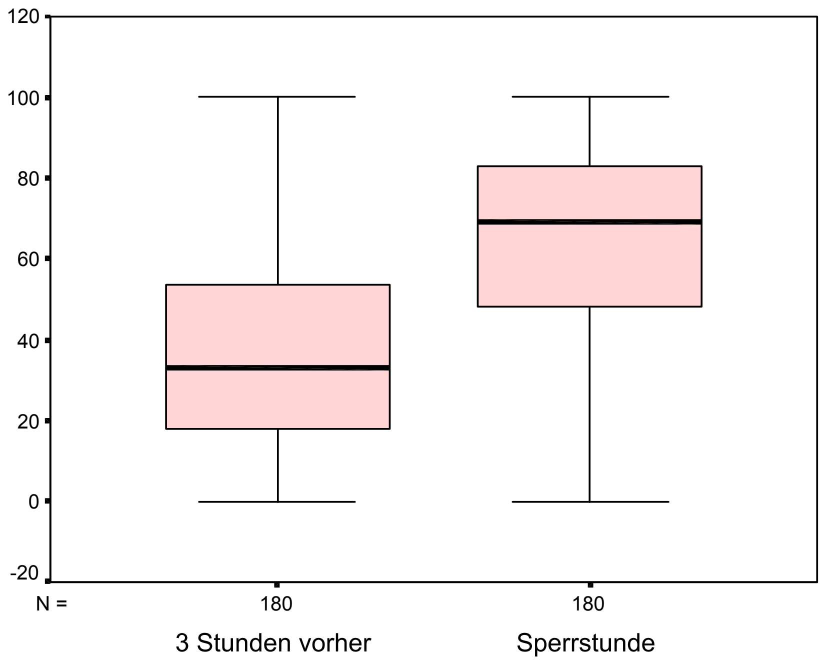
FRAGESTELLUNG 3B:

- Unterscheidet sich die Lage zweier Variablen, die an einer Beobachtungseinheit erhoben wurden ?

Bsp.: Alkohol und Beurteilung der Attraktivität

In einem Club in Ohio wurden Mitglieder gebeten, die Attraktivität der Anwesenden des jeweils anderen Geschlechts auf einer "100mm Skala" (0...extrem unattraktiv,..., 100...extrem attraktiv) zu beurteilen. Die Einschätzungen wurden 3 Stunden vor und unmittelbar vor der Sperrzeit abgegeben.

- Gibt es einen Unterschied in der Beurteilung der Attraktivität des jeweils anderen Geschlechts vor und nach Alkoholkonsum ?



abhängige Variable *Beurteilung* ist nicht intervallskaliert
wenn Voraussetzungen für t-Test nicht erfüllt sind dann →

Wilcoxon - Test
(auch Wilcoxon Matched Pairs Signed Ranks Test)

für Fragestellungen:
- unterscheiden sich 2 abhängige Stichproben bezüglich der Lage einer ordinalen (oder metrischen) Variable

oder
- wenn Voraussetzungen für *t* -Test für abhängige Stichproben nicht erfüllt sind

Voraussetzungen:
- 2 abhängige Stichproben
- ordinale Daten (oder nicht normalverteilte Differenzen)
- (nicht zuviele Bindungen)

Hypothesen:

H_0: $F(x) = G(x)$ (Lage in 2 Gruppen ist gleich)
H_A: $F(x) > G(x)$

zu Test (mittels SPSS):
- p-value (Signifikanz) bei kleinen Stichproben exakt, sonst Normalverteilungsapproximation
- in SPSS: bei einseitiger Fragestellung p-value halbieren

Ränge

	N	Mittlerer Rang	Rang-summe
Negative Ränge	39[a]	61,67	2405,00
Positive Ränge	138[b]	96,72	13348
Bindungen	3[c]		
Gesamt	180		

a. Sperrstunde < 3 Stunden vorher
b. Sperrstunde > 3 Stunden vorher
c. 3 Stunden vorher = Sperrstunde

Statistik für Test[b]

	Sperrstunde - 3 Stunden vorher
Z	-8,015[a]
Asymptotische Signifikanz (2-seitig)	,000

a. Basiert auf negativen Rängen.
b. Wilcoxon-Test

Ergebnis: Anwesende des anderen Geschlechts werden kurz vor Sperrzeit deutlich attraktiver beurteilt als 3 Stunden vorher