Statistik -3. Deskriptive Korrelations- und Regressionsanalyse

Dr. Wolfgang Böhme

Dezember 2010

Gliederung

- 1. Einführung
- 2. Deskriptive univariate Verteilungsanalyse
- **3. Deskriptive Korrelations- und Regressionsanalyse**
- 4. Deskriptive Zeitreihenanalyse
- 5. Verhältniszahlen / Indexzahlen
- 6. Einführung in ausgewählte multivariate Analysemethoden

- Statistiksoftware: SPSS

Problemstellung

mehrere metrischskalierte Merkmale X^1, X^2,...,X^K

Urlisten für jede der metrischskalierte Merkmale

$$x_1^k,\ x_2^k,\ ...x_n^k,\ \forall k = 1, 2, ...K$$

Ist zwischen den Merkmalen ein Zusammenhang statistisch nachweisbar? $\Rightarrow$ **Korrelationsanalyse**

Ableitung einer mathematischen Formel zur Beschreibung des Zusammenhanges, z. B.

$$X^1 = F(X^2, X^3, ..., X^K)$$

$\Rightarrow$ **Regressionsanalyse**

Einschränkung, zunächst K $= 2 \Rightarrow$ nur zwei Merkmale

2. Korrelations- und Regressionsanalyse - Gliederung

- **3.1 Korrelationsanalyse**
- 3.2 Einfache lineare Regression
- 3.3 Einfache nichtlineare Regression
- 3.4.Multiple Regression

Gliederung

- **3.1.1 zweidimensionale Häufigkeitsverteilung**
- 3.1.2 Kontingenzanalyse
- 3.1.3 bedingte Häufigkeiten / Unabhängigkeit
- 3.1.4 Maßkorrelationsanalyse
- 3.1.5 Rangkorrelationsanalyse
- 3.1.6 Multiple Korrelationsanalyse

zweidimensionale Häufigkeitsverteilung

Es seien X und Y zwei metrisch skalierte Merkmale mit den beiden Urlisten:

$$x_1,\ x_2,\ \ldots x_n$$

$$y_1,\ y_2,\ \ldots y_n$$

oder eine zweidimensionale Urliste:

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$$

zweidimensionale Häufigkeitsverteilung

Wir betrachten die disjunkten Merkmalsausprägungen (ξ_l *mit* $l = 1, 2, ..., m_1$ und ν_r *mit* $r = 1, 2, ..., m_2$) beider Merkmale und die zugehörenden absoluten / relativen Häufigkeiten n_{lr}/ p_{lr} . Dann können folgende Datenpaare gebildet werden:

$$\begin{aligned} (\xi_l, \nu_r) \quad & mit \quad l = 1, 2, ..., m_1 \ und \ r = 1, 2, ..., m_2 \\ (n_{lr}) \quad & mit \quad l = 1, 2, ..., m_1 \ und \ r = 1, 2, ..., m_2 \\ (p_{lr}) \quad & mit \quad l = 1, 2, ..., m_1 \ und \ r = 1, 2, ..., m_2 \end{aligned}$$

Dabei bezeichnen

$$\begin{aligned} n_{lr} \quad & - \quad Anzahl \ des \ Auftretens \ von \ (\xi_l \, \nu_r) \\ p_{lr} \quad & = \quad \frac{n_{lr}}{\sum_{l=1}^{m_1} \sum_{r=1}^{m_2} n_{lr}} = \frac{n_{lr}}{n} \end{aligned}$$

Beispiel Daten

Beispiel Kraftstoffverbrauch (in l/100 km) - X und Schadstoffausstoß (in g/km) - Y

Urlisten

i	X	Y	Typ
1	6.0	135	A
2	6.0	125	B
3	6.3	140	C
4	6.3	135	D
5	6.4	135	E
6	6.6	140	F
7	6.6	140	G
8	6.7	140	H
9	6.7	140	I
10	6.7	135	J

disjunkte Merkmale

l	ξ_l	r	v_r
1	6.0	1	125
2	6.3	2	135
3	6.4	3	140
4	6.6		
5	6.7		

Beispiel Eindimensionale Häufigkeitsverteilungen

Beispiel Kraftstoffverbrauch (in l/100 km) - X und Schadstoffausstoß (in g/km) - Y

Häufigkeitstabellen

l	ξ_l	n_l	p_l
1	6.0	2	0.2
2	6.3	2	0.2
3	6.4	1	0.1
4	6.6	2	0.2
5	6.7	3	0.3
		10	1

r	v_r	n_r	p_r
1	125	1	0.1
2	135	4	0.4
3	140	5	0.5
		10	1

zweidimensionale Häufigkeitsverteilung

gemeinsame zweidimensionale Häufigkeitsverteilung

Definition: Die absolute Häufigkeitsverteilung ist eine Zuordnung von disjunkten Merkmalsausprägungen zu deren absoluten Häufigkeiten:

$$f : (\xi_l, \nu_r) \rightarrow n_{lr} = f((\xi_l\, \nu_r))$$

Definition: Die relative Häufigkeitsverteilung ist eine Zuordnung von disjunkten Merkmalsausprägungen zu deren relativen Häufigkeiten:

$$f : (\xi_l, \nu_r) \rightarrow p_{lr} = f((\xi_l\, \nu_r))$$

Beispiel Zweidimensionale absolute Häufigkeitsverteilungen

Beispiel Kraftstoffverbrauch (in l/100 km) - X und Schadstoffausstoß (in g/km) - Y

mehrdimensionale Häufigkeitstabellen

		r	1	2	3
	X / Y		v_1	v_2	v_3
I			125	135	140
1	ξ_1	6.0	1	1	0
2	ξ_2	6.3	0	1	1
3	ξ_3	6.4	0	1	0
4	ξ_4	6.6	0	0	2
5	ξ_5	6.7	0	1	2

Beispiel Zweidimensionale relative Häufigkeitsverteilungen

Beispiel Kraftstoffverbrauch (in l/100 km) - X und Schadstoffausstoß (in g/km) - Y

mehrdimensionale Häufigkeitstabellen

		r	1	2	3
	X / Y		v_1	v_2	v_3
l			125	135	140
1	ξ_1	6.0	0.1	0.1	0
2	ξ_2	6.3	0	0.1	0.1
3	ξ_3	6.4	0	0.1	0
4	ξ_4	6.6	0	0	0.2
5	ξ_5	6.7	0	0.1	0.2

zweidimensionale Häufigkeitsverteilung

Randhäufigkeiten

Definition:

absolute Randhäufigkeiten (marginale absolute Häufigkeiten) des Merkmals X bzw. Y:

$$n_{l*} = \sum_{r=1}^{m_2} n_{lr}$$

$$n_{*r} = \sum_{l=1}^{m_1} n_{lr}$$

zweidimensionale Häufigkeitsverteilung

Definition:

relative (marginale) Randhäufigkeiten (marginale relative Häufigkeiten) des Merkmals X bzw. Y:

$$p_{l*} = \frac{n_{l*}}{n} = \frac{\sum_{r=1}^{m_2} n_{lr}}{n} = \sum_{r=1}^{m_2} p_{lr}$$

$$p_{*r} = \frac{n_{*r}}{n} = \frac{\sum_{l=1}^{m_1} n_{lr}}{n} = \sum_{l=1}^{m_1} p_{lr}$$

zweidimensionale Häufigkeitsverteilung

Randverteilungen

Definition: Die absolute Randverteilung ist eine Zuordnung von disjunkten Merkmalsausprägungen von Merkmal X bzw. Y zu deren absoluten Randhäufigkeiten:

$$f : (\xi_l) \rightarrow n_{l*} = f((\xi_l))$$
$$f : (\nu_r) \rightarrow n_{*r} = f((\nu_r))$$

Definition: Die relative Randverteilung ist eine Zuordnung von disjunkten Merkmalsausprägungen von Merkmal X bzw. Y zu deren relativen Randhäufigkeiten:

$$f : (\xi_l) \rightarrow p_{l*} = f((\xi_l))$$
$$f : (\nu_r) \rightarrow p_{*r} = f((\nu_r))$$

Beispiel absolute Randhäufigkeiten

Beispiel Kraftstoffverbrauch (in l/100 km) - X und Schadstoffausstoß (in g/km) - Y

Randverteilung 1

			n_{j*}
1	ξ_1	6	2
2	ξ_2	6.3	2
3	ξ_3	6.4	1
4	ξ_4	6.6	2
5	ξ_5	6.7	3

Randverteilung 2

	1	2	3
	v1	v2	v3
	125	135	140
n_{*r}	1	4	5

Gliederung

- 3.1.1 zweidimensionale Häufigkeitsverteilung
- **3.1.2 Kontingenzanalyse**
- 3.1.3 bedingte Häufigkeiten / Unabhängigkeit
- 3.1.4 Maßkorrelationsanalyse
- 3.1.5 Rangkorrelationsanalyse
- 3.1.6 Multiple Korrelationsanalyse

Kontingenztabelle

	r	1	2	...	m_2	$\sum_{r=1}^{m_2} n_{lr}$
l	X / Y	ν_1	ν_2	...	ν_{m_2}	
1	ξ_1	n_{11}	n_{12}	...	n_{1m_2}	n_{1*}
2	ξ_2	n_{21}	n_{22}	...	n_{2m_2}	n_{2*}
...	...	...	...	...	...	...
m_1	ξ_{m_1}	$n_{m_1 1}$	$n_{m_1 2}$	...	$n_{m_1 m_2}$	$n_{m_1 *}$
$\sum_{l=1}^{m_1} n_{lr}$		n_{*1}	n_{*2}	...	n_{*m_2}	n

Beispiel Kontingenztabelle absolut

Beispiel Kraftstoffverbrauch (in l/100 km) - X und Schadstoffausstoß (in g/km) - Y

Kontingenztabelle

		r	1	2	3	
	X / Y		v_1	v_2	v_3	n_{i*}
l			125	135	140	
1	ξ_1	6.0	1	1	0	2
2	ξ_2	6.3	0	1	1	2
3	ξ_3	6.4	0	1	0	1
4	ξ_4	6.6	0	0	2	2
5	ξ_5	6.7	0	1	2	3
n_{*r}			1	4	5	10

Kontingenztabelle Mit Intervallbildung

	r	1	2	...	m_2	$\sum_{r=1}^{m_2} n_{lr}$
l	X / Y	l_1	l_2	...	l_{m_2}	
1	l_1	n_{11}	n_{12}	...	n_{1m_2}	n_{1*}
2	l_2	n_{21}	n_{22}	...	n_{2m_2}	n_{2*}
...	...	...	...	...	...	...
m_1	l_{m_1}	$n_{m_1 1}$	$n_{m_1 2}$	...	$n_{m_1 m_2}$	$n_{m_1 *}$
$\sum_{l=1}^{m_1} n_{lr}$		n_{*1}	n_{*2}	...	n_{*m_2}	n

Beispiel

Beispiel Kraftstoffverbrauch (in l/100 km) - X und Schadstoffausstoß (in g/km) - Y
(mit Intervallbildung)

Urlisten

i	X	Y	Typ
1	6.0	135	A
2	6.0	125	B
3	6.3	140	C
4	6.3	135	D
5	6.4	135	E
6	6.6	140	F
7	6.6	140	G
8	6.7	140	H
9	6.7	140	I
10	6.7	135	J

Klassen

l	l_l	r	l_r
1	bis 6,00	1	bis 135
2	6,01bis 6,6	2	>135
3	> 6,60		

Beispiel

Beispiel Kraftstoffverbrauch (in l/100 km) - X und Schadstoffausstoß (in g/km) - Y
(mit Intervallbildung)

Häufigkeitstabellen

l	l_l	n_l	p_l
1	bis 6,00	2	0.2
2	6,01 bis 6,60	5	0.5
3	> 6,60	3	0.3
		10	1

r	l_r	n_r	p_r
1	bis 135	5	0.5
2	>135	5	0.5
		10	1

Beispiel

Beispiel Kraftstoffverbrauch (in l/100 km) - X und Schadstoffausstoß (in g/km) - Y
(mit Intervallbildung)

Kontingenztabelle

		r	1	2	
			l_1	l_2	n_{l*}
l	X / Y		bis 135	>135	
1	l_1	bis 6,00	2	0	2
2	l_2	6,01bis 6,6	2	3	5
3	l_3	> 6,60	1	2	3
n_{*r}			5	5	10

