

Erhebung und Analyse von Immobilienmarkt- informationen per Bürgerbefragung

Sitzung 5

Bachelorübung, FK Raumplanung, Modul 11

Donnerstags 09:30 bis 14:00, GB 1 335 PC-Pool

Sommersemester 2024

M.Sc. Alexander Schuhmacher

TU Dortmund, Fakultät Raumplanung

Lehrstuhl Immobilienentwicklung

- **Ziele der Präsentation**
- **Wie läuft's?**
- **Zusammenhangsanalyse**
- **Regressionsanalyse**

- **Wissensaufbau zur Zusammenhangsanalyse**
- **Wissensaufbau zur Regressionsanalyse**

- Wie läuft's mit der Befragung bei euch?
- Welche Herausforderungen gibt es?



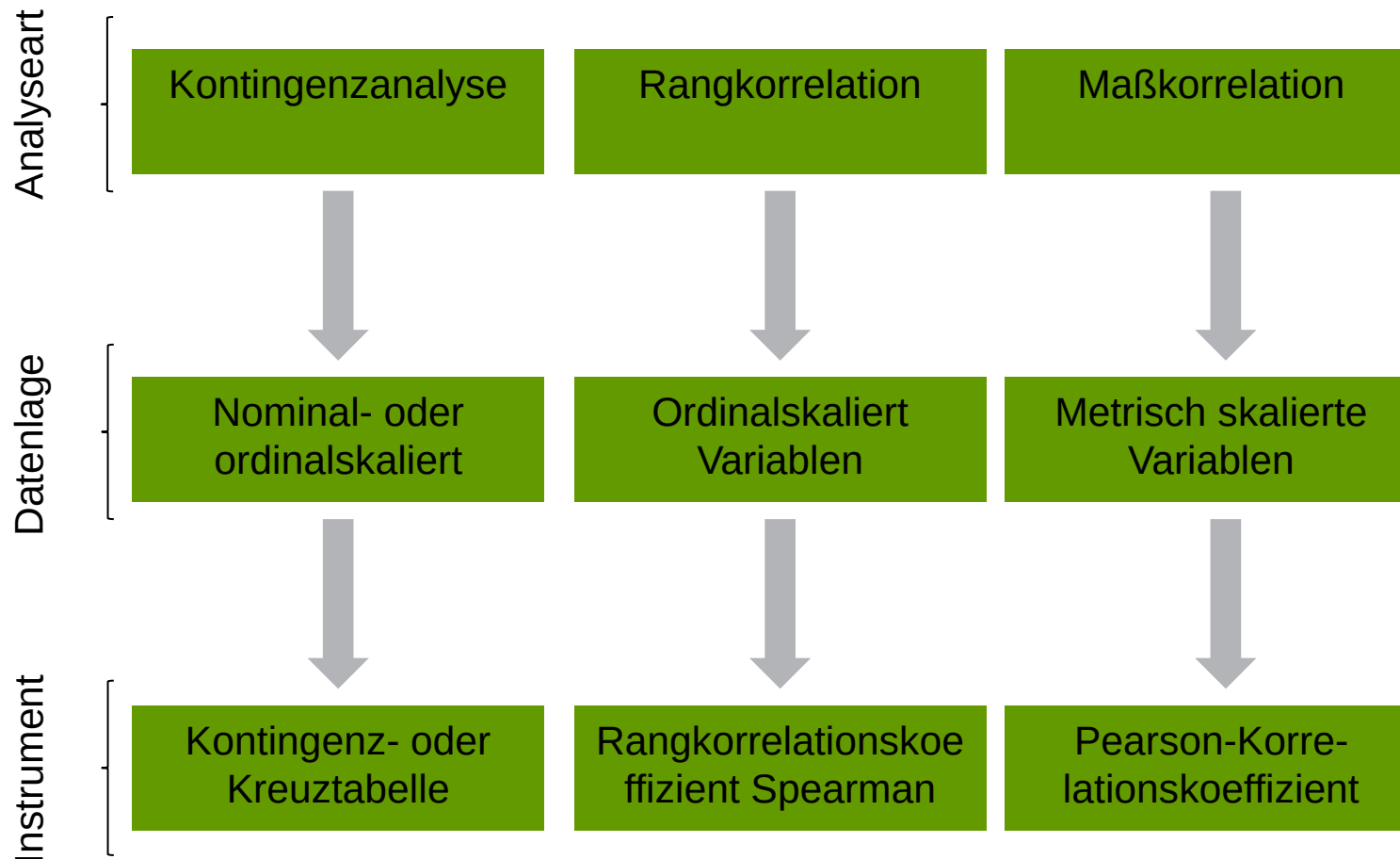
Quelle: Kramer, I. (o.J.): Befragung Interview. Abrufbar unter: <https://ingakramer.de/products/befragung-interview> [zuletzt
zugegriffen am 05.06.2024]

Zusammenhangsanalyse

- Zusammenhänge helfen uns Muster und Beziehungen in Daten zu erkennen
- Beispiel: Wie wirkt sich die Wohnfläche auf den Mietspreis aus?
- Anwendung in unterschiedlichen Fachdisziplin

- Teil der multivariaten Statistik
- **Ziel:** Messen der Intensität und der Richtung von statistischen Zusammenhängen zwischen mindestens zwei Merkmalen unter der Verwendung von Maßzahlen
- Hilft zu verstehen, ob und wie stark Veränderungen einer Variablen mit der Veränderung einer anderen Variable verknüpft ist (Eckstein 2000: 173)
- **Einteilung in:**
 - Kontingenzanalyse
 - Rangkorrelation
 - Maßkorrelationsanalyse

Anwendungsfälle



Definition

- Untersuchungsinstrument zur Bestimmung von Beziehungen und Abhängigkeiten zwischen nominal- oder ordinalskalierten Variablen auf Basis von Kontingenz-/ Kreuztabellen (synonyme Verwendung der Begriffe)

Datenbasis

- Nominal- oder ordinalskalierte Daten (z. B. Geschlecht, Wohnort, Zufriedenheit auf Skala ...)

Ziel

- Zusammenhangsanalyse von nominalen und/oder ordinalen Merkmalen mit wenigen unterschiedlichen Ausprägungen auf Basis einer Kontingenz- bzw. Kreuztabelle.

Aufbau und Inhalt einer Kontingenz-/Kreuztabelle

Kontingenztafel

Für zwei kategoriale Erhebungsmerkmale $X(\gamma_i) = x_i \in \Xi_X = \{\xi_j, j = 1, 2, \dots, r\}$ und $Y(\gamma_i) = y_i \in \Xi_Y = \{\upsilon_k, k = 1, 2, \dots, c\}$ mit r bzw. c voneinander verschiedenen Ausprägungen ξ_j bzw. υ_k , die für eine statistische Gesamtheit $\Gamma_n = \{\gamma_i, i = 1, 2, \dots, n\}$ vom Umfang n statistisch erhoben wurden, heißt die Häufigkeitstabelle, die in

Zeile	Spalte	1	2	..	k	..	c	Summe
	$X \setminus Y$	υ_1	υ_2	..	υ_k	..	υ_c	
1	ξ_1	n_{11}	n_{12}	..	n_{1k}	..	n_{1c}	$n_{1\bullet}$
2	ξ_2	n_{21}	n_{22}	..	n_{2k}	..	n_{2c}	$n_{2\bullet}$
:	:	:	:	:	:	:	:	:
j	ξ_j	n_{j1}	n_{j2}	..	n_{jk}	..	n_{jc}	$n_{j\bullet}$
:	:	:	:	:	:	:	:	:
r	ξ_r	n_{r1}	n_{r2}	..	n_{rk}	..	n_{rc}	$n_{r\bullet}$
Summe		$n_{\bullet 1}$	$n_{\bullet 2}$	..	$n_{\bullet k}$	..	$n_{\bullet c}$	n

ihrem „Inneren“ aus r Zeilen und aus c Spalten besteht und der Beschreibung der bivariaten absoluten Häufigkeitsverteilung $\{(\xi_j, \upsilon_k), n_{jk}\}$ der $(r \times c)$ - Ausprägungspaare (ξ_j, υ_k) dient, $(r \times c)$ -Kontingenztafel oder $(r \times c)$ -Kreuztafel für die kategorialen Erhebungsmerkmale X und Y .

Beispielhafte Kontingenz-/Kreuztabelle

Anzahl		Herkunftsland			Gesamt
		Ossi	Wessi	Ausland	
FKK	Fan	130	39	15	184
	kein Fan	72	56	28	156
Gesamt		202	95	43	340

Quelle: Eckstein, P. (2000³): Angewandte Statistik in SPSS – Praktische Einführung für Wirtschaftswissenschaftler. S. 177

- r (row/Zeile) = 2, c (column/Spalte) = 3 $\rightarrow r \cdot c = 2 \cdot 3 = 6$ somit Bezeichnung als rechteckige Sechsfeldertafel bzw. Kreuztabelle vom Typ $2 \cdot 3$

Beispiel Interpretation

- Bei insgesamt $n = 340$ Befragten, gaben von den Befragten $n_{12} = 39$ an, dass sie sowohl FKK-Fan sind als auch Wessi zu sein.
- Marginalverteilungen der Merkmale abzuleiten an „Gesamt“

Erstellung einer Kontingenz-/Kreuztabelle

The screenshot illustrates the steps to create a contingency table in SPSS. The 'Analysieren' (Analyze) menu is open, showing the path: **Analysieren** > **Tabellen** > **Deskriptive Statistiken** > **Kreuztabellen** > **Kreuztabellen: Zellen anzeigen**.

The 'Kreuztabellen' (Crosstabs) dialog box is shown with the following settings:

- Zeile(n):** Ortscode [Ortscode]
- Spalten:** Erhaltungszustand ETW [Erhaltungszustand ETW]
- Schicht 1 von 1:** (Empty)
- Schichtvariablen in Tabellenschichten anzeigen:** (Checked)

The 'Zellen anzeigen' (Display Cells) sub-dialog box is open, showing the following options:

- Häufigkeiten:** ☒ Begbachtet, ☐ Erwartet
- Prozentwerte:** ☐ Zeilenweise, ☐ Spaltenweise, ☐ Gesamtsumme
- Residuen:** ☐ Nicht standardisiert, ☐ Standardisiert, ☐ Angepasst standardisiert
- Nicht ganzzahlige Gewichtungen:** ☒ Anzahl in den Zellen runden, ☐ Anzahl in den Zellen kürzen, ☐ Keine Anpassungen

The 'Zellen...' button in the 'Kreuztabellen' dialog is highlighted with a red box, and an arrow points from it to the 'Zellen anzeigen' dialog.

Erstellung einer Kontingenz-/Kreuztabelle

Ortscode * Erhaltungszustand ETW Kreuztabelle

Anzahl		Erhaltungszustand ETW					Gesamt
		gut	mittelmäßig	schlecht	sehr gut	sehr schlecht	
Ortscode	Köpenick	2	0	13	7	0	22
	Charlottenburg	0	13	0	8	1	22
	Tiergarten	6	0	0	0	16	22
Gesamt		8	13	13	15	17	66

➤ Darstellungsform der absoluten Häufigkeiten

- Alternativ und oftmals anschaulicher sind jedoch die relativen Häufigkeiten auch bezeichnet als **Konditionalverteilung**
- Vorteil: „Zufällige“ **Zusammenhänge** können **identifiziert** werden

Erstellung einer Kontingenz-/Kreuztabelle

The screenshot shows the SPSS 'Kreuztabellen' (Crosstabs) dialog box. The 'Zellen...' button is highlighted with a red box. The 'Häufigkeiten' (Frequencies) section in the 'Zellen anzeigen' sub-dialog is also highlighted with a red box. A black arrow points from the 'Zellen...' button to the 'Häufigkeiten' section.

Kreuztabellen Dialog:

- Zeile(n):** Ortscode [Ortscode]
- Spalten:** Erhaltungszustand ETW [Erhaltung...]
- Schicht 1 von 1:** Zurück, Weiter
- Buttons:** Exakt..., Statistiken..., Zellen..., Format..., Stil..., Bootstrap...
- Buttons:** Zurücksetzen, Abbrechen, Hilfe

Zellen anzeigen Sub-dialog:

- Häufigkeiten:**
 - ☐ Begbachtet
 - ☐ Erwartet
 - ☐ Kleine Werte für Häufigkeiten ausblenden (Kleiner als 5)
- Prozentwerte:**
 - ☒ Zeilenweise
 - ☐ Spaltenweise
 - ☐ Gesamtsumme
- ☐ Tabelle im APA-Stil erstellen
- Z-Test:**
 - ☐ Spaltenanteile vergleichen
 - ☐ p-Werte anpassen (Bonferroni-Methode)
- Residuen:**
 - ☐ Nicht standardisiert
 - ☐ Standardisiert
 - ☐ Angepasst standardisiert
- Nicht ganzzahlige Gewichtungen:**
 - ☒ Anzahl in den Zellen runden
 - ☐ Anzahl in den Zellen kürzen
 - ☐ Keine Anpassungen
 - ☐ Fallgewichtungen runden
 - ☐ Fallgewichtungen kürzen
- Buttons:** Weiter, Abbrechen, Hilfe

Erstellung einer Kontingenz-/Kreuztabelle

Ortscode * Erhaltungszustand ETW Kreuztabelle

% von Ortscode

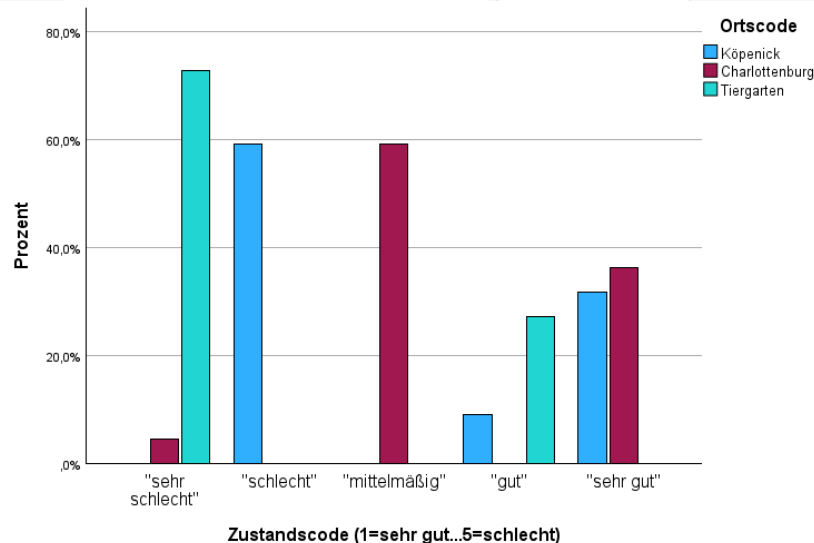
		Erhaltungszustand ETW					Gesamt
		gut	mittelmäßig	schlecht	sehr gut	sehr schlecht	
Ortscode	Köpenick	9,1%		59,1%	31,8%		100,0%
	Charlottenburg		59,1%		36,4%	4,5%	100,0%
	Tiergarten	27,3%				72,7%	100,0%
Gesamt		12,1%	19,7%	19,7%	22,7%	25,8%	100,0%

- Darstellungsform der relativen Häufigkeiten (Konditionalverteilung)

