1. Einleitung

● Uni- und bivariate Datenanalyse

Univariate Datenanalyse	**Bivariate Datenanalyse**
Auswertung eines Merkmals	Verbundene Auswertung zweier Merkmale
Beispiele:	Beispiele:
Univariate Häufigkeitsverteilungen, arithmetisches Mittel, Varianz und Standardabweichung	Bivariate Häufigkeitsverteilungen, Korrelationskoeffizient, einfache Regressionsanalyse

Merkmale (Variablen): $X_1, X_2, \ldots, X_m$

● Univariate Datenanalyse

Arithmetisches Mittel der Variablen X_j	$\bar{x}_j = \frac{1}{n} \cdot \sum_{i=1}^{n} x_{ij}$
Varianz der Variablen X_j	$s_j^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j \right)^2$
Standardabweichung der Variablen X_j	$s_j = \sqrt{s_j^2}$

12-Regionen-Beispiel

Beispiel 1.1: Wir betrachten die Werte der Variablen X_1 (=Einwohnerdichte) an 12 Untersuchungsregionen, die wir mit A, B, C, …, L bezeichnen. Es ist x_{i1} der Merkmalswert der Variablen X_1 bei Untersuchungseinheit i (i=1,2,…,n=12).

i	Region	x_{i1}	$x_{i1}-\bar{x}_1$	$(x_{i1}-\bar{x}_1)^2$
1	A	212,4	-114,158	13.032,049
2	B	623,7	297,142	88.293,368
3	C	93,1	-233,458	54.502,638
4	D	236,8	-89,758	8.056,499
5	E	412,0	85,442	7.300,335
6	F	566,7	240,142	57.668,180
7	G	331,9	5,342	28,537
8	H	111,4	-215,158	46.292,965
9	I	489,0	162,442	26.387,403
10	J	287,4	-39,158	1.533,349
11	K	166,2	-160,358	25.714,688
12	L	388,1	61,542	3.787,418
Σ	-	3.918,7	0	332.597,429

Das **arithmetische Mittel** der Einwohnerdichte beträgt

$$\bar{x}_1 = \frac{1}{n} \cdot \sum_{i=1}^{n} x_{i1} = \frac{1}{12} \cdot \sum_{i=1}^{12} x_{i1} = \frac{1}{12} \cdot 3.918{,}7 = 326{,}558 \; .$$

Für die **Varianz** erhält man den Wert

$$s_1^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_{i1} - \bar{x}_1\right)^2 = \frac{1}{11} \cdot 332.597{,}429 = 30.236{,}130$$