Gliederung

- 3.1.1 zweidimensionale Häufigkeitsverteilung
- 3.1.2 Kontingenzanalyse
- **3.1.3 bedingte Häufigkeiten / Unabhängigkeit**
- 3.1.4 Maßkorrelationsanalyse
- 3.1.5 Rangkorrelationsanalyse
- 3.1.6 Multiple Korrelationsanalyse

bedingte Häufigkeiten

bedingte Häufigkeiten: Betrachtung eines Merkmals bei Fixierung des anderen Merkmals auf einen Wert

Definition: bedingte Häufigkeiten des Merkmals X bei fixiertem Merkmal Y:

$$p_{l|r} = \frac{p_{lr}}{p_{*r}} = \frac{n_{lr}}{n_{*r}}$$

Definition: bedingte Häufigkeiten des Merkmals Y bei fixiertem Merkmal X:

$$p_{r|l} = \frac{p_{lr}}{p_{l*}} = \frac{n_{lr}}{n_{l*}}$$

Definition: Die **relative bedingte Verteilung** ist eine Zuordnung von disjunkten Merkmalsausprägungen von Merkmal X bzw. Y zu deren bedingten Häufigkeiten:

$$f : (\xi_l) \rightarrow p_{l|r} = f(\xi_l|\nu_r)$$

$$f : (\nu_r) \rightarrow p_{r|l} = f((\nu_r|\xi_l)$$

Beispiel Kontingenztabelle relativ

Beispiel Kraftstoffverbrauch (in l/100 km) - X und Schadstoffausstoß (in g/km) - Y

Kontingenztabelle (relative Häufigkeiten)

		r	1	2	3	
	X / Y		v_1	v_2	v_3	p_{l*}
l			125	135	140	
1	ξ_1	6.0	0.1	0.1	0	0.2
2	ξ_2	6.3	0	0.1	0.1	0.2
3	ξ_3	6.4	0	0.1	0	0.1
4	ξ_4	6.6	0	0	0.2	0.2
5	ξ_5	6.7	0	0.1	0.2	0.3
p_{*r}			0.1	0.4	0.5	

$$p_{l/r} = p_{1/1} = \frac{p_{l1}}{p*1} = \frac{0,1}{0,1} = 1$$

$$p_{r/l} = p_{1/1} = \frac{p_{l1}}{p1*} = \frac{0,1}{0,2} = 0,5$$

$$p_{l/r} = p_{1/2} = \frac{p_{l2}}{p*2} = \frac{0,1}{0,4} = 0,25$$

$$p_{r/l} = p_{1/2} = \frac{p_{l2}}{p1*} = \frac{0,1}{0,1} = 0,5$$

Beispiel bedingte Häufigkeiten

Beispiel Kraftstoffverbrauch (in l/100 km) - X und Schadstoffausstoß (in g/km) - Y

bedingte Häufigkeiten $p_{r,l}$

	1	2	3
	v_1	v_2	v_3
1	0.5	0.5	0
2	0	0.5	0.5
3	0	1	0
4	0	0	1
5	0	0.33333333	0.666667

bedingte Häufigkeiten $p_{l,r}$

	1	2	3
	v_1	v_2	v_3
1	1	0.25	0
2	0	0.25	0.2
3	0	0.25	0
4	0	0	0.4
5	0	0.25	0.4

Beispiel bedingte Häufigkeiten

Beispiel Kraftstoffverbrauch (in l/100 km) - X und Schadstoffausstoß (in g/km) - Y

bedingte Häufigkeiten $p_{r,l}$

	1	2	3
	v_1	v_2	v_3
1	0.5	0.5	0
2	0	0.5	0.5
3	0	1	0
4	0	0	1
5	0	0.33333333	0.6666667

bedingte Häufigkeiten $p_{l,r}$

	1	2	3
	v_1	v_2	v_3
1	1	0.25	0
2	0	0.25	0.2
3	0	0.25	0
4	0	0	0.4
5	0	0.25	0.4

Unabhängigkeit

Ausgangspunkt bedingte Häufigkeiten: Betrachtung eines Merkmals bei Fixierung des anderen Merkmals auf einen Wert

Definition: Die Merkmale X und Y heißen unabhängig, genau dann wenn gilt:

$$
\begin{aligned}
n_{lr} &= \frac{n_{l*}}{n} \cdot \frac{n_{*r}}{n} \ \textit{oder} \\
p_{lr} &= p_{l*} \cdot p_{*r} \ \textit{oder} \\
p_{l|r} &= p_{l*} \ \textit{oder} \\
p_{r|l} &= p_{*r}
\end{aligned}
$$

Beispiel

Beispiel Kraftstoffverbrauch (in l/100 km) - X und Schadstoffausstoß (in g/km) - Y
Unabhängigkeit

Kontingenztabelle (relative Häufigkeiten)

		r	1	2	3	
	X / Y		v_1	v_2	v_3	p_{l*}
l			125	135	140	
1	ξ_1	6.0	0.1	0.1	0	0.2
2	ξ_2	6.3	0	0.1	0.1	0.2
3	ξ_3	6.4	0	0.1	0	0.1
4	ξ_4	6.6	0	0	0.2	0.2
5	ξ_5	6.7	0	0.1	0.2	0.3
p_{*r}			0.1	0.4	0.5	

bedingte Häufigkeiten $p_{r/l}$

	1	2	3
	v_1	v_2	v_3
1	0.5	0.5	0
2	0	0.5	0.5
3	0	1	0
4	0	0	1
5	0	0.33333333	0.666667

bedingte Häufigkeiten $p_{l/r}$

	1	2	3
	v_1	v_2	v_3
1	1	0.25	0
2	0	0.25	0.2
3	0	0.25	0
4	0	0	0.4
5	0	0.25	0.4

Gliederung

- 3.1.1 zweidimensionale Häufigkeitsverteilung
- 3.1.2 Kontingenzanalyse
- 3.1.3 bedingte Häufigkeiten / Unabhängigkeit
- **3.1.4 Maßkorrelationsanalyse**
- 3.1.5 Rangkorrelationsanalyse
- 3.1.6 Multiple Korrelationsanalyse

empirische Kovarianz

Definition: Es seien X und Y zwei metrisch skalierte Merkmale. Die Ausprägungen dieser zwei Merkmale wurden an n statistischen Einheiten erhoben:

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

Dann ist die empirische Kovarianz d_{XY} wie folgt definiert:

$$d_{XY} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

empirische Kovarianz

$$d_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

wegen

$$\begin{aligned} d_{XY} &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{y}\frac{1}{n}\sum_{i=1}^{n} x_i - \overline{x}\frac{1}{n}\sum_{i=1}^{n} y_i + \overline{x}\cdot\overline{y} = \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\cdot\overline{y} = \\ &= \overline{xy} - \overline{x}\cdot\overline{y} \end{aligned}$$

Eigenschaften der empirische Kovarianz

- d_{XY} mißt die Stärke des linearen Zusammenhangs zwischen den Merkmalen X und Y.
- Es gilt bezüglich des Vorzeichens:
  - $d_{XY} < 0$ falls ein gegenläufiger Zusammenhand
  - $d_{XY} = 0$ falls kein Zusammenhang
  - $d_{XY} > 0$ falls gleichläufiger Zusammenhang
- d_{XY} ist dimensionsbehaftet

Eigenschaften der empirische Kovarianz - Fortsetzung

- symmetrische Maßzahl (Varianz-Kovarianzmatrix)

$$\begin{pmatrix} d_{XX} & d_{XY} \\ d_{YX} & d_{YY} \end{pmatrix}$$

mit

$$\begin{aligned} d_{XY} &= d_{YX} \\ d_{XX} &= d_X^2 \end{aligned}$$

- lineare Transformation. $\widetilde{x}_i = \alpha_1 + \beta_1 x_i$ und $\widetilde{y}_i = \alpha_2 + \beta_2 y_i$ mit $\alpha_1, \beta_1, \alpha_2, \beta_2 \in \Re$

$$d_{\widetilde{X}\widetilde{Y}} = \beta_1 \beta_2 d_{XY}$$

Korrelationskoeffizient

Definition: Es seien X und Y zwei metrisch skalierte Merkmale. Die Ausprägungen dieser zwei Merkmale wurden an n statistischen Einheiten erhoben:

$$(x_1, y_1),\ (x_2, y_2),\ \ldots, (x_n, y_n)$$

Dann ist der **empirische einfache lineare Korrelationskoeffizient** r_{XY} wie folgt definiert:

$$r_{XY} = \frac{d_{XY}}{\sqrt{d_X^2}\sqrt{d_Y^2}}$$

Synonyme:

- bivariater oder dualer Korrelationskoeffizient
- Maßkorrelationskoeffizient nach BRAVAIS und PEARSON

Eigenschaften des Korrelationskoeffizienten

- r_{XY} mißt die Stärke des linearen Zusammenhangs zwischen den Merkmalen X und Y . Wertebereich: $-1 \leq d_{XY} \leq 1$
- Es gilt:

$$
\begin{aligned}
r_{XY} = 1 &\quad \Rightarrow \quad X \textit{ und } Y \textit{ vollstaendig positiv korreliert} \\
0 \leq r_{XY} \leq 1 &\quad \Rightarrow \quad X \textit{ und } Y \textit{ positiv korreliert} \\
r_{XY} = 0 &\quad \Rightarrow \quad X \textit{ und } Y \textit{ unkorreliert} \\
-1 \leq r_{XY} \leq 0 &\quad \Rightarrow \quad X \textit{ und } Y \textit{ negativ korreliert} \\
r_{XY} = -1 &\quad \Rightarrow \quad X \textit{ und } Y \textit{ vollstaendig negativ korreliert}
\end{aligned}
$$

- d_{XY} ist dimensionslos
- normierte Maßzahl

Eigenschaften des Korrelationskoeffizienten - Fortsetzung

- symmetrische Maßzahl (Korrelationsmatrix)

$$\begin{pmatrix} r_{XX} & r_{XY} \\ r_{YX} & r_{YY} \end{pmatrix}$$

mit

$$\begin{aligned} r_{XY} &= r_{YX} \\ r_{XX} &= 1 \end{aligned}$$

- lineare Transformation. $\widetilde{x}_i = \alpha_1 + \beta_1 x_i$ und $\widetilde{y}_i = \alpha_2 + \beta_2 y_i$ mit $\alpha_1, \beta_1, \alpha_2, \beta_2 \in \Re$

$$\begin{aligned} r_{\widetilde{X}\widetilde{Y}} &= \frac{\beta_1 \cdot \beta_2}{|\beta_1| \cdot |\beta_2|} \cdot r_{XY} \Rightarrow \\ \left| r_{\widetilde{X}\widetilde{Y}} \right| &= |r_{XY}| \end{aligned}$$

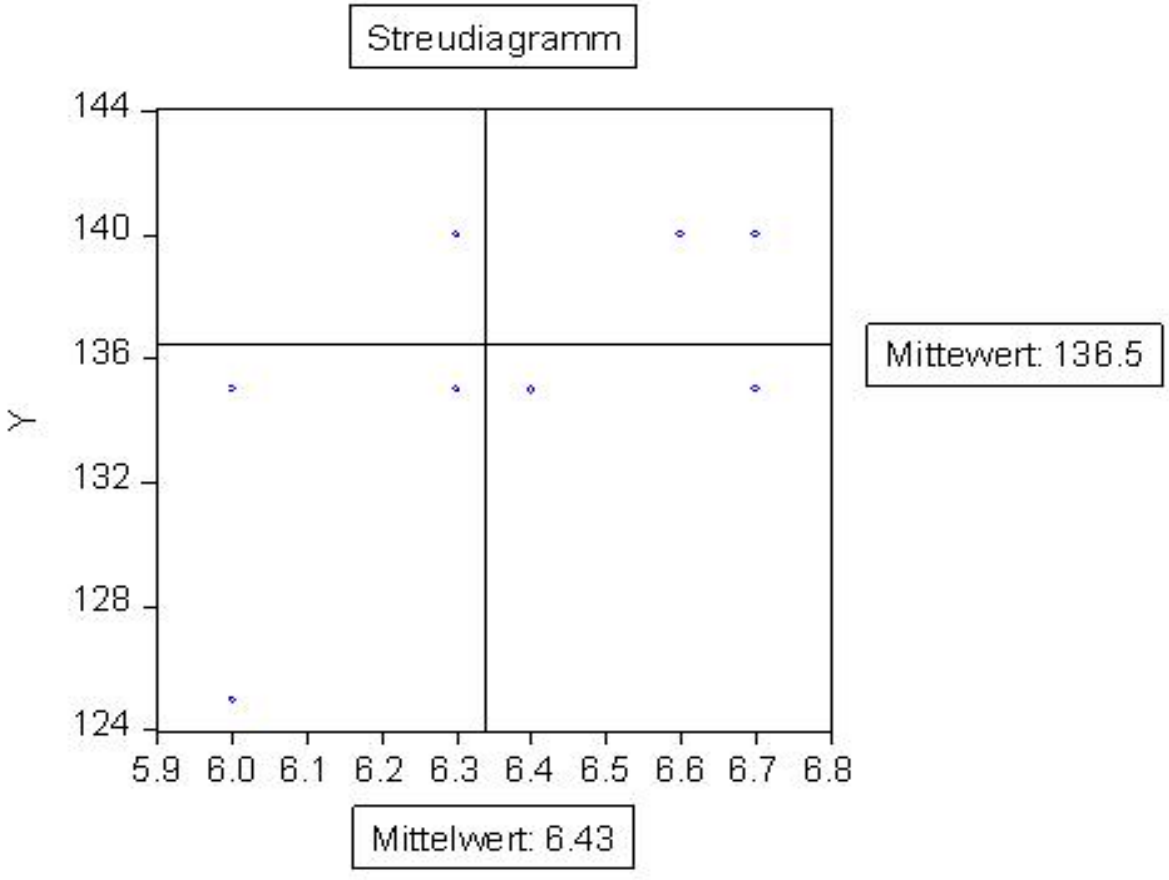
Beispiel Korrelation / empirische Kovarianz