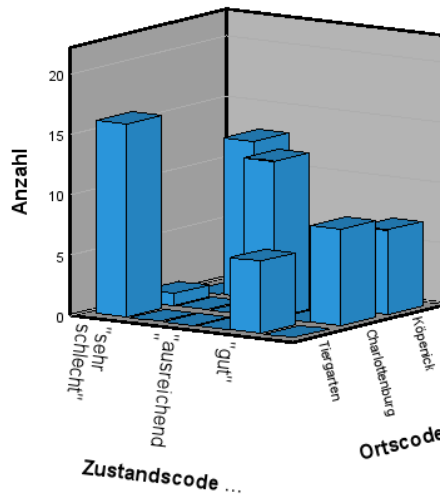
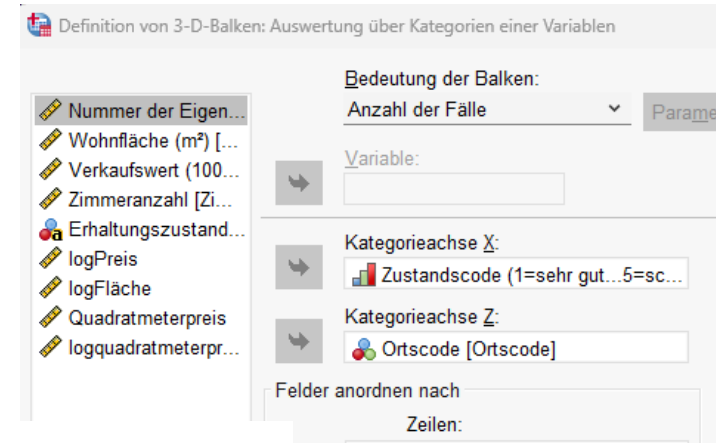
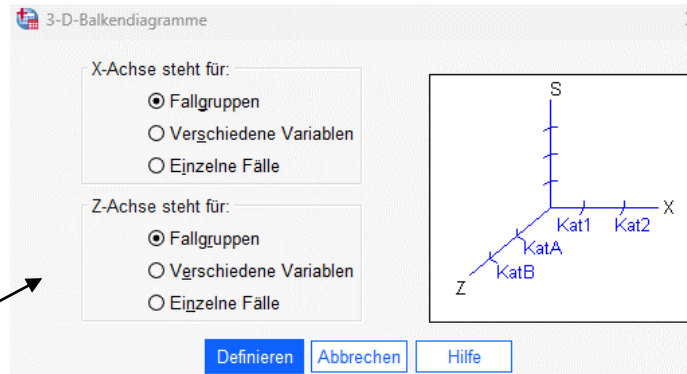
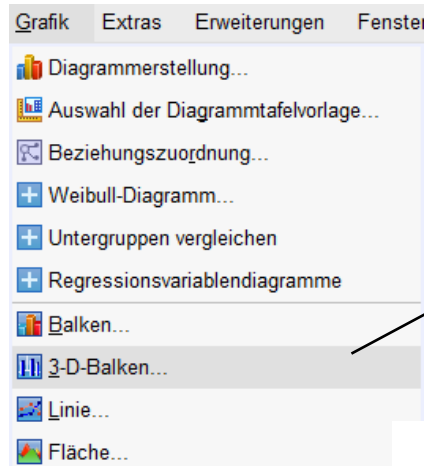
Erstellung eines Balkendiagramms

- Auch als Balkendiagramm gut darstellbar

The screenshot shows the SPSS 'Grafik' (Graphics) menu with 'Balken...' (Bar...) selected. The 'Balkendiagramme' sub-menu is open, with 'Einfach' (Simple) highlighted. The 'Gruppiertes Balkendiagramm definieren' dialog box is also shown, with 'Auswertung über Kategorien einer Variablen' (Evaluation by categories of a variable) selected. The 'Bedeutung der Balken' (Meaning of bars) section has '% der Fälle' (Percentage of cases) selected. The 'Kategorieachse' (Category axis) is set to 'Zustandscode (1=sehr gut...5=schlecht) [Zustan...]' and the 'Gruppen definieren durch' (Define groups by) is set to 'Ortscode [Ortscode]'.



3D-Balkendiagramm



Als Ergänzung zur Kontingenz-/Kreuztabelle: Chi-Quadrat-Unabhängigkeitstest

- Dient zum Prüfen der Hypothese, ob zwei Merkmale X und Y aus einer Grundgesamtheit voneinander statistisch unabhängig sind.
- **Grundlage:** (r*c)-Kontingenz-/Kreuztabelle, die aus einer Zufallsstichprobe mit einem Umfang von n-Merkmalsträgern erstellt wurde

The screenshot illustrates the steps to perform a Chi-Square test of independence in SPSS. In the 'Kreuztabelle' dialog, 'Ortscode [Ortscode]' is assigned to the 'Zeile(n):' (Rows) field and 'Erhaltungszustand ETW [Erhaltungszustand ETW]' is assigned to the 'Spalten:' (Columns) field. The 'Statistiken...' button is highlighted. In the 'Kreuztabelle: Statistik' sub-dialog, the 'Chi-Quadrat' checkbox is checked, indicating the Chi-Square test will be performed. Other statistical tests like 'Korrelationen', 'Nominal', and 'Ordinal' are listed but not selected.

Chi-Quadrat-Unabhängigkeitstest

- Testung der Unabhängigkeitshypothese mit $\alpha = 0,05$
- H_0 = Die Variablen Ortscode und Erhaltungszustand sind statistisch unabhängig
- H_1 = Die Variablen sind nicht statistisch unabhängig

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (zweiseitig)
Pearson-Chi-Quadrat	94,953 ^a	8	<,001
Likelihood-Quotient	107,685	8	<,001
Anzahl der gültigen Fälle	66		

a. 9 Zellen (60,0%) haben eine erwartete Häufigkeit kleiner 5.
Die minimale erwartete Häufigkeit ist 2,67.

- Da $\alpha^* = < 0,01$ wird die Nullhypothese abgelehnt
- Demnach ist mit einer Irrtumswahrscheinlichkeit von 0,05 davon auszugehen, dass:
 - In der Grundgesamtheit die Variablen nicht unabhängig voneinander und somit eine statistische Kontingenz (/Zusammenhang) besteht
- Fehlende Normierung von χ^2 lässt keine Aussage zur Stärke zu

Kontingenzmaße

- Alternative zu Chi-Unabhängigkeitstest
- **Vorteil:** Überprüfung der statistischen Kontingenz **UND** deskriptive Aussage über Stärke möglich
- Richtung bei nominalen Merkmalen nicht sinnvoll, aufgrund fehlender Ordnungsrelation

Anwendungsfälle

Nominale Merkmale



Cramer's V

Ordinale Merkmale



Kendall's Tau-b o.
Kendall's Tau-c

Metrisch / Kardinale
Merkmale



Cohens Kappa o.
McNemar

Unterschiedliche
Skalen



Eta

Kontingenzmaße: Beispiel Cramer's V

- Ist für n Merkmalsträger einer r*c Kontingenztafel zweier nominal skalierten Merkmale X und Y gegeben und aus der Kontingenztafel das χ^2 -Maß berechnet worden, dann heißt das Maß :

$$V = \sqrt{\frac{\chi^2}{n * (m - 1)}}$$

- Nominales Kontingenzmaß V nach Cramer

Quelle: Eckstein, P. (2000³): Angewandte Statistik in SPSS – Praktische Einführung für Wirtschaftswissenschaftler. S. 186

Faustregel nach Eckstein (2000): S. 187 für Cramer's V

- $0 < V < 0,2$ schwache Kontingenz (Zusammenhang)
- $0,2 \leq V < 0,5$ ausgeprägte Kontingenz
- $0,5 \leq V < 1$ starke Kontingenz

Kontingenzmaße: Beispiel Cramer's V + Kendall's Tau-c

- Untersuchung des Zusammenhangs zwischen dem Merkmal Standort und Erhaltungszustand mithilfe Cramer's V und Kendall's Tau-c

The screenshot displays the SPSS 'Kreuztabellen' (Crosstabs) dialog box. The 'Zeile(n):' (Row(s)) field contains 'Ortscode [Ortscode]' and the 'Spalten:' (Column(s)) field contains 'Erhaltungszustand ETW [Erhaltung...]'. The 'Statistiken...' button is highlighted with a red box. An arrow points from this button to the 'Kreuztabellen: Statistik' sub-dialog box. In this sub-dialog, under the 'Nominal' section, 'Phi und Cramer-V' is checked. Under the 'Ordinal' section, 'Kendall-Tau-c' is checked. Other options like 'Chi-Quadrat', 'Korrelationen', 'Gamma', 'Somers-d', 'Kappa', 'Risiko', 'McNemar', and 'Cochran- und Mantel-Haenszel-Statistik' are unchecked.

Kontingenzmaße: Beispiel Cramer's V

Erhaltungszustand ETW * Ortscode Kreuztabelle

Anzahl		Ortscode			Gesamt
		Köpenick	Charlottenburg	Tiergarten	
Erhaltungszustand ETW	gut	2	0	6	8
	mittelmäßig	0	13	0	13
	schlecht	13	0	0	13
	sehr gut	7	8	0	15
	sehr schlecht	0	1	16	17
Gesamt		22	22	22	66

Mithilfe von rechteckig (Ausgabetable, Tabelle umsetzen sind die Zeilen und Spalten „tauschbar“

Symmetrische Maße

		Wert	Asymptotischer Standardfehler ^a	Näherungsweise ^b	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Phi	1,199			<,001
	Cramer-V	,848			<,001
Ordinal- bzgl. Ordinalmaß	Kendall-Tau-c	,237	,142	1,669	,095
Anzahl der gültigen Fälle		66			

a. Die Null-Hypothese wird nicht angenommen.

b. Unter Annahme der Null-Hypothese wird der asymptotische Standardfehler verwendet.

Interpretation: Cramer's V

- $V = 0$ kein statistisch nachweisbare Kontingenz (Zusammenhang)
- $0 < V < 1$ = Indiz für statistische Kontingenz
- $V = 1$ totale Abhängigkeit zwischen beiden Merkmalen

Kontingenzmaße: Beispiel Cramer's V

Symmetrische Maße

		Wert	Asymptotischer Standardfehler ^a	Näherungsweise t ^b	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Phi	1,199			< ,001
	Cramer-V	,848			< ,001
Ordinal- bzgl. Ordinalmaß	Kendall-Tau-c	,237	,142	1,669	,095
Anzahl der gültigen Fälle		66			

a. Die Null-Hypothese wird nicht angenommen.

b. Unter Annahme der Null-Hypothese wird der asymptotische Standardfehler verwendet.

- $\alpha^* = < 0,01$ lediglich als ergänzendes Kontingenzmaß zum χ^2 –Unabhängigkeitstest
- **Deutung α^*** : $V = 0,848$ somit signifikant Größer als 0 und somit eine starke Kontingenz der beiden Variablen

Kontingenzmaße: Beispiel Kendall Tau-b & Kendall Tau-c

- Ist für n Merkmalsträger einer $r \times c$ Kontingenztafel zweier ordinal skalierten Merkmale X und Y gegeben und sind die Anzahl K der konkordanten, die Anzahl D der diskordanten, die Anzahl G_a der auf dem a (bhängigen) sowie die Anzahl G_u der auf dem u (nabhängigen) Merkmal gebundenen Ausprägungspaare bekannt, dann heißen die ordinalen Kontingenzmaße

$$\tau_b = \frac{K - D}{\sqrt{(K + D + G_a) * (K + D + G_u)}} \quad \text{bzw.} \quad \tau_c = \frac{2 * m * (K - D)}{n^2 * (m - 1)}$$

Quelle: Eckstein, P. (2000³): Angewandte Statistik in SPSS – Praktische Einführung für Wirtschaftswissenschaftler. S. 188

Charakteristik & Anwendung:

- τ_b nur bei quadratischen Kontingenztabellen ($r = c$)
- τ_c bei rechteckigen Kontingenztabellen ($r \neq c$) – m als die kleiner Anzahl von r und c
- Intervall: $-1 \leq \tau_b / \tau_c \leq 1$

Kontingenzmaße: Beispiel Kendall Tau-c

		Symmetrische Maße			
		Wert	Asymptotischer Standardfehler ^a	Näherungsweise t ^b	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Phi	1,199			<,001
	Cramer V	,848			,001
Ordinal- bzgl. Ordinalmaß	Kendall-Tau-c	,237	,142	1,669	,095
Anzahl der gültigen Fälle		66			

a. Die Null-Hypothese wird nicht angenommen.

b. Unter Annahme der Null-Hypothese wird der asymptotische Standardfehler verwendet.

- **Deutung Kendall:** Der Wert von 0,237 deutet einen schwachen positiven Zusammenhang auf
- **Deutung α^* :** $\alpha^* = 0,095$ bedeutet, dass das Ergebnis nicht statistisch signifikant ist
- Da der p-Wert von $0,095 > 0,05$ ist, muss die Nullhypothese angenommen werden. Das bedeutet, wir können nicht mit ausreichender Sicherheit sagen, dass ein signifikanter Zusammenhang zwischen dem Standort einer Immobilie und ihrem Erhaltungszustand besteht.

Definition

- Sachlogische begründete Analyse eines statistischen Zusammenhanges zwischen zwei ordinal skalierten bzw. unter der Verwendung von Rangwerten ordinalisierten metrischen Merkmalen mit vielen sich voneinander unterscheidenden Merkmalsausprägungen – auch als **bivariate Korrelationsanalyse** bezeichnet

Datenbasis

- ordinalskalierte bzw. Ordinalisierte metrische Daten

Ziel

- Überprüfung der Stärke und der Richtung des monotonen Zusammenhangs zwischen zwei Rangvariablen

Rangwert

Ist X ein mindestens ordinal skaliertes Merkmal, dessen n Merkmalsausprägungen x_i ($i = 1, 2, \dots, n$) für alle $i = 1, 2, \dots, n-1$ aufsteigend $x_i < x_{i+1}$ oder absteigend $x_i > x_{i+1}$ geordnet sind, dann heißt die geordnete Merkmalsausprägung x_i zugewiesene Platznummer Rangwert R_i^x der Merkmalsausprägung x_i .