und für die **Standardabweichung** den Wert

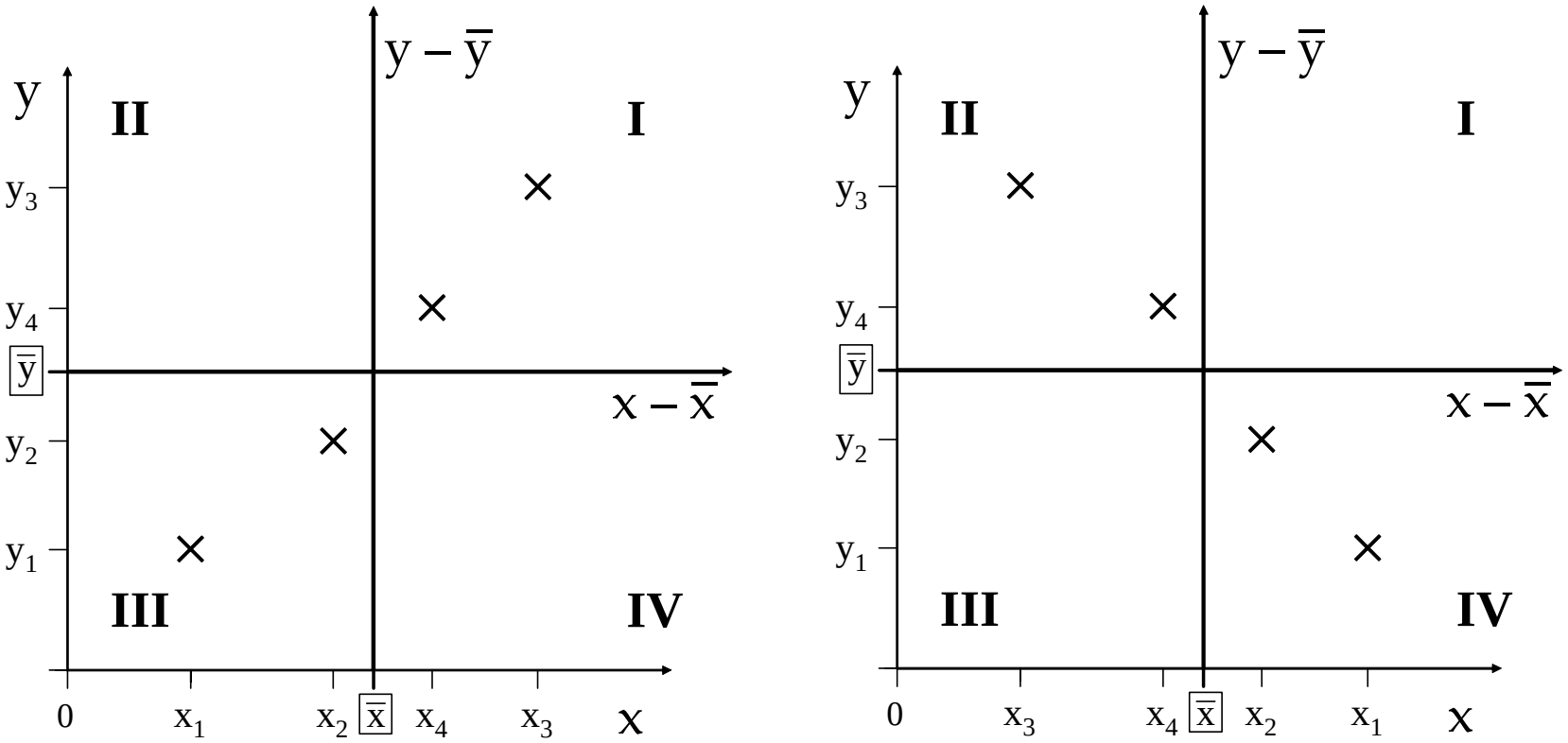
$$s_1 = \sqrt{30.236{,}130} = 173{,}885 \; .$$

Die durchschnittliche Einwohnerdichte liegt also bei 326,558 Einwohnern pro Quadratkilometer. Die durchschnittliche Abweichung der einzelnen Regionen von diesem arithmetischen Mittel beträgt 173,885 Einwohner pro km^2. ♦

● Bivariate Datenanalyse

Kovarianz Wertebereich: Menge der reellen Zahlen (keine Begrenzung)	$s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right) \cdot \left(x_{ik} - \bar{x}_k\right)$
Korrelationskoeffizient (nach Bravais und Pearson) Wertebereich: $-1 \leq r_{jk} \leq 1$	$r_{jk} = \frac{s_{jk}}{s_j \cdot s_k} = \frac{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right) \cdot \left(x_{ik} - \bar{x}_k\right)}{\sqrt{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2 \cdot \sum_{i=1}^{n} \left(x_{ik} - \bar{x}_k\right)^2}}$
Einfache Regressionsanalyse Regressionsfunktion: $\hat{x}_k = a + b \cdot x_j$	Kleinst-Quadrate-Schätzer: $\hat{b} = \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sum_{i=1}^{n} (x_{ij} - \bar{x})^2}$ $\hat{a} = \frac{\sum_{i=1}^{n} x_{ik}}{n} - \hat{b} \cdot \frac{\sum_{i=1}^{n} x_{ij}}{n}$ Determinationskoeffizient: $R^2 = r_{jk}^2$

Abbildung 1.1: Streudiagramm mit einer Unterteilung in Quadranten



a) positiver Zusammenhang

b) negativer Zusammenhang

Interpretation der **Kovarianz** s_{jk}:

- $s_{jk} > 0$: positiver Zusammenhang (positive Kreuzprodukte überwiegen, Beobachtungswerte liegen überwiegend im ersten und dritten Quadranten)
- $s_{jk} = 0$: kein (linearer) Zusammenhang (positive und negative Kreuzprodukte entsprechen sich)
- $s_{jk} < 0$: negativer Zusammenhang (negative Kreuzprodukte überwiegen, Beobachtungswerte befinden sich überwiegend im zweiten und vierten Quadranten).