i	X	Y	Typ	$x_i * y_i$	d_x	d_y
1	6.0	135	A	810	0.18	2.3
2	6.0	125	B	750	0.18	132.3
3	6.3	140	C	882	0.02	12.3
4	6.3	135	D	850.5	0.02	2.3
5	6.4	135	E	864	0.00	2.3
6	6.6	140	F	924	0.03	12.3
7	6.6	140	G	924	0.03	12.3
8	6.7	140	H	938	0.07	12.3
9	6.7	140	I	938	0.07	12.3
10	6.7	135	J	904.5	0.07	2.3
Mittelwerte	6.4	136.5		878.5	0.07	20.3

$$
\begin{aligned}
d_{XY} &= \overline{xy} - \overline{x} \cdot \overline{y} = 878.5 - 6.4 \cdot 136.5 = 0.805 \\
d_x^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 = 0.07 \\
d_y^2 &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2 = 20.3 \\
r_{XY} &= \frac{d_{XY}}{\sqrt{d_X^2}\sqrt{d_Y^2}} = \frac{0.805}{\sqrt{0.07}\sqrt{20.3}} = \frac{0,805}{1.174} = 0.686
\end{aligned}
$$

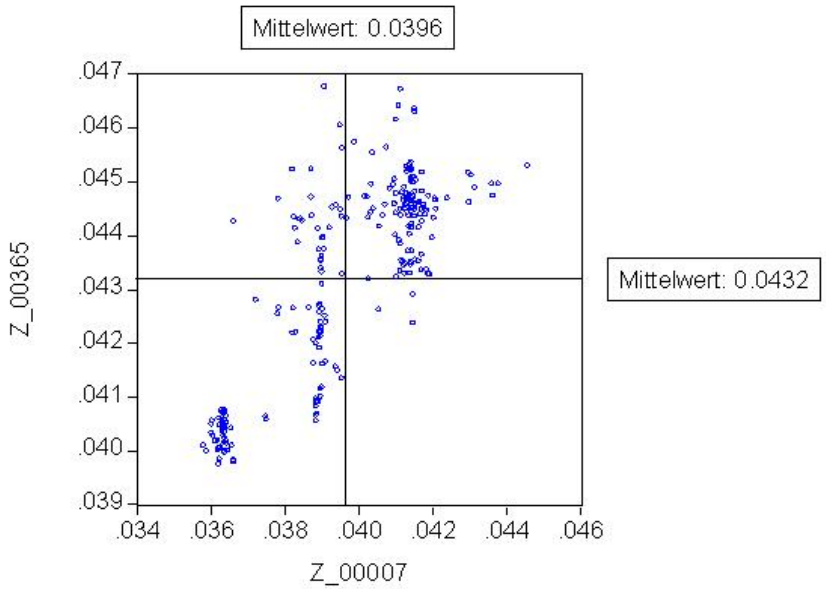
Korrelation / empirische Kovarianz - grafische Interpretation

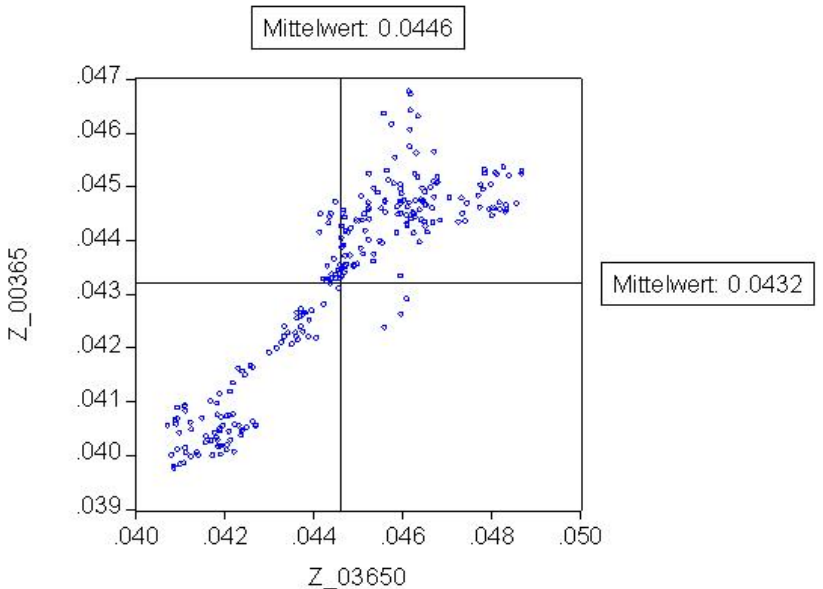
Beispiel Kraftstoffverbrauch (in l/100 km) - X und Schadstoffausstoß (in g/km) - Y



Korrelation / empirische Kovarianz - grafische Interpretation

Beispiel Zinssätze 2007





Korrelationsmatrix / Varianz-Kovarianz-Matrix

Beispiel Zinssätze 2007

Korrelationsmatrix

	Z_00007	Z_00365
Z_00007	1	0.815228902
Z_00365	0.815228902	1

Korrelationsmatrix

	Z_03650	Z_00365
Z_03650	1	0.905751
Z_00365	0.905751369	1

Varianz-Kovarianzmatrix

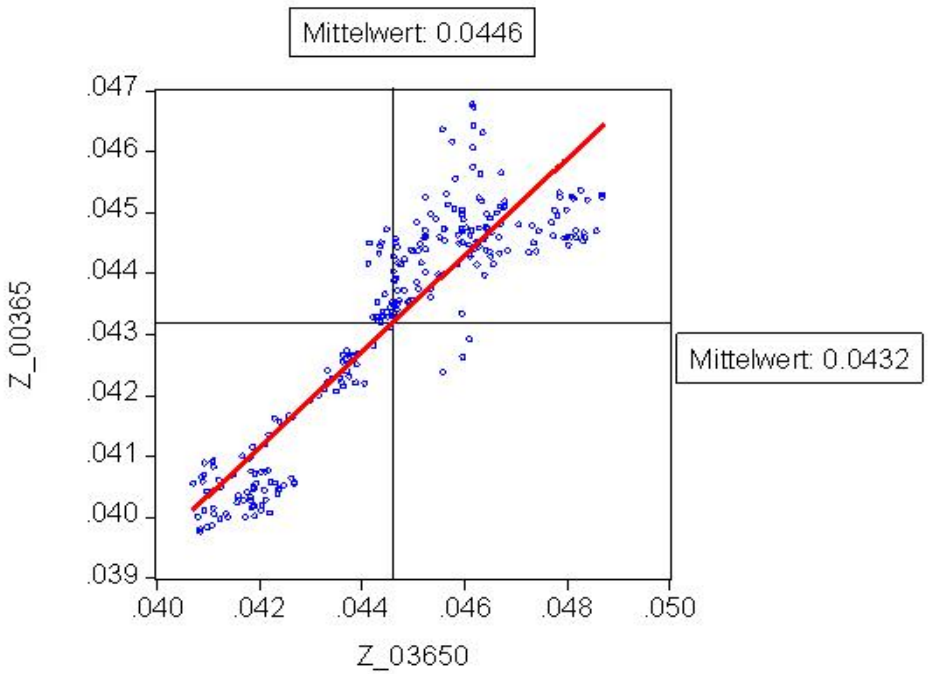
	Z_00007	Z_00365
Z_00007	0.0000043	0.0000032
Z_00365	0.0000032	0.0000035

Varianz-Kovarianzmatrix

	Z_03650	Z_00365
Z_03650	0.0000046	0.0000036
Z_00365	0.0000036	0.0000035

Korrelation / linearer Zusammenhang - grafische Interpretation

Beispiel Zinssätze 2007



$\Longrightarrow$ siehe 3.2 Einfache lineare Regression

Einführung in SPSS - Themen

Statistische Themen: Deskriptive Statistik

- Häufigkeiten - Fortsetzung
- Deskriptive Statistik -Fortsetzung
- Bivariate Korrelationen

Maßkorrelation - Beispiel

mit Excel:

Beispiel 1: BeispielW95.xls
Beispiel 2: Zinsätze 2007_2.xls

Aufgaben: Bestimmen Kovarianzen, Korrelationskoeffizienten Und der Kovarianzmatrix

mit SPSS:

Beispiel 1: World95.sav.xls
Beispiel 2: Zinsätze 2007_2.xls

Aufgaben: Bestimmen der Korrelationskoeffizienten

Gliederung

- 3.1.1 zweidimensionale Häufigkeitsverteilung
- 3.1.2 Kontingenzanalyse
- 3.1.3 bedingte Häufigkeiten / Unabhängigkeit
- 3.1.4 Maßkorrelationsanalyse
- **3.1.5 Rangkorrelationsanalyse**
- 3.1.6 Multiple Korrelationsanalyse

Rangkorrelation

Definition: Es sei X ein mindestens ordinal skaliertes Merkmal. Für das Merkmal liegt die Rangwertreihe des ordinalen Datensatzes vor:

$$x_1 \leq x_2 \leq \ ..., \leq x_n$$

Dann heißt die der Merkmalsausprägung x_i zugewiesene Platznummer i Rangzahl und wird mit R_i^X bezeichnet.

Problem: Auftreten von gleichen Merkmalsausprägungen (wird als **Rangbindung** bezeichnet).

SPSS: arithm.Mittel, max., min., fortlaufende Numerierung

Hier:

arithmetisches Mittel $\Rightarrow \frac{n\cdot(n+1)}{2} = \sum_{i=1}^{n} R_i^X$ oder

fortlaufende Numerierung $\Rightarrow \frac{n\cdot(n+1)}{2} = \sum_{i=1}^{n} R_i^X$

Rangkorrelation

Definition: Es seien X und Y zwei mindestens ordinal skalierte Merkmale. Die Ausprägungen dieser zwei Merkmale wurden an n statistischen Einheiten erhoben:

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,(x_n, y_n)$$

Jedem Wertepaar (x_i, y_i) entspricht ein Paar der Rangzahlen (R_i^X, R_i^Y). Dann heißt die Größe:

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} (R_i^X - R_i^Y)^2}{n(n^2 - 1)}$$

Rangkorrelationskoeffizient nach SPEARMAN.

Rangkorrelation Interpretation

- r_S mißt die Stärke des linearen Zusammenhangs zwischen den Merkmalen X und Y . Wertebereich: $-1 \leq r_S \leq 1$
- Es gilt bezüglich des Vorzeichens:
  - $r_S < 0$ falls ein gegenläufiger Zusammenhand
  - $r_S = 0$ falls kein Zusammenhang
  - $r_S > 0$ falls gleichläufiger Zusammenhang
- Liegt r_S nahe Null, ist kein Zusammenhang zu vermuten, liegt r_S dagegen nahe 1 oder -1, ligt ein gleich oder gegenläufiger Zusammenhang vor.