Ordnungsprinzip

- Stärke bzw. Grad der Intensität

Vorgehensweise

- Bildung von Rängen
- Zuordnung der Variablen X und Y von Werten – kleinster Wert 1, zweitkleinster Wert 2 usw.
- Anschließend Bildung der Rangdifferenz d der Datenpaare
- Mithilfe der Rangdifferenz Berechnung von Rangkorrelation nach Spearman möglich

Mathematische Grundlage – Rangkorrelationskoeffizient nach Spearman

- $$r_s = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n \cdot (n^2 - 1)}$$

Interpretation des Rangkorrelationskoeffizienten nach Spearman

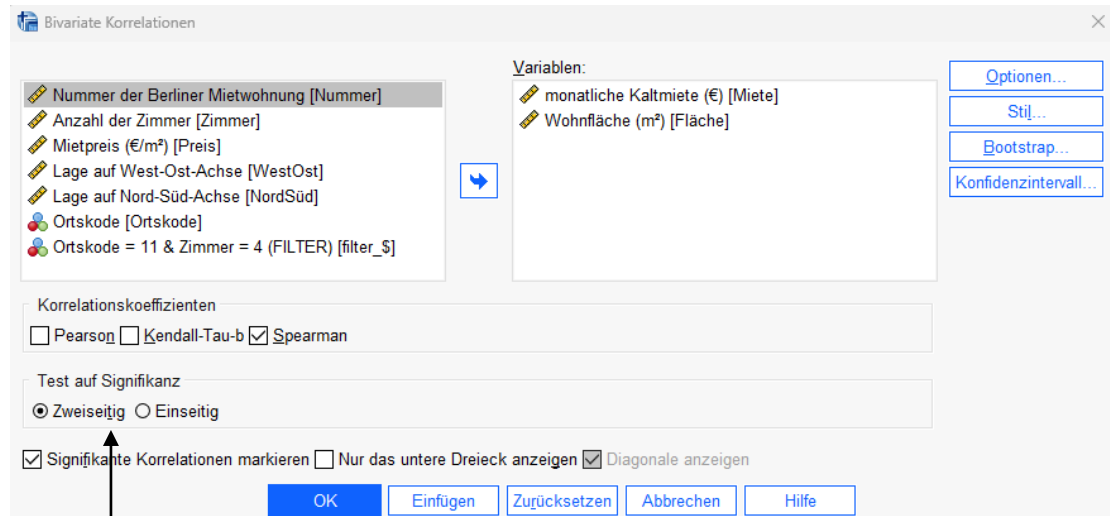
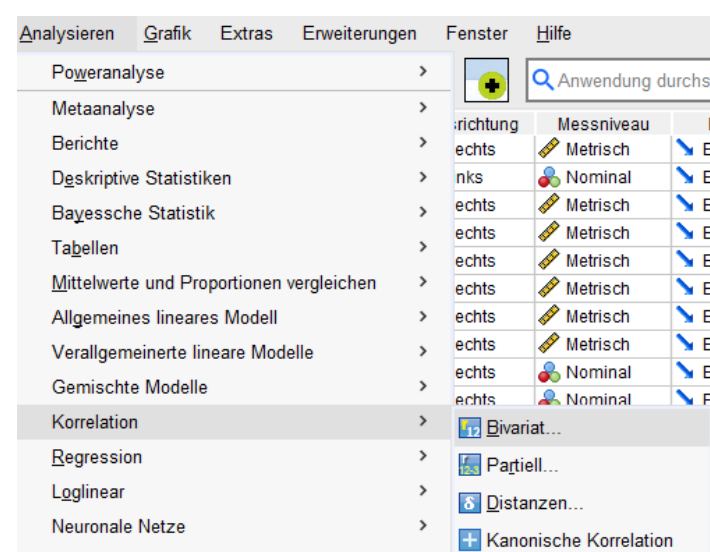
- $-1 \leq r_s \leq 1$
- Bei $r_s \leq 1$ positiver bzw. gleichläufiger statistischer Zusammenhang
- Bei $r_s \geq -1$ negativer bzw. gegenläufiger statistischer Zusammenhang
- Bei $r_s = 0$ zwischen X und Y kein statistischer Zusammenhang erkennbar

Anwendung von Spearman

- Beide Variablen ordinal skaliert
- Eine Variable ordinal, eine metrisch
- Beide Variablen metrisch, aber mindestens einer liegt nach KS-Test keine Normalverteilung vor

Übungsaufgabe 15

- Untersuchen Sie den Zusammenhang zwischen den ordinalskalierten Merkmalen der Wohnfläche und der Kaltmiete von Mietwohnungen. Nutzen Sie hierzu den Datensatz „Mietspiegel2007.sav“.



- Ohne Wirkungsvermutung „Zweiseitig“, mit „einseitig“

Übungsaufgabe 15

Korrelationen

			monatliche Kaltmiete (€)	Wohnfläche (m ²)
Spearman-Rho	monatliche Kaltmiete (€)	Korrelationskoeffizient	1,000	,778**
		Sig. (2-seitig)	.	<,001
		N	7518	7518
	Wohnfläche (m ²)	Korrelationskoeffizient	,778**	1,000
		Sig. (2-seitig)	<,001	.
		N	7518	7518

** . Die Korrelation ist auf dem 0,01 Niveau signifikant (zweiseitig).

Interpretation

- r_s ist mit 0,778 signifikant größer als null
- Da $\alpha^* < 0,001$ und somit $< 0,01$ wird (H_0 = Es gibt keinen monotonen Zusammenhang zwischen den Variablen) verworfen – Es gibt einen signifikanten Zusammenhang
- **Deshalb:** Am Berliner Wohnungsmarkt kann davon ausgegangen werden, dass zwischen der Wohnfläche und der monatlichen Kaltmiete ein gleichläufiger (positiver) monotoner linearer statistischer Zusammenhang besteht

Hinweis

- Rangkorrelation ist nicht-parametrische Korrelationsanalyse
- Demnach Einsatz auch bei metrisch skalierten ohne normalverteilten Daten sinnvoll
- Bei Normalverteilung (nach KS-Test) kann auch parametrische Korrelationsanalyse wie nach Bravais und Pearson (Maßkorrelationskoeffizienten) angewendet werden

Definition

- Sachlogische begründete Analyse von Zusammenhängen zwischen zwei oder mehr metrischen Merkmalen. Dabei ist es vorteilhaft sowohl eine explorative als auch eine graphische Datenanalyse (Streudiagramm) vorzulagern

Datenbasis

- Metrische Daten

Ziel

- Überprüfung eines **linearen**, statistischen Zusammenhanges

Interpretation des (zu berechnenden) Korrelationskoeffizienten (r) nach Bravais-Pearson

- Einheit zwischen -1 und 1 dabei:
 - $r = -1$ perfekte negative Korrelation
 - $r = 1$ perfekte positive Korrelation
 - $r = 0$ keine Korrelation

Fortführung der Interpretation von r

- $r = 0$ Keine Korrelation
- $0 < r < 0,5$ schwache Korrelation
- $0,5 \leq r < 0,8$ mittlere Korrelation
- $0,8 \leq r < 1$ starke Korrelation
- $r = 1$ perfekte Korrelation

Voraussetzung zur Anwendung Bravais-Pearson

- Beide zu korrelierenden Variablen metrisch skaliert
- Vorliegen einer bivariaten Normalverteilung
 - In SPSS nicht prüfbar, deshalb einzelne Variablen auf über KS auf Normalverteilung testen

Mathematische Vorgehensweise

- Berechnung von r (Pearson-Korrelationskoeffizient) nach:

$$r = \frac{\sum(x_i - \bar{x}) - (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 - \sum(y_i - \bar{y})^2}}$$

Perfekte negative Korrelation

- Korrelationskoeffizient = $r = -1$
- Eine Variable steigt, während die andere Variable im (exakt) selben Verhältnis sinkt
- Die Wertepaare liegen auf einer fallenden Gerade

Beispiel: Variablen Entfernung zum Stadtzentrum und Immobilienpreise

- Sofern die Entfernung zum Stadtzentrum zunimmt, sinken die Immobilienpreise proportional
- D.h. Verdoppelt sich die Entfernung zum Stadtzentrum, so halbieren sich die Immobilienpreise

Perfekte positive Korrelation

- Korrelationskoeffizient = $r = 1$
- Beide Variablen steigen im (exakt) selben Verhältnis zueinander an.
- Die Wertepaare liegen auf einer steigenden Gerade

Beispiel: Variablen Quadratmeterpreis und Immobilienwert einer Immobilie

- Sofern der Quadratmeterpreis einer Immobilie steigt, so steigt auch der Immobilienwert einer Immobilie proportional
- D.h. Verdoppelt sich der Quadratmeterpreis, so verdoppelt sich der Immobilienwert

Keine Korrelation

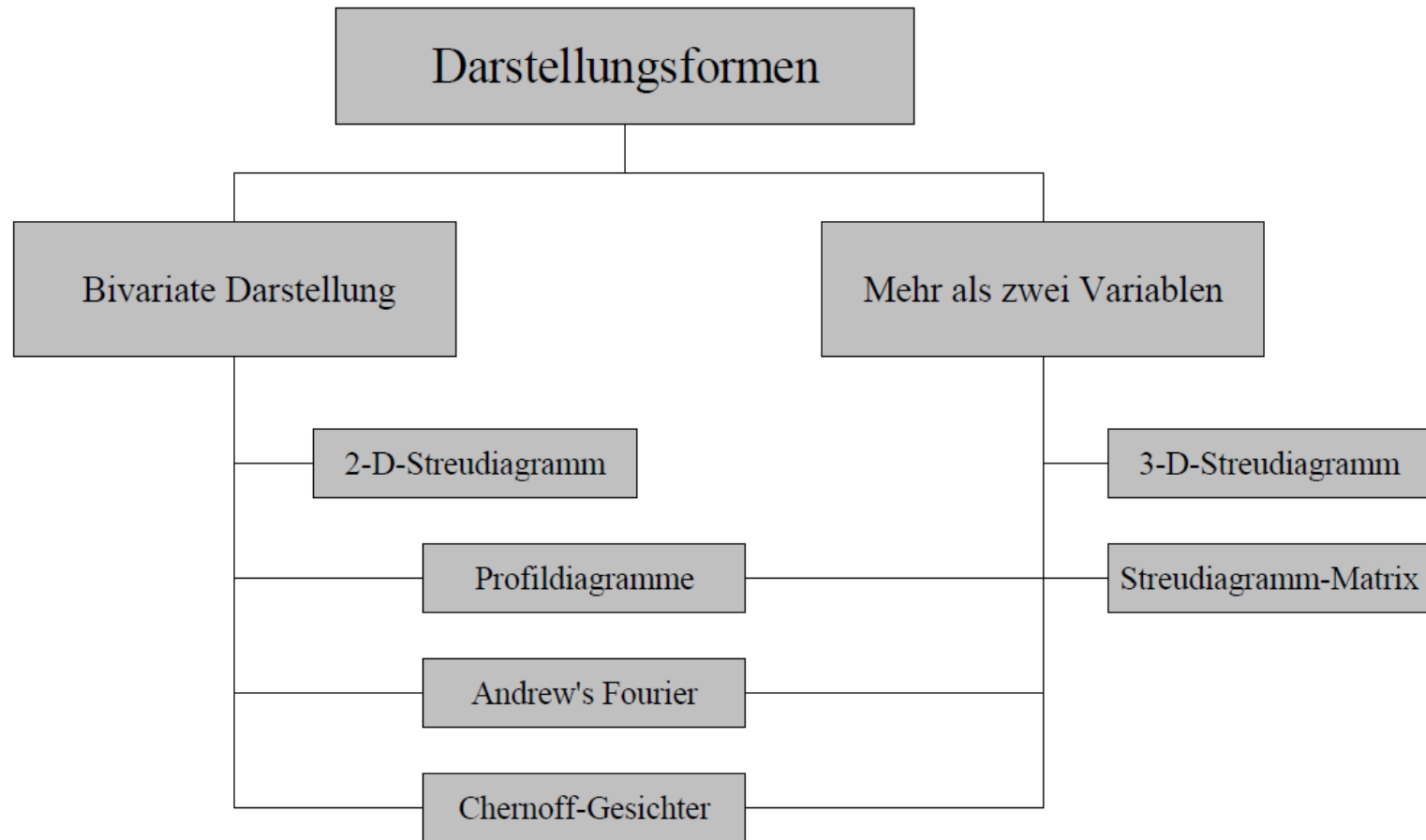
- Korrelationskoeffizient = $r = 0$
- Zwischen den beiden Variablen gibt es keinen linearen Zusammenhang somit keine Tendenz auf Gerade erkennbar

Beispiel: Variablen Anstrich des Hauses und Mietpreis

- Der Anstrich / die Farbe eines Hauses hat i. d. R. keine Einfluss auf den Mietpreis
- D.h. Zwischen den beiden Variablen gibt es keinen systematischen Zusammenhang

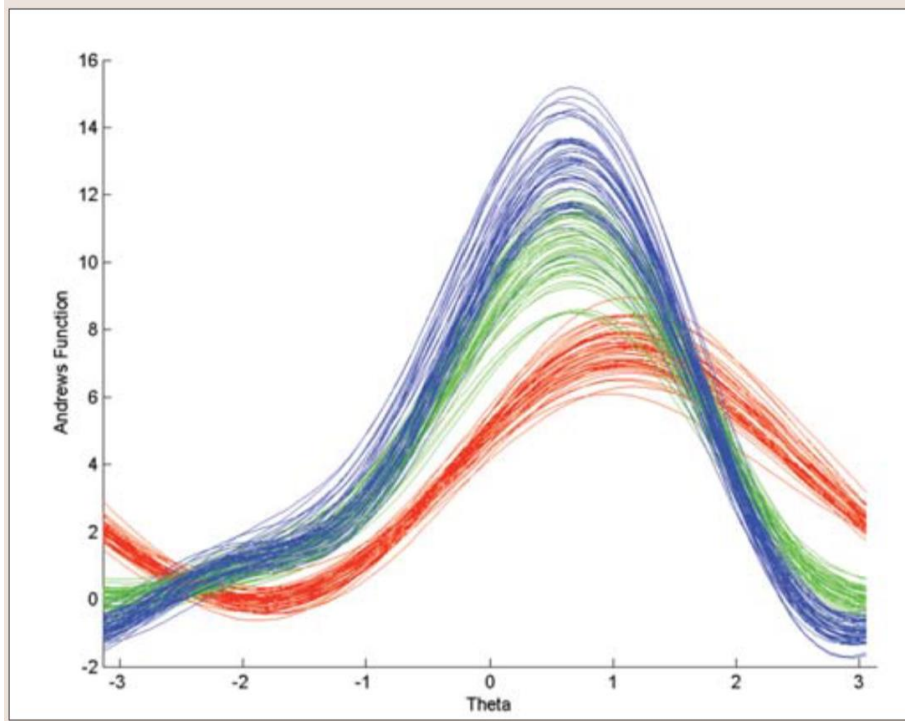
Zusammenhangsanalyse: Maßkorrelation

Darstellungsformen in Diagramme



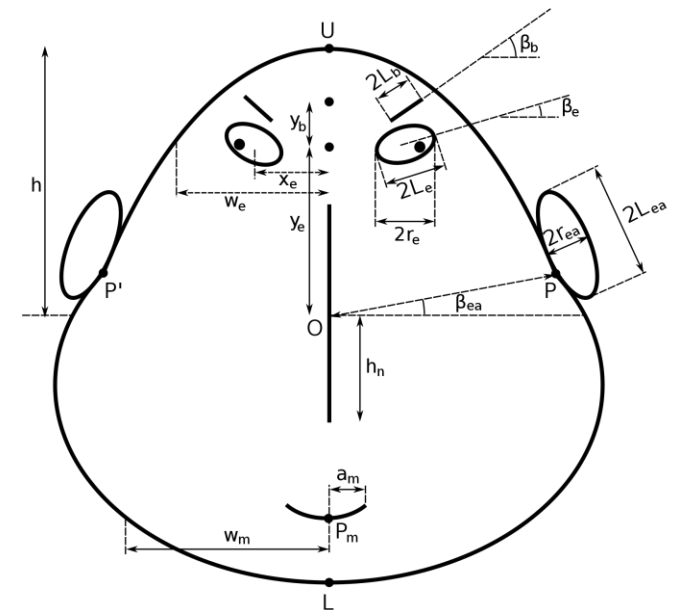
Quelle: Nadler, M. (o.J.)