Interpretation des **Korrelationskoeffizienten** r_{jk}

- $-1 \leq r_{jk} < -0{,}8$: Starker negativer Zusammenhang
- $-0{,}8 \leq r_{jk} < -0{,}3$: Mittlerer negativer Zusammenhang
- $-0{,}3 \leq r_{jk} < 0$: Schwacher negativer Zusammenhang
- $0 < r_{jk} \leq 0{,}3$: Schwacher positiver Zusammenhang
- $0{,}3 < r_{jk} \leq 0{,}8$: Mittlerer positiver Zusammenhang
- $0{,}8 < r_{jk} \leq 1$: Starker positiver Zusammenhang

Beispiel 1.2: Wir nehmen als weitere Variable das BIP je Einwohner auf und wollen den Zusammenhang der Einwohnerdichte (X_1) und dem BIP je Einwohner (X_2) überprüfen. Allgemein ist nun x_{ij} der Wert von Untersuchungseinheit i (i=1,2,…, n; hier n=12) der Variablen j (j=1,2,…,m; hier j=2). Die benötigten Summen lauten:

i (Region)	x_{i1}	x_{i2}	$(x_{i1}-\bar{x}_1)^2$	$(x_{i2}-\bar{x}_2)^2$	$(x_{i1}-\bar{x}_1)\cdot(x_{i2}-\bar{x}_2)$
1 (A)	212,4	20.116	13.032,049	5.837.457,063	275.815,200
2 (B)	623,7	24.966	88.293,368	5.923.951,963	723.218,970
3 (C)	93,1	19.324	54.502,638	10.291.796,535	748.952,640
4 (D)	236,8	23.113	8.056,499	337.464,561	-52.141,950
5 (E)	412,0	23.076	7.300,335	295.845,703	46.473,360
6 (F)	566,7	24.516	57.668,180	3.935.926,663	476.421,800
7 (G)	331,9	22.187	28,537	119.082,277	-1.843,430
8 (H)	111,4	20.614	46.292,965	3.679.042,395	412.690,900
9 (I)	489,0	25.006	26.387,403	6.120.265,323	401.868,030
10 (J)	287,4	23.136	1.533,349	364.715,743	-23.648,180
11 (K)	166,2	20.707	25.714,688	3.330.927,957	292.666,660
12 (L)	388,1	23.624	3.787,418	1.192.282,735	67.198,760
Σ	3.918,7	270.385	332.597,429	41.428.758,918	3.367.672,760

Das durchschnittliche BIP pro Einwohner liegt damit bei

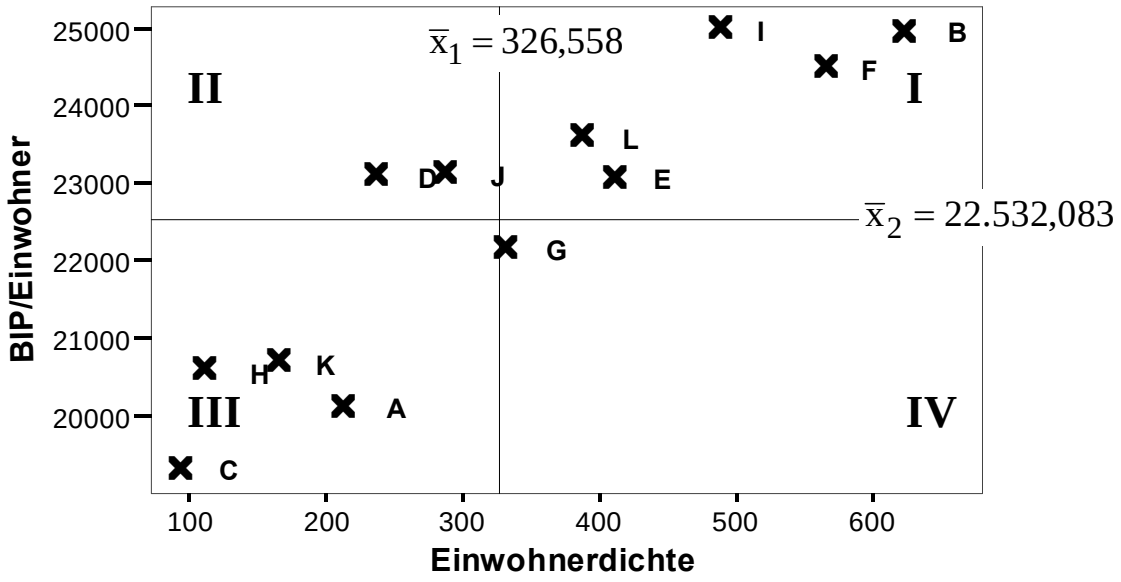
$$\bar{x}_2 = \frac{1}{n} \cdot \sum_{i=1}^{n} x_{i2} = \frac{1}{12} \cdot \sum_{i=1}^{12} x_{i2} = \frac{1}{12} \cdot 270.385 = 22.532{,}083 \ .$$