Rangkorrelation - Beispiel

Beispiel Benzinverbrauch / Schadstoffausstoss 2007

i	X	Y	R^X	R^Y	$(R^X-R^Y)^2$
1	6,0	135	1	2	1
2	6,0	125	2	1	1
3	6,3	140	3	6	9
4	6,3	135	4	3	1
5	6,4	135	5	4	1
6	6,6	140	6	7	1
7	6,6	140	7	8	1
8	6,7	140	8	9	1
9	6,7	140	9	10	1
10	6,7	135	10	5	25
					42

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} (R_i^X - R_i^Y)^2}{n(n^2-1)} = 1 - \frac{6 \cdot 42}{10(100-1)} = 0.745$$

SPSS Themen

Statistische Themen: Deskriptive Statistik

- Bivariate Korrelationen - Fortsetzung

Rangkorrelation - Beispiel

mit Excel:

Beispiel 1: BeispielW95.xls
Beispiel 2: Zinsätze 2007_2.xls

Aufgaben: Bestimmen der Rangkorrelationskoeffizienten

mit SPSS:

Beispiel 1: World95.sav.xls
Beispiel 2: Zinsätze 2007_2.xls

Aufgaben: Bestimmen der Rangkorrelationskoeffizienten

Gliederung

- 3.1.1 zweidimensionale Häufigkeitsverteilung
- 3.1.2 Kontingenzanalyse
- 3.1.3 bedingte Häufigkeiten / Unabhängigkeit
- 3.1.4 Maßkorrelationsanalyse
- 3.1.5 Rangkorrelationsanalyse
- **3.1.6 Multiple Korrelationsanalyse**

Problemstellung

Bisher: X und Y zwei metrisch skalierte Merkmale mit den beiden Urlisten:

$$x_1,\ x_2,\ \ldots x_n$$
$$y_1,\ y_2,\ \ldots y_n$$

oder eine zweidimensionale Urliste:

$$(x_1, y_1),\ (x_2, y_2),\ \ldots, (x_n, y_n)$$

Jetzt:: X_j mit $j = 1, 2, \ldots, m$ - metrisch skalierte Merkmale mit m Urlisten:

$$x_{11},\ x_{12},\ \ldots, x_{1m}$$
$$x_{21},\ x_{22},\ \ldots, x_{2m}$$
$$\ldots$$
$$x_{n1},\ x_{n2},\ \ldots, x_{nm}$$

oder eine m-dimensionale Urliste:

$$(x_{11}, x_{12}, \ldots, x_{1m}),\ (x_{21}, x_{22}, \ldots, x_{2m}),\ \ldots, (x_{n1}, x_{n2}, \ldots, x_{nm})$$

empirische Kovarianzen

Definition: X_j mit $j = 1, 2, ..., m$ - metrisch skalierte Merkmale. Die Ausprägungen dieser Merkmale wurden an n statistischen Einheiten erhoben:

$$(x_{11}, x_{12}, ..., x_{1m}),\ (x_{21}, x_{22}, ..., x_{2m}),\ ...,(x_{n1}, x_{n2}, ..., x_{nm})$$

Dann ist die empirische Kovarianz d_{jk} wie folgt definiert:

$$\begin{aligned} d_{jk} &= \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})(x_{ik} - \overline{x_k}) = \\ &= \overline{x_j x_k} - \overline{x_j} \cdot \overline{x_k} \end{aligned}$$

Kovarianzmatrix:

$$\Sigma = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ d_{m1} & d_{m2} & \cdots & d_{mm} \end{pmatrix}$$

empirische Korrelationskoeffizienten

Definition: X_j mit $j = 1, 2, ..., m$ - metrisch skalierte Merkmale. Die Ausprägungen dieser Merkmale wurden an n statistischen Einheiten erhoben:

$$(x_{11}, x_{12}, ..., x_{1m}),\ (x_{21}, x_{22}, ..., x_{2m}),\ ...,(x_{n1}, x_{n2}, ..., x_{nm})$$

Dann ist der empirische Korrelationskoeffizient r_{jk} wie folgt definiert:

$$r_{jk} = \frac{d_{jk}}{\sqrt{\overline{d_{jj}^2}}\sqrt{\overline{d_{kk}^2}}}$$

Korrelationsmatrixmatrix:

$$P = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{pmatrix}$$

Einführung in SPSS - Themen

Statistische Themen: Deskriptive Statistik

- Bivariate Korrelationen - Fortsetzung

multiple Korrelation - Beispiel

mit Excel:

Beispiel 2: Zinsätze 2007_2.xls

Aufgaben: Bestimmen der Korrealtionsmatrix und der Kovarianzmatrix

mit SPSS:

Beispiel 2: Zinsätze 2007_2.xls

Aufgaben: Bestimmen der Korrelationskoeffizienten

2. Korrelations- und Regressionsanalyse - Gliederung

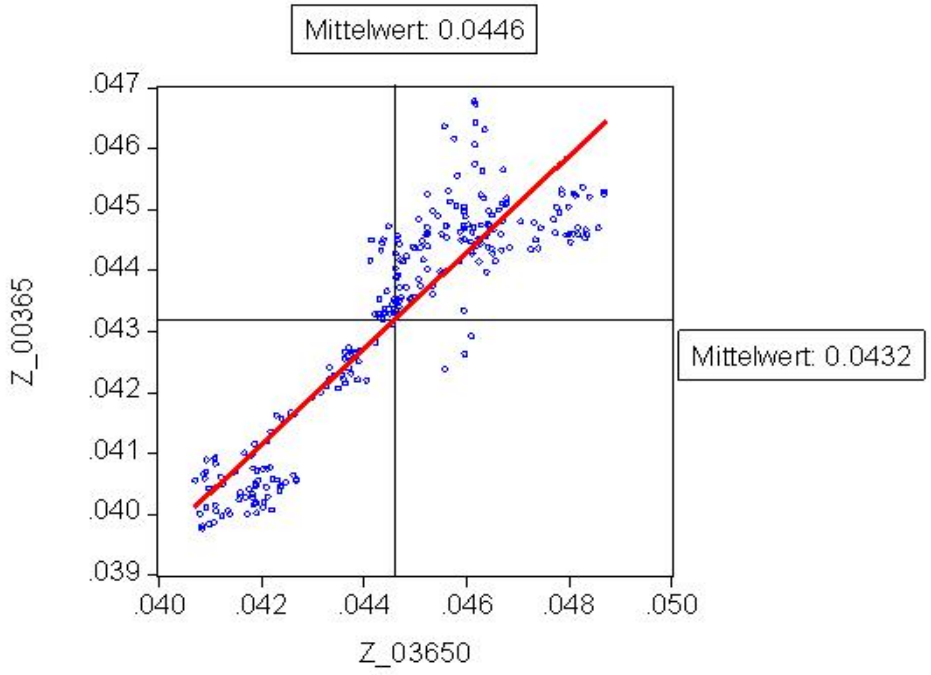
- 3.1 Korrelationsanalyse
- 3.2 **Einfache lineare Regression**
- 3.3 Einfache nichtlineare Regression
- 3.4 Multiple Regression

Gliederung

- **3.2.1 Problemstellung und Herleitung**
- 3.2.2 Interpretation und Beispiel
- 3.2.3 Elastizitäten
- 3.2.4 Gütemaße

Korrelation / linearer Zusammenhang - grafische Interpretation

Beispiel Zinssätze 2007



$\Longrightarrow$ siehe 3.2 Einfache lineare Regression

Problemstellung

Definition: Es seien X und Y zwei metrisch skalierte Merkmale. Die Ausprägungen dieser zwei Merkmale wurden an n statistischen Einheiten erhoben:

$$(x_1, y_1),\ (x_2, y_2),\ ...,(x_n, y_n)$$

Eine lineare Funktion

$$y^* = f(x) = b_0 + b_1 \cdot x$$

zur Beschreibung der statistischen Abhängigkeit des Merkmals Y vom Merkmal X heißt **einfache lineare Regression** von Y auf X.

Problemstellung

$$y^* = f(x) = b_0 + b_1 \cdot x$$

Bezeichnungen:

X - Regressor, unabhängige oder exogene Variable
Y - Regressand, abhängige oder endogene Variable
x - Realisation des Merkmals X ($x \in \Re$)
y - Realisation des Merkmals Y ($x \in \Re$)
$y^* = f(x) = b_0 + b_1 \cdot x$ Regresswert (Realisation von y Bei vorgegebenen x)
b_0 - Regressionskonstante (intercept)
b_1 - linearer Regressionskoeffizient (slop)
$u = y - y^*$ Fehler (Residuum)

Problemstellung

lineare Regressionsfunktion: $y^* = f(x) = b_0 + b_1 \cdot x$ mit $(x \in \Re)$

x - beobachteter Wert des Merkmals X
$y^* = y^*(x)$ berechneter (geschätzter) Wert des Merkmals Y
y - beobachteter Wert des Merkmals Y

$$\begin{aligned} u_i &= y_i - y_i^* = y_i - y_i^*(x_i) \ \textit{Residuum (Fehler)} \\ y_i &= y_i^* + u_i = b_0 + b_1 \cdot x_i + u_i \end{aligned}$$

Forderung:

$$\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - y_i^*)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 \cdot x_i))^2 = MIN$$

$\Rightarrow$ **Methode der kleinsten Quadrate**

Herleitung

Herleitung der **Methode der kleinsten Quadrate**

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 \cdot x_i))^2 = MIN$$

Extremwertproblem einer Funktion mit zwei Variablen (b_0, b_1)

notwendige Bedingung:

$$\begin{aligned} \frac{\partial S(b_0, b_1)}{\partial b_0} &= 0 \\ \frac{\partial S(b_0, b_1)}{\partial b_1} &= 0 \end{aligned}$$

Herleitung - Fortsetzung

notwendige Bedingung:

$$
\begin{aligned}
\frac{\partial S(b_0, b_1)}{\partial b_0} &= -2 \sum_{i=1}^{n} (y_i - (b_0 + b_1 \cdot x_i)) = 0 \\
&= \sum_{i=1}^{n} y_i - b_0 \sum_{i=1}^{n} 1 - b_1 \sum_{i=1}^{n} x_i = 0 \\
&= \sum_{i=1}^{n} y_i - b_0 n - b_1 \sum_{i=1}^{n} x_i = 0 \\
\frac{\partial S(b_0, b_1)}{\partial b_1} &= -2 \sum_{i=1}^{n} (y_i - (b_0 + b_1 \cdot x_i)) \cdot x_i = 0 \\
&= \sum_{i=1}^{n} y_i x_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0
\end{aligned}
$$

Herleitung - Fortsetzung

Gleichungssystem:

$$
\begin{aligned}
b_0 n + b_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\
b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i
\end{aligned}
$$

$$
\begin{aligned}
b_0 A + b_1 B &= C \\
b_0 D + b_1 E &= F
\end{aligned}
$$

Lösung des Gleichungssystems

$$b_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \overline{y} - b_1 \overline{x}$$

$$b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$= \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{\frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{d_{YX}}{d_X^2}$$

Herleitung - Fortsetzung

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 \cdot x_i))^2 = MIN$$

Extremwertproblem einer Funktion mit zwei Variablen (b_0, b_1)

hinreichende Bedingung:

$$\frac{\partial^2 S(b_0, b_1)}{\partial b_0^2} = 2 \sum_{i=1}^{n} 1 = 2n > 0$$

$$\frac{\partial^2 S(b_0, b_1)}{\partial b_1^2} = 2 \sum_{i=1}^{n} x_i^2 > 0$$

Herleitung - Fortsetzung

Ableitungen:

$$\frac{\partial^2 S(b_0, b_1)}{\partial b_0^2} = 2\sum_{i=1}^{n} 1 = 2n > 0$$

$$\frac{\partial^2 S(b_0, b_1)}{\partial b_1^2} = 2\sum_{i=1}^{n} x_i^2 > 0$$

$$\frac{\partial^2 S(b_0, b_1)}{\partial b_0 b_1} = 2\sum_{i=1}^{n} x_i$$

$$\frac{\partial^2 S(b_0, b_1)}{\partial b_0 b_1} = 2\sum_{i=1}^{n} x_i$$

$$det \begin{pmatrix} \frac{\partial^2 S(b_0,b_1)}{\partial b_0^2} & \frac{\partial^2 S(b_0,b_1)}{\partial b_0 b_1} \\ \frac{\partial^2 S(b_0,b_1)}{\partial b_1 b_0} & \frac{\partial^2 S(b_0,b_1)}{\partial b_1^2} \end{pmatrix} = 4n\sum_{i=1}^{n} x_i^2 - 4(\sum_{i=1}^{n} x_i)^2 > 0$$

$\Longrightarrow$ hinreichende Bedingung ist erfüllt.