Beispiel Andrew's Fourier Diagramm / Andrew's curve



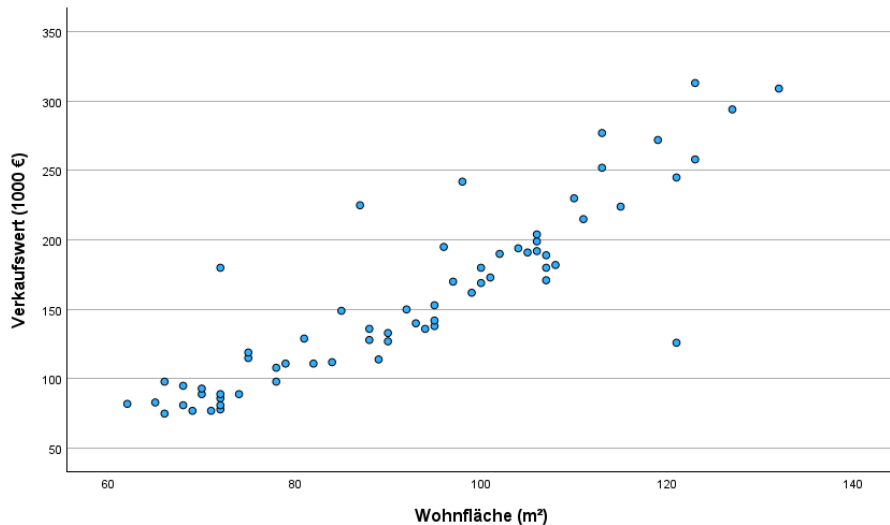
Quelle: Wallet, B. C. & Marfurt, K. (2008): A grand tour of multispectral components: A tutorial. S. 335

Beispiel Chernoff-Gesichter



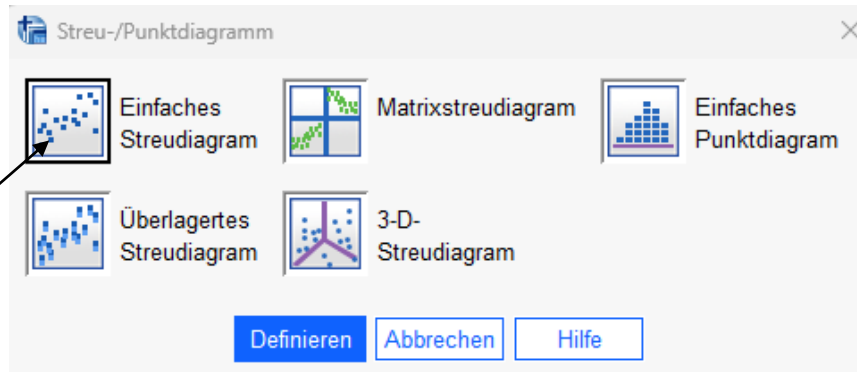
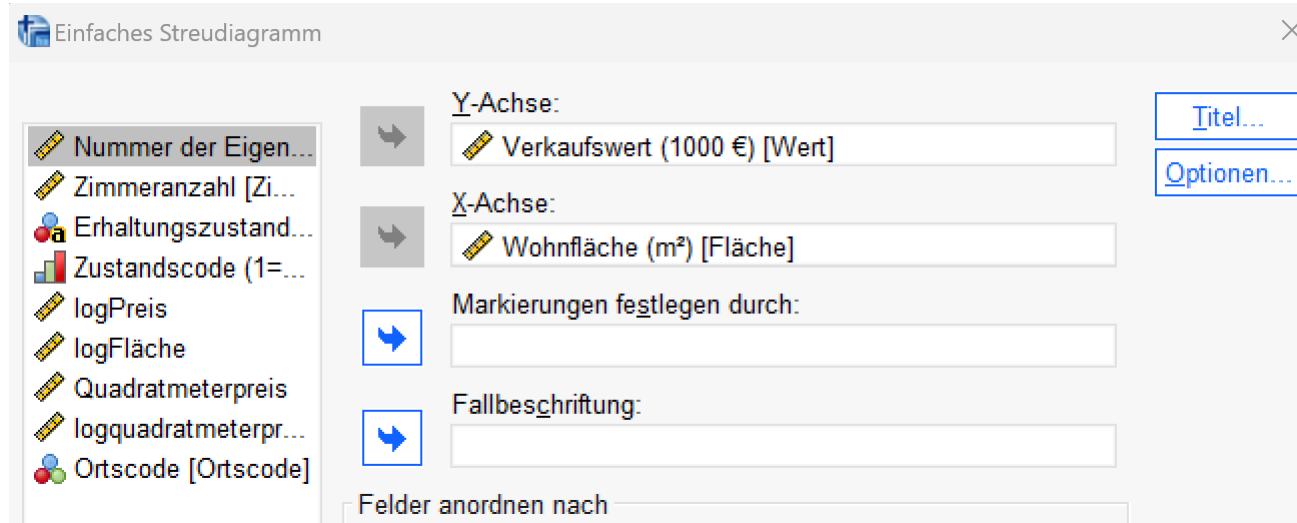
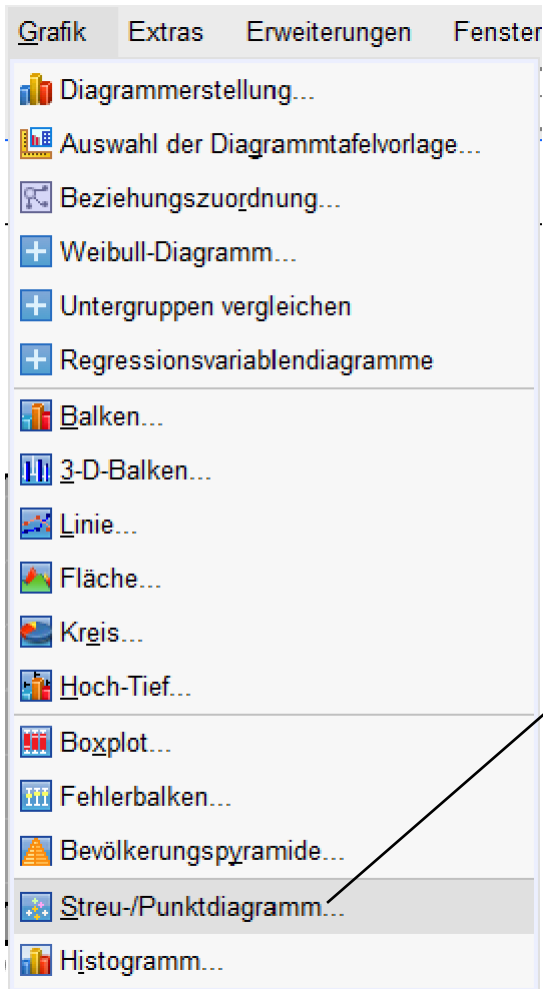
Quelle: Wikipedia.org Abrufbar unter: https://de.wikipedia.org/wiki/Chernoff-Gesichter#/media/Datei:Chernoff_faces_construction.svg zuletzt zugegriffen am 01.06.2024

Beispiel Streudiagramm



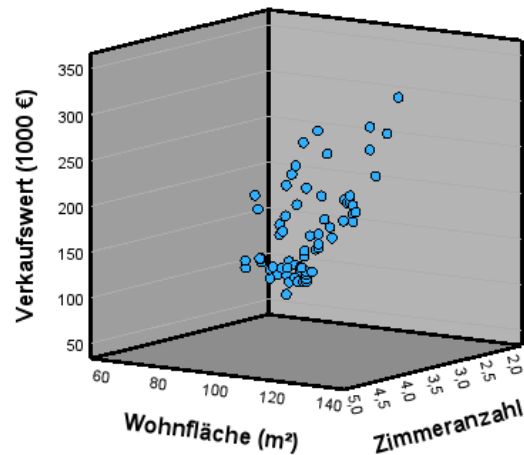
- Darstellung gemeinsame Verteilung der Werte zweier Variablen
- Durch Lage und Verteilung sind Rückschlüsse auf Zusammenhänge möglich
- Richtung ist nicht ablesbar, d.h. ob Variable A Variable B beeinflusst oder umgekehrt
- Ggfs. besteht Scheinzusammenhang

Beispiel Streudiagramm

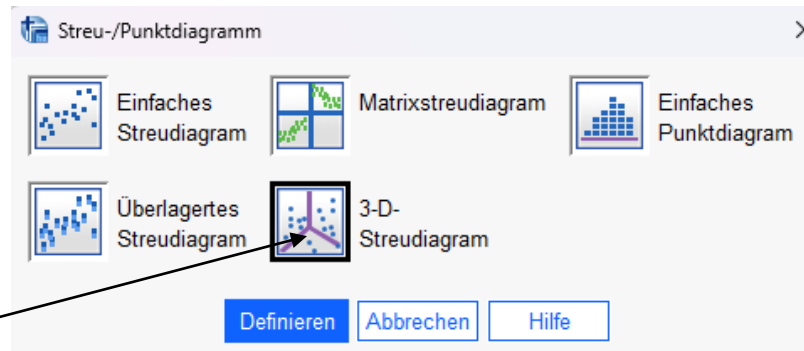
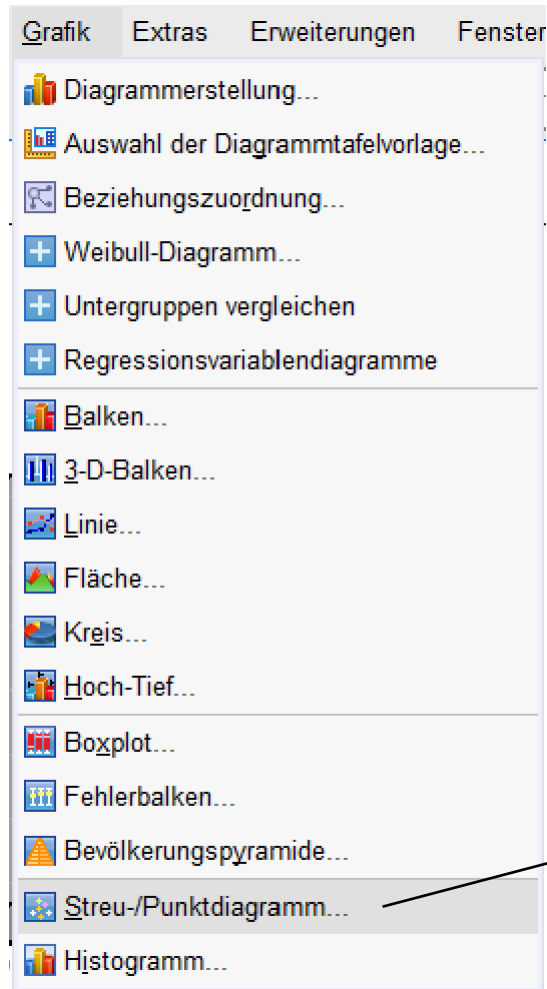


Beispiel 3D-Streudiagramm

- Einsatz bei drei Variablen



Beispiel 3D-Streudiagramm



Beispiel: 1. Test auf Normalverteilung

mithilfe von K-S-Test

H_0 = Daten sind normalverteilt

H_1 = Daten sind nicht normalverteilt

Kolmogorov-Smirnov-Test bei einer Stichprobe

		Wohnfläche (m ²)	Verkaufswert (1000 €)
N		66	66
Parameter der Normalverteilung ^{a,b}	Mittelwert	92,41	157,95
	Std.-Abweichung	18,180	63,483
Extremste Differenzen	Absolut	,104	,099
	Positiv	,104	,099
	Negativ	-,061	-,096
Teststatistik		,104	,099
Asymp. Sig. (2-seitig) ^c		,076	,176
Monte-Carlo-Signifikanz (2-seitig) ^d	Sig.	,078	,109
	99% Konfidenzintervall		
	Untergrenze	,071	,101
	Obergrenze	,084	,117

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

c. Signifikanzkorrektur nach Lilliefors.

d. Lilliefors-Methode auf der Basis von 10000 Monte-Carlo-Stichproben mit Startwert 1314643744.

Beispiel: 2. Korrelations-Koeffizienten

nach Pearson (r) berechnen um

Kontingenz/Zusammenhang zu prüfen

Interpretation:

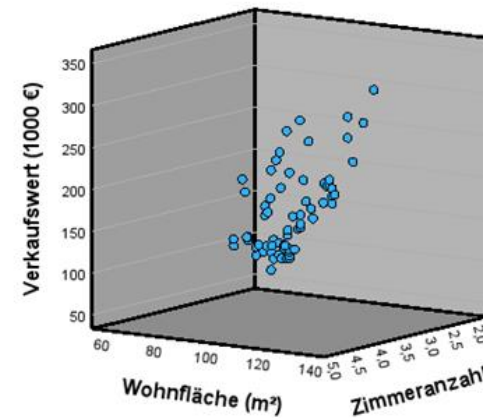
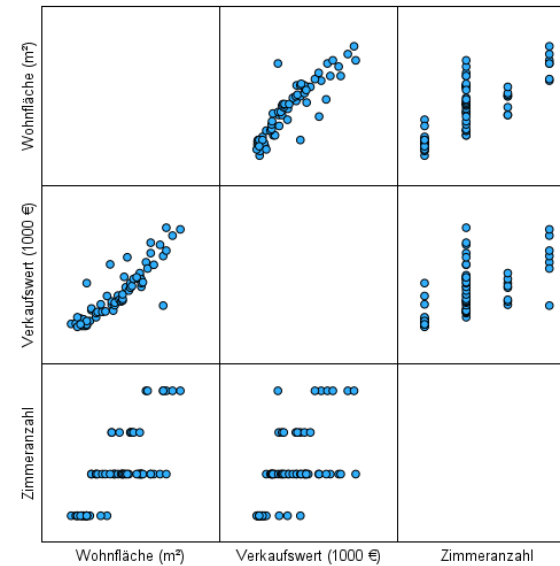
- Mit 0,886 positive Korrelierung der Variablen
- Zudem eine sehr wahrscheinlich nicht zufällige Signifikanz (Sternchen + Sig.)