Die Kovarianz

$$s_{12} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_{i1} - \bar{x}_1\right) \cdot \left(x_{i2} - \bar{x}_2\right) = \frac{1}{11} \cdot 3.367.672{,}760 = 306.152{,}069$$

zeigt einen positiven Zusammenhang an. Mit steigenden Werten der Einwohnerdichte erhöht sich also tendenziell das BIP je Einwohner. Das Ergebnis wird durch das **Streudiagramm** bestätigt. 10 Regionen befinden sich im ersten oder dritten Quadranten. Die Regionen D und J liegen im zweiten Quadranten ($\bar{x}_1 = 326{,}558$), die Region G liegt im vierten Quadranten.

Streudiagramm zwischen ED und BIP



Wie stark der positive Zusammenhang ist, kann mit den **Korrelationskoeffizienten** herausgefunden werden. Er weist mit einem Wert von (n=12):

$$r_{12} = \frac{\sum_{i=1}^{n} \left(x_{i1} - \bar{x}_1\right) \cdot \left(x_{i2} - \bar{x}_2\right)}{\sqrt{\sum_{i=1}^{n} \left(x_{i1} - \bar{x}_1\right)^2 \cdot \sum_{i=1}^{n} \left(x_{i2} - \bar{x}_2\right)^2}} = \frac{3.367.672,760}{\sqrt{332.597,429 \cdot 41.428.758,918}} = 0,907$$

einen starken positiven Zusammenhang aus.

Für eine **Regressionsanalyse** müssen eine abhängige Variable (Regressand) und eine unabhängige Variable (Regressor) festgelegt werden. Nimmt man z.B. die Einwohnerdichte (X_1) als Indikator für Agglomerationen (Ballungen) mit gut ausgebildeter Bevölkerung, innovativen Unternehmen und einer guten Infrastruktur an, lässt sich hieraus ein Einfluss auf das BIP je Einwohner (X_2) begründen:

$$X_1 \rightarrow X_2$$

Mit den Kleinst-Quadrate-Schätzern (n=12)

$$\hat{b} = \frac{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2} = \frac{3.367.672,760}{332.597,429} = 10,125382$$