Herleitung Matrizenschreibweise

Die Gleichung $y = f(x) = b_0 + b_1 \cdot x + u$ gilt für allle Beobachtungen der Urliste:

$$\begin{array}{lcllcl} y_1 & = & b_0 & + & b_1 x_1 & + & u_1 \\ y_2 & = & b_0 & + & b_1 x_2 & + & u_2 \\ \cdots & & & & & & \\ y_n & = & b_0 & + & b_1 x_n & + & u_n \end{array}$$

oder in Matrizenschreibweise

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_n \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \cdots \\ u_n \end{pmatrix}$$

oder kompakt: $\mathbf{y} = \mathbf{X} \cdot \mathbf{b} + \mathbf{u}$

Operationen mit Matrizen

- $(A \cdot B)^T = B^T \cdot A^T$
- $\frac{\partial (A \cdot B)}{\partial A} = B$
- $\frac{\partial (A^T \cdot B)}{\partial A} = B$
- $\frac{\partial (A(x) \cdot B(x))}{\partial x} = \frac{\partial A(x)}{\partial x} \cdot B(x) + \frac{\partial B(x)}{\partial x} \cdot A(x)$

Herleitung

Ausgangspunkt:

$y = X \cdot b + u$
Regeln: $(A + B)^T = A^T + B^T$ und $(AB)^T = B^T A^T$

Extremwertproblem:

$$\begin{aligned} S(b) &= u^T u = (y - Xb)^T (y - Xb) = (y^T - (Xb)^T)(y - Xb) \rightarrow MIN \\ &= y^T y - (XB)^T y - y^T Xb + (XB)^T Xb = y^T y - 2b^T X^T y + (Xb)^T Xb \end{aligned}$$

$$\begin{aligned} \frac{\partial S(b)}{\partial b} &= \frac{\partial y^T y}{\partial b} - 2\frac{\partial b^T X^T y}{\partial b} + \frac{\partial (Xb)^T Xb}{\partial b} = \\ &= 0 - 2X^T y + 2X^T Xb = 0 \end{aligned}$$

Normalgleichungssystem:

$X^T Xb = X^T y \implies$

Lösung:

$b = (X^T X)^{-1} X^T y$

Gliederung

- 3.2.1 Problemstellung und Herleitung
- **3.2.2 Interpretation und Beispiel**
- 3.2.3 Elastizitäten
- 3.2.4 Gütemaße

Realistisches Beispiel

In einer Studie des ADAC wurden für 269 Personkraftwagen unterschiedlicher Fabrikate und Modelle Daten zum Kraftstoffverbrauch (in Liter je 100 km) und zum Kohlendioxidausstoß (in Gramm pro km) erhoben (ADACmotorwelt , 5/1998).
Aus dieser Statistik ist Ihnen nachfolgend ein Auszug aus der gesamten Urliste gegeben (10 Kleinwagen mit einem Hubraum von höchstens 1300 ccm, die den Kraftstoff Super oder Normal verbrauchen):

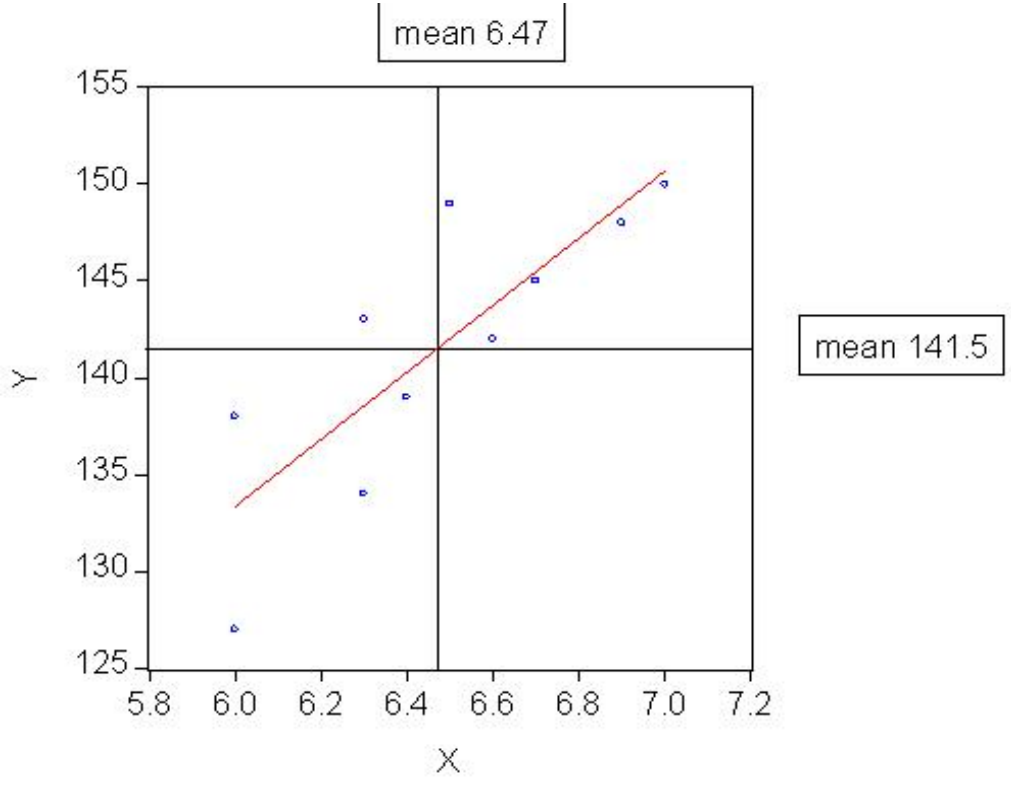
PKW-Nummer	Kraftstoffverbrauch l / 100 km	Kohlendioxidausstoß g / km	Fabrikat / Modell
Sp. 1	Sp. 2	Sp. 3	Sp. 4
1	6,0	138	Opel Corsa
2	6,0	127	Daihatsu Cuore
3	6,3	143	Renault Twingo
4	6,3	134	Suzuki Alto
5	6,4	139	Seat Arosa
6	6,5	149	Peugot 106
7	6,6	142	VW Polo
8	6,7	145	Fiat Cinquecento 0.9 i.e
9	6,9	148	Fiat Cinquecento Sporting
10	7,0	150	Lancia Y

Für die zwei kardinalskalierten Erhebungsmerkmale (zweidimensionales Merkmal)

X : Kraftstoffverbrauch (in l / 100 km
Y : Kohlendioxidausstoß (in g / km)

grafische Interpretation

realistisches Beispiel Kraftstoffverbrauch (in l/100 km) - X und Schadstoffausstoß (in g/km) - Y



$\Longrightarrow$ siehe 3.2 Einfache lineare Regression

i	X	Y	X*X	X*Y
1	6.0	138	36.0	828
2	6.0	127	36.0	762
3	6.3	143	39.7	900.9
4	6.3	134	39.7	844.2
5	6.4	139	41.0	889.6
6	6.5	149	42.3	968.5
7	6.6	142	43.6	937.2
8	6.7	145	44.9	971.5
9	6.9	148	47.6	1021.2
10	7.0	150	49.0	1050
Summe	64.7	1415.0	419.7	9173.1

Berechnungen

$$b_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} =$$

$$= \frac{1415.0 \cdot 419.7 - 64.7 \cdot 9173.1}{10 \cdot 419.7 - 4186.09} = \frac{305.18}{10.41} = 29.31$$

$$b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} =$$

$$= \frac{10 \cdot 9173.1 - 64.7 \cdot 1415.0}{10 \cdot 419.7 - 4186.09} = \frac{180.5}{10.41} = 17.34$$

Interpretation

Maßeinheiten: X -Kraftstoffverbrauch $[\frac{l}{100km}]$ und Y Schadstoffausstoß $[\frac{g}{km}]$

$y^* = f(x) = b_0 + b_1 \cdot x = 29.31 + 17.34 \cdot x$

$$\begin{array}{ccccccc} y^* & = & b_0 & + & b_1 & & x \\ [\frac{g}{km}] & = & [\frac{g}{km}] & + & [\frac{l00g}{l}] & & [\frac{l}{100km}] \end{array}$$

$y = f(x) = b_0 + b_1 \cdot x + u = 29.31 + 17.34 \cdot x + u$

$$\begin{array}{ccccccccc} y & = & b_0 & + & b_1 & & x & + & u \\ [\frac{g}{km}] & = & [\frac{g}{km}] & + & [\frac{l00g}{l}] & & [\frac{l}{100km}] & + & [\frac{g}{km}] \end{array}$$

Interpretation

Maßeinheiten: X -Kraftstoffverbrauch $[\frac{l}{100km}]$ und Y Schadstoffausstoß $[\frac{g}{km}]$

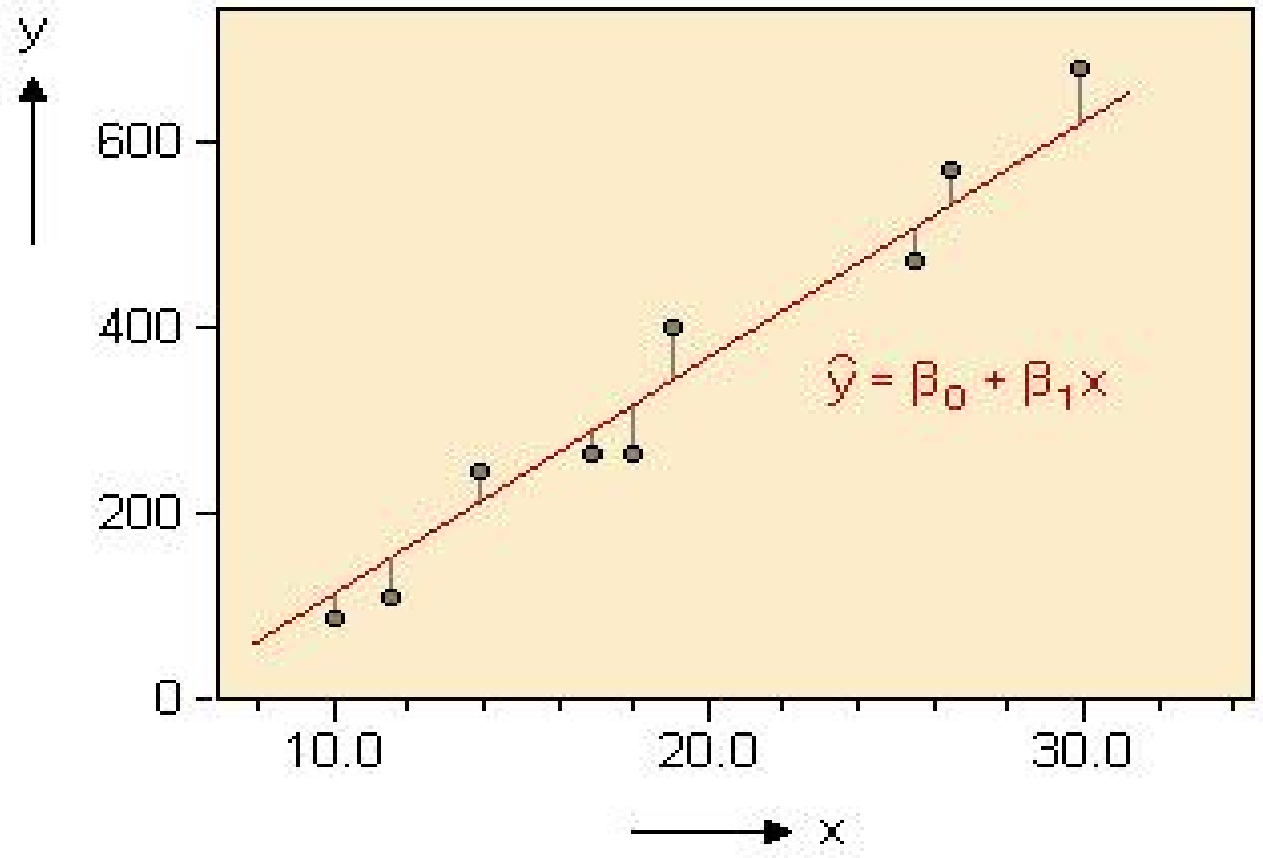
$y^* = f(x) = b_0 + b_1 \cdot x = 29.31 + 17.34 \cdot x$

$$\begin{array}{ccccccc} y^* & = & b_0 & + & b_1 & & x \\ [\frac{g}{km}] & = & [\frac{g}{km}] & + & [\frac{l00g}{l}] & & [\frac{l}{100km}] \end{array}$$