Korrelationen

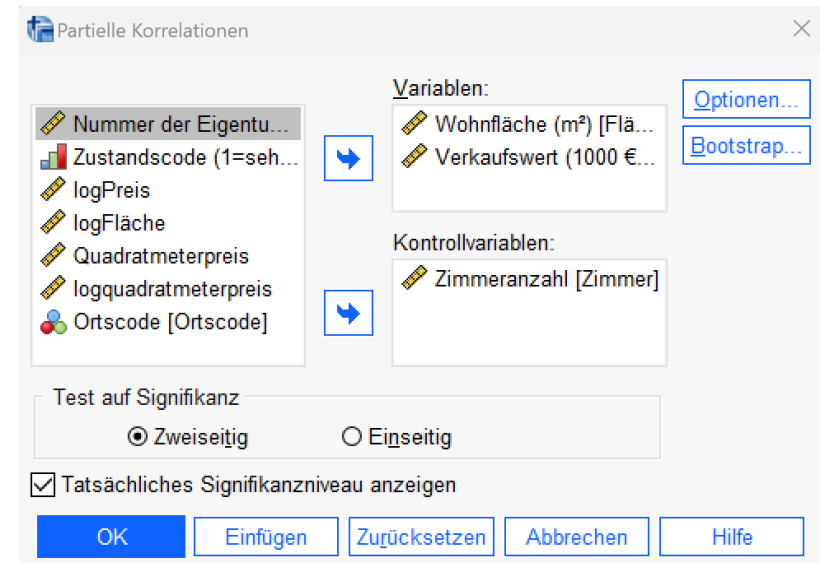
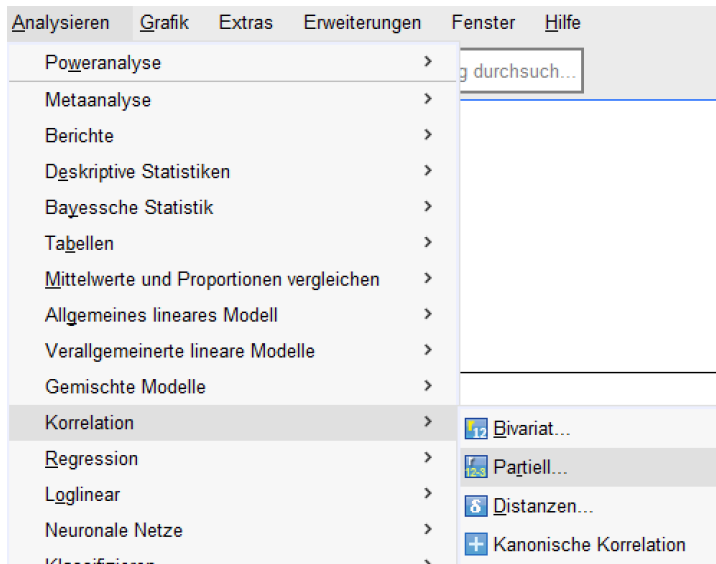
		Wohnfläche (m ²)	Verkaufswert (1000 €)
Wohnfläche (m ²)	Pearson-Korrelation	1	,886**
	Sig. (2-seitig)		<,001
	N	66	66
Verkaufswert (1000 €)	Pearson-Korrelation	,886**	1
	Sig. (2-seitig)	<,001	
	N	66	66

** . Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Beispiel: 3. Ergänzung der Analyse durch z.B. Streudiagramm-Matrix oder 3D-Streudiagramm



Beispiel: 4. Weitere Ergänzung der Analyse durch partielle Korrelation



Korrelationen

Kontrollvariablen			Wohnfläche (m²)	Verkaufswert (1000 €)
Zimmeranzahl	Wohnfläche (m²)	Korrelation	1,000	,819
		Signifikanz (zweiseitig)	.	<,001
		Freiheitsgrade	0	63
	Verkaufswert (1000 €)	Korrelation	,819	1,000
		Signifikanz (zweiseitig)	<,001	.
		Freiheitsgrade	63	0

- Mit $r = 0,819$ gibt es einen starken positiven Zusammenhang zwischen WF und VW, unabhängig von Zimmer
- Mit $\alpha^* = <0,001$ Zusammenhang ist hochsignifikant, somit sehr unwahrscheinlich, dass dies zufällig ist

Übungsaufgabe 16

- Berechnen Sie bitte mithilfe des Datensatzes „Eigentum neu.sav“ die Korrelation zwischen der Variable Quadratmeterpreis und dem Vermögenswert der Immobilie.
- a) Führen Sie eine Korrelationsanalyse wie im zuvor genannten Beispiel durch und interpretieren Sie die Ergebnisse.

Arbeitsschritte

1. Test auf Normalverteilung
2. Durchführung der Korrelationsanalyse
3. Interpretation der Ergebnisse
4. Ergänzend Partielle Korrelation

Übungsaufgabe 16

Kolmogorov-Smirnov-Test bei einer Stichprobe

		Verkaufswert (1 000 €)	Quadratmeterpreis
N		66	66
Parameter der Normalverteilung ^{a, b}	Mittelwert	157,95	1656,6353
	Std.-Abweichung	63,483	405,09112
Extremste Differenzen	Absolut	,099	,085
	Positiv	,099	,085
	Negativ	-,096	-,064
Teststatistik		,099	,085
Asymp. Sig. (2-seitig) ^c		,176	,200 ^e
Monte-Carlo-Signifikanz (2-seitig) ^d	Sig.	,105	,269
	99% Konfidenzintervall	Untergrenze	,257
		Obergrenze	,280

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

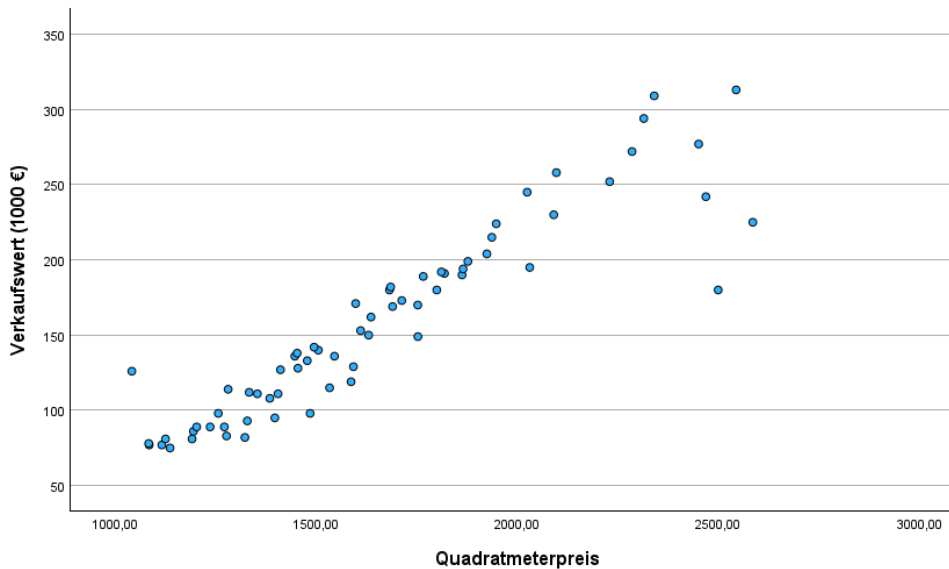
c. Signifikanzkorrektur nach Lilliefors.

d. Lilliefors-Methode auf der Basis von 10000 Monte-Carlo-Stichproben mit Startwert 2000000.

e. Dies ist eine untere Grenze der echten Signifikanz.

- Da $\alpha^* > 0,05$ liegt eine Normalverteilung für beide Variablen vor
- Somit kann Bravais-Pearson's-Korrelationskoeffizient berechnet werden

Übungsaufgabe 16



Korrelationen		Verkaufswert (1000 €)	Quadratmeterp reis
Verkaufswert (1000 €)	Pearson-Korrelation	1	,926**
	Sig. (2-seitig)		<,001
	N	66	66
Quadratmeterpreis	Pearson-Korrelation	,926**	1
	Sig. (2-seitig)	<,001	
	N	66	66

** . Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

- Aus Streudiagramm lässt sich Zusammenhang bereits ableiten
- Da Pearson = 0,926 korrelieren die Variablen positiv miteinander
- Aufgrund Sig. und ** wird eine hoch Signifikanz deutlich
- Somit besteht ein starker positiver Zusammenhang zwischen den beiden Variablen

Übungsaufgabe 16

Korrelationen			Verkaufswert (1000 €)	Quadratmeterpreis
Kontrollvariablen				
Wohnfläche (m²)	Verkaufswert (1000 €)	Korrelation	1,000	,966
		Signifikanz (zweiseitig)	.	<,001
		Freiheitsgrade	0	63
	Quadratmeterpreis	Korrelation	,966	1,000
		Signifikanz (zweiseitig)	<,001	.
		Freiheitsgrade	63	0

- Mithilfe der partiellen Korrelation und der Kontrollvariable WF wird weiterhin eine starke positive Korrelation deutlich
- Diese ist mit $\alpha^* = <0,001$ hochsignifikant, sodass es sehr wahrscheinlich ist, dass der Zusammenhang nicht zufällig ist

Regressionsanalyse

Regressionsanalyse

- Wie Zusammenhangsanalyse ist die Regression keine Kausalanalyse, d.h. die möglichen statistischen Zusammenhänge sind stets sachlogisch zu begründen („theoretisches Modell aufstellen/nutzen“)!
- Als Fortführung der Maßkorrelations- und der Varianzanalyse zu verstehen

Ziel

- Eine abhängige Variable (y) durch eine oder mehrere unabhängige Variablen (x_i) zu erklären, wobei dies nie perfekt gelingt (= Störterm U oder ε). Dabei soll eine Schätzung der Regressionsparameter gelingen und der Störterm U möglichst minimiert werden. Oder auch Vorhersage einer Variable durch eine oder mehrere Variablen.

Regressionsanalyse: Grundlagen

Unabhängige Variable entweder metrisch, ordinal oder nominal

Einfache lineare Regression



Eine unabhängige Variable um auf abhängige Variable zuschließen



Bsp.: Rückschluss auf Verkaufswert durch Anzahl qm, Erhaltungszustand **ODER** Mietpreis

Multiple lineare Regression



Mehrere unabhängige Variable um auf abhängige Variable zuschließen



Bsp.: Rückschluss auf Verkaufswert durch Anzahl qm, Erhaltungszustand **UND** Mietpreis

Logistische Regression



Bei „Ja-Nein“-Antworten bei abhängigen Variablen



Bsp.: Hat Mietpreis und die Anzahl der qm Einfluss auf die Wahrscheinlichkeit auf den Verkaufswert?

Abhängige Variable ist metrisch

Kategorische Abh. Variable (nominal o. ordinal)

(einfache) Lineare Regression

- Bei zwei metrischen Variablen gilt folgender Zusammenhang:

Sind $X(y_i) = x_i$ und $Y(y_i) = y_i$ die Werte zweier metrischer Erhebungsmerkmale X und Y einer statistischen Gesamtheit $\Gamma = \{y_i, i = 1, 2, \dots, n\}$ vom Umfang n Merkmalsträger y_i , dann heißt das Modell $Y = b_0 + b_1 \cdot X + U$ zur funktionalen Beschreibung der linearen statistischen Abhängigkeit des Merkmals Y vom Merkmal X bivariate oder einfache lineare Regression von Y und X .

- Abhängiges Merkmal Y = Regressand/Kriterium, unabhängiges Merkmal X = Regressoren/Prädiktoren, U = Zufallsbedingte Restvariable (= Störterm), b_i = (unbekannte) Regressionsparameter

Anzuwendenden Methodik

- Methode der kleinsten Quadrate (Schätzung der Parameter)

Zielstellung einfache lineare Regression

- Wert einer abhängigen Variable auf Basis **einer** unabhängigen Variable vorherzusagen
- Je größer der lineare Zusammenhang zwischen einer abhängigen und einer unabhängigen Variable, desto genauer ist die Vorhersage

Darstellungsmöglichkeiten

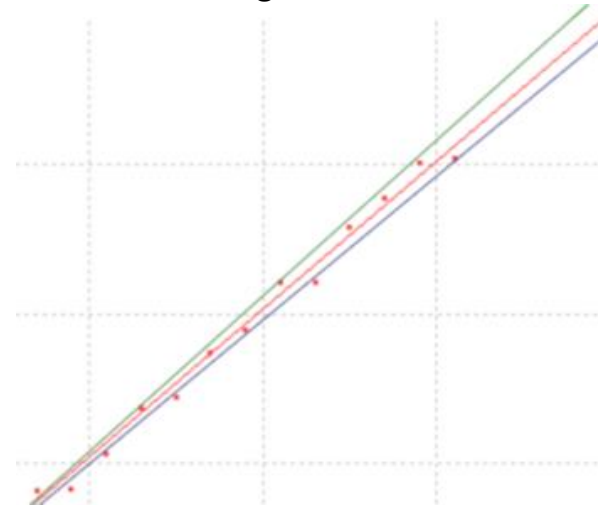
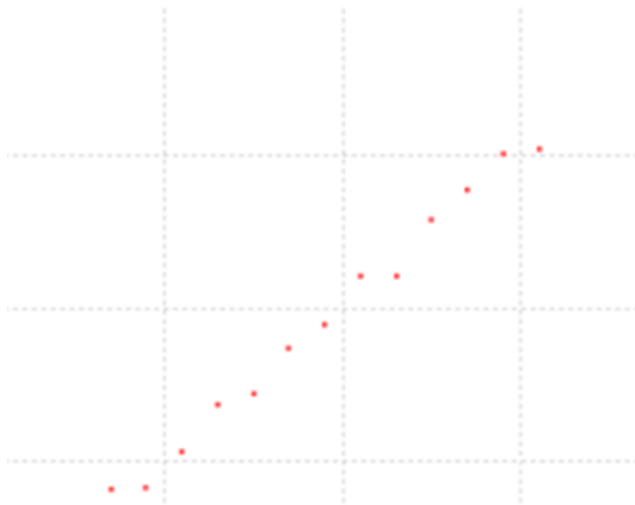
- Streudiagramm
- Je größer der Zusammenhang ist, desto mehr befinden sich die Datenpunkte auf einer Geraden

Aufgabe der Regression

- Die Gerade identifizieren, welchen den Zusammenhang am besten beschreibt

Beispiel & Grundidee

- Streudiagramm zeigt, dass der Zusammenhang beider Variablen nicht perfekt ist
- Tendenziell jedoch in die gleiche Richtung, sodass linearer Trend erkennbar ist
- Mehrere Geraden, die zum nachzeichnen der Gerade in Frage kommen würden



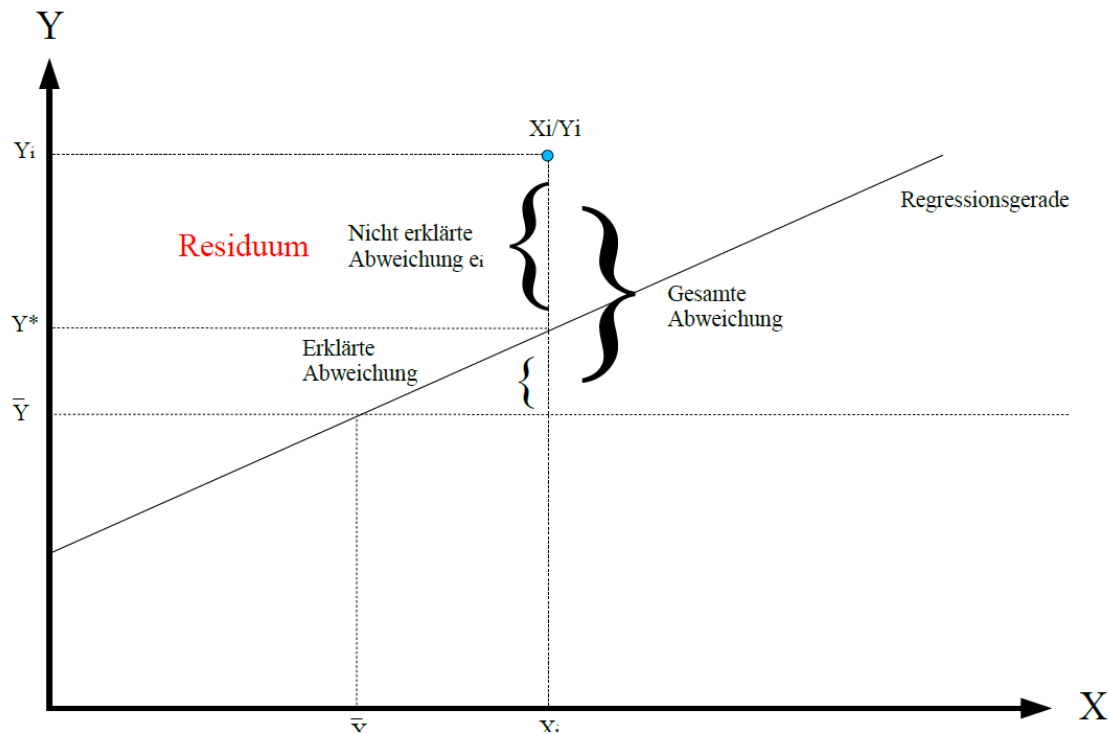
- Frage: Welche der möglichen Geraden beschreibt den Zusammenhang am besten?