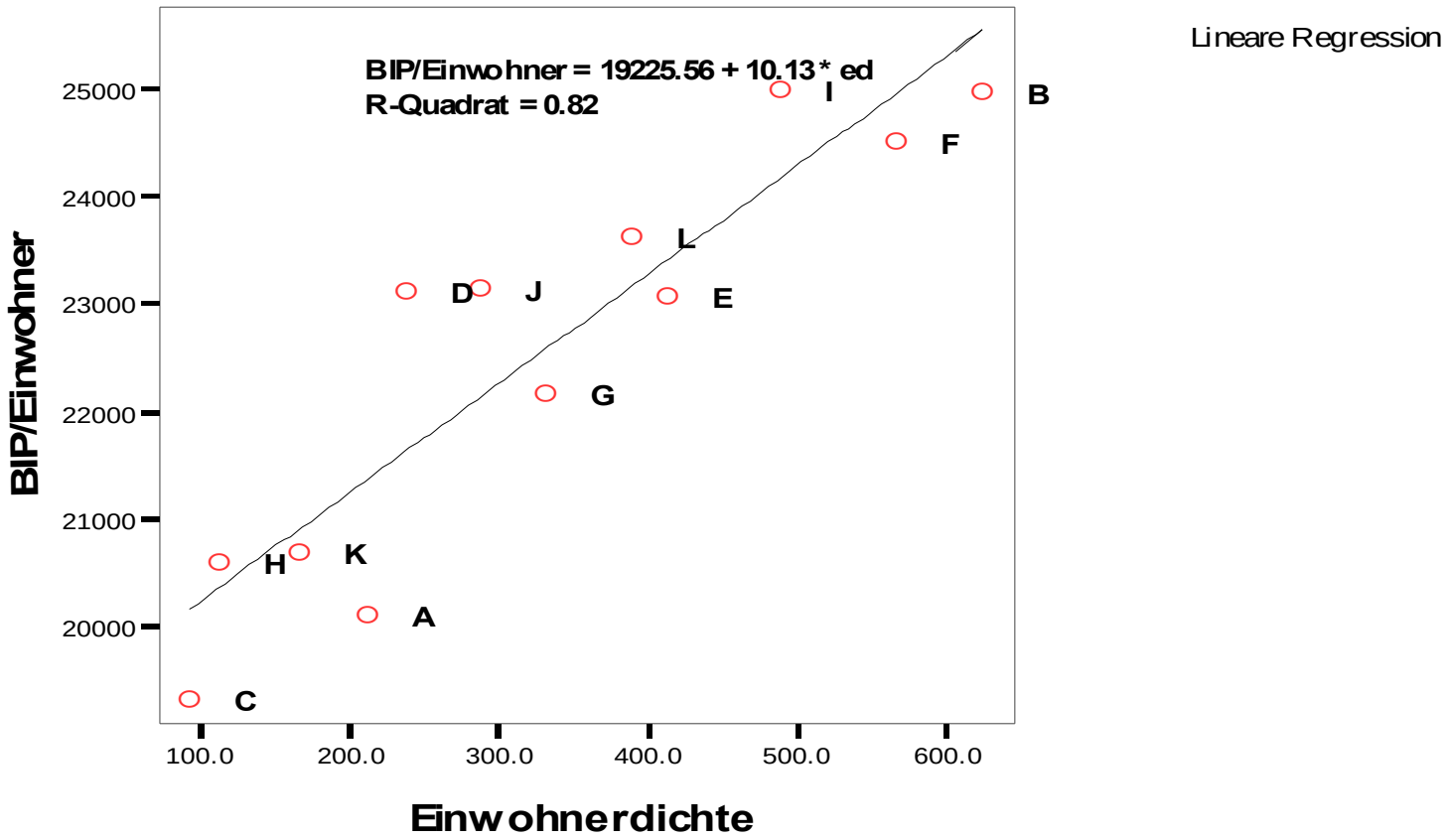
$$\hat{a} = \frac{\sum_{i=1}^{n} x_{i2}}{n} - \hat{b} \cdot \frac{\sum_{i=1}^{n} x_{i1}}{n} = \frac{270.385}{12} - 10{,}125372 \cdot \frac{3918{,}7}{12} = 19.225{,}55873$$