Interpretation der Koeffizienten b_0 und b_1

- b_0 - wenn x = 0 (kein Kraftstoffverbrauch !) dann wird trotzdem Schadstoff in der Höhe von 29,31 $[\frac{g}{km}]$ ausgestoßen. $\Longrightarrow$ Kraftstoffverbrauchsunabhängiger Schadstoffausstoß.
- b_1 - wegen $\frac{\partial(b_0+b_1x)}{\partial x} = b_1 \Longrightarrow$
  bei einem Kraftstoffverbrauch von $1\frac{l}{100km}$ entsteht ein Schadstoffausstoß von $17.34\frac{g}{km}$ oder
  bei einer Erhöhung /Verringerung des Kraftstoffverbrauchs um $1\frac{l}{100km}$ steigt / fällt der Schadstoffausstoß um $17.34\frac{g}{km}$.

grafische Interpretation



Eigenschaften lineare Regression

1. $y^*(\overline{x}) = \overline{y}$
2. $\overline{y^*} = \frac{1}{n}\sum_{i=1}^{n} y_i^* = \overline{y}$
3. $d^2_{\overline{y^*}} = \frac{1}{n}\sum_{i=1}^{n}(y_i^* - \overline{y^*}) = \frac{1}{n}\sum_{i=1}^{n}(y_i^* - \overline{y}) = b_1^2 \cdot d_x^2$
4. $\sum_{i=1}^{n} u_i = 0$
5. Zusammenhang zwischen b_1 und r_{xy}: $b_1 \cdot \frac{d_x}{d_y} = r_{xy}$

Gliederung

- 3.2.1 Problemstellung und Herleitung
- 3.2.2 Interpretation und Beispiel
- **3.2.3 Elastizitäten**
- 3.2.4 Gütemaße

Grenzfunktion / absolute Veränderung

Definition: Es sei $y^*(x)$ eine stetige, bivariate Regressionsfunktion. Dann heißt die erste Ableitung $y'^*(x)$ **Grenzfunktion der Regressionsfunktion**:

$$y'*(x) = \frac{dy^*(x)}{dx}$$

Der Wert an der Stelle x_0 heißt **durchschnittlich absolute Veränderung der exogenen Variablen y** oder **durchschnittliche absolute Elastizität**

$$y'*(x_0) = \frac{dy^*(x)}{dx} \|_{x=x_0}$$

Wenn die bivariate Regressionsfunktion linear ist, d. h. $y^* = f(x) = b_0 + b_1 \cdot x$, dann gilt:

$$y'*(x_0) = \frac{dy^*(x)}{dx} \|_{x=x_0} = b_1 \ \forall x = x_0$$

Elastizität / Elastizitätsfunktion

Definition: Es sei $y^*(x)$ eine stetige, bivariate Regressionsfunktion. Dann heißt die Funktion $\epsilon(x)$ **Elastizitätsfunktion der Regressionsfunktion**:

$$\epsilon(x) = \frac{x}{y^*(x)} \cdot \frac{dy^*(x)}{dx} = x \cdot \frac{y'^*(x)}{y^*(x)}$$

Der Wert an der Stelle x_0 heißt **Elastizität der exogenen Variablen Y** oder **durchschnittlich relative Veränderung** oder **durchschnittliche relative Elastizität**

$$\epsilon(x) = \epsilon(x_0) = \frac{x}{y^*(x)} \cdot \frac{dy^*(x)}{dx} \|_{x=x_0} = x_0 \cdot \frac{y'^*(x_0)}{y^*(x_0)}$$

Wenn die bivariate Regressionsfunktion linear ist, d. h. $y^* = f(x) = b_0 + b_1 \cdot x$, dann gilt:

$$\epsilon(x) = \frac{x}{y^*(x)} \cdot \frac{dy^*(x)}{dx} = \frac{b_1 \cdot x}{b_0 + b_1 \cdot x}$$

Elastizität / Interpretation

Bewertung der Nachgiebigkeit der Variablen Y bei Veränderung der Variablen X:

$$\begin{array}{lll} |\epsilon(x_0)| & > 1 & \textit{ueberproportional elastisch} \\ |\epsilon(x_0)| & = 1 & \textit{proportional elastisch} \\ |\epsilon(x_0)| & < 1 & \textit{unterproportional elastisch} \end{array}$$

$$y^* = f(x) = b_0 + b_1 \cdot x = 29.31 + 17.34 \cdot x$$

$$\begin{aligned} y'*(x_0) &= \frac{dy^*(x)}{dx} \|_{x=x_0} = b_1 = 17.34 \ \forall x = x_0 \\ \epsilon(x) &= \frac{b_1 \cdot x}{b_0 + b_1 \cdot x} = \frac{17.34 \cdot x}{29.31 + 17.34 \cdot x} \\ \epsilon(x_0) &= \frac{17.34 \cdot x_0}{29.31 + 17.34 \cdot x_0} = \frac{17.34 \cdot 6.5}{29.31 + 17.34 \cdot 6.5} = 0.7936 \quad \textit{wenn} \quad x_0 = 6,5 \\ &\Longrightarrow \quad |\epsilon(x_0)| < 1 \Longrightarrow \ \textit{unterproportional elastisch} \end{aligned}$$

Elastizität / Interpretation

Maßeinheiten: X -Kraftstoffverbrauch $[\frac{l}{100km}]$ und Y Schadstoffausstoß $[\frac{g}{km}]$

$y^* = f(x) = b_0 + b_1 \cdot x = 29.31 + 17.34 \cdot x$

$$\begin{array}{ccccccc} y^* & = & b_0 & + & b_1 & x \\ [\frac{g}{km}] & = & [\frac{g}{km}] & + & [\frac{l00g}{l}] & [\frac{l}{100km}] \end{array}$$

$$\begin{aligned} y'*(x_0) &= \frac{dy^*(x)}{dx}\|_{x=x_0} = b_1 = 17.34\frac{100g}{l} \ \forall x = x_0 \\ \epsilon(x_0) &= \frac{17.34 \cdot x_0}{29.31 + 17.34 \cdot x_0} = \frac{17.34 \cdot 6.5}{29.31 + 17.34 \cdot 6.5} = 0.7936 \frac{\frac{g}{km}}{\frac{g}{km}} \end{aligned}$$

$\Longrightarrow$ dimensionslos

Elastizität / Interpretation

$$y^* = f(x) = b_0 + b_1 \cdot x = 29.31 + 17.34 \cdot x$$

$$\begin{aligned} y'*(x_0) &= \frac{dy^*(x)}{dx}\|_{x=x_0} = b_1 = 17.34 \ \forall x = x_0 \\ \epsilon(x_0) &= \frac{17.34 \cdot x_0}{29.31 + 17.34 \cdot x_0} = \frac{17.34 \cdot 6.5}{29.31 + 17.34 \cdot 6.5} = 0.7936 \quad \textit{wenn} \quad x_0 = 6,5 \end{aligned}$$

- $y'*(x_0)$ - **durchschnittlich absolute Veränderung der exogenen Variablen y** bei einer Erhöhung /Verringerung des Kraftstoffverbrauchs um $1 \frac{l}{100km}$ steigt / fällt der Schadstoffausstoß um $17.34 \frac{g}{km}$.
- $\epsilon(x_0)$ - **Elastizität der exogenen Variablen Y** bei einer Erhöhung /Verringerung des Kraftstoffverbrauchs um 1 Prozent steigt / fällt der Schadstoffausstoß um 0,79 Prozent.

Gliederung

- 3.2.1 Problemstellung und Herleitung
- 3.2.2 Interpretation und Beispiel
- 3.2.3 Elastizitäten
- **3.2.4 Gütemaße**

Varianz / Standardabweichung der Residuen

Definition: Die Varianz der Residuen (Mittlerer quadratischer Fehler der Regression) d_u^2 wird analog zur empirischen Varianz berechnet:

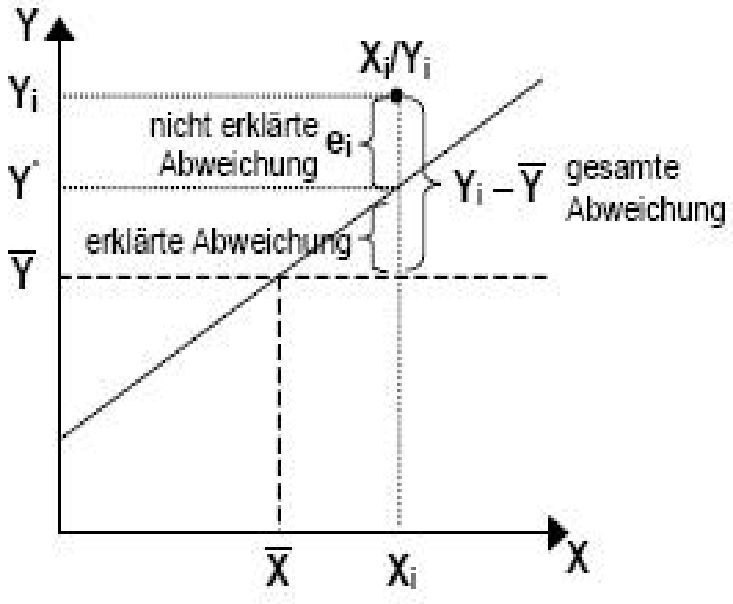
$$d_u^2 = \frac{1}{n}\sum_{i=1}^{n}(u_i - \overline{u})^2 = \frac{1}{n}\sum_{i=1}^{n} u_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2$$

Definition: Die Standardabweichung der Residuen (Standardabweichung der Regression) d_u^2 wird analog zur empirischen Standardabweichung berechnet:

$$d_u = \sqrt{d_u^2}$$

Bestimmtheitsmaß

Interpretation:

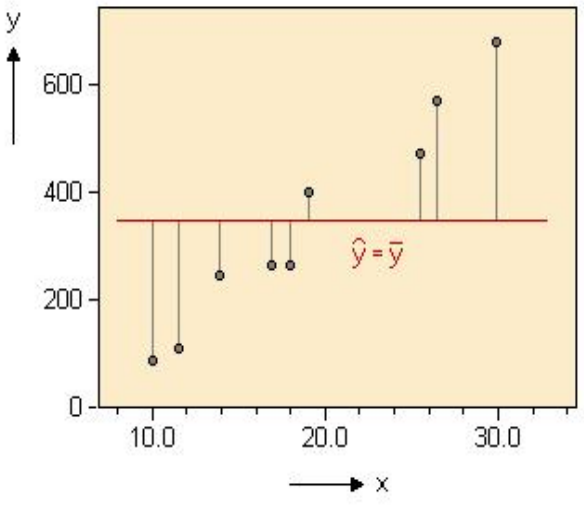


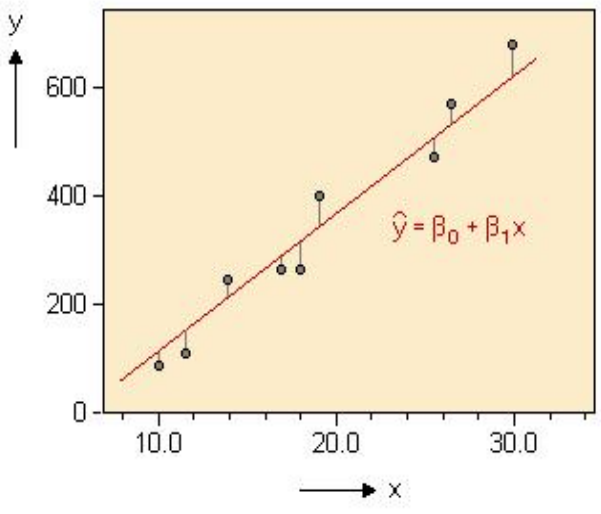
$$R^2 = \frac{erklaerte\ Abweichung}{Gesamte\ Abweichung}$$

Bestimmtheitsmaß

$$
\begin{aligned}
R^2 &= \frac{\textit{erklaerte Abweichung}}{\textit{Gesamte Abweichung}} = \\
&= \frac{\sum_{i=1}^{n}(y_i^* - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n}(y_i^* - \overline{y^*})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{d_{y^*}^2}{d_y^2} \\
&\quad (\textit{aus der Grafik}) \\
&= \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2 - \sum_{i=1}^{n}(y_i - y_i^*)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \\
&= 1 - \frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n} u_i^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{d_u^2}{d_y^2}
\end{aligned}
$$

Bestimmtheitsmaß





$$(y_i - \overline{y}) = (y_i - \overline{y^*})$$

$$u_i^2 = (y_i - y_i^*)^2$$

Anforderungen an Gütemaße

Um Gütemaße zum Vergleich verschiedener Modellansätze heranziehen zu können, müssen folgende Bedingungen erfüllt sein:

1. die endogene Variable in zu vergleichenden Modellen muß identisch sein (numerische Werte müssen übereinstimmen).
2. die Anzahl der exogenen Parameter muß in denzu vergleichenden Modellen identisch sein.
3. Die Modelle besitzen einen Niveauparameter (b_0).