Beantwortung / Methode

- Minimierung der Summe der Abweichungsquadrate mithilfe der Methode der kleinsten Quadrate

Methode der kleinsten Quadrate

- Verwendung der senkrechten Abständen der realen Werte von der Gerade
- Quadrierung der Abstände zum Wegfall negativer Vorzeichen
- Auswahl der Gerade, bei der die Summe der quadrierten Abstände minimal ist



Regressionsgerade

- Gibt Zusammenhänge, die nicht perfekt linear sind nur imperfekt wieder
- Deshalb nur selten möglich alle Veränderungen in Y durch Koeffizienten zu erklären
- Nur ein Teil der Veränderung erklärbar, andere Teil bleibt unaufgeklärt
- Verhältnis von erklärter Streuung zur Gesamtstreuung als Maß für die Güte des Regressionsmodells

Berechnung des Gütemaßes R^2

- TSS = Total Sum of Squares = Summe aller quadrierten Abweichungen
 - ESS = Explained Sum of Squares = Summe aller erklärten quadrierten Abweichungen
 - RSS = Residual Sum of Squares = Summe aller nicht erklärten quadrierten Abweich.
- Relation zwischen erklärter Streuung und Gesamtstreuung (R^2) = $\frac{ESS}{TSS}$

Gütemaß R^2 Interpretation

- R^2 als Prozentwert zu verstehen, daher stets zwischen 0 und 1
- $R^2 = 1$, gesamte Streuung wird erklärt, sodass ein perfekter linearer Zusammenhang besteht
- Umso kleiner R^2 ausfällt, desto mehr weicht der Fall vom linearen Zusammenhang ab

Bestimmtheitsmaß

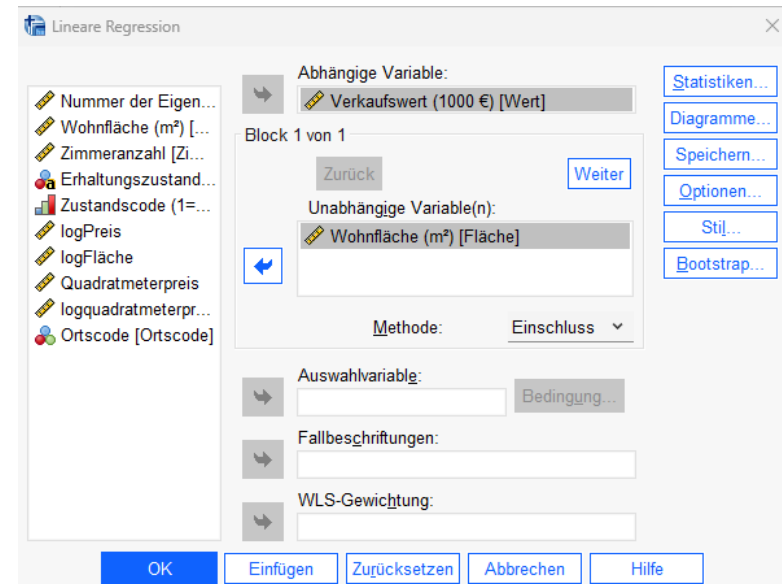
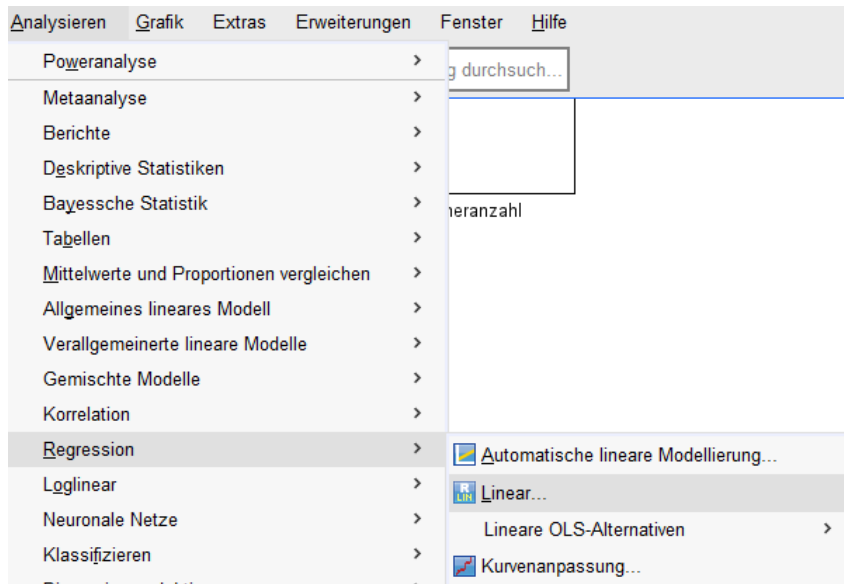
Ist $y^*(x) = y^*$ eine Regression von Y auf X, dann heißt die Größe

$$R^2 = 1 - \frac{d_x^2}{d_y^2} = 1 - \frac{\frac{1}{n} \cdot \sum_{i=1}^n (y_i - y_i^*)^2}{\frac{1}{n} \cdot \sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - y_i^*)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

empirisches Bestimmtheitsmaß einer Regression von Y auf X.

Übungsaufgabe 17

Analysieren Sie den Zusammenhang zwischen der Wohnfläche und dem Verkaufswert von Eigentumswohnungen in Berlin mithilfe der linearen Regression in SPSS anhand der Beispielsdatei „Eigentum_neu“.



Übungsaufgabe 17

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,886 ^a	,786	,782	29,625

a. Einflußvariablen : (Konstante), Wohnfläche (m²)

Koeffizienten^a

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.
		Regressionskoeffizient B	Std.-Fehler	Beta		
1	(Konstante)	-128,045	19,030		-6,729	<,001
	Wohnfläche (m ²)	3,095	,202	,886	15,313	<,001

a. Abhängige Variable: Verkaufswert (1000 €)

- $H_0: \beta_1 \leq 0$ [$H_1: \beta_1 > 0$] auf einem Signifikanzniveau $\alpha = 0.05$
- Da $\alpha=0.00 \leq 0.05$, ist H_0 zu verwerfen, d.h. es besteht ein signifikanter Zusammenhang und es gilt: Jeder qm Wohnfläche ist $3,095 \cdot 1000$ €, also 3095 € wert! Zusammen mit Korrigiertes R^2 kann mit 0.782 (= 78,2%iger) Wahrscheinlichkeit der Verkaufspreis über die Wohnfläche geschätzt werden.
- Der bivariate Korrelationskoeffizient ist über R gegeben.
- Lineare Funktion ableitbar über Regressionkoeffizienten: $Y = -128,045 + 3,095 \cdot X(\text{WF})$

Übungsaufgabe 17

ANOVA^a

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	205785,359	1	205785,359	234,482	<,001 ^b
	Nicht standardisierte Residuen	56167,504	64	877,617		
	Gesamt	261952,864	65			

a. Abhängige Variable: Verkaufswert (1000 €)

b. Einflußvariablen : (Konstante), Wohnfläche (m²)

- Findet auch Berücksichtigung für den Hypothesentest hierbei wäre:
 - H0 = Wohnfläche hat keinen Einfluss auf den Verkaufswert
 - H1 = Wohnfläche hat Einfluss auf den Verkaufswert
 - Mit $\alpha^* = 0,001$ kann H0 verworfen werden
 - Somit hat die Wohnfläche Einfluss auf den Verkaufswert

Übungsaufgabe 18

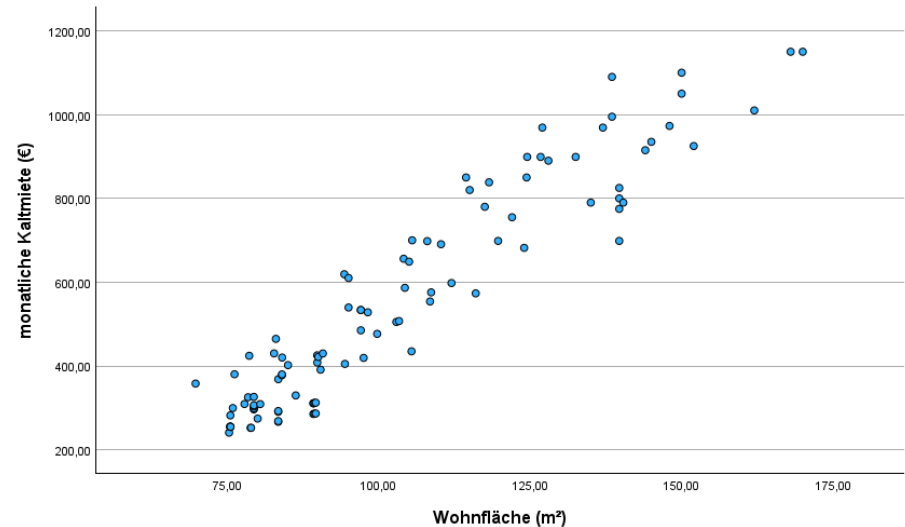
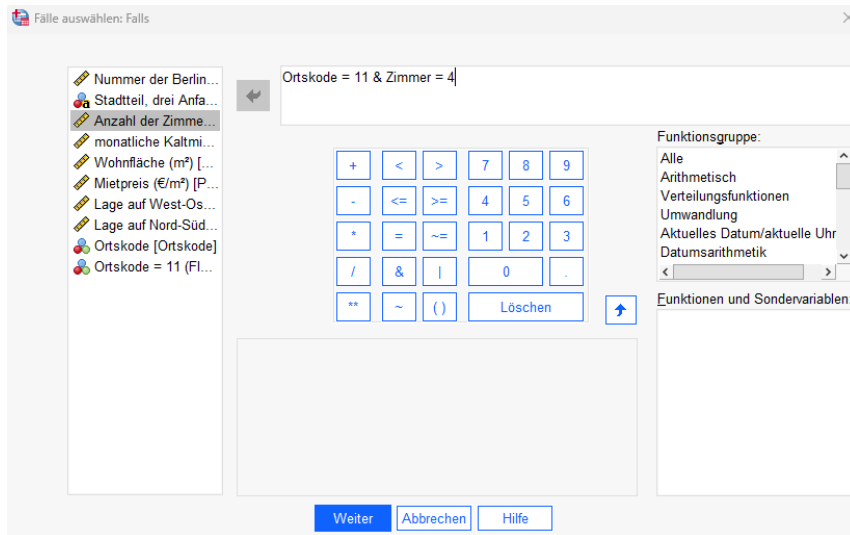
Untersuchen Sie anhand der Daten des Berliner Mietspiegels 2007 (Mietspiegel2007.sav) das Marktsegment der 4-Zimmerwohnungen im Stadtbezirks Pankow.

- a) Überprüfen Sie grafisch und explorativ die Voraussetzungen für eine lineare Regression!
- b) Konstruieren Sie mit Hilfe der Methode der kleinsten Quadrate Summe ein Modell, das die statistische Abhängigkeit des Mietwertes von der Wohnfläche zum Gegenstand hat. Verwenden Sie dazu einen linearen Modellansatz. Stellen Sie das numerisch bestimmte Modell in seiner funktionalen Form explizit dar.
- c) Interpretieren Sie die Güte des Prognosemodells! Wie hoch ist der Anteil der nicht erklärten Varianzen?
- d) Stellen Sie die lineare Mietfunktion für den Stadtbezirk Pankow auf!
- e) Zeigen Sie grafisch die Qualität der linearen Regression.
- f) Wie hoch prognostizieren Sie die Miete für eine 100qm Wohnung in Pankow?

Übungsaufgabe 18

a) Überprüfen Sie grafisch und explorativ die Voraussetzungen für eine lineare Regression!

1. Filterung der Daten nach Pankow & 4-Zimmerwohnung
2. Graphische Darstellung durch Streudiagramm



Punktwolke zeigt, dass gleich große Mietwohnungen unterschiedliche Kaltmieten aufweisen (Streuung um durchschnittliche Kaltmiete – Anzeichen für Normalverteilung)

Übungsaufgabe 18

a) Überprüfen Sie grafisch und explorativ die Voraussetzungen für eine lineare Regression!

3. (explorativ) K-S-Test zur Prüfung, ob Normalverteilung vorliegt

Kolmogorov-Smirnov-Test bei einer Stichprobe

		Wohnfläche (m ²)	monatliche Kaltmiete (€)
N		99	99
Parameter der Normalverteilung ^{a,b}	Mittelwert	104,0067	559,4145
	Std.-Abweichung	25,26889	262,57674
Extremste Differenzen	Absolut	,154	,157
	Positiv	,154	,157
	Negativ	-,118	-,113
Teststatistik		,154	,157
Asymp. Sig. (2-seitig) ^c		<,001	<,001
Monte-Carlo-Signifikanz (2-seitig) ^d	Sig.	<,001	<,001
	99% Konfidenzintervall		
	Untergrenze	,000	,000
	Obergrenze	,000	,000

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

c. Signifikanzkorrektur nach Lilliefors.

d. Lilliefors-Methode auf der Basis von 10000 Monte-Carlo-Stichproben mit Startwert 1502173562.

Übungsaufgabe 18

- b), c), d) und f)

4. Durchführung der linearen Regression

Modellzusammenfassung^b

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,945 ^a	,892	,891	86,69645

a. Einflußvariablen : (Konstante), Wohnfläche (m²)

b. Abhängige Variable: monatliche Kaltmiete (€)

ANOVA^a

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	6027682,602	1	6027682,602	801,951	<,001 ^b
	Nicht standardisierte Residuen	729078,550	97	7516,274		
	Gesamt	6756761,152	98			

a. Abhängige Variable: monatliche Kaltmiete (€)

b. Einflußvariablen : (Konstante), Wohnfläche (m²)

Koeffizienten^a

Modell		Nicht standardisierte Koeffizienten	Standardisierte Koeffizienten	T	Sig.
1	(Konstante)	-461,377		-12,441	<,001
	Wohnfläche (m²)	9,815	,945	28,319	<,001

a. Abhängige Variable: monatliche Kaltmiete (€)

- Ableitung einer Mietpreisfunktion mittels Regressionskoeffizienten möglich = $9,815 \cdot x \text{ (qm)} - 461,377$
- Zusammenhang ist hoch signifikant, sodass H_0 abgelehnt wird
- Hohes Bestimmtheitsmaß (R^2) zeigt Prognosewahrscheinlichkeit von 89,2 %
- Standardfehler bei $\pm 86,70\text{€}$ somit Sigma-Prognose = $433,42\text{€}; 606,82\text{€}$

Übungsaufgabe 18

e) Zeigen Sie grafisch die Qualität der linearen Regression.

5. Streudiagramm mit Regressionsgerade

The screenshot shows the 'Kurvenanpassung' (Curve Fitting) dialog box in a statistical software application. The 'Abhängige Variable(n):' (Dependent Variable(s)) field contains 'monatliche Kaltmiete (...)'. The 'Unabhängige Variable' (Independent Variable) section has 'Variable:' selected, and 'Wohnfläche (m²) [Fläch...]' is added. The 'Modelle' (Models) section has 'Linear' selected. The 'Fallbeschriftungen:' (Case Labels) field is empty. The 'ANOVA-Tabelle anzeigen' (Show ANOVA Table) checkbox is unchecked. The 'Speichern...' (Save...) button is visible in the top right.