erhält man die Regressionsfunktion

$$\hat{y} = 19.225{,}559 + 10.125 \cdot x \ ,$$

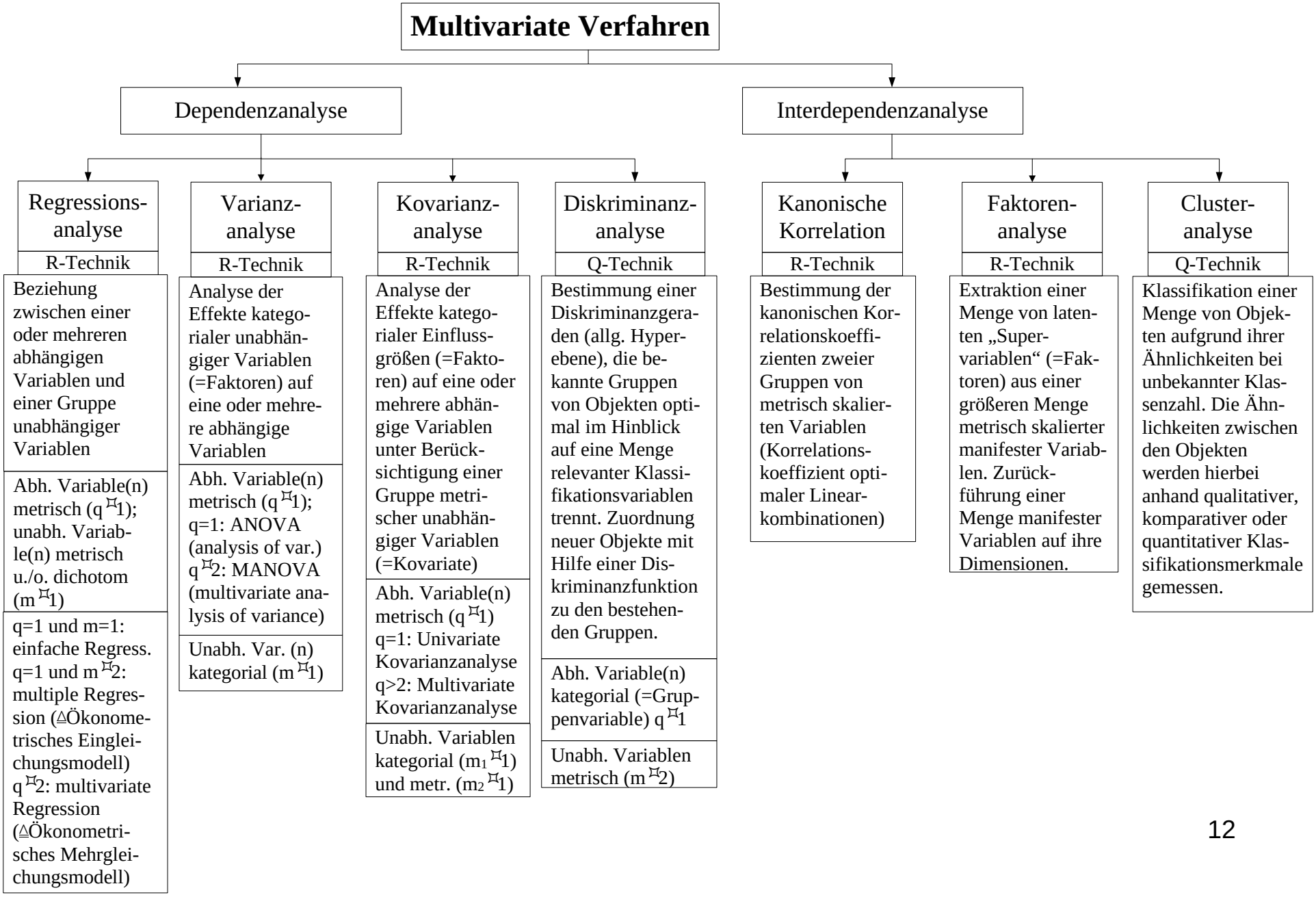
die den Einfluss der Einwohnerdichte (unabhängige Variable X_1) auf das BIP je Einwohner (abhängige Variable X_2) wiedergibt. Die Regressionsgerade ist im Stützbereich [(Einwohnerdichte X_1: 93,1 (Region C) bis 623,7 (Region B)] in einem Streudiagramm eingezeichnet.

Streudiagramm mit Regressionsgerade



Der **Determinationskoeffizient** (Bestimmtheitsmaß) ergibt sich aus dem Quadrat des Korrelationskoeffizienten:

$$R^2 = r_{12}^2 = 0{,}907^2 = 0{,}823$$

Danach erklärt die Einwohnerdichte (X_1) 82,3 % der Varianz des BIP pro Einwohner (X_2).

Systematisierung der multivariaten Verfahren