Einführung in SPSS - Themen

Statistische Themen: Deskriptive Statistik

- Lineare Regression
- Diagramme (Sequenz) - Fortsetzung

Lineare Regression - Beispiel

mit Excel:

Beispiel 2: Zinsätze 2007_2.xls

Aufgaben: Bestimmen der Regressionsgleichung und der Gütemaße

mit SPSS:

Beispiel 2: Zinsätze 2007_2.xls

Aufgaben: Schätzen der Koeffizienten, Auswertung der ANOVA und anderer Tabellen, grafische darstellung der Ergebnisse

2. Korrelations- und Regressionsanalyse - Gliederung

- 3.1 Korrelationsanalyse
- 3.2 Einfache lineare Regression
- **3.3 Einfache nichtlineare Regression**
- 3.4 Multiple Regression

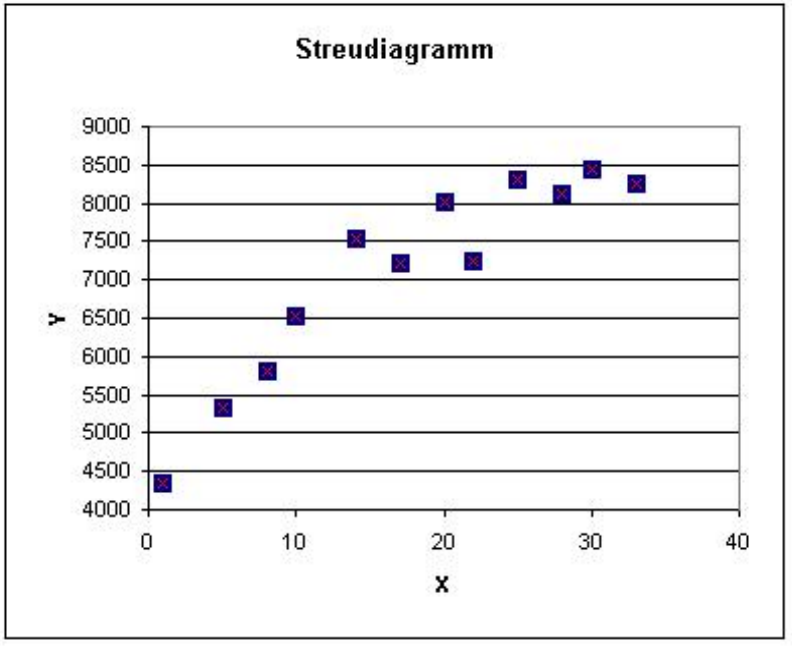
Gliederung

- **3.2.1 Problemstellung**
- 3.2.2 Linearisierung
- 3.2.3 Elastizitäten
- 3.2.4 Nichtlineare Regressionstechniken - Alternativen

einführendes Beispiel

Beispiel : Kraftfutter (in 50 kg/Jahr) X und Milchleistung (l / Jahr) Y für 12 Jahre

	X	Y
1	10	6525
2	30	8437
3	20	8019
4	33	8255
5	5	5335
6	22	7236
7	8	5821
8	14	7531
9	25	8320
10	1	4336
11	17	7225
12	28	8112



Quelle: von Auer, Ökonometrie Berlin, Heidelberg, New York 1999

Gliederung

- 3.2.1 Problemstellung
- **3.2.2 Linearisierung**
- 3.2.3 Elastizitäten

Linearisierbare Funktionen

Transformation von nichtlinearen Funktionen in lineare Funktionen: Linarisierung

Funktionstyp	**Funktionsgleichung**	**Linearisierte Funktion**
logarithmisch	$e^y = e^{(b_0+u)}x^{b_1}$ $y = b_0 + b_1 ln(x) + u$	$y = b_0 + b_1 ln(x) + u$
exponential	$y = e^{b_0+b_1 x+u}$	$ln(y) = b_0 + b_1 x + u$
invers	$y = b_0 - b_1 \frac{1}{x} + u$	$y = b_0 + b_1 z + u \ mit \ z = \frac{1}{x}$
Potenzfunktion	$y = b_0 x^{b_1} e^u$	$ln(y) = ln(b_0) + b_1 ln(x) + u$

Linearisierbare Funktionen

verwendete transformierte Größen

Funktionstyp	**Linearisierte Funktion**	**Verwendete Beobachtungen**
logarithmisch	$y = b_0 + b_1 ln(x) + u$	y, $ln(x)$
exponential	$ln(y) = b_0 + b_1 x + u$	$ln(y)$, x
invers	$y = b_0 + b_1 z + u$ *mit* $z = \frac{1}{x}$	y, $\frac{1}{x}$
Potenzfunktion	$ln(y) = ln(b_0) + b_1 ln(x) + u$	$ln(y)$, $ln(x)$

Probleme

Die Technik der Linearisierung wirft folgende Probleme auf:

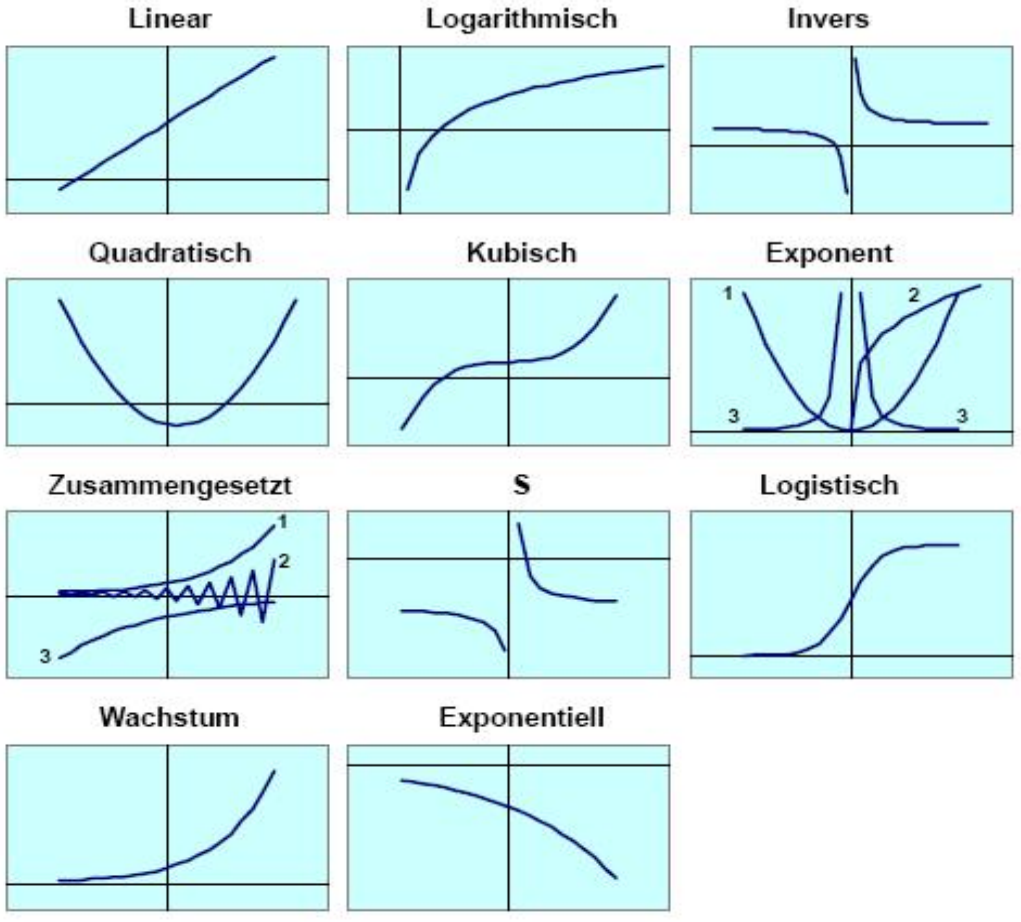
- Die Fehler u haben für jede linearisierte Funktion unterschiedliche Maßeinheiten. Damit ist der Vergleich der Modellgüte mit Hilfe der Kriterien Varianz der Residuen und Besrtimmtheitsmaß nicht zulässig.
- Nicht alle nichtlinere Funktionen lassen sich linearisieren.
- Es ist zu überprüfen, ob alle statistischen Eigenschaften durch die lineare Transformation erhalten bleiben.

Linearisierbare Funktionen in SPSS

Kurventyp	Funktion	Lineare Form
Linear	$y = b_0 + b_1 \cdot x$	$y = b_0 + b_1 \cdot x$
Logarithmisch	$y = b_0 + b_1 \cdot \ln(x)$	$y = b_0 + b_1 \cdot \ln(x)$
Invers	$y = b_0 + \frac{b_1}{x}$	$y = b_0 + \frac{b_1}{x}$
Quadratisch	$y = b_0 + b_1 \cdot x + b_2 \cdot x^2$	$y = b_0 + b_1 \cdot x + b_2 \cdot x^2$
Kubisch	$y = b_0 + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3$	$y = b_0 + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3$
Exponent	$y = b_0 \cdot x^{b_1}$	$\ln(y) = \ln(b_0) + b_1 \cdot \ln(x)$
Zusammengesetzt	$y = b_0 \cdot b_1^x$	$\ln(y) = \ln(b_0) + \ln(b_1) \cdot x$
S	$y = e^{b_0 + \frac{b_1}{x}}$	$\ln(y) = b_0 + \frac{b_1}{x}$
Logistisch	$y = \frac{1}{\frac{1}{o} + b_0 \cdot b_1^x}$	$\ln\left(\frac{1}{y} + \frac{1}{o}\right) = \ln(b_0) + \ln(b_1) \cdot x$
Wachstum	$y = e^{b_0 + b_1 \cdot x}$	$\ln(y) = b_0 + b_1 \cdot x$
Exponentiell	$y = b_0 \cdot e^{b_1 \cdot x}$	$\ln(y) = \ln(b_0) + b_1 \cdot x$

$\Longrightarrow$ Berechnung mit SPSS (Kurvenanpassung)

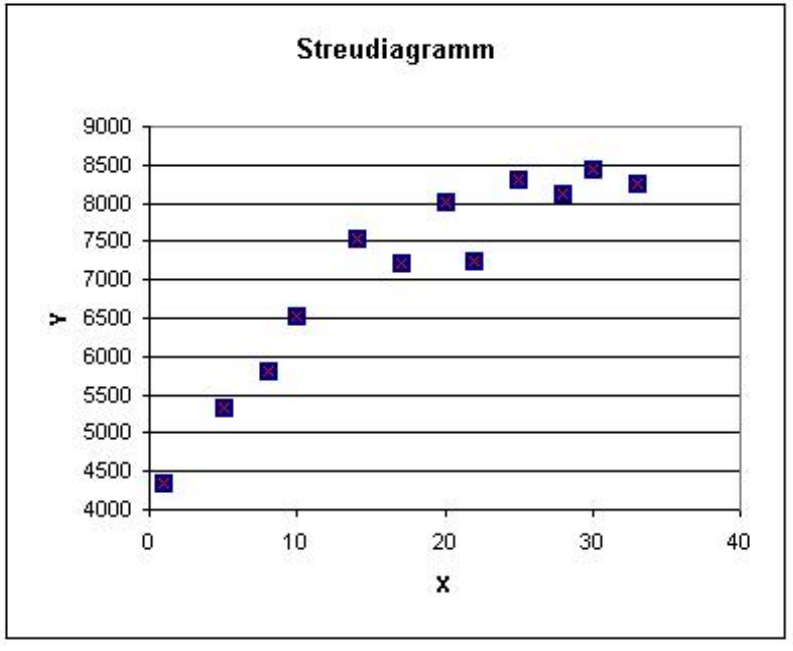
Kurvenverlauf



Beispiel

Beispiel : Kraftfutter (in 50 kg/Jahr) X und Milchleistung (l / Jahr) Y für 12 Jahre