Übungsaufgabe 18

e) Zeigen Sie grafisch die Qualität der linearen Regression

5. Streudiagramm mit Regressionsgerade - Output

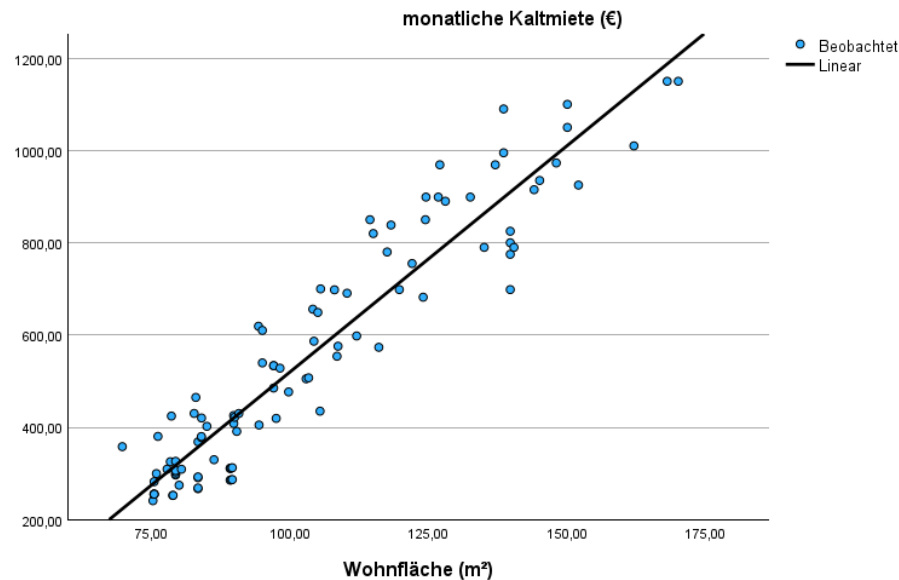
Durch die Parameterschätzungen lässt sich eine lineare Funktion (Regressionsgerade) ableiten

Modellzusammenfassung und Parameterschätzer

Abhängige Variable: monatliche Kaltmiete (€)

Gleichung	R-Quadrat	F	Modellübersicht			Sig.	Parameterschätzungen	
			df1	df2			Konstante	b1
Linear	,892	801,951	1	97		<,001	-461,377	9,815

Die unabhängige Variable ist Wohnfläche (m²).



Ausreißer

- Gemessenen oder erhobenen Wert, der nicht den Erwartungen entspricht bzw. nicht zu den restlichen Werten der Verteilung passt
- Keine klare Definition

Mögliche Ursachen für Ausreißer

- Verfahrenstechnische Fehler z.B. bei Dateneingabe
- Außergewöhnlichen Wert, jedoch erklärbar (z.B. einzig befragter Millionär) ggfs. Indiz für falsche angelegte Befragung und nicht repräsentativ
- Korrekt erfassten außergewöhnlichen Wert, welcher nicht erklärbar ist

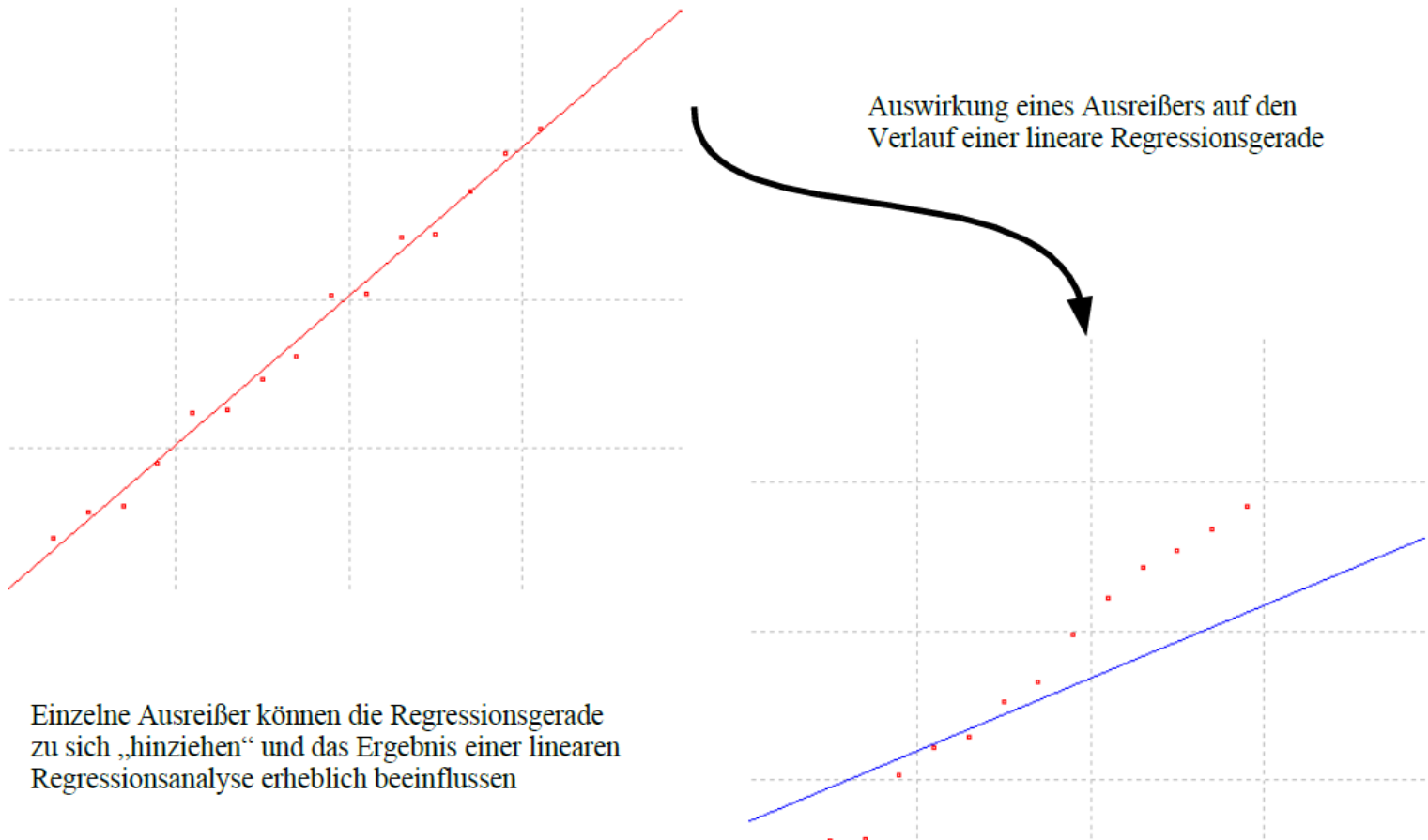
Unterscheidung

- Normale Ausreißer = außergewöhnlich großer/kleiner Wert (z.B. Einkommen)
- Multivariarer Ausreißer = für sich normal zu betrachtender Einzelwerte, in Kombination mit weiteren Variablen aber einzigartig (z.B. Ü90 Person mit Internetanschluss)

Frage in Ausreißeranalyse:

Werden Ausreißer beibehalten oder verworfen?

Beispiel: Regression mit Leverage(Hebel)-Effekt



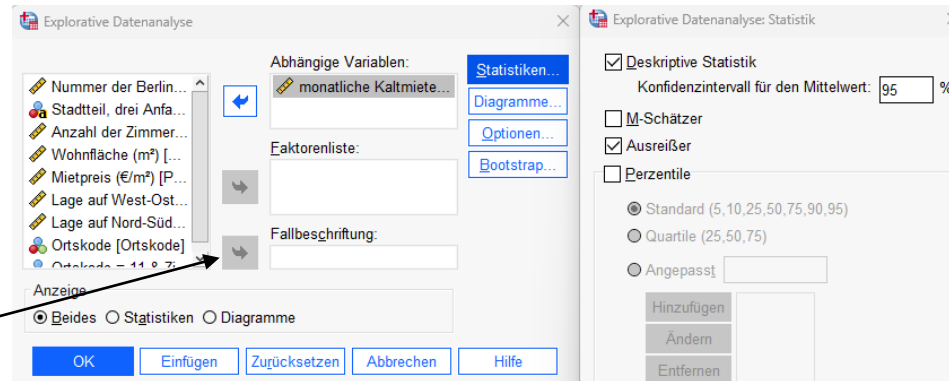
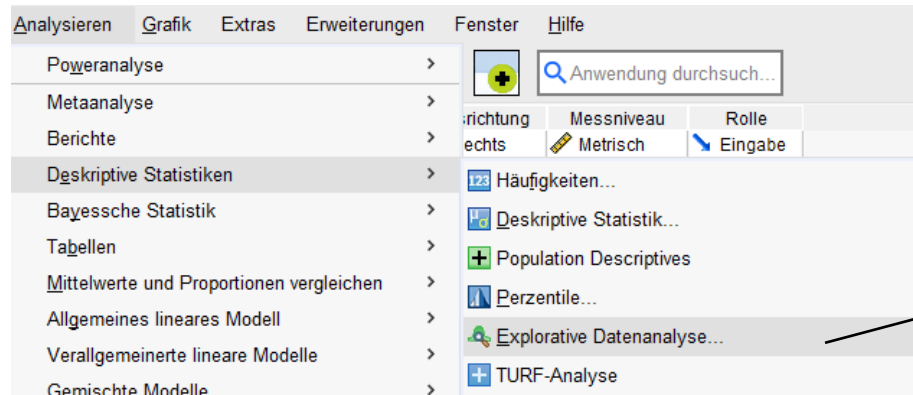
Umgang mit Ausreißern

1. Ausschluss aus Analyse
2. Eingang in Analyse
3. Kennzeichnung als fehlende Werte

Zustellende Fragen

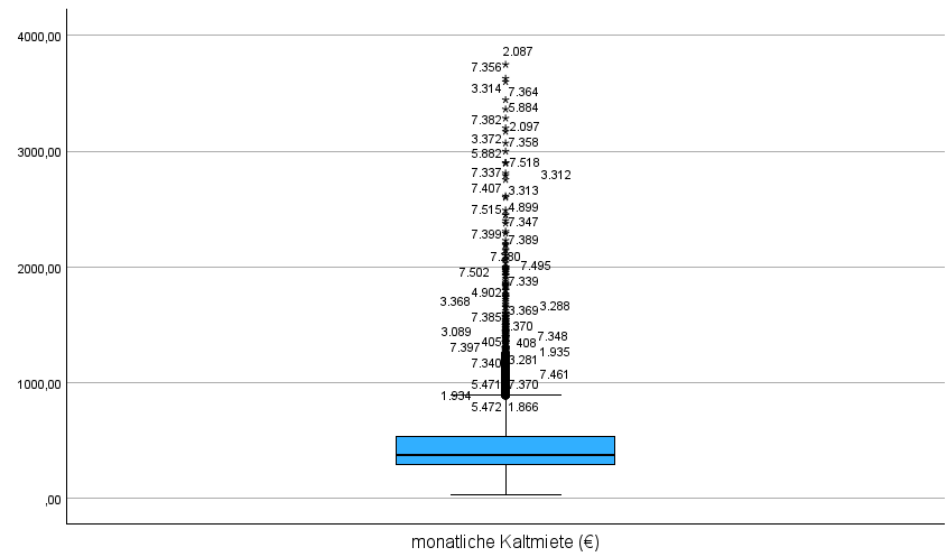
- **Wie ist das Auftreten des Ausreißer zu erklären?**
 - Eingabefehler, Bereinigung möglich?
 - Sind Rückschlüsse auf Anlage und Durchführung der Erhebung möglich?
- **Welche Auswirkungen haben die Ausreißer auf die Ergebnisse der Datenanalyse?**
 - Beeinflussen die Ausreißer die Regressionsgerade? (siehe Leverage-Effekt)
 - Führen die Ausreißer zu einer Verzerrung der Analyse, sodass diese entfernt werden müssen?
- **Entstehen mögliche Datenverluste bei Entfernung der Ausreißer?**

Identifikation von Ausreißer: Extremwerte und Boxplot

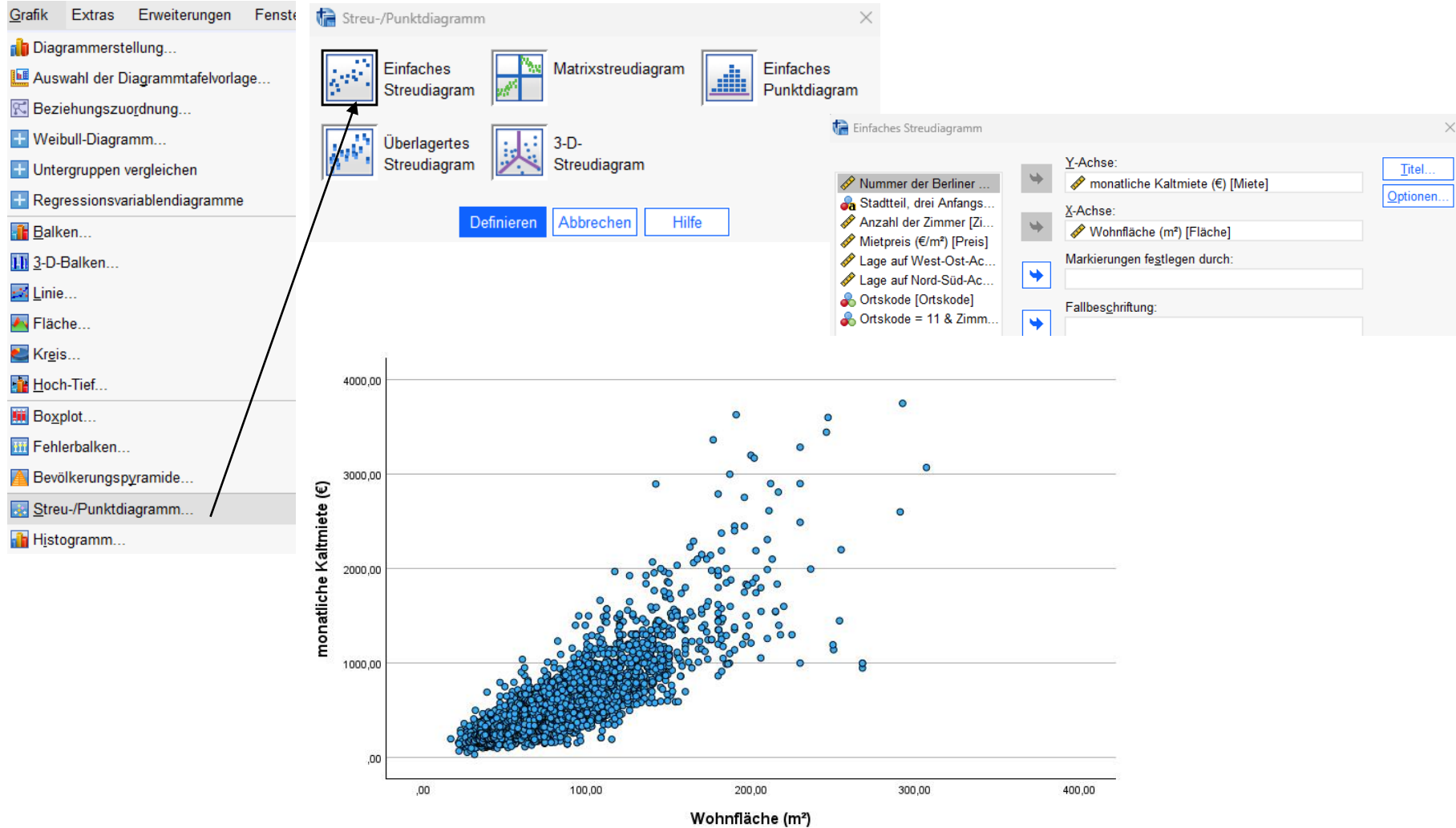


Extremwerte

		Fallnummer		Wert
monatliche Kaltmiete (€)	Größte Werte	1	2087	3749,21
		2	7356	3629,00
		3	7364	3600,00
		4	3314	3444,00
		5	7382	3363,00
	Kleinste Werte	1	6072	32,63
		2	6064	55,21
		3	3832	69,26
		4	6067	76,69
		5	3835	97,16



Identifikation von Ausreißer: Streudiagramm



Voraussetzung zur Regression

- Erklärende Variablen sind sich ähnlich – keine Multikollinearität
 - Überprüfung: $VIF < 5$
- Einzelne Daten sind unterschiedlich vertrauenswürdig – keine Heteroskedastizität
 - Überprüfung: t-Werte (Koeffizient/Standardfehler) > 2
- Störterme beeinflussen sich gegenseitig – keine serielle Korrelation
 - Überprüfung: Durbin-Watson Statistik ca. 2 (0 oder 4 schlecht)

Übungsaufgabe 19

- Überprüfen Sie mithilfe des Datensatzes „Eigentum neu.sav“ den Zusammenhang zwischen dem Verkaufspreis, der Wohnfläche und der Zimmeranzahl von Eigentumswohnungen und prüfen Sie auf Regressionsfehler.