	X	Y
1	10	6525
2	30	8437
3	20	8019
4	33	8255
5	5	5335
6	22	7236
7	8	5821
8	14	7531
9	25	8320
10	1	4336
11	17	7225
12	28	8112



Quelle: von Auer, Ökonometrie Berlin, Heidelberg, New York 1999

Beispiel Vorgehen

Schritte:

1. Wahl der linearisierbaren Funktion anhand der Grafik
   $\Rightarrow$ logarithmisch, exponential
2. Umformen der Ausgangswerte
3. Lösung des linearisierten Problems
4. Formulierung der nichtlinearen Schätzgleichung

Beispiel Linearisierung und Lösung des linearen Problems

logarithmisch: $e^y = e^{(b_0+u)}x^{b_1} \Rightarrow y = b_0 + b_1 ln(x) + u$

Transformation: $x \rightarrow ln(x) \Rightarrow \ z = ln(x)$

Lösung des linearen Problems:

$$b_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} z_i^2 - \sum_{i=1}^{n} z_i \sum_{i=1}^{n} z_i y_i}{n \sum_{i=1}^{n} z_i^2 - (\sum_{i=1}^{n} z_i)^2}$$

$$b_1 = \frac{n \sum_{i=1}^{n} z_i y_i - \sum_{i=1}^{n} z_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} z_i^2 - (\sum_{i=1}^{n} z_i)^2}$$

Beispiel Umformen der Ausgangswerte

Beispiel : Kraftfutter (in 50 kg/Jahr) X und Milchleistung (l / Jahr) Y für 12 Jahre

	X	Y	ln(X)	ln(y)	1/x
1	10	6525	2.302585	8.783396	0.1
2	30	8437	3.401197	9.040382	0.033333
3	20	8019	2.995732	8.989569	0.05
4	33	8255	3.496508	9.018574	0.030303
5	5	5335	1.609438	8.582044	0.2
6	22	7236	3.091042	8.886824	0.045455
7	8	5821	2.079442	8.669227	0.125
8	14	7531	2.639057	8.926783	0.071429
9	25	8320	3.218876	9.026418	0.04
10	1	4336	0	8.374708	1
11	17	7225	2.833213	8.885303	0.058824
12	28	8112	3.332205	9.0011	0.035714

$\Longrightarrow$ hier nur ln(y) notwendig

Berechnungen

$$b_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} z_i^2 - \sum_{i=1}^{n} z_i \sum_{i=1}^{n} z_i y_i}{n \sum_{i=1}^{n} z_i^2 - (\sum_{i=1}^{n} z_i)^2} = 3818.334$$

$$b_1 = \frac{n \sum_{i=1}^{n} z_i y_i - \sum_{i=1}^{n} z_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} z_i^2 - (\sum_{i=1}^{n} z_i)^2} = 1268.803$$

lineare Regressionsgleichung: $y = b_0 + b_1 ln(x) + u = 3818.334 + 1268.803 ln(x) + u$

nichtlineare Regressionsgleichung: $e^y = e^{(b_0+u)} x^{b_1} = e^{(3818.334+u)} x^{1268.803}$

Beispiel Linearisierung und Lösung des linearen Problems

exponential: $y = e^{b_0+b_1x+u} \Rightarrow ln(y) = b_0 + b_1x + u$
Transformation: $y \rightarrow ln(y) \Rightarrow \ z = ln(y)$ Lösung des linearen Problems:

$$b_0 = \frac{\sum_{i=1}^{n} z_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i z_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = 8,523600651$$

$$b_1 = \frac{n \sum_{i=1}^{n} x_i z_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} z_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = 0,018315116$$

lineare Regressionsgleichung:
$ln(y) = b_0 + b_1x + u = 8,523600651 + 0,018315116x + u$

nichtlineare Regressionsgleichung:$y = e^{b_0+b_1x+u} = e^{8,523600651+0,018315116x+u}$

Gliederung

- 3.2.1 Problemstellung
- 3.2.2 Linearisierung
- **3.2.3 Elastizitäten**

Elastizitäten

marginale Veränderungen: $\frac{dy}{dx}$

Funktionstyp	**Funktionsgleichung**	**Grenzfunktionen**
logarithmisch	$y = b_0 + b_1 ln(x) + u$	$b_1 x^{-1}$
exponential	$y = e^{b_0 + b_1 x + u}$	$b_1 e^{b_0 + b_1 x}$
invers	$y = b_0 - b_1 \frac{1}{x} + u$	$-b_1 x^{-2}$
Potenzfunktion	$y = b_0 x^{b_1} e^u$	$b_0 b_1 x^{b_1} x^{-1}$

Elastizitäten

marginale Veränderungen: $\frac{dy}{dx} \cdot \frac{x}{y}$

Funktionstyp	**Funktionsgleichung**	**Grenzfunktionen**	**Elastizitäten**
logarithmisch	$y = b_0 + b_1 ln(x) + u$	$b_1 x^{-1}$	$b_1 [b_0 + b_1 ln(x)]^{-1}$
exponential	$y = e^{b_0 + b_1 x + u}$	$b_1 e^{b_0 + b_1 x}$	$b_1 x$
invers	$y = b_0 - b_1 \frac{1}{x} + u$	$-b_1 x^{-1}$	$-b_1 [b_0 x + b_1]^{-1}$
Potenzfunktion	$y = b_0 x^{b_1} e^u$	$b_0 b_1 x^{b_1} x^{-1}$	b_1

Einführung in SPSS - Themen

Statistische Themen: Deskriptive Statistik

- Kurvenanpassung
- nichtlineare Regression

Nichtlineare Regression - Beispiel

mit Excel:

Beispiel : ???

Aufgaben: Bestimmen der Regressionsgleichung und der Gütemaße

mit SPSS:

Beispiel : ???

Aufgaben: Schätzen der Koeffizienten, Auswertung ANOVA und anderer Tabellen, grafische darstellung der Ergebnisse

2. Korrelations- und Regressionsanalyse - Gliederung

- 3.1 Korrelationsanalyse
- 3.2 Einfache lineare Regression
- 3.3 Einfache nichtlineare Regression
- **3.4 Multiple Regression**

Gliederung

- **3.4.1 Definition und Herleitung**
- 3.2.2 Gütemaße
- 3.2.3 Beispiel

Definition

Definition: Y und X_j mit $j = 1, 2, ..., m$ - metrisch skalierte Merkmale. Die Ausprägungen dieser Merkmale wurden an n statistischen Einheiten erhoben:

$$(y_1, x_{11}, x_{12}, ..., x_{1m}),\ (y_2, x_{21}, x_{22}, ..., x_{2m}),\ ..., (y_n, x_{n1}, x_{n2}, ..., x_{nm})$$

Eine lineare Funktion $y^* = f(x_1, x_2, ..., x_m) = b_0 + \sum_{j=1}^{m} b_j \cdot x_j$ zur Beschreibung der statistischen Abhängigkeit des Merkmals Y von den Merkmalen X_j heißt **einfache lineare Regression**.

X_j - Regressoren, unabhängige oder exogene Variablen

Y - Regressand, abhängige oder endogene Variable

x - vorgebener Vektor der Realisationen der Merkmale X_j mit $\mathbf{x} = (x_1, x_2, ..., x_m)^T$ *und* $x_j \in \Re$)

$y^* = y^*(\mathbf{x})$ Regresswert (Realisation von y Bei vorgegebenen **x**)

b_0 - Regressionskonstante (intercept)

b_j - linearer Regressionskoeffizienten, slop, $\mathbf{b} = (b_1, b_2, ..., b_m)$

Es gilt: $y^* = y^*(\mathbf{x}) = b_0 + \mathbf{bx}$

Definition

Die Gleichung $y = f(x_1, x_2, ..., x_m) = b_0 + \sum_{j=1}^{m} b_j \cdot x_j + u$ gilt für allle Beobachtungen der Urliste:

$$\begin{array}{lllllllllll} y_1 & = & b_0 & + & b_1 x_{11} & + & \cdots & + & b_m x_{1m} & + & u_1 \\ y_2 & = & b_0 & + & b_1 x_{21} & + & \cdots & + & b_m x_{2m} & + & u_2 \\ \cdots y_n & = & b_0 & + & b_1 x_{n1} & + & \cdots & + & b_m x_{nm} & + & u_n \end{array}$$

oder in Matrizenschreibweise

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ b_1 \\ \cdots \\ b_m \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \cdots \\ u_n \end{pmatrix}$$

oder kompakt: $\mathbf{y} = \mathbf{X} \cdot \mathbf{b} + \mathbf{u}$

Herleitung

Ausgangspunkt:

$\mathbf{y} = \mathbf{X} \cdot \mathbf{b} + \mathbf{u}$

Extremwertproblem:

$S(b) = u^T u = (y - Xb)^T (y - Xb) = y^T y - 2bX^T y + b^T X^T Xb \rightarrow MIN$

$\ldots$

Normalgleichungssystem:

$X^T Xb = X^T y$

$\Longrightarrow$

Lösung:

$b = (X^T X)^{-1} X^T y$

Gliederung

- 3.4.1 Definition und Herleitung
- **3.2.2 Gütemaße**
- 3.2.3 Beispiel

Gütemaße

Residualquadratsumme:

$u^T u = (y - Xb)^T (y - Xb)$

Bestimmtheitsmaß:

$R^2 = 1 - \frac{u^T u}{y^T y - n\bar{y}}$

korrigiertes Bestimmtheitsmaß:

$\overline{R^2} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$ mit k - Anzahl der exogenen Variablen

Gliederung

- 3.4.1 Definition und Herleitung
- 3.2.2 Gütemaße
- **3.2.3 Beispiel**

Beispiel

	daxhchst	daxtief	daxende
	X1	X2	Y
02.Jun.98	5589	5554	5583
03.Jun.98	5626	5607	5614
04.Jun.98	5618	5575	5592
05.Jun.89	5689	5636	5689
08.Jun.98	5788	5755	5780
09.Jun.98	5783	5745	5760
10.Jun.98	5754	5719	5754
12.Jun.98	5739	5642	5671
15.Jun.98	5597	5513	5527
16.Jun.98	5604	5569	5592
17.Jun.98	5712	5675	5709
18.Jun.98	5756	5697	5718
19.Jun.98	5707	5683	5703
22.Jun.98	5655	5616	5655
23.Jun.98	5719	5680	5719
24.Jun.98	5780	5722	5780
25.Jun.98	5877	5840	5867
26.Jun.98	5893	5849	5870
29.Jun.98	5921	5896	5915
30.Jun.98	5901	5858	5897

Quelle: dax.sav

Vorlesungsbeispiel 3

Einführung in SPSS - Themen

Statistische Themen: Deskriptive Statistik

- Lineare Regression
- Methoden zur Auswahl der unabhängigen Variablen

Einführung in SPSS - Themen

1. Einschluß (Regression). Eine Prozedur für die Variablenauswahl, bei der alle Variablen eines Blocks in einem einzigen Schritt aufgenommen werden.
2. Schrittweise Selektion. Bei jedem Schritt wird die noch nicht in der Gleichung enthaltene unabhängige Variable, wenn das Auswahlkriterium erfüllt ist
3. Entfernen (Variable Selection). Ein Verfahren zur Variablenselektion, bei dem alle Variablen eines Blocks in einem Schritt ausgeschlossen werden.
4. Rückwärtselimination. Eine Methode zur Variablendselektion, bei der alle Variablen in die Gleichung aufgenommen und anschließend sequentiell ausgeschlossen werden.
5. Vorwärtsselektion. Ein Verfahren zur schrittweisen Variablenauswahl, in dem die Variablen nacheinander in das Modell aufgenommen werden.

Multiple lineare Regression - Beispiel

mit Excel:

Beispiel 2: Zinsätze 2007_2.xls:

Aufgaben: Bestimmen der Regressionsgleichung und der Gütemaße

mit SPSS:

Beispiel 2: Zinsätze 2007_2.xls:

Aufgaben: Schätzen der Koeffizienten, Auswertung ANOVA und anderer Tabellen, grafische darstellung der Ergebnisse

Multiple nichtlineare Regression - Beispiel

mit Excel:

Beispiel 2: Produktionsfunktion Cobb-Douglas

Aufgaben: Bestimmen der Regressionsgleichung und der Gütemaße, ökonomische Interpretation

mit SPSS:

Beispiel 2: Produktionsfunktion Cobb-Douglas

Aufgaben: Schätzen der Koeffizienten, Auswertung ANOVA und anderer Tabellen, grafische darstellung der Ergebnisse, ökonomische Interpretation