The screenshot displays the SPSS interface for performing a linear regression analysis. The main window shows the 'Lineare Regression' dialog box with the following settings:

- Abhängige Variable:** Verkaufswert (1000 €) [Wert]
- Unabhängige Variable(n):** Wohnfläche (m²) [Fläche], Zimmeranzahl [Zimmer]
- Methode:** Einschluss
- Auswahlvariable:** (empty)
- Fallbeschriftungen:** (empty)
- WLS-Gewichtung:** (empty)

The 'Lineare Regression: Statistiken' sub-dialog box is also visible, showing the following options:

- Regressionskoeffizienten:**
 - ☒ Schätzungen
 - ☐ Konfidenzintervalle
 - Stufe (%): 95
 - ☐ Kovarianzmatrix
 - ☒ Anpassungsgüte des Modells
 - ☐ Änderung in R-Quadrat
 - ☐ Deskriptive Statistik
 - ☐ Teil- und partielle Korrelationen
 - ☒ Kollinearitätsdiagnose
 - ☐ Auswahlkriterien
- Residuen:**
 - ☐ PRESS
 - ☒ Durbin-Watson
 - ☐ Fallweise Diagnose
 - ☒ Ausreißer außerhalb 3 Standardabweichungen
 - ☐ Alle Fälle

Buttons at the bottom of the sub-dialog include 'Weiter', 'Abbrechen', and 'Hilfe'. The main dialog box also has buttons for 'OK', 'Einfügen', 'Zurücksetzen', 'Abbrechen', and 'Hilfe'.

Übungsaufgabe 19

Modellzusammenfassung^b

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Durbin-Watson-Statistik
1	,892 ^a	,795	,789	29,178	1,431

a. Einflußvariablen : (Konstante), Zimmeranzahl, Wohnfläche (m²)

b. Abhängige Variable: Verkaufswert (1000 €)

ANOVA^a

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	208316,636	2	104158,318	122,342	<,001 ^b
	Nicht standardisierte Residuen	53636,227	63	851,369		
	Gesamt	261952,864	65			

a. Abhängige Variable: Verkaufswert (1000 €)

b. Einflußvariablen : (Konstante), Zimmeranzahl, Wohnfläche (m²)

Toleranz sollte über 0,1/0,2 liegen
VIF max. 10 besser unter 5

Koeffizienten^a

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	Kollinearitätsstatistik	
		Regressionskoeffizient B	Std.-Fehler	Beta			Toleranz	VIF
1	(Konstante)	-133,180	18,978		-7,018	<,001		
	Wohnfläche (m ²)	3,504	,310	1,004	11,313	<,001	,413	2,421
	Zimmeranzahl	-10,677	6,192	-,153	-1,724	,090	,413	2,421

a. Abhängige Variable: Verkaufswert (1000 €)

Übungsaufgabe 19

Modellzusammenfassung^b

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Durbin-Watson-Statistik
1	,886 ^a	,786	,782	29,625	1,420

a. Einflußvariablen : (Konstante), Wohnfläche (m²)

b. Abhängige Variable: Verkaufswert (1000 €)

ANOVA^a

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	205785,359	1	205785,359	234,482	<,001 ^b
	Nicht standardisierte Residuen	56167,504	64	877,617		
	Gesamt	261952,864	65			

a. Abhängige Variable: Verkaufswert (1000 €)

b. Einflußvariablen : (Konstante), Wohnfläche (m²)

- Wie t-Werte zeigen, kann die Eliminierung einer Variable z.B. Zimmer sinnvoll sein
- Er hilft dabei die Bedeutung der unabhängigen Variable(n) zu verstehen
- (absolut) hoher t-Wert und kleine Sig. Deutet auf signifikanten Einfluss der Unabhängigen auf Abhängigen

Koeffizienten^a

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		Kollinearitätsstatistik	
		Regressionskoeffizient B	Std.-Fehler			Toleranz	VIF
1	(Konstante)	-128,045	19,030		-6,729	<,001	
	Wohnfläche (m ²)	3,095	,202	,886	15,313	<,001	1,000

a. Abhängige Variable: Verkaufswert (1000 €)

Modellfehler

- **wichtige erklärende Variablen fehlen im Modell**
 - R^2 ist niedrig
 - R^2 sollte $\geq 0,80$ also 80% sein
- **Überflüssige Variablen befinden sich im Modell**
 - Die Effizienz des Modells wird niedriger
 - Durch Einzelsignifikanzprüfungen der Variablen kann Einfluss auf R^2 geprüft werden
- **Falsche funktionale Form des Modells**
 - Erklärungsgehalt von R^2 vermindert und Verzerrung der Schätzwerte
 - Bsp. Wohnraumnachfrage abhängig vom Einkommen: Linear (fester Anteil von Wohnraum für jeden Euro) oder exponentiell (kein fester Anteil, sondern mit zunehmenden einkommen sinkender Anteil)

Deshalb: Neben bivariate lineare Regression auch nichtlineare (logarithmisch o. logistische) Regression

Multiple lineare Regression

- Mehrere unabhängige Variablen

$$Y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \dots + \beta_k * x_k + \varepsilon_i$$

Nicht lineare Regressionen

- z.B. logarithmisch, quadratisch, kubisch, exponentiell...

$$Y = \beta_0 + \beta_1 * (\ln x)$$

$$Y = \beta_0 + \beta_1 * X + \beta_2 * X^2$$

$$Y = \beta_0 + \beta_1 * X + \beta_2 * X^2 + \beta_3 * X^3$$

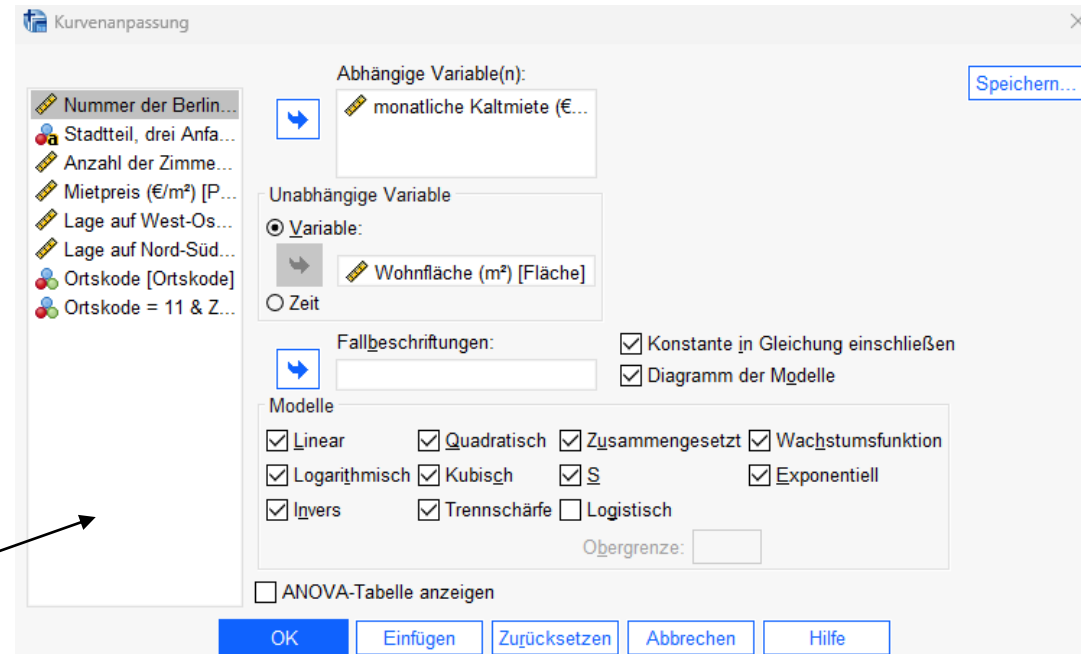
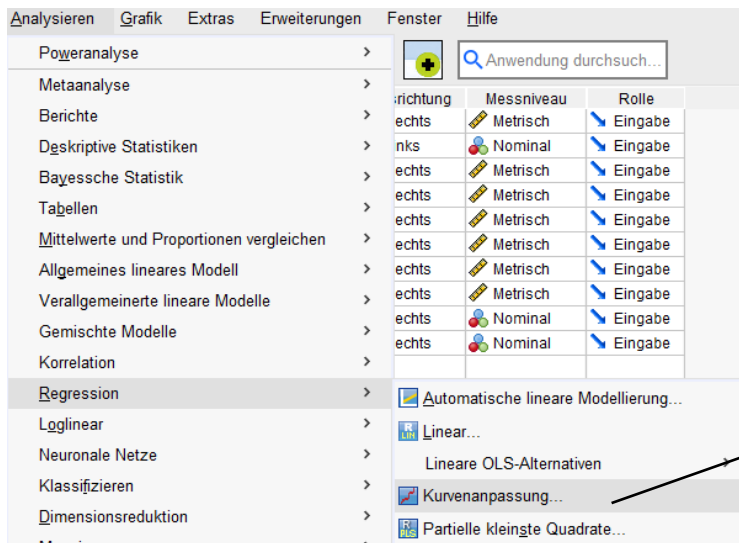
$$Y = \beta_0 * e^{(\beta_1 * X)}$$

Funktionstyp	funktionale Form
Logarithmisch	$Y = \beta_0 + \beta_1 \cdot (\ln X)$
Invers	$Y = \beta_0 + \beta_1 \cdot X^{-1}$
Quadratisch	$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2$
Kubisch	$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \beta_3 \cdot X^3$
Power bzw. Potenz	$Y = \beta_0 \cdot X^{\beta_1}$
Zusammengesetzt	$Y = \beta_0 \cdot \beta_1^X$
S (förmig)	$Y = \exp(\beta_0 + \beta_1 \cdot X^{-1})$
Logistisch	$Y = (1/c + \beta_0 \cdot \beta_1^X)^{-1}$
Wachstum	$Y = \exp(\beta_0 + \beta_1 \cdot X)$
Exponentiell	$Y = \beta_0 \cdot \exp(\beta_1 \cdot X)$

Quelle: Eckstein, P. (2000³): Angewandte Statistik in SPSS – Praktische Einführung für Wirtschaftswissenschaftler. S. 227

Identifikation der geeigneten Regression

- Einsatz der Kurvenanpassung
- Mithilfe von R^2 (höchster Wert) findet sich die passende Funktion



Identifikation der geeigneten Regression

- Somit ist nicht die lineare Funktion am sinnvollsten, sondern die kubische.
- Ableitung durch R^2

Modellzusammenfassung und Parameterschätzer

Abhängige Variable: monatliche Kaltmiete (€)

Gleichung	Modellübersicht					Parameterschätzungen			
	R-Quadrat	F	df1	df2	Sig.	Konstante	b1	b2	b3
Linear	,666	14998,185	1	7516	<,001	-171,694	8,516		
Logarithmisch	,525	8320,600	1	7516	<,001	-2075,548	598,140		
Invers	,354	4121,231	1	7516	<,001	925,791	-30223,708		
Quadratisch	,689	8325,906	2	7515	<,001	29,288	3,761	,024	
Kubisch	,700	5837,124	3	7514	<,001	262,008	-4,280	,103	,000
Zusammengesetzt	,660	14619,624	1	7516	<,001	140,035	1,014		
Netzstrom	,655	14288,413	1	7516	<,001	3,581	1,112		
S	,546	9030,306	1	7516	<,001	6,952	-62,456		
Wachstumsfunktion	,660	14619,624	1	7516	<,001	4,942	,014		
Exponentiell	,660	14619,624	1	7516	<,001	140,035	,014		

Die unabhängige Variable ist Wohnfläche (m²).

Ableitung der Formel

$$Y(\text{Miete pro qm}) = 262,008 + (-4,280X) + 0,104X^2 + 0,000X^3$$

Bei einer 100qm Wohnung würde man mit einer Wahrscheinlichkeit von 70% eine Miete von 874,008 € prognostizieren.

- **Faktorenanalyse**
- **Zuverlässigkeitsanalyse**
- **Get Started: Datensatz aufbereiten und Analysebeginn**

- Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2016¹⁴): Multivariate Analysemethoden – Eine anwendungsorientierte Einführung. Springer-Verlag, Berlin Heidelberg. DOI:10.1007/978-3-662-46076-4
- Eckstein, P. (2000³): Angewandte Statistik mit SPSS – Praktische Einführung für Wirtschaftswissenschaftler. Gabler Verlag, Wiesbaden. DOI: 10.1007/978-3-322-93157-3
- Eckstein, P. (2017⁶): Datenanalyse mit SPSS – Realdatenbasierte Übungs- und Klausuraufgaben mit vollständigen Lösungen. Springer Fachmedien, Wiesbaden. 10.1007/978-3-658-18039-3
- Kohn, W. & Öztürk, R. (2013²): Statistik für Ökonomen – Datenanalyse mit R und SPSS. Springer-Verlag, Berlin Heidelberg. DOI: 10.1007/978-3-642-37352-7
- Schnell, R. (2019²): Survey-Interviews – Methoden standardisierter Befragungen. Springer Fachmedien, Wiesbaden. DOI: 10.1007/978-3-531-19901-6

- „Statistik am PC“
- „DATAtab“



Kontakt:

Lehrstuhl Immobilienentwicklung, GB1 323

M.Sc. Miriam Bretz

miriam.bretz@tu-dortmund.de

M.Sc. Alexander Schuhmacher

alexander.schuhmacher@tu-dortmund.de